Genome Annotation Errors and How to Fix Them

Michael Peter Dunne

Worcester College
University of Oxford

A thesis submitted for the degree of Doctor of Philosophy

Hilary 2018

Abstract

Many inferences about the biological properties of an organism depend on the completeness and accuracy of its genome annotation. Advances in sequencing technologies and the associated decreased costs have brought whole genome sequencing projects into the reach of individual laboratories, precipitating a huge acceleration in the publication of draft genome assemblies and annotations. While genome assembly quality metrics have received substantial attention, adequate frameworks for quantifying and controlling errors in genome annotations are lacking, and thus the completeness and accuracy of published genome annotations are unknown. Moreover, genome annotations are frequently taken at face value by those using them, and any errors present are propagated in downstream inferences and analyses. Despite underpinning much of comparative genomic research, little attention has been paid to quantifying the extent of genome annotation inaccuracies and the majority of attempts to systematically rectify such errors have relied either on manual input, which is impractical on a large scale, or on universally conserved gene sets, which only account for a small percentage of genes. The aim of the research described in this thesis is to provide methods to assess and rectify two main classes of genome annotation errors (missing genes and incorrect gene models) at a phylogenetically local level, by mutually improving genome annotations for sets of related species, in the absence of extrinsic experimental data. I introduce several non-extrinsic metrics for assessing genome annotation completeness and the accuracy of the gene models contained therein, and provide two self-contained methods that improve genome annotation accuracy. In summary, this thesis reveals that genome annotation errors are widespread, even in widely studied community annotated genomes, and that many of these errors can be identified and corrected using automated phylogenetically local approaches.

Genome Annotation Errors and How to Fix Them



Michael Peter Dunne
Worcester College
University of Oxford

A thesis submitted for the degree of
Doctor of Philosophy
Hilary 2018

For Triangle.

Abstract

Many inferences about the biological properties of an organism depend on the completeness and accuracy of its genome annotation. Advances in sequencing technologies and the associated decreased costs have brought whole genome sequencing projects into the reach of individual laboratories, precipitating a huge acceleration in the publication of draft genome assemblies and annotations. While genome assembly quality metrics have received substantial attention, adequate frameworks for quantifying and controlling errors in genome annotations are lacking, and thus the completeness and accuracy of published genome annotations are unknown. Moreover, genome annotations are frequently taken at face value by those using them, and any errors present are propagated in downstream inferences and analyses. Despite underpinning much of comparative genomic research, little attention has been paid to quantifying the extent of genome annotation inaccuracies and the majority of attempts to systematically rectify such errors have relied either on manual input, which is impractical on a large scale, or on universally conserved gene sets, which only account for a small percentage of genes. The aim of the research described in this thesis is to provide methods to assess and rectify two main classes of genome annotation errors (missing genes and incorrect gene models) at a phylogenetically local level, by mutually improving genome annotations for sets of related species, in the absence of extrinsic experimental data. I introduce several non-extrinsic metrics for assessing genome annotation completeness and the accuracy of the gene models contained therein, and provide two self-contained methods that improve genome annotation accuracy. In summary, this thesis reveals that genome annotation errors are widespread, even in widely studied community annotated genomes, and that many of these errors can be identified and corrected using automated phylogenetically local approaches.

Acknowledgements

There are many people without whom this document would not have existed, and many more without whom the process of its creation would have been several shades less enjoyable. I would like to use these pages to thank some of those people. If your name isn't here, that's probably because I like you enough to have hidden it in the main text: you'd better read the whole thing.

First and foremost, I would like to extend my sincere gratitude to my supervisor, Steve Kelly. The completion of this thesis would not have been possible without your support and guidance. Thanks for taking me on and putting up with me, for all your ideas and suggestions and occasional cattle-prodding. A heroic feat, well done and thank you. Thanks also to Ben, for showing me the ropes and teaching me the joys of bash one-liners, and to Martin, who taught me how fun it can be to play with colourless liquids. Especial thanks also to Dave G, Elspeth, and Antonis for their support, counsel, and patience, and Gem, who always knew what to do. I would also like to thank all those who helped fund my existence for the duration of my studies: the EPSRC, the Ludwig Institute, and Worcester College, without whom I would have been very hungry and probably quite cold.

I'd be a big fat liar if I said I didn't complain during the process of producing this thesis. I would therefore like to present a special medal to all those who endured my grumbling, offered perspective, and generally listened and supported. Unfortunately I can't afford a medal, so I'll use words instead. Firstly, thanks to my family, especially my parents Joan and Pete and my sister Anna, for providing warmth, shelter, food, and drink in times of need, travelling hundreds of miles to buy me lunch, and for hopping on board this crazy train with me. I think they sell refreshments at the next station. It's my round. Thanks also to my feline pals PTL and Audrey, for your reassuring purrs and for the supportive gifts. You tried.

Same goes to everyone who's made the brave decision to spend time with me over the course of my degree, especially everyone I've lived with, be it permanently or just sometimes, years, weeks, or weekends. It's been a great pleasure getting repeatedly injured with you all: bouncing on poorly constructed trampolines, diving into dubious murky hot tubs, and getting burnt by trolley barbecues, not to mention the less dangerous pursuits of cat sushi, spoon dancing, and Patrick Stewarts (spicy or otherwise). I haven't decided yet whether frequent trips to Plush and Thirst

should be included in the dangerous or not dangerous category. The long-suffering housemates: Joshua, Lottie, Jo, Leah, Jemma, and Caroline all deserve a mention. Thanks to you all for letting me moan at you, and for telling me to stop being such an idiot. You were right. One day I'll make you that croquembouche. Most of all, thanks to our guiding light, Triangle. You were the very best of us 🐾. Thanks also to all the many groups who distracted me and kept me sane, especially everyone in Mama Brass, Temple Funk Collective, and the Dutty Moonshine Big Band. Without you I probably would have submitted this six months earlier. Many thanks.

The 9-to-5 would have been very lonely if I'd spent these years sitting alone in a cold, grey office, perhaps with just a yo-yo and a spider for company. That didn't happen, and I would therefore like to thank all of the real and honorary members of the Kelly and Langdale labs who helped bring colour and light to my working days. In particular, all my office buddies: Emily, wise owl, fount of reassurance and general force for good, who bravely helped wade through this tome searching for ytpos; Jessie, appreciator of caffeine and all things generally nice, who was there to help trek through the wild forests of madness; Ross, trouser visionary and brave companion of the spicy jalapeño corner; Micky Nikky, master of abbrevs and appreciator of lamellophones; Rona, who understands that mulled wine improves code; David, who remains calm when I crash his processes; Squactus, the trailblazer, an inspiration; and Peng, who gave me two packets of Pom-Bears yesterday. *Two packets*. Thank you, they are my favourite.

Finally, the unsung heroes. The waterfowl of University Parks: your eternal wisdom shows no bounds. Thank you for helping me finish my bread. Even you, Goose. Thanks to the Alternative Tuck Shop for remembering my face even when I'd forgotten it. Thanks to Rudy for the perspective. Thanks to Hassan, purveyor of dreams. In ten years, nobody's done it better. Thanks to Crayola: may you follow your heart. And thanks to everyone I bumped into in the street who asked "Are you still here?" It was very motivational.

Contents

List of figures **xvii**

List of abbreviations and symbols **xxi**

1 General introduction **1**
1.1 Introduction 1
1.2 Draft genome assemblies are quick to produce 5
1.2.1 High-throughput methods yield fragmented assemblies . . . 5
1.2.2 Automated annotation pipelines require little human input . 7
1.2.3 Published annotations are often poorly maintained 11
1.3 Annotation errors lead to false inferences 13
1.3.1 Absent annotations can be mistaken for gene loss 13
1.3.2 Inaccurate gene models lead to transcript quantification errors 15
1.3.3 Inaccurate gene models inhibit domain identification 16
1.3.4 Subcellular localisation predictions require accurate translation start sites 17
1.3.5 False inferences from inaccurate annotations are propagated 18
1.4 Assessments of annotation quality are scarce 19
1.4.1 No gene finder is 100% accurate 20
1.4.2 Existing assessments of annotation quality rely on universally conserved features or extrinsic data 22
1.5 Thesis aims 24
1.5.1 Thesis plan 25

2 Phylogenetically local measures of genome completeness **27**
2.1 Chapter introduction 27
2.1.1 Background 27
2.1.2 Chapter aims and outline 29
2.2 Two test species sets: plants and fungi 31
2.2.1 Species list and sources 32
2.2.2 Plant genomes are larger than fungal genomes 33
2.2.3 Most of the plant assemblies are highly unfinished 34

2.2.4 Plants have larger gene sets than fungi 35
2.2.5 Introns are more common in plants than in fungi 36
2.2.6 Plants contain more non-coding material than fungi 37
2.2.7 Gene distributions affect gene finder accuracy 39
2.3 Analysis of plant and fungal orthogroups 39
2.3.1 Orthogroups formalise the notion of conserved genes 40
2.3.2 Plants have larger and more numerous orthogroups than fungi 41
2.3.3 Counting absent orthogroup members is not enough 42
2.3.4 *C. payaya* is probably under-annotated 44
2.4 An evidence-based method for estimating the extent of missing genes 45
2.4.1 Hidden Markov models identify potential missing genes . . . 46
2.4.2 An adjusted bit score avoids E-value biases 47
2.4.3 HMM hits can be classified by comparison to existing annotations 48
2.4.4 A further filtering step removes spurious hits 53
2.4.5 Many genomes show signs of annotation incompleteness . . 56
2.5 Examples of missing genes identified using HMMs 58
2.5.1 Missing members of plant orthogroups 59
2.5.2 Missing members of fungal orthogroups 63
2.6 Comparison of candidate missing genes to RNA-seq data 65
2.6.1 Accepted hits have higher read coverage than rejected hits . 65
2.6.2 Accepted hits have higher read coverage than intergenic regions 67
2.7 Methods 70
2.7.1 Genome algebra and feature file manipulation 70
2.7.2 Orthogroup calculation 70
2.7.3 Acquisition of phylogenetic trees 70
2.7.4 Alignment of protein and nucleotide sequences 70
2.7.5 Acquisition and implementation of HMMs 71
2.7.6 Score distribution estimation 71
2.7.7 RNA-seq data sources 71
2.7.8 Alignment and quantification of RNA-seq data 72
2.8 Chapter discussion 72
2.8.1 Different species sets display markedly different genome characteristics 73
2.8.2 Incomplete orthogroups can indicate annotation issues . . . 73
2.8.3 Filtered HMM searches inform incompleteness estimates . . 74
2.8.4 RNA-seq evidence supports putative missing genes 75
2.8.5 Summary 76

3 Quantifying gene model errors **77**
3.1 Chapter introduction . . . 77
3.1.1 Background . . . 77
3.1.2 Chapter aims and outline . . . 79
3.2 Gene model errors come in several flavours . . . 81
3.2.1 Inaccurate exon boundaries . . . 81
3.2.2 Exon omission/inclusion errors . . . 82
3.2.3 Intron omission/inclusion errors . . . 85
3.2.4 Inaccurate start codons . . . 87
3.2.5 Split and fused genes . . . 89
3.2.6 Complex errors . . . 91
3.2.7 Accurate gene model annotation is not always possible . . . 91
3.3 Extrinsic metrics reveal gene model errors . . . 94
3.3.1 Splice junction recall identifies erroneous splice junctions . . . 95
3.3.2 Clear differences in splice junction recall scores between species 96
3.3.3 Coverage score identifies missing introns and exons . . . 100
3.3.4 Coverage and splice junction recall are correlated . . . 102
3.4 A non-extrinsic metric for gene model assessment . . . 104
3.4.1 A scoring system for amino acid alignments . . . 104
3.4.2 Poor alignment scores reflect low sequence similarity . . . 106
3.4.3 Fungal orthogroups have lower alignment scores . . . 108
3.4.4 Pools, bridges, and discrepant genes . . . 110
3.4.5 Many orthogroups contain clear discrepancies . . . 111
3.4.6 Discrepant plant genes have poor splice junction recall . . . 113
3.4.7 Discrepant plant genes have suboptimal coverage . . . 115
3.5 Subcellular localisation predictions indicate start codon accuracy . 120
3.5.1 Extrinsic analyses agree with localisation inconsistencies . . . 121
3.5.2 Initial discrepancies correlate with localisation inconsistencies 122
3.6 Gene model scores can inform primary transcript choice . . . 126
3.6.1 Alternate transcript choices can improve alignment scores . 127
3.6.2 Alternate transcript choices can improve localisation consistency 127
3.6.3 Many primary transcripts have suboptimal coverage and splice junction recall . . . 128
3.7 Methods . . . 130
3.7.1 Obtaining mapping junctions . . . 130
3.7.2 Subcellular localisation prediction . . . 130
3.7.3 Calculation of entropy score . . . 130
3.8 Chapter discussion . . . 131
3.8.1 Coverage and splice junction assessments are correlated . . . 131

3.8.2 *C. papaya* and *T. cacao* show strong error signals 132
3.8.3 Discrepant genes are evidence for gene model errors 132
3.8.4 Non-extrinsic methods reveal otherwise hidden errors 133
3.8.5 Primary transcript choices are frequently suboptimal 133
3.8.6 Summary 134

4 OrthoFiller: Utilising data from multiple species to improve the completeness of genome annotations 135
4.1 Chapter introduction 135
4.1.1 Background 135
4.1.2 Chapter aims and outline 136
4.2 Results 137
4.2.1 Algorithm overview and evaluation criteria 137
4.2.2 Evaluation of OrthoFiller on *A. thaliana* after removal of 10% of gene annotations 140
4.2.3 Evaluation of OrthoFiller on *A. thaliana* after removal of 90% of gene annotations 144
4.2.4 Evaluation of OrthoFiller on *S. cerevisiae* after removal of 10% of gene annotations 147
4.2.5 Evaluation of OrthoFiller on *S. cerevisiae* after removal of 90% of gene annotations 152
4.2.6 OrthoFiller mainly predicts simple gene models 154
4.2.7 OrthoFiller behaves consistently under addition and removal of species 156
4.2.8 OrthoFiller recovers mostly verified fungal genes 158
4.2.9 OrthoFiller detects hundreds of conserved genes not present in the reference genome annotations 158
4.2.10 Novel genes are supported by RNA-seq data 163
4.3 Methods 168
4.3.1 Acquisition and evaluation of HMM hits 168
4.3.2 Acquisition of putative gene models 168
4.3.3 Evaluation of putative gene models: the hint filter 169
4.3.4 Orthogroup membership evaluation 171
4.3.5 Obtaining simulated data sets 172
4.3.6 Assessment of recovered genes 173
4.4 Chapter discussion 174
4.4.1 OrthoFiller detected hundreds of new genes 175
4.4.2 OrthoFiller filters out spurious gene predictions 175
4.4.3 More careful gene model construction could improve recall . 176
4.4.4 Summary 177

5 OMGene: Mutual improvement of gene models through optimisation of evolutionary conservation 179
5.1 Chapter introduction . . . 179
5.1.1 Background . . . 179
5.1.2 Chapter aims and outline . . . 181
5.2 Results . . . 182
5.2.1 Algorithm overview and evaluation criteria . . . 182
5.2.2 OMGene changed thousands of existing annotations . . . 184
5.2.3 *De novo* annotations are strongly affected by OMGene . . . 188
5.2.4 Complex gene models are improved by OMGene . . . 194
5.2.5 OMGene improves subcellular localisation consistency . . . 197
5.3 Methods . . . 199
5.3.1 Cross-aligning genes to find missing exonic regions . . . 199
5.3.2 Adjacency group calculation . . . 200
5.3.3 A choice function for efficient best subset acquisition . . . 202
5.3.4 Resulting genes are double-checked against the input . . . 207
5.3.5 A final filtering step accounts for over-correction . . . 207
5.3.6 An additional species set for algorithm verification . . . 208
5.3.7 Acquisition and choice of orthogroups . . . 212
5.3.8 Simulated *de novo* data sets . . . 212
5.3.9 RNA-seq data sources . . . 212
5.3.10 Alignment and quantification of RNA-seq data . . . 213
5.3.11 Coverage and splice junction scoring metrics . . . 214
5.4 Chapter discussion . . . 215
5.4.1 OMGene results reflect differences in gene model complexity 215
5.4.2 OMGene works well on complex gene models . . . 216
5.4.3 OMGene improves start site identification . . . 217
5.4.4 Summary . . . 218

6 General discussion 219
6.1 Overview . . . 219
6.2 Some annotations are better than others . . . 220
6.2.1 A novel orthogroup HMM search and filtration system identifies under-annotated genomes . . . 221
6.2.2 Putative missing genes are supported by RNA-seq evidence 223
6.2.3 Extrinsic gene model scores identify poor annotations . . . 224
6.2.4 Alignment discrepancies betray erroneous gene models . . . 226
6.2.5 Non-extrinsic methods reveal hidden errors . . . 227
6.2.6 Many primary transcripts are suboptimal . . . 228

6.2.7 Discrepancy score likely underestimates the problem 229
6.3 OrthoFiller and OMGene improve annotations 229
6.3.1 OrthoFiller detects missing genes 229
6.3.2 OrthoFiller proceeds with caution 230
6.3.3 Strategies for improving OrthoFiller 231
6.3.4 OMGene improves gene models 233
6.3.5 Strategies for improving OMGene 233
6.4 The future 235
6.4.1 An OrthoFiller/OMGene hybrid could increase recall 235
6.4.2 OrthoFiller/OMGene will reach singularities 236
6.4.3 A combined approach could identify non-canonical features . 236
6.4.4 What to do with all the data? 238
6.4.5 The future of sequencing 238
6.5 Overall summary 239

Appendices

A OrthoFiller (published version) 243

B OMGene (published version) 261

References 281

List of figures

1.1 Decreased sequencing costs have driven a huge influx of genome data 5
1.2 Species with larger genomes have only recently become sequenceable 6
1.3 Absent annotations can be mistaken for gene loss 14
1.4 Inaccurate gene models lead to transcript quantification errors . . . 16
1.5 Inaccurate gene models can inhibit domain identification 17
1.6 Inaccurate start codons can prevent localisation prediction 18
1.7 Gene prediction algorithms vary in number of genes predicted . . . 21
1.8 Gene prediction algorithms vary in gene model choice 21
1.9 CEGMA and BUSCO scores can indicate annotation completeness 23

2.1 Plants have larger genomes than fungi 33
2.2 Plants have more non-coding DNA than fungi 39
2.3 Examples of various types of orthogroup 40
2.4 Species trees for plants and fungi 43
2.5 Quantities of SCU-1 orthogroups that omit each species 44
2.6 Hits can be classified by comparison to existing annotations 48
2.7 HMM hit distribution for plant orthogroups 51
2.8 HMM hit distribution for fungal orthogroups 52
2.9 A missing gene from *A. thaliana* 60
2.10 A missing gene from *B. rapa* 60
2.11 A missing gene from *C. papaya* 61
2.12 A missing gene from *C. rubella* 61
2.13 A missing gene from *T. cacao* 62
2.14 A missing gene from *E. gossygpii* 63
2.15 A missing gene from *D. hansenii* 63
2.16 A missing gene from *K. lactis* 64
2.17 A missing gene from *S. cerevisiae* 64
2.18 A missing gene from *Y. lipolytica* 64
2.19 Accepted hits have higher read coverage than rejected hits 66
2.20 Accepted hits have higher coverage than simulated intergenic hits . 69

3.1 Example of an extended exon 82

3.2 Example of a contracted exon 83
3.3 Example of an omitted exon 83
3.4 Example of extraneous exons 84
3.5 Example of a missing intron 85
3.6 Example of an extraneous intron 86
3.7 Example of an erroneous start codon 87
3.8 Example of a split gene 88
3.9 Example of fused genes 90
3.10 Example of a complex gene model error 92
3.11 Assembly errors can cause annotation errors 93
3.12 Illustration of splice junction recall score 96
3.13 Distributions of imperfect splice junction recall scores 99
3.14 Illustration of coverage score 101
3.15 Coverage and splice junction recall scores are correlated 103
3.16 Alignment scores of simulated perfect alignments 107
3.17 Alignment score rapidly decreases for broken alignments 107
3.18 Fungal orthogroups have lower alignment scores 108
3.19 Larger orthogroups have lower alignment scores 109
3.20 Illustration of pools and bridges 111
3.21 Some species have more discrepancies than others (all orthogroups) 112
3.22 Some species have more discrepancies than others (SCU orthogroups) 112
3.23 Mean splice junction recall score is lower for discrepant genes . . . 116
3.24 Discrepant gene models commonly have imperfect splice junction recall 117
3.25 Mean coverage score is lower for discrepant genes 118
3.26 Imperfect coverage scores are common for discrepant genes 119
3.27 Low coverage scores are common for discrepant genes 120
3.28 Localisation inconsistencies suggest poor start site annotation . . . 122
3.29 Most orthogroups have consistent localisation 123
3.30 Mean entropy score is lower for discrepant orthogroups 124
3.31 Inconsistent localisation is more common for discrepant orthogroups 125

4.1 Workflow diagram for the OrthoFiller algorithm 138
4.2 OrthoFiller vs *de novo* recovery statistics, plants, 10% depletion . . 141
4.3 Gene recoveries at each OrthoFiller stage, plants, 10% depletion . . 143
4.4 OrthoFiller vs *de novo* recovery statistics, plants, 90% depletion . . 145
4.5 Gene recoveries at each OrthoFiller stage, plants, 90% depletion . . 148
4.6 OrthoFiller vs *de novo* recovery statistics, fungi, 10% depletion . . 150
4.7 Gene recoveries at each OrthoFiller stage, fungi, 10% depletion . . 151
4.8 OrthoFiller vs *de novo* recovery statistics, fungi, 90% depletion . . 153
4.9 Gene recoveries at each OrthoFiller stage, fungi, 90% depletion . . 155

4.10 Phylogenetic trees for additional species sets 157
4.11 OrthoFiller behaves consistently for expanded plant species set . . . 159
4.12 OrthoFiller behaves consistently for expanded fungal species set . . 160
4.13 A missing gene recovered in *A. thaliana* 165
4.14 A missing gene recovered in *B. rapa* 165
4.15 A missing gene recovered in *C. papaya* 165
4.16 A missing gene recovered in *C. rubella* 166
4.17 A missing gene recovered in *T. cacao* 166
4.18 A missing gene recovered in *K. lactis* 166
4.19 A missing gene recovered in *S. cerevisiae* 167
4.20 A missing gene recovered in *Y. lipolytica* 167
4.21 Hint F-score distributions for hits to plant genes 170
4.22 Hint F-score distributions for hits to fungal genes 171

5.1 Simplified OMGene workflow 183
5.2 Number of changes made by OMGene for each species 189
5.3 Average magnitudes of changes made by OMGene 190
5.4 Distribution of types of changes made in the three data sets 191
5.5 Examples of individual gene model changes for genes in *A. thaliana* 192
5.6 Assembly issues can prevent accurate gene model formation 195
5.7 Calculation of adjacency groups 201
5.8 Illustration of multipartite choice function 204
5.9 Mammals have large genomes 210
5.10 Most mammal DNA is non-coding 211

List of abbreviations and symbols

aa . . . . . amino acid

bp . . . . . base pair

$\boldsymbol{C}$ . . . . . coverage score

CDS . . . . . coding sequence

$\boldsymbol{\bar{d}}$ . . . . . mean read depth

DG . . . . . discrepant gene

DNA . . . . . deoxyribonucleic acid

DOG . . . . . discrepant orthogroup

EST . . . . . expressed sequence tag

Gbp . . . . . giga base pair

GFF2 . . . . . General Feature Format 2

GFF3 . . . . . General Feature Format 3

GTF . . . . . Generalised Transfer Format

HMM . . . . . hidden Markov model

$\boldsymbol{hF}$ . . . . . hint F-score

$\boldsymbol{hP}$ . . . . . hint precision score

$\boldsymbol{hR}$ . . . . . hint recall score

iDG . . . . . initial discrepant gene

iDOG . . . . . initial discrepant orthogroup

JGI . . . . . Joint Genome Institute

$\boldsymbol{jF}$ . . . . . junction F-score

$\boldsymbol{jP}$ . . . . . splice junction precision score

$\boldsymbol{jR}$ . . . . . splice junction recall score

kbp . . . . . kilo base pair

LCA . . . . . last common ancestor

lincRNA . . . long intergenic non-coding RNA

Mbp . . . . . mega base pair

miRNA . . . . micro-RNA

mRNA . . . . messenger RNA

ncRNA . . . . non-coding RNA

NCBI . . . . . National Center for Biotechnology Information

niDG . . . . . non-initial discrepant gene

niDOG . . . . non-initial discrepant orthogroup

OG . . . . . orthogroup

OMGene . . . Optimise My Gene

oF . . . . . orthogroup F-score

oP . . . . . orthogroup precision score

oR . . . . . orthogroup recall score

ORF . . . . . open reading frame

pF . . . . . protein F-score

pP . . . . . protein precision score

pR . . . . . protein recall score

RNA . . . . . ribonucleic acid

SCU . . . . . single copy, ubiquitous

SCU-1 . . . . . single copy, ubiquitous except for one species

s.d. . . . . . standard deviation

SGD . . . . . Saccharomyces Genome Database

TAIR . . . . . The Arabidopsis Information Resource

UTR . . . . . untranslated region

WGS . . . . . whole genome sequencing

1 General introduction

1.1 Introduction

Creating an accurate reference genome assembly for a given species has become a relatively standard procedure. Whole genome sequencing projects for complex organisms, once the preserve of large consortia of experts, are now within the capabilities of individual laboratories. As such, the speed at which novel genomes are published has increased rapidly. For example, on NCBI GenBank (Benson et al. 2013), both the quantity (in terms of bases sequenced) of whole genome sequencing data and the number of species represented have continued to increase roughly exponentially with each release since 2002 (GenBank 2017; NCBI 2018a).

A genome sequence is of little practical use in the absence of an *annotation*. In the general sense, a genome annotation contains coordinates for specific features encoded in the genome sequence, such as genes, pseudogenes, and non-coding RNA such as ribosomal RNA, transfer RNAs, microRNAs, and long intergenic non-coding RNAs. In the majority of this thesis, the term *genome annotation* will be taken to mean the set of annotated *protein-coding genes* only, and any non-coding annotations mentioned will be specifically referred to as such. Knowledge of the precise locations and structures of protein-coding genes is vital for carrying

out any large-scale biological investigation into the properties of an organism. In many cases, the genome assembly and its annotation are taken at face value by those who utilise these resources, and thus any errors caused by absent or incorrect gene annotations can be propagated, both between genome projects and through layering of false inferences from flawed analyses of a single species (Salzberg 2007; Schnoes et al. 2009). Given the ever-increasing quantity of genomic data available, it is of critical importance that measures be taken to ensure the accompanying annotations are as complete and accurate as possible.

The best-studied model genomes have been extensively and painstakingly manually curated over many years by large, dispersed teams of experts. In particular, most model organisms will have had many of their protein-coding genes characterised *before* their genome sequences became available. Thus, the wealth of pre-existing knowledge that could be transferred to these genomes will have been vastly superior to that of a newly sequenced non-model organism. Newly sequenced genomes today are rarely afforded the luxury of such manual curation, and therefore must rely on advanced comparative and predictive automated genome annotation techniques. These techniques range from purely computational *ab initio* methods such as those implemented by Augustus (Stanke and Morgenstern 2005) or GENSCAN (Burge 1997) to fully automated prediction pipelines such as the NCBI Eukaryotic Genome Annotation Pipeline (Thibaud-Nissen et al. 2013), which typically incorporate various sources of homology and experimental evidence to improve the accuracy of their predictions. Each approach has its own merits and flaws, and all computationally driven genome annotation methods are prone to some degree of error (Mudge and Harrow 2016). Furthermore, due to the nature of second-generation sequencing technologies, many gene predictions are performed on fragmented, incomplete genome assemblies, compounding these errors (Denton et al. 2014). All newly sequenced genomes are subjected to some level of computational prediction, and therefore annotation errors are likely to be common.

"Incorrect and incomplete annotations poison every experiment that makes use of them" (Yandell and Ence 2012). Gene annotation errors are bad news for biological analyses. Errors fall into two main categories. Firstly, the gene finding problem: identifying the precise set of protein-coding genes in a genome sequence, and avoiding spurious annotations. Genome annotation algorithms must strike a balance between caution and quantity: in filtering out spurious gene predictions, many genuine protein-coding genes will inevitably fail to be annotated. Absent annotations for biologically present genes can lead to false inferences in the biochemical properties of an organism, as well as inaccurate assertions regarding its evolutionary history (Zallot et al. 2016; Casola and Koralewski 2017; Han et al. 2013). Secondly, even when a gene is present in an annotation, predictions of the precise sequence of introns and coding exons – the *gene models* – can vary wildly in terms of their accuracy (Nasiri et al. 2013; Schnoes et al. 2009), especially for higher eukaryotic genomes with complex multi-exon gene models. Inaccurate gene models can have serious downstream effects including transcript quantification errors and erroneous protein function prediction (Triant and Pearson 2015; Robert and Watson 2015; Promponas et al. 2015).

While several methods for assessing *genome assembly* quality have been developed (Rahman and Pachter 2013; Gurevich et al. 2013), less attention has been paid to assessing and improving the completeness and accuracy of *annotations*, and the quality of a genome assembly and its annotation are not necessarily correlated. Similarly, there are currently no tools available to smaller-scale laboratories for the automated improvement of genome annotations. Given the vital importance of ensuring genome annotations are of as high a quality as possible, methods both to assess the correctness of genome annotations and to systematically improve them are both necessary and underdeveloped, and are thus in high demand. Existing efforts to assess the quality of genome annotations have appealed either to highly conserved gene or protein domain sets (Eilbeck et al. 2005; Veeckman et al. 2016;

Holt and Yandell 2011) or assessed genes based on the presence of EST, RNA-seq or protein evidence (Eilbeck et al. 2009; Eilbeck et al. 2005; Zhao et al. 2013). Given that universally conserved genes and domains are likely to have been used in the initial curation of a gene set, any assessment of genome annotation completeness or accuracy that uses them is inherently circular, and likely to underestimate the extent of the problem. Similarly, highly expressed genes from which reliable RNA-seq reads can be obtained are likely to have already been used to inform gene model procurement in some way. For these reasons, it is desirable to seek methods for annotation assessment and improvement that do not rely on these resources. One potential solution that has inspired much of the work in this thesis is to restrict analyses to small, close subsets of the phylogenetic tree, and to leverage multiple local sources of homology to assess accuracy of gene annotations without relying on experimental data. These *phylogenetically local*, *experiment-free* assessment and reannotation approaches can help minimise the effects of some of the problems outlined above.

In this thesis I tackle the dual problems of genome annotation assessment and improvement at phylogenetically local scales, in the absence of extrinsic experimental data. Presented over four results chapters, I develop approaches for assessing the quality of genome annotations under these constraints, and present two methods designed to improve the annotations with the same restrictions. In this introduction to the thesis, I will discuss the current state of affairs regarding approaches to genome annotation. I will then proceed to discuss some of the effects that inaccurate annotations can have on downstream analyses, and provide an overview of recent assessments of genome annotation completeness and accuracy. The introduction finishes with a brief outline of the aims and structure of the thesis as a whole.

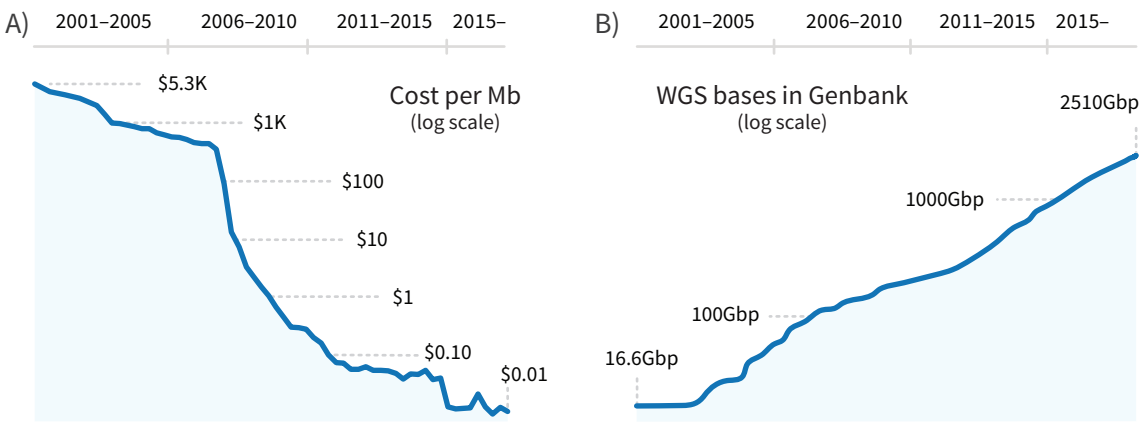


Figure 1.1: Decreased sequencing costs have driven a huge influx of genome data. A) Sequencing costs have fallen dramatically since 2001; B) This has precipitated a huge increase in genomic material available on sequence databases such as GenBank. Data: (GenBank 2017; Wetterstrand 2016)

1.2 Draft genome assemblies are quick to produce

The costs and effort involved in producing a whole genome assembly have fallen dramatically since the completion of the first human genome project (Figure 1.1A). As such, the amount of genome sequencing data available has increased considerably (Figure 1.1B). With every advancement of sequencing technologies, another class of organisms becomes accessible (Figure 1.2): in particular, the sequencing of larger genomes has only recently become practical. The first genome sequence ever determined was that of Bacteriophage MS2, a viral genome comprising 3,569 bases (Fiers et al. 1976). The 2001 release of the human genome, at 3 giga base pairs (Gbp), marked the sequencing of the first gigabase-scale genome (Lander et al. 2001). Now, obtaining whole genome sequencing (WGS) samples of human cells is a standard research procedure, and draft assemblies of multi-gigabase novel genomes for new species can be produced in a matter of weeks.

1.2.1 High-throughput methods yield fragmented assemblies

The decreasing cost of whole genome sequencing owes in part to the emergence of second generation sequencing techniques. For the acquisition of genome sequences,

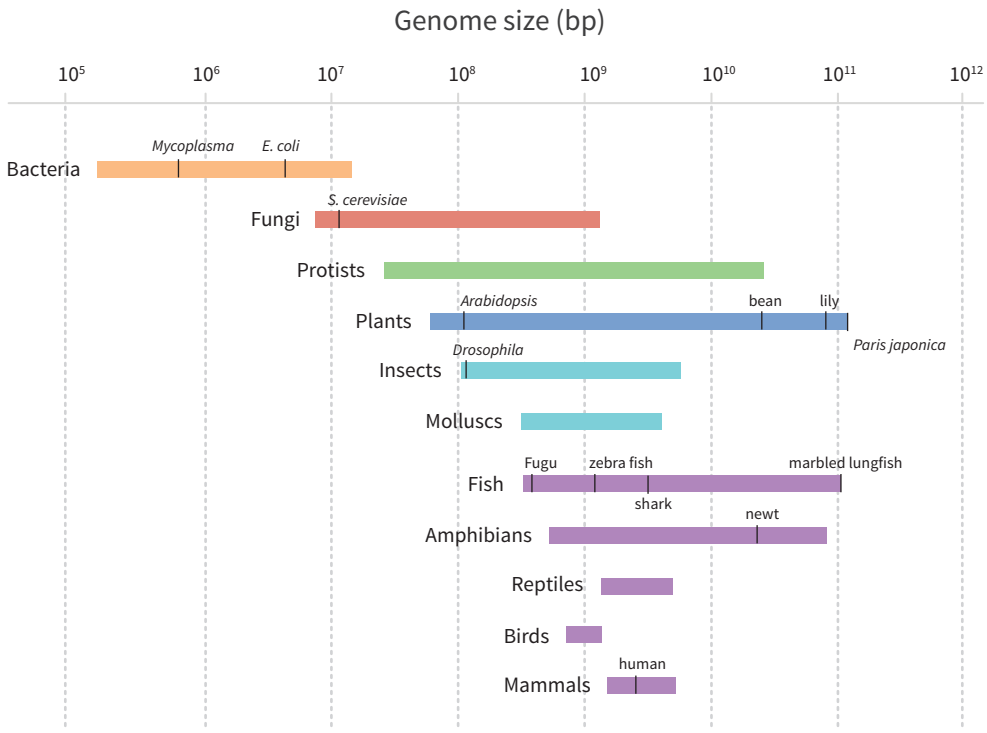


Figure 1.2: Species with larger genomes have only recently become sequenceable. Up until recently, it was infeasible to obtain WGS data for species with larger geonmes, such as mammals, fish, amphibians, and some plants. Source: (Milo and Phillips 2018)

traditional Sanger sequencing methods have been largely superseded by short-read, high-throughput approaches such as the near-ubiquitous Illumina platform (Bentley et al. 2008). These approaches are cheaper, faster, and much less labour-intensive than Sanger sequencing, however they are also less accurate (Schatz et al. 2010). In particular, assembly contiguity and quality are often sacrificed in favour of speed and cost, and the resulting genome assemblies are typically highly fragmented, comprising many disconnected contigs and scaffolds. This is mainly due to the short-read nature of high-throughput technologies, which make repetitive genomic regions of low complexity difficult to resolve. Larger genomes often contain high proportions of highly repetitive regions, compounding this problem. In addition, large, highly polyploid genomes are difficult to represent in a linear format – this is especially a problem for plant genomes (Tang 2017). As a result, much of the

publicly available genome data today consists of works in progress – fragmented genome assemblies of varying quality.

More recent "third generation" sequencing technologies such as those provided by PacBio (Rhoads and Au 2015), BioNano (Cao et al. 2014), 10X Genomics (Kitzman 2016) and Oxford Nanopore (Bentley et al. 2008) can mitigate the problem of assembly discontiguity by either obtaining long reads spanning many kilo bases (kb), or by providing localisation information to link collections of short reads or contigs together. Currently these technologies are not sufficiently accurate to facilitate full *de novo* genome assembly, but can be hybridised with second-generation techniques to improve contiguity (English et al. 2012; Thomma et al. 2016). However, these techniques are not used as standard, and most assemblies today are still produced using purely second-generation technology (Greenleaf and Sidow 2014).

Poor-quality assemblies produce poor-quality annotations: analyses conducted on fragmented assemblies are usually less accurate than those carried out on genome sequences of a more contiguous nature (Alkan et al. 2011; Thomma et al. 2016; Denton et al. 2014). Genome annotations can be improved markedly by even small assembly improvements (Elsik et al. 2007). Often, however, the annotations for a new assembly are simply transferred from the old one using a tool such as the Rapid Annotation Transfer Tool (RATT – Otto et al. 2011), and thus annotation errors can be propagated between assembly versions. Thus, it is of critical importance that an assembly be finished to as high a degree as possible before embarking on an annotation project.

1.2.2 Automated annotation pipelines require little human input

Approaches to genome annotation fall into three principle categories. The first category, *ab initio* or *de novo* prediction refers to the identification and model prediction of genes through purely computational methods, using no information

other than local sequence context. Many tools have been developed to serve this purpose, some of which are listed in Table 1.1. These tools typically use probabilistic frameworks such as HMMs (Augustus, FGENESH, GENSCAN), machine learning / conditional random fields (mGene, CRAIG, Conrad) or dynamic programming (GAZE), typically by analysing genomic patterns such as GC content, open reading frames, and poly-adenylation addition sites. Some of these gene prediction tools, or *gene finders*, can leverage multiple information sources, including parallel *de novo* prediction for multiple species (SLAM, SGP-1, TWINSCAN/ CONTRAST, and the most recent version of Augustus), and incorporation of extrinsic data such as protein and EST hints.

Ab initio gene finders all require training of some kind, be it from existing data sources or by providing small sets of already annotated genes on which to base subsequent searches. Any training set inherently biases the results of a gene finder in favour of what is already known about a particular genome. For example, different regions of a single genome may have different properties such as varying GC content and increased incidence of repetitive regions: gene finders trained on universal characteristics of the genome may fall down in such cases. Gene finders are typically highly proficient at locating and modelling genes, but none of them is perfect, and none can be relied on completely to reliably return the complete accurate set of protein-coding genes contained within a genome assembly (Mudge and Harrow 2016).

Ab initio gene prediction programs are rarely used in isolation for serious genome projects, especially for higher eukaryotes with complex gene models and large non-coding regions, for which gene prediction is notoriously difficult (Wang et al. 2004). The second approach to genome annotation is to run multiple gene predictors and combine the results to obtain the best consensus gene set. Such approaches allow evidence from many different sources to be considered, improving the likelihood that a given prediction is correct. Several tools exist to combine the results of

multiple gene finders, for example JIGSAW (Allen and Salzberg 2005), which was used on the rice genome, EvidenceModeler (Haas et al. 2008), GLEAN (Elsik et al. 2007), and EVIGAN (Liu et al. 2008), all of which have been shown to improve the quality of predicted genes compared with isolated gene finders.

For the larger, more complex genomes being published today, a common avenue for accurate genome annotation is to submit a genome to a fully automated genome annotation pipeline, such as PASA (Florea et al. 1998), MAKER (Cantarel et al. 2008; Holt and Yandell 2011), or the NCBI Eukaryotic Genome Annotation Pipeline (Thibaud-Nissen et al. 2013) (See Table 1.2). These pipelines accept many different sources of homology and experimental evidence, and automatically compare genome sequences to large databases of existing annotations, domains, and protein sequences to maximise the likelihood of gene set and gene model accuracy. Homology evidence can be provided by EST and domain mapping, or gene projection from closely related organisms. Experimental data can come in the form of RNA sequencing (RNA-seq), mass spectrometry, or direct protein sequencing. Post-processing steps to determine UTR locations and transcript variants may also be incorporated. These pipelines are resource-intensive and their setup may require careful tailoring, however they are usually considered the most accurate approach to genome annotations (Yandell and Ence 2012), barring additional manual curation steps (for example manual comparison to RNA-seq or protein data using a gene model editor such as Artemis (Rutherford et al. 2000) or Apollo (Lewis et al. 2002)).

Manual curation is usually only carried out genome-wide for the best-studied genomes, such as *M. musculus* (mouse) and *H. sapiens* (human), and small genomes such as viruses. The typical approach is to compare directly the annotations of individual loci to mapped RNA-seq reads, EST, or protein evidence. Manually curated gene annotations are the "true" models to which automated gene finders are usually compared. For mouse and human, concentrated annotation optimisation efforts are in constant progress: for example the Consensus Coding Sequence

Software name	Target group	Description	References
Augustus	Eukaryotes	Uses generalised hidden Markov models. Generally considered to be accurate. Can be trained to specific data sets.	(Stanke and Morgenstern 2005)
Conrad	General	Conditional random field model with a semi-Markov structure	(Decaprio et al. 2007)
CONTRAST/TWINSCAN	General	Alignment of multiple informant genomes for parallel *de novo* prediction.	(Gross et al. 2007)
CRAIG	Eukaryotes	Conditional random field model with a semi-Markov structure	(Bernal et al. 2007)
EUGENE	General	Modular gene prediction package utilising multiple sources of information	(Foissac et al. 2008)
FGENESH	Eukaryotes	Fast and accurate HMM-based gene prediction tool.	(Salamov et al. 2000)
GAZE	General	Dynamic programming approach. Highly configurable.	(Howe et al. 2002)
Geneid	Eukaryotes	Uses position-weight arrays. Old.	(Parra et al. 2000)
Genemark	General	Self-training HMM-based gene finder.	(Borodovsky and Mcininch 1993)
GeneSeqer	Plants	Based on spliced alignment of ESTs.	(Usuka et al. 2000)
GENSCAN	General	Generalised HMM-based gene predictor with web server.	(Yeh et al. 2001)
GLIMMERHMM	General	Uses generalised hidden Markov models and decision tree-based model optimisation.	(Majoros et al. 2004)
Gnomon	Eukayotes	Based on GENSCAN. Uses HMMS and ESTs/protein evidence	(Souvorov 2018)
mGene	Eukaryotes	Machine-learning based gene finder.	(Schweikert et al. 2009)
SGP-1	Eukaryotes	Parallel *ab initio* prediction	(Wiehe and Gebauer-Jung 2001)
SLAM	General	Parallel *ab initio* prediction	(Alexandersson et al. 2003)
SNAP	General	Self-training semi-HMM based gene finder allowing EST and mRNA evidence incorporation.	(Korf 2004)

Table 1.1: List of selected *ab initio* gene prediction tools.

(CCDS) project is a curated consensus gene list that aims to provide extremely high confidence, "canonical" gene annotations (Pruitt et al. 2009). The CCDS regularly updates its gene set: since 2005 there have been 8 *M. musculus* CCDS releases and 13 *H. sapiens* CCDS releases (CCDS 2018). Genome annotation improvement by manual curation is out of reach of most genome projects, and so most of them rely heavily on the computational methods described above. There is still no one unified standard approach to genome annotation in use today, and thus publicly available genome annotations can vary considerably in quality.

1.2.3 Published annotations are often poorly maintained

Many genomes are annotated once and left static for many years (Salzberg 2007). In contrast, and as noted above, well-studied genomes such as *Homo sapiens* and *Mus musculus* are frequently re-annotated, and their revised annotations are made public several times a year. In the most recent CCDS update in 2017, annotations for 1,229 mouse genes were changed, added or removed, and 1,781 human gene annotations were changed, added, or removed (based on changed gene entries in the CCDS current data sets – CCDS 2018). Considering that these genome projects have essentially been in progress for 30 years, with thousands of contributors, one can imagine that many of the newer genome projects have a long road ahead.

Re-annotation projects are not undertaken systematically for most species. In part, this is because of the difficulty of updating genome data on public databases – the authorisation to update a genome annotation tends to lie with the authors of the original submission (NCBI 2018b), and those submitting genomes often simply move on after the first submission. When database updates are possible, there are a number of existing options for improving genome annotations. Manual reannotation is the obvious choice, either through trickles of individual informant experiments over many years, or through concentrated manual curation projects, such as annotation "jamborees" (Elsik et al. 2006). These events involve gathering interested people

Software name	Target group	Description	References
PASA	Eukaryotes	Uses spliced-aligned RNA-seq data to identify genes and improve gene models.	(Florea et al. 1998; Haas et al. 2011)
MAKER	General	Designed for small-scale genome projects. Combines multiple information sources.	(Cantarel et al. 2008; Holt and Yandell 2011)
NCBI	Eukaryotes	Automated Eukaryotic gene prediction pipeline. Combines Gnomon predictions and various compiled information such as RNA-seq and EST libraries.	(Thibaud-Nissen et al. 2013)
Ensembl	Chordates	Predicts protein-coding and non-coding genes by leveraging experimental evidence (RNA-seq, ESET libraries), and combines predicted and curated evidence from multiple public databases (RefSeq, EMBL, UniProtKB).	(Curwen et al. 2004; Aken et al. 2016)

Table 1.2: List of selected automated genome annotation pipelines.

together, either physically or virtually, to mass-reannotate sequences based on extrinsic data. Naturally, these are impractical to carry out for most genomes. Alternatively, newer versions of genome annotation pipelines utilising updated and more specialised experimental and homology data sets can be re-run, particularly upon the publication of an improved genome assembly. This is the approach employed by NCBI (NCBI 2018a) on selected data sets in their RefSeq and GenBank annotation libraries (Benson et al. 2013; Pruitt et al. 2007). The process for which species are selected for re-annotation is not systematic, and reannotation is usually initiated by the laboratories who first submitted the data. Standardised, regular reannotation of genomes will be the key to ensuring that genome annotations are reliable and well maintained, and the framework to do so is currently lacking.

1.3 Annotation errors lead to false inferences

The complete and accurate determination of an organism's protein-coding genes is of critical importance to any study of that species' biological properties. Moreover, genome annotations are typically taken at face value by those using them, and as such any errors present can be propagated. These errors can cause all manner of implications for downstream analyses, and as such should be considered seriously and accounted for in any *in silico* experiment that makes use of them. In this section I provide examples of such errors, indicating some of the negative effects that they can have on individual and cross-species analyses.

1.3.1 Absent annotations can be mistaken for gene loss

Inaccurate inferences about the biological and metabolic properties of an organism can be drawn from unannotated but biologically present genes (Zallot et al. 2016; Han et al. 2013; Casola and Koralewski 2017). Such genes are readily confused with gene loss events: cases where a gene has disappeared from an organism's genome sequence due to functional redundancy. The apparent presence or absence of a

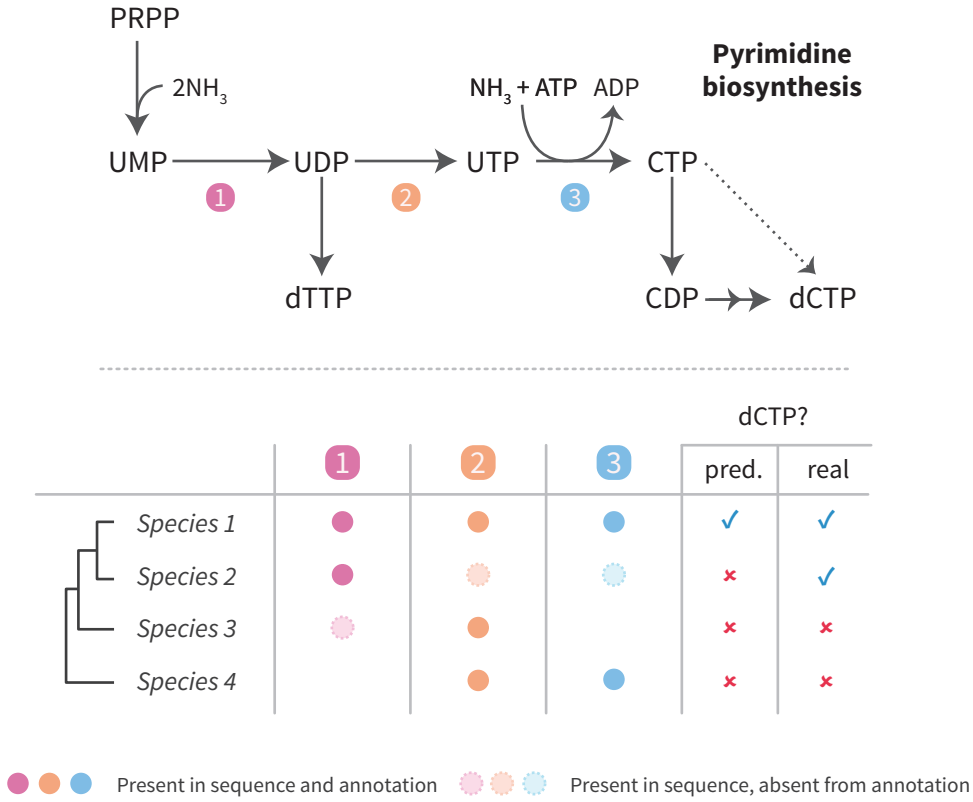


Figure 1.3: Absent annotations can be mistaken for gene loss. The biochemical pathway shown is a subset of the pyrimidine biosynthesis pathway. Enzymes 1, 2, and 3 (uridine-cytidine kinase 2, nucleoside diphosphate kinase, CTP synthase) are all required to convert PRPP into dCTP, a necessary cellular product. Absence of any one of these would prohibit the organism from producing dCTP. False annotation of absence of any of these can lead to inaccurate assertions, for example, analysis of the annotated proteomes here would suggest that Species 2 cannot synthesise dCTP, contrary to the biological truth.

given gene can have a considerable impact on the analysis of biochemical pathways present in a cell. To illustrate this, Figure 1.3 shows an example of a case where false inferences can be drawn from missing gene annotations. The biochemical pathway shown is a subset of a pyrimidine biosynthesis pathway. Enzymes 1, 2, and 3 (respectively uridine-cytidine kinase 2, nucleoside diphosphate kinase, CTP synthase) are all required to convert phosphoribosyl pyrophosphate (PRPP) into deoxycytidine triphosphate (dCTP), which is necessary for the synthesis of cytosine. Absence of any one of these would prohibit the organism from producing

dCTP, and would mean that the organism must scavenge it from the environment. False annotation absence of any of these enzymes would suggest this to be the case, even if it were not biologically true. In the example, Species 1 and 2 are both biologically able to produce dCTP, but Species 2 lacks annotations for both nucleoside diphosphate kinase and CTP synthase. Thus any biochemical analysis of this species using this annotation would inaccurately assume that it is unable to make dCTP. In contrast, the missing gene in Species 3 does not affect this particular inference (though would suggest inability to produce UDP), due to a genuinely missing gene, and no false inferences would be drawn for Species 4.

1.3.2 Inaccurate gene models lead to transcript quantification errors

The presence of a gene in an annotation is not enough. For any given gene locus, there can be many possible viable *gene models* for that gene, i.e. combinations of start codon choice and exon start and end points. Although translationally valid, the vast majority of viable gene models will not actually ever be produced by the organism in question, and such gene models, if annotated, are considered to be erroneous. This is particularly true for higher eukaryotes such as plants and animals. Accurate gene models are important for a number of reasons. For example, popular transcript quantification methods such as Salmon (Patro et al. 2015; Bray et al. 2016) and Kallisto (Bray et al. 2016) typically map RNA-seq reads to predicted representative RNA transcripts. Inaccurate gene models generate inaccurate representative transcripts. Even when multiple alternative transcripts are biologically real, the choice of gene model for a given gene can have a considerable impact on expression analysis (Conesa et al. 2016; Zhao and Zhang 2015). In particular, inaccurate gene models can prohibit mapping for large proportions of reads, causing transcript quantities to be vastly underestimated (Robert and Watson 2015), as illustrated in Figure 1.4. Thus, inaccurate gene models can cause

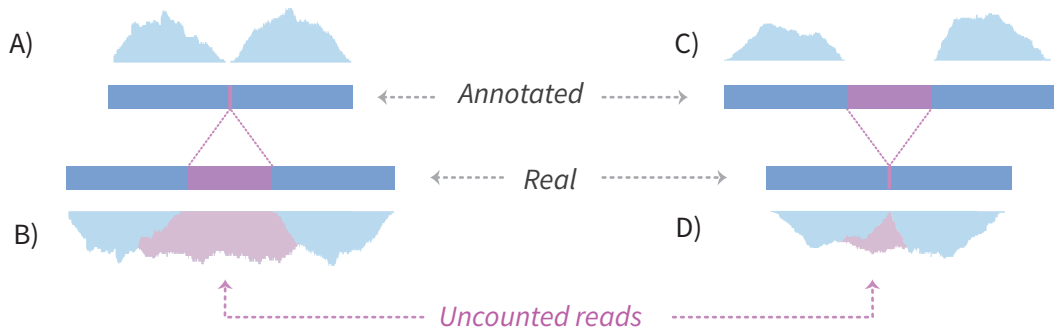


Figure 1.4: Inaccurate gene models lead to transcript quantification errors. A) An incorrectly predicted gene model that yields a transcript that is too short (e.g. missing an exon): reads cannot align to the region where the error occurs. B) The transcript generated from the true gene model, recovering large quantities of read data; C) An incorrectly predicted gene model that yields a transcript that is too long (e.g. containing intronic sequence that should have been removed) : reads are missed from the point of error. D) The transcript generated from the true gene model, recovering large quantities of read data. In both cases, read density highlighted in purple should have been counted but was discarded because of gene model error.

inaccurate transcript quantity calculations, affecting any experiment that relies on precise measurements of relative protein abundance.

1.3.3 Inaccurate gene models inhibit domain identification

Many aspects of protein function can be inferred through the identification of conserved domains (Mills et al. 2015; Marchler-Bauer et al. 2011). Putative protein domains for a given gene are compared to large databases of previously characterised domains from other species. Inaccurate gene models can cause domains that are biologically present to be missed due to incomplete predicted transcripts. For example, many domains in publicly available gene annotations appear to be partial or truncated when compared to domains in the Pfam database (Finn et al. 2014): the majority of these partial domains are simply errors due to inaccurate underlying gene models (Triant and Pearson 2015). Furthermore, domains can be missed completely: an example of this is shown in Figure 1.5. Figure 1.5A shows the biologically true gene model, consisting of four domains in the translated protein. Figure 1.5B shows an inaccurately predicted gene model for which an exon is

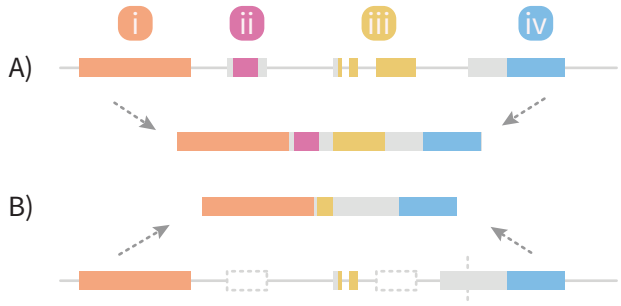


Figure 1.5: Inaccurate gene models can inhibit domain identification. A) The true gene model has four domains, marked i-iv; B) an inaccurate gene model annotation causes domain ii to be omitted, and domain iii to be truncated.

missing, consequently omitting domain ii from the translated protein. Domain iii has also been truncated in this example. Finally, it should be noted that partial domains present in inaccurate gene models can compound the problem: databases such as Pfam rely on gene annotations to obtain gene information. Conversely, domain information is often used to aid gene annotation, and if the domain information is wrong then a vicious cycle ensues. Thus, the accurate annotation of gene models is of critical importance to inferring their biological function and minimising the spread of inaccurate information.

1.3.4 Subcellular localisation predictions require accurate translation start sites

Many proteins are localised to specific organelles or cellular compartments through the presence of target peptides at the N-terminus of their amino acid sequences (Parton et al. 2014; Emanuelsson et al. 2000; Schneider and Fechner 2004; Dönnes and Höglund 2004). The canonical start motif for a gene is a methionine codon (ATG), the choice of which is not always clear: multiple methionine codons can exist in any given open reading frame (ORF), and therefore the precise locations of start codons are often ambiguous. Bioinformatic tools that utilise the beginnings of *in silico* translated genes can be tripped up by inaccurate start codons, and thus false inferences can be made about the cellular location characteristics of

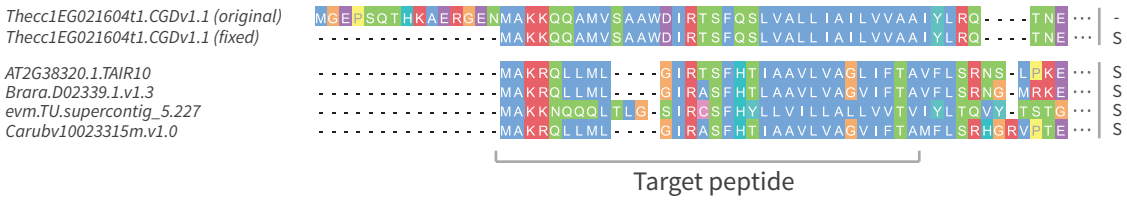


Figure 1.6: Inaccurate start codons can prevent localisation prediction. Targeting peptides at the start of genes can be overlooked due to inaccurate start codons, such as for the *T. cacao* gene *Thecc1EG021604.t1.CGDv1.1* here. An amended version of this gene restores its targeting prediction.

the protein in question.

As an example, Figure 1.6 shows the start of the *T. cacao* (cocoa tree) gene *Thecc1EG021604.t1.CGDv1.1*, in a multiple sequence alignment alongside homologs from other species. The target peptide for the genes is marked, and is 15 amino acids downstream of the annotated start for the *T. cacao* gene. The predicted subcellular localisation for all the other homologs is the secretory pathway, obtained using TargetP (Emanuelsson et al. 2000); the prediction for *Thecc1EG021604.t1.CGDv1.1* is nuclear. If the *T. cacao* gene is amended such that the evolutionarily conserved start codon is used, then the predicted localisation is correct. Thus, any experiment relying on the accurate prediction of subcellular localisation is highly dependent on accurately annotated start codons.

1.3.5 False inferences from inaccurate annotations are propagated

Often, information from one or more already annotated genomes is used to "project" gene information onto a new genome. This can occur at two levels. Firstly, basic annotation gene projection involves utilising gene model and gene location information from one species to find genes in another, closely related species, for example with the Rapid Annotation Transfer Tool (RATT – Otto et al. 2011), or more commonly using BLAST (Altschul et al. 1990) or Exonerate (Slater and Birney 2005). Secondly, functional annotation projection involves transferring the

knowledge about the function or context of a particular gene onto its homologs (Brenner 1999; Mudge and Harrow 2016). Inaccurate annotations can propagate errors in both cases – inaccurate genes projected between species perpetuate any gene model errors present, and may keep alive spurious annotations; similarly inaccurate information relating to domain structure and protein function can be propagated to other species (Bork and Koonin 1998; Green and Karp 2005; Devos and Valencia 2001). Once the errors are free to roam across the landscape of genome annotations, it is very difficult to find their source, moreover they are very difficult to hunt down and fix (Promponas et al. 2015). Thus, gene annotation errors are a communicable disease for genome projects, spreading throughout bio-databases with reckless abandon.

1.4 Assessments of annotation quality are scarce

Considerable work has been done to compare the proficiency of various gene finding algorithms and pipelines. These typically work by comparing the output of the various tools to some static high-quality annotation such as that of a model organism. It is much more difficult, however, to assess the completeness and accuracy of a genome annotation in the absence of a fixed reference. Objective analyses of the accuracy and completeness of published or theoretical genome annotations are few and far between.

The scope for annotation errors can be appreciated by considering the inherent differences between gene finders. I begin this section by citing examples of comparisons between different gene prediction programs. Usually one or more of these algorithms will have been used to inform gene predictions for a species, and the differences between them demonstrate that it is difficult to know which one to trust. This would suggest that errors are inevitable. I then outline a few recent approaches to assessing the quantity and quality of annotated genes.

1.4.1 No gene finder is 100% accurate

Gene predictors vary considerably in the frugality of their predictions. For highly fragmented genomes especially, gene predictions often overestimate the quantity of genes present in a genome sequence (Denton et al. 2014; Guigó et al. 2000). In such cases, genes are often split over the ends of a contig, causing many partial genes to be predicted. In accounting for over-prediction, many genuine genes will be lost, and for less fragmented genomes, preference for precision over recall can mean many genes are simply overlooked. Figure 1.7 shows the quantities of genes predicted by a selection of *ab initio* predictors on simulated fragmented *Drosophila melanogaster* genomes. The quantities of genes predicted in this example vary wildly, both between the methods and across the different simulated assemblies. It is difficult to tell from these data which contributions gene-splitting errors and false positive/negative gene predictions have made to the predicted gene count differences for each method, although gene-splitting does, unsurprisingly, appear to increase for more fragmented scaffolds. However, given the clear differences between the results for different gene finders, it can be expected that each method will produce both some false negative and false positive results. Any genome annotation created using one of these tools can therefore not *a priori* be assumed to be complete.

In a similar vein, gene predictors can differ considerably in the *models* that they assign to genes. Figure 1.8 shows a result from a study of bread wheat genes (Nasiri et al. 2013) comparing the results from multiple different *ab initio* gene prediction tools. In this example, none of the gene models agree. Furthermore, none of them predicts the annotated (and presumed correct) gene model, although several get close. Such cases exemplify the difficulty in trusting gene finders. Given that in many genome annotation projects one or more of these gene finders will have been used, one can infer from this that errors and inconsistencies are likely to be common.

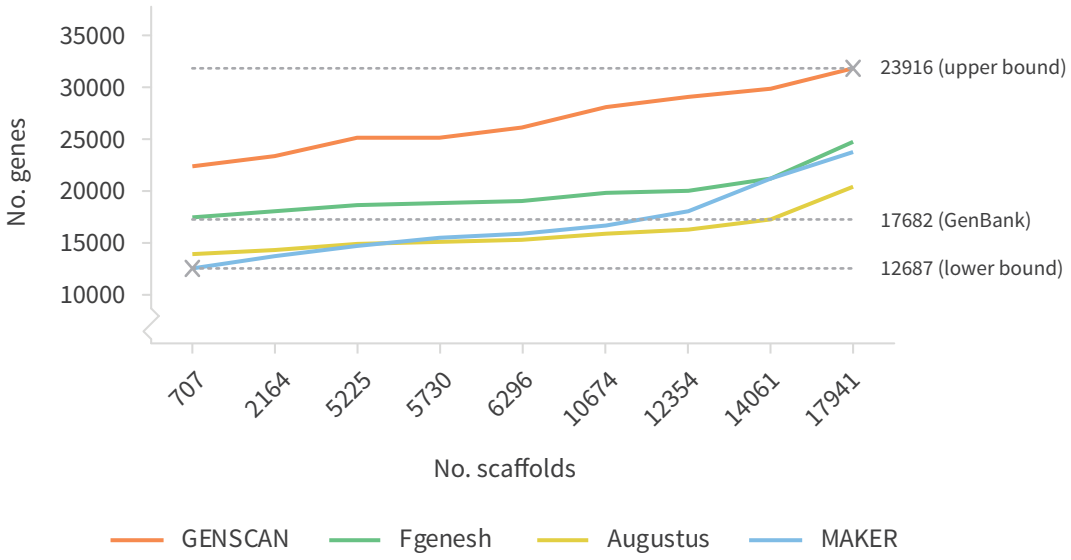


Figure 1.7: Different gene prediction algorithms vary considerably in the quantity of genes that they predict. In particular, the number of genes predicted by all algorithms is greater for fragmented assemblies (Denton et al. 2014).

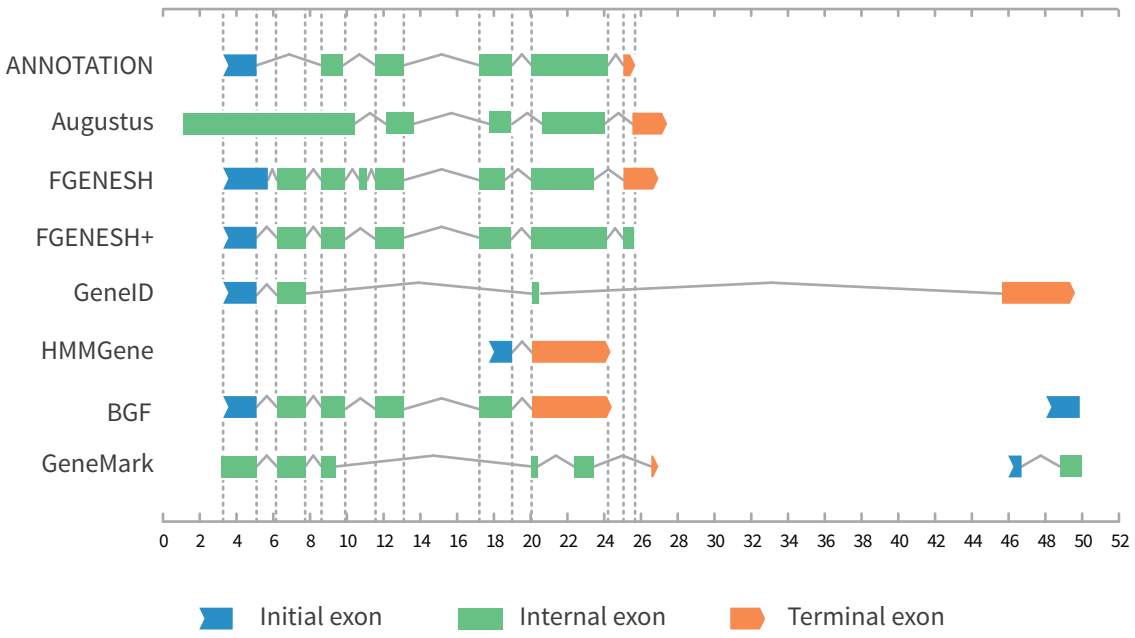


Figure 1.8: Different gene prediction algorithms vary considerably in the gene models that they predict for a given gene. This example (Nasiri et al. 2013) shows a number of predictions for a wheat gene, none of which match the annotated gene model.

1.4.2 Existing assessments of annotation quality rely on universally conserved features or extrinsic data

Direct assessments of the scope of missing gene type errors in published genomes are scarce. One recent study (Veeckman et al. 2016) compared genome sequences for 12 established or recently published plant genomes, assessing them in terms of their completeness relative to several conserved gene families: CEGMA (Parra et al. 2007), BUSCO (Simão et al. 2015), and core gene families (coreGFs) from the PLAZA plant database (see Figure 1.9), as well as mapped ESTs from *Arabidopsis* and rice. The study found that many of the species were clearly lacking in even these assumedly universal gene sets, indicating a larger scale deficiency in annotation completeness. The model organisms *Arabidopsis thaliana* and *Oryza sativa* performed best under this particular analysis. This method is likely to underestimate the extent of missing genes in these species, given that the universal genes assessed should have been used for the initial annotation of the plant species in question. Indeed, BUSCO/CEGMA are recommended for obtaining default training sets for the gene finders Augustus (Stanke and Morgenstern 2005) and MAKER2 (Holt and Yandell 2011), and thus are much more likely to be present in the annotation than less well conserved genes. Furthermore, it is conventional for researchers to report BUSCO and CEGMA scores for an annotation and the annotations that maximise these scores are often selected as the best annotations (Austin et al. 2017; Wingfield et al. 2016; Xiong et al. 2016). These genes should therefore be the best predicted and least likely to be absent in all genome annotations. Genes with more restricted phylogenetic distributions were not assessed: it is likely that these genes are more affected by gene prediction errors, as they may not have BLAST hits to conserved genes to help train the gene finders. Thus, while this method succeeded in highlighting annotation completeness deficiencies for a number of organisms, it is likely to be unable to detect annotation errors for non-universally conserved genes.

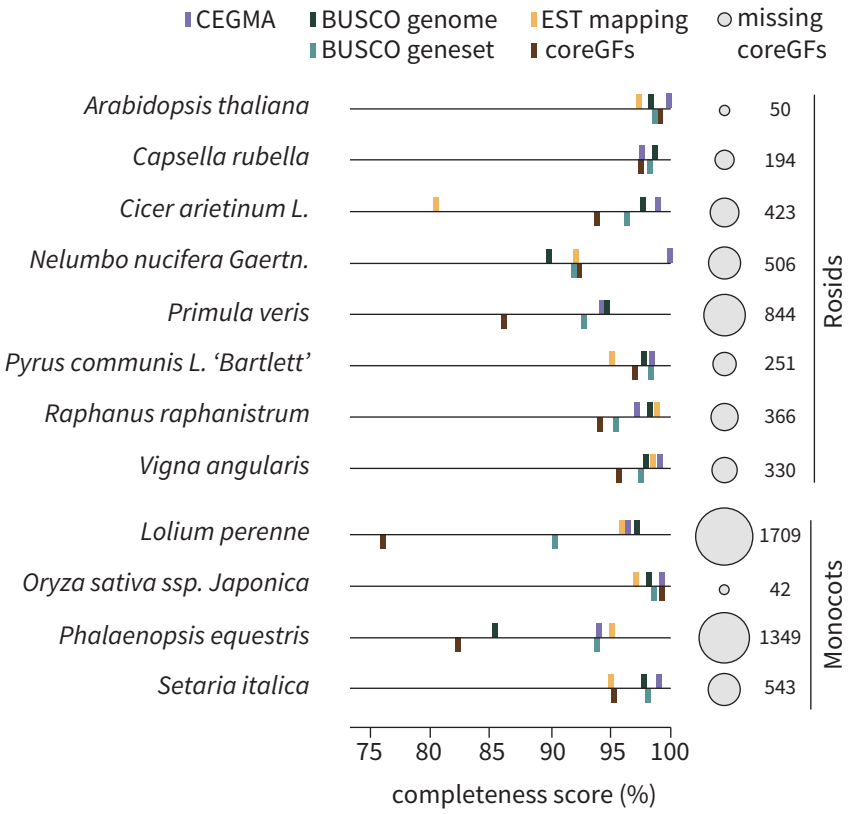


Figure 1.9: CEGMA and BUSCO scores can indicate annotation completeness. Shown are annotation completeness scores for various species under various different metrics. Large numbers of presumed universal gene families appear to be missing from many recently sequenced genomes. Model organisms *A. thaliana* and *O. sativa* have the highest completeness scores by all metrics. From (Veeckman et al. 2016).

For assessing gene model accuracy, the MAKER2 pipeline (Holt and Yandell 2011) adapted the annotation edit distance (AED), as developed by the Sequence Ontology (Eilbeck et al. 2005), to directly compare predicted sequences to EST and protein homology evidence, thus generating scores for the gene model quality of particular genes. In addition, the study compared the percentages of found genes which contained a Pfam domain (Finn et al. 2014). It has been shown that the proportion of genes containing various categories of Pfam domains is relatively consistent between species (Zmasek and Godzik 2011). The analyses in the MAKER2 paper consider those Pfam families which are under-represented in a genome annotation to generate a rough measure for the extent of inaccurate gene

models. This analysis was performed in the context of assessing the success of the MAKER2 gene prediction software. This approach to assessing gene model accuracy is once again limited in that it only assesses genes containing highly conserved domains, which are likely to have more accurate annotations than those which do not: such analyses may overlook genes that have a narrower phylogenetic distribution. A similar argument applies to RNA-seq based methods of gene model determination and assessment: RNA-seq data are not always available for a given organism, and many genes are expressed in a tissue- or cell-cycle-specific manner. Thus, it is not possible to assess all genes by leveraging conserved domain information or RNA-seq data, and for many genes alternative evidence must be sought.

1.5 Thesis aims

A genome is only as good as its annotation. Given the rapidity with which novel genome sequences are being produced and published, it is beyond the realms of feasibility for each genome sequence to undergo careful manual analysis to maximise the completeness and accuracy of its annotated protein-coding gene set. This is particularly true for smaller genome projects produced by individual laboratories with limited resources. Moreover, many genomes remain static for many years after the point of initial annotation, thus allowing propagation of errors into subsequent analyses and spreading inaccurate gene models to other annotation projects. Strategies for assessing the extent of genome annotation errors are scarce. Furthermore, efforts to stem the flow of inaccurate annotations by actively seeking to improve existing gene sets are usually only carried out for the best-studied genomes. There is therefore a high demand for software packages that systematically assess and improve the quality of genome annotations.

In this thesis I aim to address two important facets of genome annotation error: gene set completeness and gene model accuracy. Together, these comprise two of the main problems plaguing the quest for accurate annotations of genomes today, thus it

is of paramount importance that methods both to measure the extent of these errors and to offer ways to improve them are considered. As noted, existing approaches typically rely on comparison to universally conserved gene sets or incorporation of extrinsic data such as RNA-seq or EST evidence, or by analysing distributions of conserved domains. Such approaches are limited in that they often assess genome annotations in comparison to the data used to inform their construction, and are not useful for the analysis of lineage-specific genes. Furthermore, experimental data are frequently either unavailable or not comprehensive or high-quality enough to be useful in assessing the full complement of a species' gene set. Thus the work of this thesis is designed to operate under two specific constraints: absence of experimental data, and non-reliance on universally conserved gene sets.

In the absence of extrinsic data, and when considering non-universally conserved genes, I propose that phylogenetically local homology-based approaches to genome annotation could provide a solution to assessing and improving both gene set incompleteness and gene model inaccuracies. In this thesis I provide such methods, and demonstrate their performance on several groups of species. The goals of this thesis can be summarised as follows:

1. Develop phylogenetically local frameworks for assessing the completeness and accuracy of genome annotations in the absence of extrinsic data;

2. Develop methods to systematically improve genome annotations with respect to their completeness and accuracy.

1.5.1 Thesis plan

This thesis comprises two principle parts, each corresponding to one of the aims laid out above. The first part, consisting of Chapters 2 and 3, discusses methods to assess the extent of gene model errors in existing publicly available annotations, using multiple related species to determine relative measures of completeness and

accuracy at a phylogenetically local scale. This part also introduces the two main species sets that are used to evaluate the methods of the remainder of the thesis. The second part of the thesis, consisting of Chapters 4 and 5, comprises the two software methods designed specifically to address the problems raised in the first part, namely OrthoFiller, a method to improve the completeness of genome annotations, and OMGene, a method to improve the accuracy of individual gene models. These two methods have been prepared as two independent papers. The first, OrthoFiller, was accepted for publication in May 2017 and is in included in its original format in Appendix A. The second, OMGene, was accepted for publication in April 2018 and is included in its original format in Appendex B. In summary, the chapters contained in this thesis are as follows:

- **Chapter 1: General introduction.**
- **Chapter 2: Phylogenetically local measures of genome annotation completeness.** Here I introduce the main test species sets of the thesis and discuss a novel evidence-based method for assessing the relative completeness of genome annotations.
- **Chapter 3: Quantifying gene model errors.** Here I discuss several extrinsic and homology-based methods to identify deficiencies in gene model accuracy, and demonstrate the agreement between them.
- **Chapter 4: OrthoFiller: Utilising data from multiple species to improve genome annotations.** Here I present *OrthoFiller*, a method to improve the completeness of publicly available genome annotations.
- **Chapter 5: OMGene: Mutual improvement of gene models through optimisation of evolutionary conservation.** Here I present *OMGene*, a method to improve gene models in publicly available genome annotations.
- **Chapter 6: General discussion.**

2

Phylogenetically local measures of genome completeness

2.1 Chapter introduction

This chapter considers the problem of assessing and quantifying the extent of missing gene errors, specifically in the context of small groups of closely related species, and in the absence of RNA-seq evidence or other experimental data. The introduction to this thesis covers the important points motivating this chapter. However, to facilitate the reading of this chapter as a stand-alone entity, the most relevant points are briefly summarised in Section 2.1.1.

2.1.1 Background

All genome annotations are incomplete to some degree (Mudge and Harrow 2016; Muñoz et al. 2013). Annotations for even the best-studied model genomes are still works in progress. In particular, genomes annotated either entirely or partly using computational gene prediction tools are prone to annotation absences. Gene finders such as Augustus (Stanke and Morgenstern 2005), MAKER (Cantarel et al. 2008), and GENSCAN (Burge 1997), and annotation pipelines such as the NCBI Eukaryotic Genome Annotation Pipeline (Thibaud-Nissen et al. 2013) and

Ensembl (Aken et al. 2016; Curwen et al. 2004) are typically highly proficient at locating genuine protein-coding genes, however none of them has a 100% success rate. Genome annotation algorithms must strike a balance between caution and quantity: in filtering out spurious gene predictions, many genuine protein-coding genes will inevitably fall through the net.

The complete and accurate determination of an organism's protein-coding genes is of critical importance to any study of that species' biological properties. For example, absent annotations can be mistakenly interpreted as gene loss events, or independent gene gain events on a separate branch of the phylogenetic tree. In many cases, the genome assembly and its annotation are taken at face value by those who utilise these resources, and thus any such errors caused by absent gene annotations are propagated. Given that many protein-coding genes are missed by computational gene finders, missing gene errors are likely to be common for computationally annotated genomes. Thus, to help limit the negative downstream consequences of missing gene annotations, methods to measure the extent of these errors and to highlight potentially under-annotated genomes are in high demand.

Existing approaches to assessing genome annotation completeness are scarce, and there is no unified standard tactic. Moreover, existing methods typically rely on either universally conserved genes (Veeckman et al. 2016) or domains (Holt and Yandell 2011) or appeal to extrinsic experimental information. Methods that make use of universally conserved features are limited in that these features are typically used as base sets to train the gene finding algorithms used to obtain the initial annotations. For example, CEGMA (Parra et al. 2007) and BUSCO (Simão et al. 2015) are typically used to train the popular gene finders Augustus (Stanke and Morgenstern 2005) and MAKER2 (Holt and Yandell 2011). Such approaches are therefore inherently circular. Furthermore, any approach that appeals to extrinsic information, such as mapped RNA-seq reads, by definition relies on experimental data that may be unavailable or incomplete. In particular

RNA-seq data sets that represent the full complement of an organism's protein-coding genes are very difficult to obtain, especially for complex organisms with many distinct tissue types and life stages with distinct expression profiles. Finally, genes with strong read evidence are again more likely to have already been picked up by evidence-guided gene finding algorithms, and therefore this approach too will suffer from some degree of circularity.

A recent paper (Veeckman et al. 2016) that assessed genomes in terms of CEGMA and BUSCO gene representation acknowledged the limitations of utilising universally conserved features. In particular, the paper called for approaches that are based at less generic phylogenetic levels, using a fine-grained approach to assess annotations based on genes conserved at the level of single sets of species. This chapter introduces a novel method that fits this bill, proposing an experiment-free, orthogroup-based approach for assessing relative genome annotation completeness, which avoids the reliance on extrinsic data and on universally conserved features. The method is evaluated for distinct sets of species, and the performance of the approach is assessed by comparing the results to extrinsic experimental data in the form of mapped RNA-seq reads.

2.1.2 Chapter aims and outline

In this chapter I describe a phylogenetically local homology-based method for assessing the completeness of genome annotations, as well as providing estimate relative completeness measures for two biologically distinct sets of species. This is distinct from the assessment of genome assembly completeness: an assembly can be complete and the annotation of that assembly incomplete, and *vice versa*. The method I describe here provides the basis for Chapter 4, which will leverage the ideas behind the completeness metric to actively improve the completeness of genome annotations. This chapter, along with the rest of the thesis, leverages the concept of *orthogroups*, defined in Section 2.3. The goal of this chapter is to

demonstrate that orthogroups can be used to identify potential unannotated genic regions in a group of species, and thus provide a rough measure of relative genome annotation completeness levels. Briefly, the structure of the chapter is as follows:

- **Section 2.1: Chapter introduction**
- **Section 2.2: Two test species sets: plants and fungi**. In this section I introduce the two main sets of test species that will be used for this and subsequent chapters.
- **Section 2.3: Analysis of plant and fungal orthogroups**. Here I introduce the concept of an orthogroup and demonstrate how orthogroup absences can sometimes provide clues toward genome underannotation.
- **Section 2.4: An evidence-based method for estimating the extent of missing genes**. This is the main result of this chapter, a method that uses filtered hidden Markov model (HMM) hits to provide estimated measures of relative genome annotation completeness.
- **Section 2.5: Examples of missing genes identified using HMMs**. This section provides examples of genes located using the method developed in Section 2.4, in the context of amino acid alignments of other members of their orthogroups.
- **Section 2.6: Comparison of candidate missing genes to RNA-seq data**. Here I demonstrate that the HMM hits identified in Section 2.4 typically have greater RNA-seq read evidence than the intergenic backgrounf, providing further evidence for their worth as potential missing genes.
- **Section 2.7: Methods**.
- **Section 2.8: Chapter discussion**.

2.2 Two test species sets: plants and fungi

This section introduces the two main biologically distinct species sets that will be used for analysis in this chapter, as well as serving as test sets for the rest of the thesis. Before beginning any analysis of these genomes or using them for method development, it is helpful to describe briefly some important characteristics of each set, highlighting the differences between them. Aside from gaining a greater familiarity with the nature of the data being used, this helps inform an appreciation of the breadth of the challenge involved in obtaining genome annotations, and in subsequently assessing their quality.

Disparate taxa can display considerable variation in genome size and in the distribution and composition of their protein-coding genes, even among closely related species (Hong et al. 2006; Kapusta et al. 2017; Boulesteix et al. 2006; Deutsch and Long 1999). Notwithstanding individual differences, however, there are certain trends that distinguish certain groups of organisms. For example, prokaryotes almost always have smaller genomes and smaller genes than eukaryotes (Lynch and Conery 2003). Plants and animals have larger genomes and more complex gene models than fungi (Gregory et al. 2007), and subgroups of these kingdoms have their own individual general characteristics regarding average exon count, intron and exon size, genome size, and gene density.

The species sets described below represent two groups with highly distinct characteristics: five fungal species, with small genomes, simple gene models and densely packed genes, and five land plant species, with much larger genomes, complex intron-containing gene models, and large non-coding regions. What follows will detail briefly some properties of the species comprising these groups, highlighting the overarching differences between them.

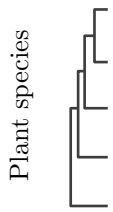
	Tree	Species name	Source	Version / Strain	Tax. ID	Refs
Plant species		*Arabidopsis thaliana*	JGI	TAIR10	3702	[1,2]
		Capsella rubella	JGI	183 v1.0	81985	[1,3]
		Brassica rapa	JGI	FPsc v1.3	3711	[1]
		Carica papaya	JGI	ASGPBv0.4	3649	[1,4]
		Theobroma cacao	JGI	Matina 1-6 v1.1	3641	[1,5]
Fungal species		*Eremothecium gossypii*	JGI	ATCC10895	284811	[6]
		Kluyveromyces lactis	JGI	CLIB210	284590	[7]
		Saccharomyces cerevisiae	SGD	S288C	559292	[8]
		Debaryomyces hansenii	JGI	CBS767	284592	[7,9]
		Yarrowia lipolytica	JGI	CLIB122	284591	[7]

Table 2.1: List of species used, a phylogenetic species tree for each group, sources, and references. All data sets were downloaded from the Joint Genome Institute (JGI) website – specifically Phytozome (Goodstein et al. 2012) and Mycocosm (Grigoriev et al. 2014) for plants and fungi respectively – with the exception of *S. cerevisiae*, which was downloaded from the Saccharomyces Genome Database (SGD) (NHGRI 2004). Strain and cultivar information is listed alongside the genome version, where applicable. References: [1]Goodstein et al. 2012, [2]Lamesch et al. 2012, [3]Slotte et al. 2013, [4]Ming et al. 2008, [5]Motamayor et al. 2013, [6]Dietrich et al. 2004, [7]Dujon et al. 2004, [8]Gnerre et al. 2011, [9]Sacerdot et al. 2008

2.2.1 Species list and sources

The full set of chosen species is listed in Table 2.1, along with relevant information regarding strain/cultivar, assembly version, source, and references. Briefly, the fungal species set (Species Set F) comprises *Eremothecium* (also known as *Ashbya*) *gossypii*, *Debaryomyces hansenii*, *Kluyveromyces lactis*, ***Saccharomyces cerevisiae*** and *Yarrowia lipolytica*; the plant species set (Species Set P) comprises ***Arabidopsis thaliana***, *Brassica rapa*, *Carica papaya*, *Capsella rubella*, and *Theobroma cacao*. Each group contains one well-studied model organism, listed here in bold. The purpose of including a well-studied species is to compare the other members of the group to what can be assumed to be a well-constructed, mostly complete genome assembly whose annotation has been subject to extensive manual curation.

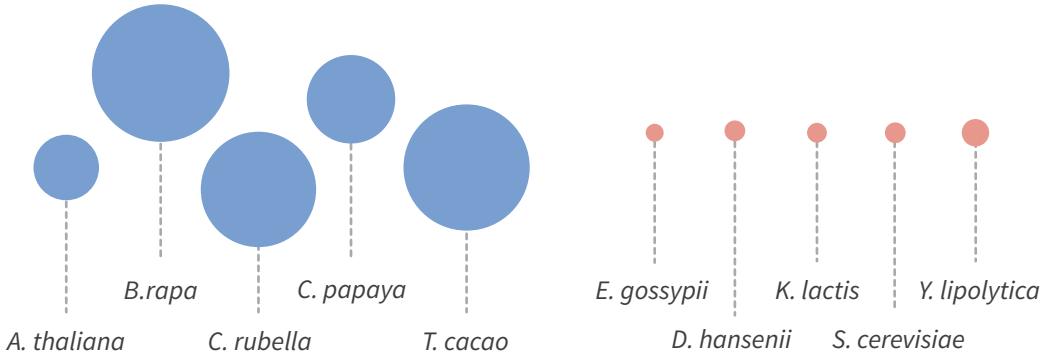


Figure 2.1: Plants have larger genomes than fungi. Shown are relative sizes of genomes for the ten species described in table 2.2. Ball sizes represent estimated assembly size (area). The plant genomes are on average an order of magnitude larger than their fungal counterparts.

2.2.2 Plant genomes are larger than fungal genomes

Information regarding the assemblies and genomes for each of the species described above can be found in Table 2.2. The data demonstrate a stark contrast between the two groups, with respect to both the size of the genomes and the completion level of the assemblies. The relative estimated sizes of these genomes are illustrated in Figure 2.1. The size estimates here are obtained from sources as noted in the table legend and have been calculated using combined statistical and experimental procedures such as renaturation kinetics (Sulston and Brenner 1974) and flow cytometry (Doležel and Bartoš 2005). Estimates can vary considerably, and where multiple possible sizes were given by a source listed here, the smallest estimated size is shown. There is a stark contrast between the size of the plant assemblies here compared to their fungal counterparts: plant genomes are on average an order of magnitude larger than fungal genomes, with the average fungal genome here comprising 13 mega base pairs (Mbp), compared to the average plant genome size of 337Mbp.

	Species name	Chromosomes	Ploidy	Est. size	Regions	Assembly
Plants	*A. thaliana*	$2 \times 5 + C + M$	Diploid	120 Mbp[1]	7	119.7 Mbp
	B. rapa	$2 \times 10 + C + M$	Diploid	529 Mbp[2]	5713	315.1 Mbp
	C. papaya	$2 \times 9 + C + M$	Diploid	372 Mbp[3]	5901	331.5 Mbp
	C. rubella	$2 \times 8 + C + M$	Diploid	219 Mbp[4]	853	134.9 Mbp
	T. cacao	$2 \times 10 + C + M$	Diploid	445 Mbp[5]	712	346.2 Mbp
Fungi	*E. gossypii*	$7 + M$	Haploid	9 Mbp[6]	7	9.1 Mbp
	D. hansenii	$7 + M$	Haploid	13 Mbp[7]	7	12.2 Mbp
	K. lactis	$(1 \text{ or } 2) \times 6 + M$	Hap-/diploid	11 Mbp[6]	6	10.7 Mbp
	S. cerevisiae	$(1 \text{ or } 2) \times 16 + M$	Hap-/diploid	12 Mbp[8]	17	12.2 Mbp
	Y. lipolytica	$(1 \text{ or } 2) \times 6 + M$	Hap-/diploid	21 Mbp[9]	6	20.5 Mbp

Table 2.2: Genome and assembly information for test species. Chromosomes "C" and "M" refer to chloroplast and mitochondrial genomes respectively. For many of the fungal species, mitochondrial genomes are not annotated. Some species described here exhibit multiple modes of ploidy, and are marked as such. "Regions" refers to contiguous entries in the genome FASTA file, and may be scaffolds, contigs, supercontigs, or chromosomes. Genome size estimates are taken from the following sources: [1]Bennett et al. 2003; The Arabidopsis Genome Initiative 2000, [2]Tang and Lyons 2012, [3]Ming et al. 2008, [4]Slotte et al. 2013, [5]Motamayor et al. 2013, [6]Grigoriev et al. 2014, [7]Jacques et al. 2010, [8]Wilhelm 2003, [9]Casaregola et al. 2000.

2.2.3 Most of the plant assemblies are highly unfinished

Most complex eukaryotic genome assemblies, including the human genome, remain unfinished to some degree, in part due to difficult-to-assemble highly repetitive regions (Muñoz et al. 2013). This is especially apparent in the plant genomes here, all of which – except the well-studied *A. thaliana* – exhibit a clear shortfall between the estimated genome size and the assembly size (the absolute number of bases in the assembly). Plant genomes often exhibit large repetitive regions of low complexity, a phenomenon that is more acute for larger plant genomes (Claros et al. 2012). The shortfall between the estimated genome size and the assembled genome size is therefore in part due to the genomes themselves being difficult to assemble. Compare this to the fungal genomes, each of whose assembly size matches its estimated size and should be considered essentially complete. This is

expected: fungal genomes are smaller as noted, and contain less repetitive material, as such their genomes are somewhat easier to assemble.

The most important aspect of the unfinished assemblies to note is the extent to which they are fragmented: the genomic data are presented in many more contiguous regions than the 7-12 chromosomes marked. These take the form of disconnected contigs, supercontigs, and scaffolds instead of full chromosomes. Fragmented genomes such as these are difficult to annotate correctly: full-length genes may be split over multiple different contigs, making their annotation difficult, and resulting either in split or partial gene models or genes that have failed to be annotated at all (Denton et al. 2014; Zhang et al. 2016). The quality of an assembly and an annotation are not necessarily correlated, however, and a higher-quality assembly does not necessarily mean a higher-quality anntotation.

2.2.4 Plants have larger gene sets than fungi

Another principle difference between the plant and fungal data sets can be seen in the distribution and composition of their annotated gene models. The number of genes annotated in each of the species is shown in Table 2.3. In keeping with the differences in genome size, plant genomes typically have an order of magnitude more genes than fungal genomes, with an average of 30,315 genes per plant genome and 5,832 genes per fungal genome in these species sets. Furthermore, unlike the fungal genes, many of the plant genes have multiple alternate transcripts listed, although the figures vary hugely within the species set. *C. papaya* in particular has very few alternate transcripts marked. Alternative splicing is common in plants (Syed et al. 2012), and this lack of alternate transcripts is an artefact of the annotation, not a reflection of the biology, of *C. papaya*.

	Species name	No. genes	No. of genes w/ alternate transcripts
Plants	*A. thaliana*	27416	5804 (21.2%)
	B. rapa	40492	2437 (6.0%)
	C. papaya	27695	34 (0.1%)
	C. rubella	26521	1742 (6.6%)
	T. cacao	29452	7630 (25.9%)
Fungi	*E. gossypii*	4768	N/A
	D. hansenii	6270	2 (<0.1%)
	K. lactis	5076	N/A
	S. cerevisiae	6600	N/A
	Y. lipolytica	6447	N/A

Table 2.3: Numbers of annotated genes for each species. The plant genomes here have an average of 30,315 genes per genome; the fungi have on average 5,832. Plant species often have many alternate transcripts marked, although the number of these varies considerably due to differences annotation quality.

2.2.5 Introns are more common in plants than in fungi

Introns are rare in fungal genes, but common in plant genes. Table 2.4 shows summarised data on the typical gene composition characteristics for the two species sets. The percentage of multi-exon genes is dramatically larger for plant genes: on average 78.1% of plant genes contain introns, compared with just 6.9% of fungal genes. This is reflected in the considerably smaller exon-count average for fungi than plants, at 1.08 and 5.17 exons respectively. The mean intron size for both sets is of the same order across all the species, however the range is considerably greater for plants: in the annotations considered here the largest nuclear intron in a fungal genome is found in the *Y. lipolytica* gene *YALI0E14058g*, with 1,550 base pairs; the longest annotated intron in a plant gene occurs in the *T. cacao* gene *Thecc1EG011092t1.CGDv1.1*, and encompasses 1,284,698 base pairs. This is one of many ultra-long introns annotated in *T. cacao*, and it is likely that these are genome annotation errors

		Mean				
	Species name	Exon size	Intron size	No. exons	Transcript size	% multi-exon genes
Plants	*A. thaliana*	221	157	5.57	1231	79.1%
	B. rapa	194	231	5.04	1163	79.5%
	C. papaya	220	479	4.05	892	70.2%
	C. rubella	240	170	5.22	1254	79.1%
	T. cacao	220	493	5.97	1315	82.4%
Fungi	*E. gossypii*	1390	101	1.06	1471	5.7%
	D. hansenii	1340	132	1.07	1440	6.7%
	K. lactis	1408	293	1.03	1457	3.3%
	S. cerevisiae	1267	353	1.05	1329	4.3%
	Y. lipolytica	1255	275	1.17	1464	14.7%

Table 2.4: Gene distribution and composition statistics. Plant genes generally have smaller but more numerous exons than fungal genes, with far more introns. Multi-exon genes are relatively rare in the fungal species described here.

2.2.6 Plants contain more non-coding material than fungi

In keeping with the higher prevalence of introns, plant genomes contain far more intergenic material than fungi (see Table 2.5). Fungal gene sets are denser (68.1% average genic in fungi versus 37.3% in plants). This statistic incorporates untranslated regions (UTRs) for most of the plant data genes, but not for the fungal genes as this information is not provided in the genome annotation file that was downloaded from JGI.

The difference is even more striking when comparing coding sequence (CDS) regions and intronic regions: on average 16.9% of bases in the plant genomes listed here constitute CDS regions, compared with 67.4% for fungal genomes (see Table 2.5 and Figure 2.2). Conversely, 16.3% of bases in the plant genomes are intronic, compared with a mere 0.66% of fungal genomes (this refers to regions which are purely intronic, i.e. discounting from a gene's intron any coding sequences contained therein that belong to another gene.)

		% genome which is		
	Species name	Genic	Coding	Intronic
Plants	*A. thaliana*	50.1%	27.8%	15.8%
	B. rapa	28.5%	14.8%	10.5%
	C. papaya	19.7%	7.4%	12.2%
	C. rubella	43.9%	24.5%	14.3%
	T. cacao	44.3%	10.0%	28.5%
Fungi	*E. gossypii*	77.4%	77.0%	0.3%
	D. hansenii	74.8%	74.2%	0.5%
	K. lactis	69.6%	69.1%	0.5%
	S. cerevisiae	71.2%	70.7%	0.6%
	Y. lipolytica	47.4%	46.0%	1.4%

Table 2.5: Percentages of each genome that are annotated as genic, coding, and intronic. Intronic regions are not straightforward to define, given that introns of one transcript can contain coding sequences for another: this number therefore describes only regions which are purely intronic, i.e. the parts of each intron which do not contain coding sequences for any other gene. It should also be noted that the sum of the coding and intronic percentages does not always equal the genic percentages: this is because "genic" includes untranslated regions (UTRs), which are neither coding nor intronic. UTRs were not annotated for either *C. papaya* or for the fungal species.

It should be noted that the gene quantities and densities here are likely to be somewhat intertwined with varying levels of genome annotation completeness, however the data provide a basic indication of the differences between the two species sets. For example, the *C. papaya* and *B. rapa* genomes, both of whose assemblies consist of thousands of scaffolds representing only a few chromosomes, have much smaller proportions of genic regions. This is at least partly because genes cannot be annotated across contigs: if a genome sequence is split into multiple parts then information will necessarily be lost at the points of incision (Zhang et al. 2016; Denton et al. 2014). This has the dual effect of reducing the annotated gene count and distorting the apparent gene density, due to the reduced representation of repetitive, harder-to-resolve genomic regions, which typically contain fewer genes.

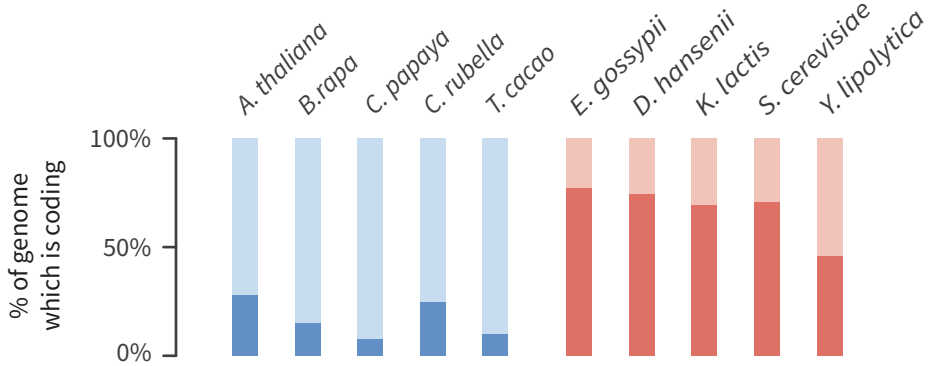


Figure 2.2: Plants have more non-coding DNA than fungi. Shown are the percentages of plant and fungal genomes which are annotated as coding. Clearly a much larger proportion of the average fungal genome is coding, compared to the average plant genome.

2.2.7 Gene distributions affect gene finder accuracy

Fundamental differences between gene distribution, density, and structure can affect the success of computational gene finding algorithms. For example, genomes with high gene density are prone to merging of neighbouring genes (Yandell and Ence 2012); conversely sparse genomes with smaller exons and large introns can be harder to annotate accurately, and so are likely to require much more care and attention when constructing gene models.

The differences in genome and predicted proteome composition described here highlight the considerable variety that can exist between different groups of species, and in turn the difficulty in assessing the levels of completeness of the data. In particular, given that different species sets can display such markedly different characteristics, homology-based methods carried out from a universal perspective are limited, supporting the idea of carrying out assessments at a phylogenetically local level.

2.3 Analysis of plant and fungal orthogroups

The main result of this chapter will be discussed in Section 2.4, and assessed in Sections 2.5–2.6. The method developed in Section 2.4 will rely on assessing

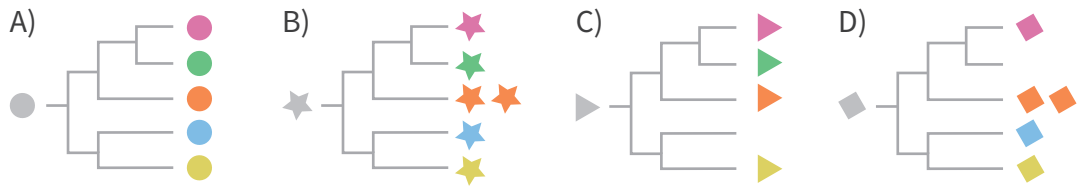


Figure 2.3: Examples of various types of orthogroup. An orthogroup is the set of genes descended from a single inferred ancestral gene in the last common ancestor of a set of species. Here, shapes represent related genes, colours represent species. The tree is a phylogenetic species tree. A) An orthogroup may contain exactly one gene for each species; B) An orthogroup may include multiple paralogs for a species, due to gene duplication events; C) An orthogroup may have missing members, either due to gene loss events or absent gene annotations; D) Combinations of gene duplication and gene loss events can be present in an orthogroup.

conserved gene families across multiple species and searching the genomes for potential missing members. The concept of conserved gene families is formalised in this section, in the form of *orthogroups*. Here, orthogroups are defined, and a brief analysis of the orthogroups formed from genes in Species Sets P and F is carried out. In addition, this section discusses a naïve approach to assessing missing genes based on quantifying orthogroup absences, and demonstrates why it is insufficient for revealing missing gene errors.

2.3.1 Orthogroups formalise the notion of conserved genes

The notion of sets of genes that are evolutionarily related across multiple species can be formalised with the concept of an *orthogroup*. For a given set of species, an orthogroup is the set of genes descended from a single ancestral gene in the last common ancestor (LCA) of those species (Emms and Kelly 2015). This is a direct extension of the pairwise concept of gene orthology. By definition, orthogroups may contain paralogous as well as homologous genes. Similarly, genes can be absent from an orthogroup due to genuine gene loss or gain events. An illustration of this important concept is shown in Figure 2.3.

2.3.2 Plants have larger and more numerous orthogroups than fungi

Orthogroups for each complete set of annotated protein-coding genes in Species Sets P and F were calculated using OrthoFinder (Emms and Kelly 2015) (See Methods Section 2.7.2. Data on the quantities of orthogroups calculated can be seen in Table 2.6. For the plant species set, 81.9% of genes were classified into orthogroups, the remaining genes having insufficient similarity to genes in other species to facilitate grouping. Similarly, for the fungal data set, 84.3% of genes were classified into orthogroups. For the plants, the average orthogroup size was greater (6.8 genes on average per orthogroup, versus 4.9 for fungi), indicating a higher incidence of gene duplication in line with previous observations (Lee et al. 2013).

Much of the analysis of this thesis will focus on *single-copy genes*. These are genes that are present exactly once in a species' genome, i.e. with no paralogs present. Single-copy genes have been shown to be better conserved than non-single-copy genes, and thus they provide a good test set for evolutionary analyses (Gu et al. 2004; Jiang et al. 2011; De Smet et al. 2013). A direct extension of single-copy genes are *single-copy orthogroups*, i.e. orthogroups containing only single copy genes, and *single copy ubiquitous* (SCU) orthogroups, i.e. orthogroups containing exactly one single-copy gene from each species under consideration. These constitute large percentages of each orthogroup set: 40.4% and 80.0% of plant and fungal orthogroups respectively are single-copy; 21.2% and 55.4% of plant and fungal orthogroups respectively are SCU. The lower percentage in plants further reflects the higher incidence of gene duplication in plants than fungi (Lee et al. 2013). Finally, for the purposes of assessing missing genes it is helpful to consider orthogroups that have exactly one gene for all but one species. These will be denoted SCU-1 orthogroups. As will be discussed next, these can sometimes be indicative of a missing gene annotation. In Species sets P and F above, 4.0% and 8.4% of orthogroups respectively are SCU-1.

	Plants	Fungi
No. orthogroups	18330	4994
Total assigned genes	124181 (81.9%)	24569 (84.3%)
Total unassigned genes	27451 (18.1%)	4564 (15.7%)
Mean no. genes per OG	6.8	4.9
No. single-copy orthogroups	7407 (40.4%)	3998 (80.0%)
No. SCU orthogroups	3885 (21.2%)	2771 (55.4%)
No. SCU-1 orthogroups	735 (4.0%)	418 (8.4%)

Table 2.6: Data on the partition of the two gene sets into orthogroups. In this analysis, plant orthogroups were much more numerous, although orthogroups containing paralogs were much more common than in fungi. For both plants and fungi, the majority of genes were classified into orthogroups. Orthogroups were generally larger for plants, due to a higher incidence of paralogous genes. Single-copy orthogroups have at most one gene in each species, single-copy ubiquitous (SCU) orthogroups have exactly one gene in each species (i.e. no gene duplication events), and SCU-1 orthogroups have exactly one gene in all but one species.

2.3.3 Counting absent orthogroup members is not enough

The most basic assessment of absent orthogroup members would be to simply consider which species are missing from otherwise well-conserved orthogroups. For species with highly under-annotated genomes, one would expect that a large number of orthogroups do not contain a gene from that species: the data simply won't be there. Thus, levels of species representation in orthogroups can provide a very basic indication of annotation completeness. Unfortunately, this approach is fraught with danger. If an orthogroup does not contain a gene for a particular species, one of three events has occurred: a genuine gene loss or gain event, a missing gene annotation, or an inaccurate or ambiguous orthogroup assignment. *A priori*, it is impossible to tell which of these events has occurred: however, some patterns of gene loss and gain events are more likely than others. The gene loss/gain landscape is largely dependent on the shape of the underlying phylogenetic tree. Crucially, this means that some patterns of absent annotations cannot be readily explained by gene loss and gain events, and unusual patterns can provide clues

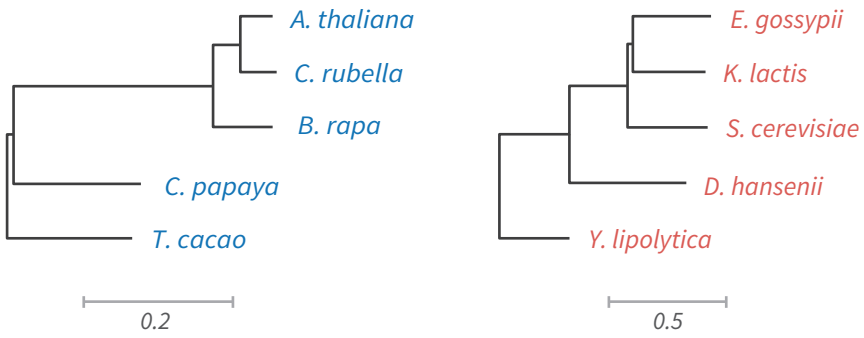


Figure 2.4: Species trees for the two species sets. Species trees were calculated using OrthoFinder and rooted using Phytozome and Mycocosm respectively (see Methods Section 2.7.3).

that particular genomes are under-annotated.

Figure 2.4 shows species trees calculated for the plant and fungal species sets (see Methods Section 2.7.3). These trees are calculated using SCU orthogroups and so should be largely unaffected by missing gene errors. Simple inspection of the species trees shows why some orthogroup shapes are more likely than others, i.e. why some sets of species are likely to be absent from a given orthogroup. Take the plant data set as an example, and single copy orthogroups with exactly two members. Assuming no missing annotations, a two-member orthogroup can be the result of a gene loss event on the long branch connecting *A. thaliana*, *B. rapa*, and *C. rubella* to *C. papaya* and *T. cacao*, or the result of a single gene gain event specific to the *A. thaliana* / *C. rubella* subtree. Alternatively, it could be the result of some combination of gene loss and gene gain events across the tree. Single events are more likely than multiple events, and so the first two choices are much more likely. This is reflected in the data: out of 1,508 two-member single-copy orthogroups, 797 (52.9%) contain *C. papaya* and *T. cacao*, and 374 (24.8%) contain *A. thaliana* and *C. rubella*. The other eight possible two-species combinations combined make up the remaining 22.3%. Similarly, 1,167 out of 1,279 (91.2%) of three-species single-copy orthogroups are made up of *A. thaliana*, *B. rapa*, and *C. rubella* genes, indicating a gene gain event on the long branch. Similar

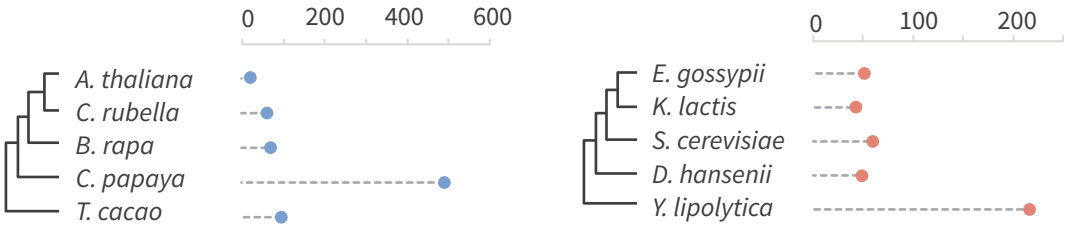


Figure 2.5: Quantities of SCU-1 orthogroups that omit each species. For the plants, most orthogroup omissions take place in *C.papaya*. For fungi, most omissions occur for *Y. lipolytica*. For the fungi, this can be readily explained by gene loss and gain events due to the structure of the species tree; this is not the case for plants, and the large number of absences for *C. papaya* here indicate an under-annotated genome.

distributions can be seen for the fungi, however these are less clear due to the more evenly distributed nature of the fungal phylogenetic tree.

Finally, it should be noted that orthogroup inference becomes more difficult as dissimilarity between amino acid sequence increases. In particular, amino acid sequences become progressively less similar with evolutionary distance, making orthogroup inference less accurate for more distant groups. The distribution of orthogroup absences may, therefore, be affected by unevenly distributed orthogroup inference errors. Given that the two sets here are internally relatively well related, this should have only a minor effect.

2.3.4 *C. payaya* is probably under-annotated

With the above caveats in mind, it is still plausible that particularly strong missing gene signals can be observed by counting absent species, and so a basic analysis was carried out for Species Sets P and F. For clarity, the analysis was restricted to SCU-1 orthogroups. For one of these to be real, either a single gene loss event has to have taken place on a leaf branch, or, in the case of the trees above, a gene gain event has to have taken place on the branch separating the outgroup from the other species. Figure 2.5 shows the number of SCU-1 orthogroups with each species missing.

According to the plant species tree shown in Figure 2.4, *T. cacao* and *C. papaya* have the longest individual leaf branches and essentially diverged from the other

species at the same time. They should, therefore, show similar amounts of gene loss. Similarly, the small branch connecting the last common ancestor to the non-*T. cacao* subtree is very short, and thus a gene gain event that occurs on this branch, though quite possible, is unlikely. Such an event would result in a gene that was present in all species except *T. cacao*.

Based on the above observations, one would expect that the two plant outgroups should show similar levels of orthogroups absences resulting from gene loss and gain events. This is however not reflected in the data – *C. papaya* has far more absences than *T. cacao*. The most likely explanation for this is that *C. papaya* is lacking many gene annotations. In the fungal data set, *Y. lipolytica* has the largest number of absences, however it is likely that this can be explained as either gene gain events along the branch separating it from the other species, isolated gene loss events along the *Y. lipolytica* leaf branch, or orthogroup classification errors due to *Y. lipolytica* being an outgroup. Thus, although there are many underlying factors preventing orthogroup absences from being used to accurately determine missing genes, here a strong signal for an under-annotated genome can be seen from *C. papaya*. In general, however, a more fine-grained approach is required: looking for the missing genes themselves.

2.4 An evidence-based method for estimating the extent of missing genes

Although quantification of absent orthogroup members can provide some basic clues to the prevalence of missing genes in some species, it still provides little *concrete* evidence for their existence. In order to more concretely quantify the scope of missing genes, genomes must be searched for evidence of unannotated members of conserved families. This is similar to the approach used by Veeckman et al. (2016), using universally conserved genes. However, as stated above these universally conserved genes are almost always used to train gene finders and thus

are likely over-detected by genome annotation methods. Proposed here is a more fine-grained approach that instead utilises any orthogroup that contains two or more orthologous genes in the input species set. The method employs a novel filtering system for assessing hits from hidden Markov models (HMMs), and uses these to provide estimates for the prevalence of missing genes in each species set.

2.4.1 Hidden Markov models identify potential missing genes

Hidden Markov models (HMMs) provide a statistical framework for assessing the likelihood of a particular sequence of states being produced by an unknown, inferred generating procedure (Henderson and Salzberg 1997). In the case of gene sequences, HMMs can be used to check the similarity of a particular nucleotide sequence to a given multiple sequence alignment. The HMM is inferred as a probabilistic model based on aligned nucleotide sequences, such as an aligned orthogroup of coding sequences, assuming some basic, unobserved generating state from which all sequences are produced. In the case of an orthogroup, this base state may be the ancestral protein sequence from which all members of the orthogroup are descended, in which case the HMM can be interpreted as a model for the ancestral sequence. Upon comparison to a given HMM, stretches of a given nucleotide sequence are scored based on how likely it is that they are drawn from some part of the HMM. This is typically assessed using measures such as the bit score and the E-value, discussed below.

For the purposes of finding missing orthogroup members, evidence can be sought by generating HMMs from orthogroup alignments and searching a genome for potential candidates. Note that this evidence alone is not sufficient to demonstrate that a gene exists: sequence similarity can be present but the coordinates of the proposed new gene may not be amenable to the construction of a viable protein.

HMM hits can, however, provide the basis for finding high-confidence potential genic regions, a good proportion of which should provide evidence for missing genes.

2.4.2 An adjusted bit score avoids E-value biases

For each orthogroup in each of Species Sets P and F, the nucleotide sequences of its constituent genes were aligned and used to form a hidden Markov model (see Methods Sections 2.7.4 and 2.7.5). These HMMs were then used to search each genome in the relevant species set, using Hmmer (Eddy 2011) (Methods Section 2.7.5). Hmmer identifies stretches of genomic sequence which have a high likelihood of being drawn from the same probabilistic distribution as sequences from within the alignment. It reports two main scores, the *E-value* and the related *bit score*.

The bit score is a log-odds ratio score based on the Blosum62 matrix (Henikoff and Henikoff 1992), and compares the likelihood of the HMM matching the target sequence to the likelihood of the HMM matching an independent, identically distributed (i.i.d.) random sequence model (Eddy 2011). This is the same scoring metric used by BLAST (Altschul et al. 1990). The E-value, or *expect value* is the number of hits expected to attain this score for a certain constant number of sequences drawn from an independent and identically distributed random sequence model.

The E-value, while a good measure of the validity of a hit, is affected by two major restrictions. Firstly, it is heavily biased by the length of a sequence: shorter sequences necessarily tend to have higher E-values (Emms and Kelly 2015). Secondly, Hmmer implements a cut-off of around 10^{-200}, below which scores are rounded down to zero. This reduces the fidelity of the scoring system and prevents accurate re-adjustment of scores based on their length. To circumvent these issues with the E-value, the following analyses will make use of an alternative scoring system, using the adjusted bit score defined by

$$score_{adj}\left(h\right) = \frac{score\left(h\right)}{length\left(h\right)}$$

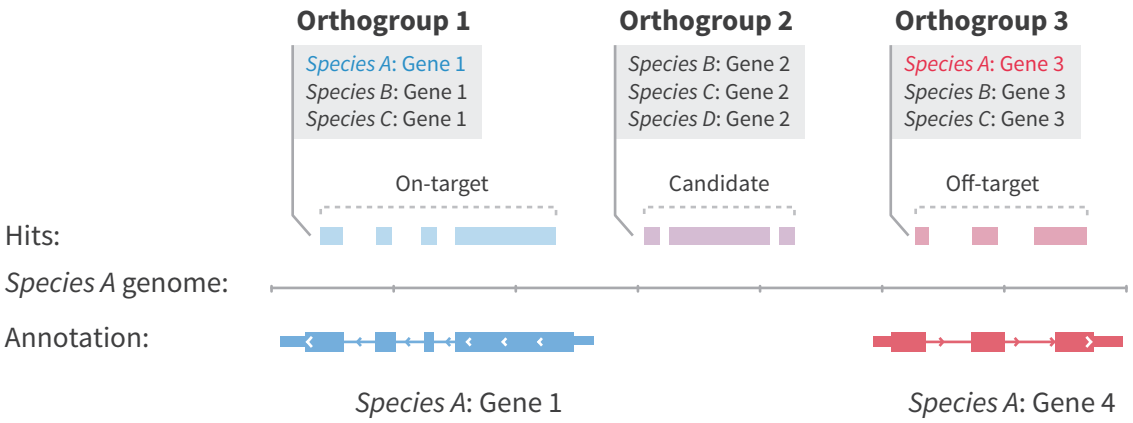


Figure 2.6: Hits can be classified by comparison to existing annotations. On-target hits are hits to regions of the genome containing a gene from the orthogroup used to construct the HMM; candidate hits are hits which match a region of the genome where no gene has been annotated; off-target hits are hits which match a region of the genome containing a gene belonging to an orthogroup other than the one used to construct the HMM.

for a given hit h to the genome.

2.4.3 HMM hits can be classified by comparison to existing annotations

For each orthogroup, adjusted bit scores were calculated for all hits to the genome outputted by Hmmer. Adjusted scores ranged between -0.220 and 1.368. In order to distinguish inevitable spurious matches from hit regions that genuinely resembled the input sequences, the adjusted scores for known correct hits (i.e. where the HMM returned a hit to a gene used to construct the multiple sequence alignment that generated that HMM) were compared with the general landscape of all adjusted hit scores returned by Hmmer. Hits were grouped into three categories: *on-target* hits, those which have matched to the location of one of the genes present in the orthogroup used to construct the HMM; *off-target* hits, those which match to a location of a gene present in the annotation but which do not belong to the orthogroup used to construct the HMM; and *candidate* hits, those which do not match to any gene already present in the genome annotation. An illustration of the different hit types can be seen in Figure 2.6.

The nomenclature described above distinguishes hits based purely on their relationship to the gene set used to construct the HMMs. It makes no assertion as to the validity of the hit, i.e. whether it is a true or false positive in terms of whether the hit should be deemed a genuine member of the HMM – and therefore orthogroup – that was used to find it. For example, off-target hits could potentially arise from erroneous orthogroup inference, and the hit gene could in fact be a perfectly valid member of the orthogroup that matched it, but which failed to be placed in the correct group during orthogroup inference. In particular, there is bound to be some fuzziness regarding orthogroup membership, given that orthology is not always a transitive relation; furthermore, OrthoFinder does not achieve perfect precision and recall on benchmark datasets (Emms and Kelly 2015) and thus there is likely to be some error in orthogroup inference. One can, however, assume that the majority of on-target hits are true positives and the majority of off-target hits are false positives. It should therefore be possible to analyse candidate hits based on the hit score distributions for on-target and off-target hits.

The default reporting threshold for Hmmer was not changed (E-value ≤ 10) and all reported hits were analysed. A breakdown of the number of hits seen for each species can be found in Table 2.7. As expected from the relaxed inclusion threshold, most hits are off-target: on-target hits feature quite rarely in the landscape of all hits, with an average of 7.2% of hits to plant genomes and 5.6% of hits to fungal genomes matching genes that are present in the source orthogroup. Table 2.7 also shows the number of genes "recovered" by the HMMs: the majority of recovered genes in all cases correspond to on-target hits. Finally, and most importantly, there is a clear difference between the average adjusted hit score for all hits and the average score for on-target hits, with average on-target hit scores of 0.71 and 0.66 for plants and fungi respectively, and average off-target hit scores of 0.20 and 0.11 for plants and fungi respectively.

		No. hits		No. genes recovered		Mean score	
	Species	All	On-target	All	By on-target hit	All	On-target
Plants	*A. thaliana*	1668274	131361 (7.9%)	26500	24814 (93.6%)	0.20	0.74
	B. rapa	2330114	171022 (7.3%)	39924	34857 (87.3%)	0.21	0.73
	C. papaya	1133670	80850 (7.1%)	24172	17522 (72.5%)	0.21	0.68
	C. rubella	1722447	129410 (7.5%)	26368	24919 (94.5%)	0.20	0.74
	T. cacao	1744125	111318 (6.4%)	26261	21091 (80.3%)	0.20	0.66
Fungi	*E. gossypii*	58521	4846 (8.3%)	4671	4633 (99.2%)	0.13	0.67
	D. hansenii	93826	5334 (5.7%)	5526	5057 (91.5%)	0.11	0.63
	K. lactis	94320	4991 (5.3%)	4916	4852 (98.7%)	0.11	0.68
	S. cerevisiae	110340	5566 (5.0%)	6074	5297 (87.2%)	0.11	0.67
	Y. lipolytica	147891	5331 (3.6%)	5809	4741 (81.6%)	0.10	0.63

Table 2.7: Hit scores from all HMMs run across the genomes from their respective species sets. The vast majority of hits are not on-target. Recovered genes usually correspond to on-target hits. On-target hits typically exhibit much higher scores than the hit background.

Distributions of adjusted scores for the various categories can be seen in Figures 2.7 and 2.8 for the plant and fungal data sets respectively. Parts A-C of Figures 2.7 and 2.8 show the density distributions of adjusted hit scores for off-target, on-target, and candidate hits respectively, each line representing the distribution of scores from a single species. It is clear that the on- and off-target hits follow distinct distributions. For both species sets, the distribution of scores for the candidate hits more closely matches the distribution of scores for the off-target hits, suggesting that this group of candidate hits contains more spurious hits than hits to real missing genes. Worth noting is the slightly perturbed distribution seen in Figure 2.8C, where one of the distributions has a slightly shifted peak on the left hand side and a small bump at the right. This is the distribution for *E. gossypii*: the two perturbations are due to a prolific retrotransposon repeat and a separate dense region of repeated CAT repeats, both causing more hits for their respective HMMs than one would expect and slightly skewing the distribution.

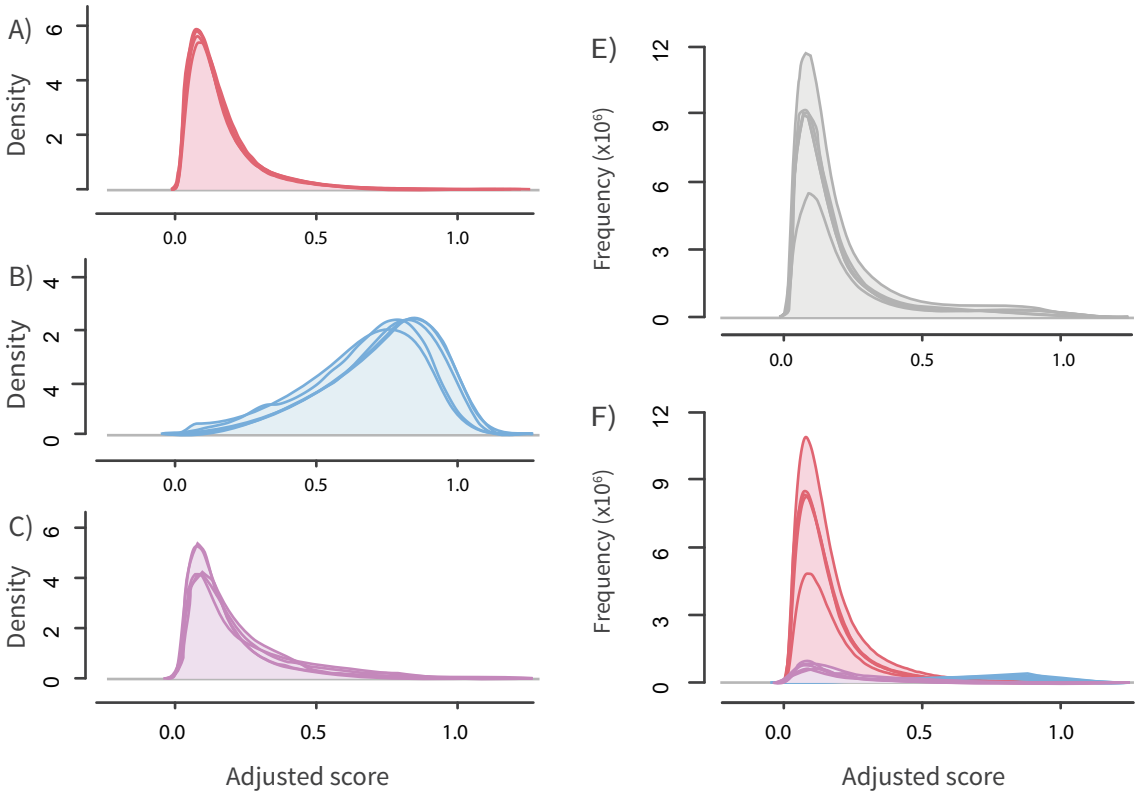


Figure 2.7: HMM hit distribution for plant orthogroups. A-C) Off-target, on-target, and candidate hits respectively. The off-target and on-target hits follow markedly distinct distributions. E) Frequency distributions of all hits; F) Frequency distributions of each hit category, overlaid. The on-target and candidate hits are dwarfed by the off-target hits.

Parts E and F of Figures 2.7 and 2.8 show the frequency distributions of all hits and the superimposed frequency distributions for the individual categories. From this it can clearly be seen that, in all cases, the off-target hits vastly outnumber both the on-target and candidate hits. This has implications for deciding whether a candidate hit is a true or false positive: the prior probability for a hit being spurious would appear to be much greater than for a hit being genuine. This needs to be taken into account when deciding how to categorise a candidate hit based on its adjusted score.

In order to decide whether or not to trust a candidate hit, i.e. to assert that it is the result of an HMM matching a truly unannotated gene rather than a false positive or spurious hit, the distributions of hit scores for on- and off-target hits are

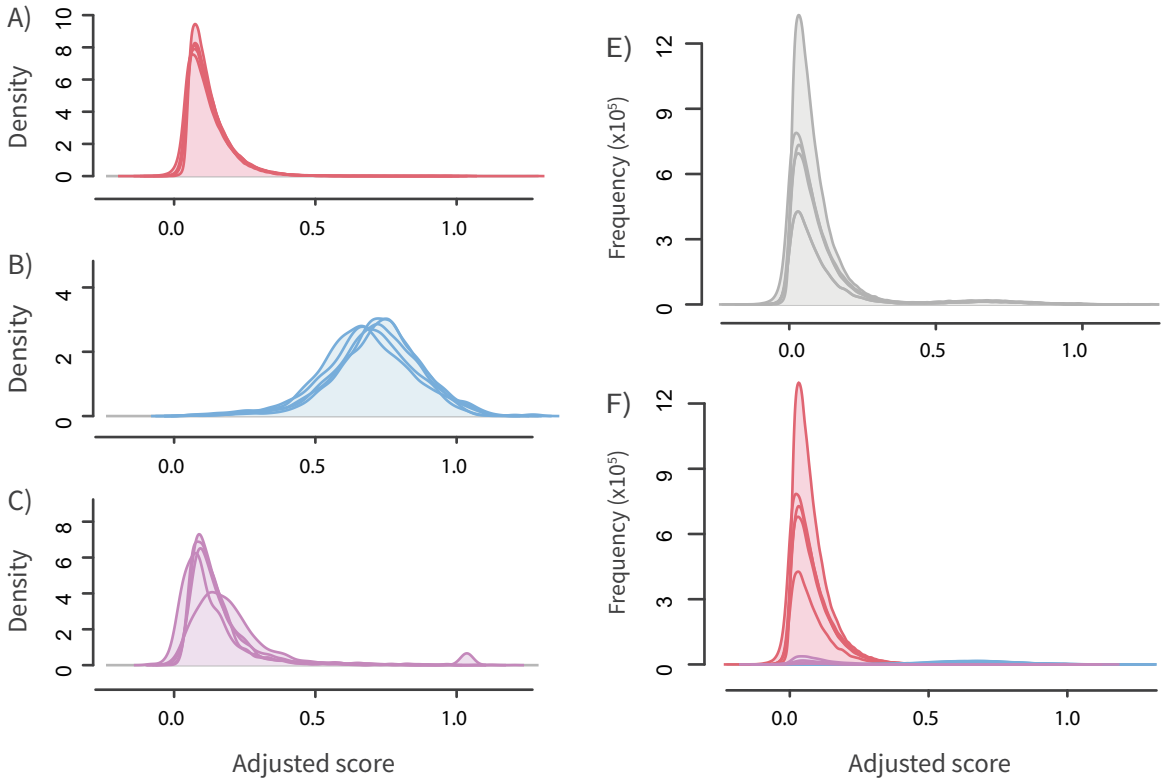


Figure 2.8: HMM hit distribution for fungal orthogroups. A-C) Off-target, on-target, and candidate hits respectively. The off-target and on-target hits follow markedly distinct distributions. Slight discrepancies for *E. gossypii* are caused by repetitive regions in the genome; E) Frequency distributions of all hits; F) Frequency distributions of each hit category, overlaid. The on-target and candidate hits are dwarfed by the off-target hits.

estimated, and a conditional threshold function is calculated incorporating these and the relative absolute probabilities of categorising a hit as on- or off-target. For any given species and set of hits, denote the estimated probability density functions for the adjusted scores of on- and off-target hits by p_{on} and p_{off} respectively. These distributions are estimated by fitting skew-T distributions to the on-target and off-target hit data (see Methods Section 2.7.6). For hits h and adjusted hit scores x, the distributions of true (T) and false (F) positive hits are then estimated using:

$$\mathbb{P}\left(x|T\right) \sim p_{\text{on}}; \quad \mathbb{P}\left(x|F\right) \sim p_{\text{off}}$$

$$\mathbb{P}\left(T\right) \sim \frac{|\{h : h \text{ on-target}\}|}{|\{h : h \text{ on- or off-target}\}|}; \quad \mathbb{P}\left(F\right) = 1 - \mathbb{P}\left(T\right)$$

These conditional distribtions are inverted using:

$$\mathbb{P}(T|x) = \frac{\mathbb{P}(x|T)\,\mathbb{P}(T)}{\mathbb{P}(x|T)\,\mathbb{P}(T) + \mathbb{P}(x|F)\,\mathbb{P}(F)}$$

$$\mathbb{P}(F|x) = \frac{\mathbb{P}(x|F)\,\mathbb{P}(F)}{\mathbb{P}(x|T)\,\mathbb{P}(T) + \mathbb{P}(x|F)\,\mathbb{P}(F)}$$

This provides a guide for finding the threshold value for each species to categorise candidate hits as either likely true or false positives based only on their adjusted bit scores. To adjust for behaviour at the tails of the distributions, where very high and very low scores can cause the result to flip, define the choice functions as

$$\tau(x) = [\mathbb{P}(T|x) < \mathbb{P}(F|x) \text{ or } x \leq \min(\{x : \mathbb{P}(T|x) \leq \mathbb{P}(F|x)\})]$$

$$\phi(x) = [\mathbb{P}(T|x) \geq \mathbb{P}(F|x) \text{ or } x \geq \max(\{x : \mathbb{P}(T|x) \geq \mathbb{P}(F|x)\})]$$

that is, any hit with a score higher than the highest true positive hit must also be classified as a true positive, and any hit with a score less than the lowest false positive hit must also be a false positive. Exactly one of ϕ and τ will be true, thus one can define the filtration criterion for an adjusted score x as pass if $\tau(x)$, and fail if $\phi(x)$.

2.4.4 A further filtering step removes spurious hits

Using this system to assess HMM hits to the plant and fungal data sets, a large number of candidate missing genic regions from each species under observation was identified, as listed in Table 2.8. There were a large number of genome hits for each plant species, and comparatively fewer hits to the fungal genomes. This is in part due to the fact that fungal genomes are far smaller, with far less intergenic material than the plant genomes and therefore less genomic material to search. Some species, however, appeared to have a considerable imbalance between the number of hits which passed the thresholding criteria described above and the number of orthogroups to which those hits belonged.

In many cases these multiple hits are due to a single HMM matching multiple regions of the same, biologically real gene, or to multiple paralogs of a given gene.

		Candidates that passed hit score analysis	
	Species	No. hits	No. orthogroups
Plants	*A. thaliana*	3212	1498
	B. rapa	13653	5330
	C. papaya	6008	2279
	C. rubella	5072	2038
	T. cacao	24530	1116
Fungi	*E. gossypii*	64	17
	D. hansenii	90	40
	K. lactis	42	30
	S. cerevisiae	16	13
	Y. lipolytica	78	62

Table 2.8: Number of accepted candidate hits to each genome, the number of merged hits, and an adjusted estimate of the number of missing genes. Many hits were obtained to plants genomes that passed the hit score analysis above and did not match to an existing gene. Fewer hits were obtained to fungal genomes. *T. cacao* has an unusually high number of hits considering how many orthogroups it has: this is due to ambiguous HMMs causing multiple erroneous but nevertheless high-scoring hits.

In some cases, however, this is due to single HMMs returning large numbers of spurious hits to multiple regions of the genome. In these cases, the hits are of sufficient quality to pass the threshold criteria, and the proliferation of hits is due to retrotransposons or repetitive regions of the HMM or target sequence. For example, one plant orthogroup already containing 41 paralogous genes in *T. cacao* (OG0000086 in this analysis) returned 9,360 threshold-passing hits to the *T. cacao* genome, significantly skewing the data. It is unlikely that these hits are real. A further filtering step is therefore required to remove these spurious hits.

Orthogroups that hit an unusually high number of genomic regions can be filtered out as follows. For each species, define L_{95} to be the 95^{th} percentile value of gene length, from start codon to stop codon, of all annotated genes. L_{95} values for each species are listed in Table 2.9. For each orthogroup O, expand each

			Accepted candidate hits	
	Species	L_{95}	No. hits	No. OGs
Plants	*A. thaliana*	4532	2444	1434
	B. rapa	4856	10150	4979
	C. papaya	7771	3675	2154
	C. rubella	4658	3742	1950
	T. cacao	8599	1627	995
Fungi	*E. gossypii*	3459	25	16
	D. hansenii	3417	71	38
	K. lactis	3360	42	30
	S. cerevisiae	3366	16	13
	Y. lipolytica	3537	78	62

Table 2.9: L_{95}-adjusted hit counts and orthogroup quantities. After filtering out obvious spurious hits, far fewer hits remain. Notably, large numbers of hits remain for *B. rapa*. This is due to high levels of historical genome duplication in this species. *T. cacao* has had very large numbers of hits removed. Many of the original hits to *T. cacao* appeared to be due to highly duplicated sequences that were unique to this species and represented in an orthogroup. GTF coordinates for all accepted candidate hits can be found in the Supplemental Material to this thesis on the Zenodo database, at `https://doi.org/10.5281/zenodo.1198798`, directory `ch2/2-passed_hits`, along with a list of orthogroups (`ch2/1-orthogroups`).

hit from its accompanying HMM by L_{95} bases in each direction, and merge all resulting overlapping regions (see Methods Section 2.7.1). This step is intended to consider together any separate hits that match to the same gene, which should be kept, and to flag up large regions of potentially spurious hits, which should be discarded. If any of the resulting regions are longer than $2 \cdot L_{95}$, reject all hits comprising those regions. Further, if the total number of genomic regions resulting from applying this process to hits from O is greater than 3, reject all hits from O. This allows for a small number of paralogs, but aims to protect against the proliferation of large numbers of potentially spurious hits.

Hits that passed both the hit score filtration step and the additional filtration step described above will be referred to as *accepted hits*. Table 2.9 shows the

number of accepted hits and orthogroups from each species. For *T. cacao* especially, the number of hits was considerably reduced by carrying out the additional filtering step. Thus, many more of the hits marked here are likely to be to real genes than before the additional filtration.

It is important to note here that this process is not immune to the identification of *pseudogenes*, and some accepted hits may well be just these. This may be the case for *B. rapa*, which has experienced a large number of genomic rearrangements, including several whole-genome duplication events, in its evolutionary history. These rearrangements are likely contribute to a higher number of pseudogenes. Many of the hits are, however, likely to be real: in particular, hits from orthogroups not already containing a gene for the hit species are strong candidates.

2.4.5 Many genomes show signs of annotation incompleteness

To form a robust estimate for the extent of missing gene annotations, the hits are reduced further to consider only hits to species not represented in their source orthogroup, that is, whose orthogroup does not already contain a gene from the genome that was hit. Table 2.10 shows the numbers of orthogroups of various types for which an accepted hit was found in a species not contained in the orthogroup. The categories are, respectively: all orthogroups for which the given species was missing; all single-copy orthogroups for which the species was missing; and all SCU-1 orthogroups for which the species was the one that was missing. The motivation for these latter categories is that single-copy genes typically show stronger signs of conservation than non-single-copy genes (Gu et al. 2004; Jiang et al. 2011; De Smet et al. 2013), and so provide good test sets for analysis. In particular, for SCU-1 orthogroups the gene is present in all four of the other species, and inclusion of a gene arising from the high-quality hit would yield an SCU orthogroup, indicating the highest level of evolutionary conservation.

For each category and for each species, define the completeness score as $\frac{H}{H+B}$, where H is the number of orthogroups not containing a representative for that species, and B is a category-specific background parameter. For the "All orthogroups" category, B is the number of orthogroups containing a representative for the species in question. For the "Single-copy orthogroups" category, B is the number of single-copy orthogroups containing a representative for the species in question. For the "SCU-1 OGs" category, B is the total number of SCU orthogroups for the species set in question. In each case, $B + H$ represents for each category the number of orthogroups that would result if the genes found here were added to each species' annotations. The scores therefore express what is annotated as a percentage of what *could be* annotated based on the evidence found here.

For all sets, the distribution across species is roughly consistent. Generally speaking, all of the scores are fairly high. Consistent with what was observed in Section 2.3, *C. papaya* has the lowest completeness score in all orthogroup categories. *B. rapa* and *C. rubella* also have lower scores, especially when considering only single copy orthogroups. The model species *A. thaliana* and *S. cerevisiae* both have high scores in all cases. Notably *S. cerevisiae* and *E. gossypii* both have perfect scores in all categories, despite each of them displaying some accepted hits. this is because the hits in question are from orthogroups already containing genes for these species, and so the identified hits are likely to be paralogs. *T. cacao* scores highly all round. This could indeed be real, or it could be an artefact: this species being an outgroup, meaning high quality hits are less likely. Furthermore, observation shows that *T. cacao* has a complex genome annotation, with many overlapping genes and large numbers of genes contained within introns. These large numbers of overlapping genes may have reduced the number of candidate hits. Finally, *D. hansenii* scores highly in all classes, despite having some candidate missing genes, and *K. lactis* and *Y. lipolytica* both score slightly lower in the *SCU-1* category.

		All OGs		Single-copy OGs		SCU-1 OGs	
	Species	No. OGs	Score	No. OGs	Score	No. OGs	Score
Plants	*A. thaliana*	85	99.5%	45	99.3%	6	99.9%
	B. rapa	142	99.1%	109	98.2%	23	99.7%
	C. papaya	481	96.5%	206	96.1%	170	97.8%
	C. rubella	276	98.4%	147	97.7%	35	99.5%
	T. cacao	56	99.6%	23	99.6%	8	99.9%
Fungi	*E. gossypii*	0	100.0%	0	100.0%	0	100.0%
	D. hansenii	5	99.9%	4	99.9%	3	99.3%
	K. lactis	14	99.7%	11	99.7%	10	97.7%
	S. cerevisiae	0	100.0%	0	100.0%	0	100.0%
	Y. lipolytica	7	99.8%	7	99.8%	6	98.6%

Table 2.10: Numbers of orthogroups with hits to species not already represented by the orthogroup, and completeness scores. Completeness scores are calculated as $\frac{H}{H+B}$, where H is the number of hit orthogroups and B is a background count, which is different for each category. For the "All orthogroups" category, B is the number of orthogroups containing a representative for the species in question. For the "Single-copy OGs" category, B is the number of single-copy orthogroups containing a representative for the species in question. For the "SCU-1 OGs" category, B is the total number of SCU orthogroups. Most species have good completeness with respect to these scores. *C. papaya* stands out as having a lower score, as does *K. lactis* for the fungi. GTF coordinates for all hits from each category can be found in the Supplementary Material at `https://doi.org/10.5281/zenodo.1198798`, directory `ch2/3-og_categories`.

It is likely that there are more missing genes than were found here, due to the stringent filtering criteria of the method used. However, this approach works as a basic indication. It revealed that most of the fungal species here are largely complete, but at least three of the plant species, *B. rapa*, *C. papaya*, and *C. rubella* all had lower scores, and thus show room for improvement.

2.5 Examples of missing genes identified using HMMs

High-fidelity hits to the genome, despite demonstrating sequence similarity, don't necessarily imply the existence of *bona fide* translatable genes. To demonstrate

briefly the genuine presence of viable genes arising from identified missing gene candidates, an example from each species was picked, by taking the genome hit and manually building a gene model around it, according to the canonical rules (ATG start codon, GT-AG splice junctions, TGA, TAA, or TAG stop codons). Example alignments from these manually constructed gene models can be seen below, in the context of some or all of the other genes from the orthogroup used to find them. Amino acid alignments and gene trees were constructed as described in Methods Sections 2.7.3 and 2.7.4. Where applicable, descriptions of *S. cerevisiae* and *A. thaliana* homologs have been taken from the Saccharomyces Genome Database (SGD) (Dwight et al. 2004) and The Arabidopsis Information Resource (TAIR) (Lamesch et al. 2012) respectively.

2.5.1 Missing members of plant orthogroups

Figures 2.9–2.13 show plant genes recovered using accepted HMM hits, for which gene models have been manually built, in the context of some or all of the other members of their orthogroup in a multiple sequence alignment. Plant gene names follow the patterns *AT...* (*A. thaliana*), *Brara...* (*B. rapa*), *evm.TU...* (*C. papaya*), *Carub...* (*C. rubella*) and *Thecc...* (*T. cacao*).

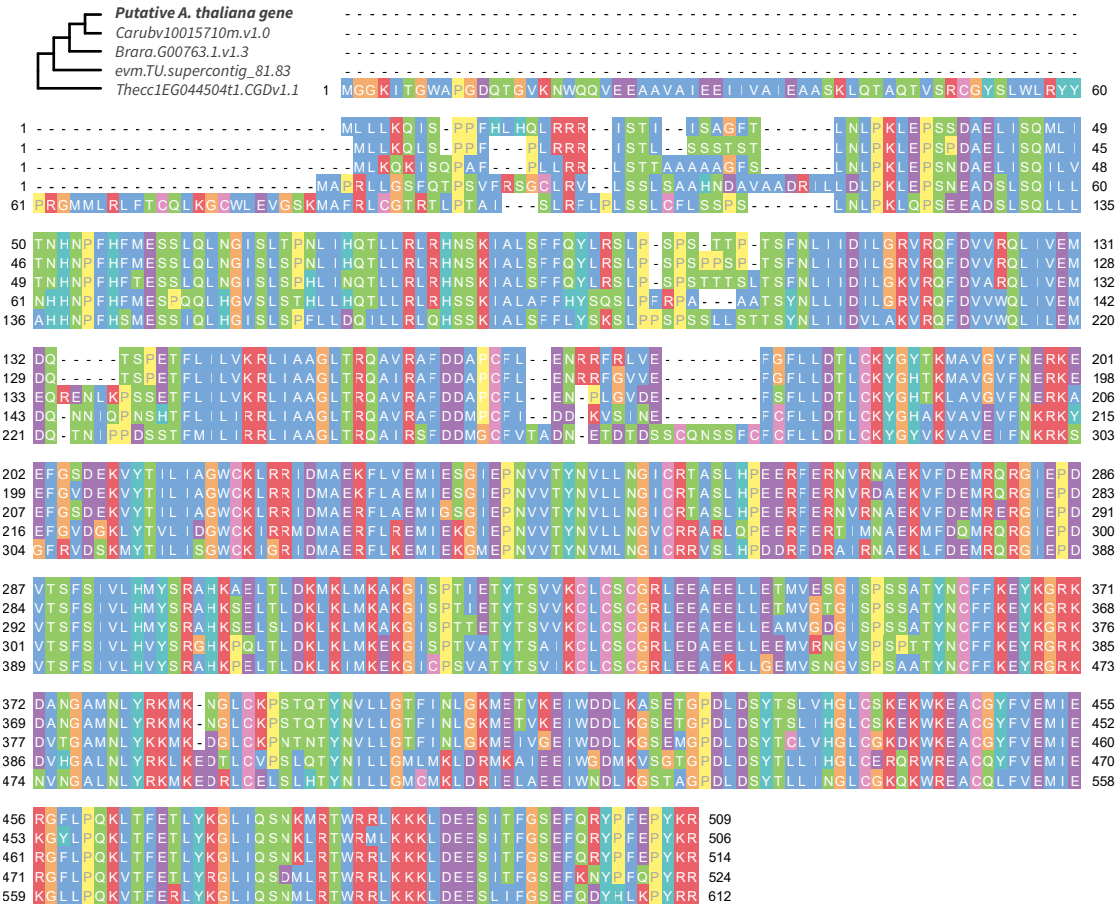


Figure 2.9: A missing gene from *A. thaliana*. This gene is located on chromosome 2, region 5586kbp–55881kbp, negative strand. The gene is almost identical to its homolog in *C. rubella*.

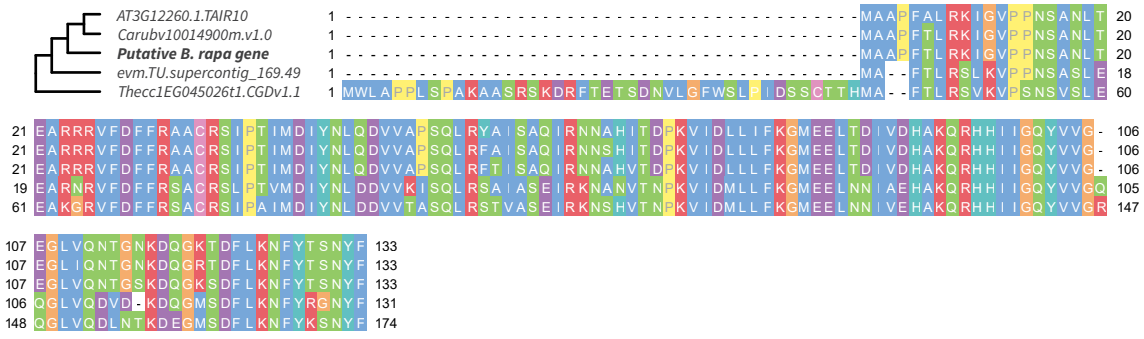


Figure 2.10: A missing gene from *B. rapa*. This gene is located on chromosome A03, region 17184kbp–17186kbp, negative strand. The gene is almost identical to its homolog in *C. rubella*. The *A. thaliana* homolog to this gene is in the LYR family of Fe/S cluster biogenesis proteins.

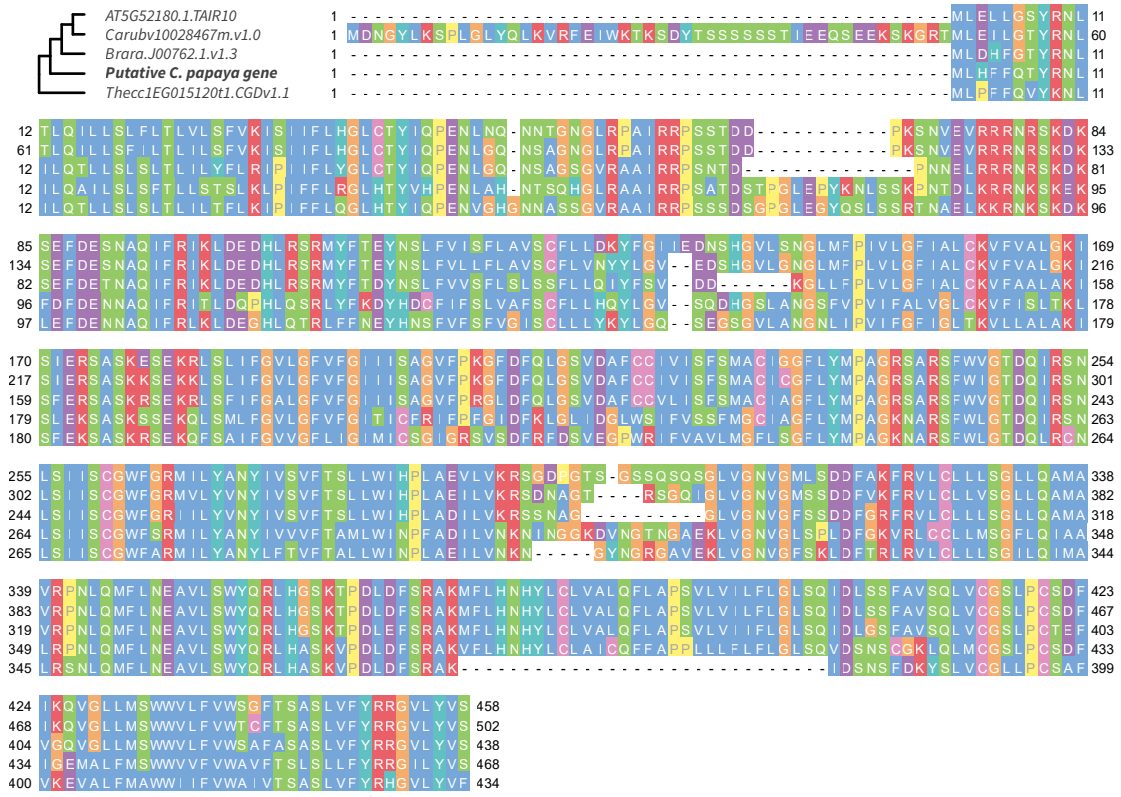


Figure 2.11: A missing gene from *C. papaya*. This gene is located on supercontig 107, region 159kbp–161kbp, negative strand. The gene is similar to all other genes in the orthogroup, and exhibits points of low conservation in the same regions as the others. The *A. thaliana* homolog to this gene is marked in TAIR as transmembrane protein 161AB.

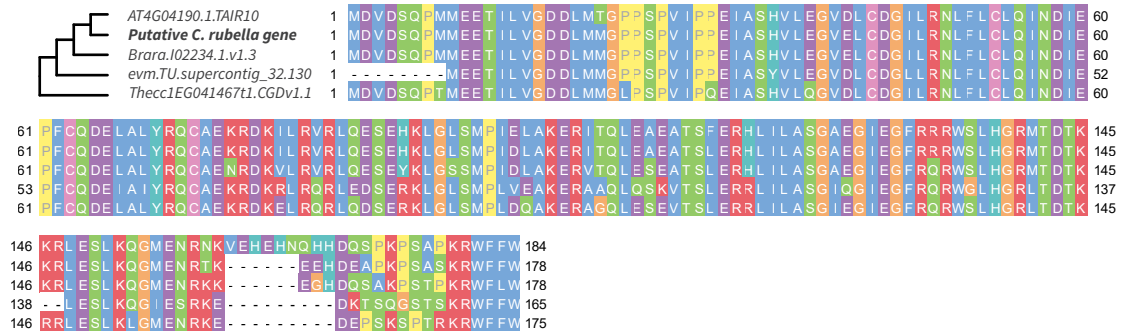


Figure 2.12: A missing gene from the *C. rubella* genome. This gene is located on scaffold 6, region 14552kbp–14554kbp. The gene is almost identical to its *B. rapa* homolog. The *A. thaliana homolog* to this gene is marked in TAIR as a transmembrane protein.



Figure 2.13: A putative missing gene in the *T. cacao* genome. This gene is located on scaffold 1, region 2201kbp–2204kbp, negative strand. Except for the 3′ end of the gene, sequence similarity is high compared with other genes in the orthogroup. The *A. thaliana* homolog to this gene is marked in TAIR as being in the pentatricopeptide repeat (PPR) superfamily.

2.5.2 Missing members of fungal orthogroups

Figures 2.14–2.18 show manually constructed gene models built around accepted HMM hits to fungal genomes, in the context of some or all of the other members of their orthogroup in a multiple sequence alignment. Fungal gene names follow the patterns *AGOS*...(*E. gossypii*), *DEHA*...(*D. hansenii*), *KLLA*...(*K. lactis*), *YXX###*...(*S. cerevisiae*, where *X* is a letter and # is a number) and *YALI*...(*Y. lipolytica*).

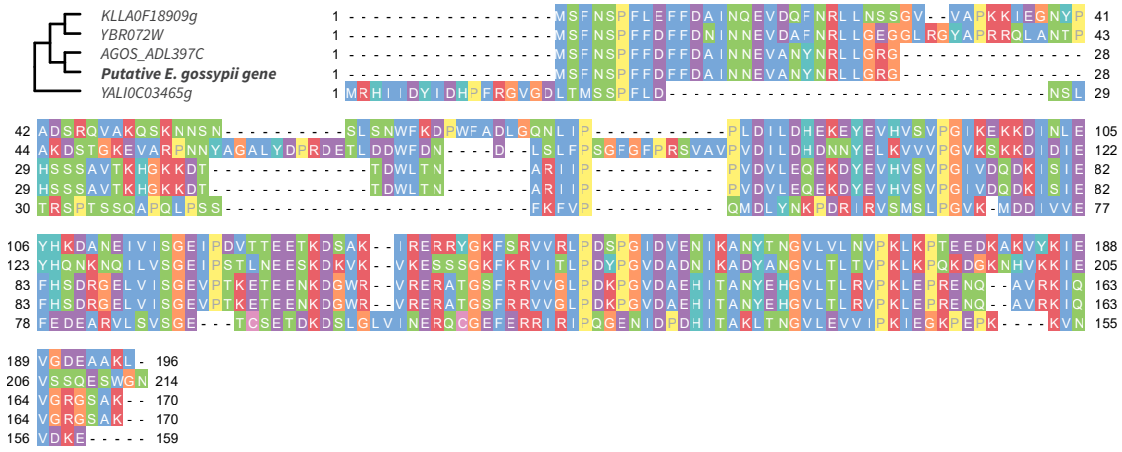


Figure 2.14: A missing gene from *E. gossypii*. This gene is located on chromosome 6, region 1820kbp–1821kbp, positive strand. The gene is a paralog to an existing *E. gossypii* annotation, *AGOS_ADL397C*, and is identical in sequence. The *S. cerevisiae* homolog to this gene is HSP26, a small heat shock protein with chaperone activity.



Figure 2.15: A short missing gene in *D. hansenii*. This gene is located on chromosome Deha2D_4, region 1282kbp–1284kbp, negative strand. The gene is almost identical to its homologs in *E. gossypii* and *Y. lipolytica*, but with an alanine substituted for an aspartic acid residue. The two *S. cerevisiae* homologs to this gene are RPL41A and RPL41B, both ribosomal 60S subunit proteins.



Figure 2.16: A missing gene from *K. lactis*. This gene is located on chromosome Klla0D_4, region 72kbp–73kbp, negative strand. The gene shows high sequence similarity to the other genes in the orthogroup. The *S. cerevisiae* homolog to this missing gene is ATP15, the Epsilon subunit of the F1 sector of mitochondrial F1F0 ATP synthase.



Figure 2.17: A missing gene in *S. cerevisiae*. This gene is located on chromosome XII, region 463kbp–465kbp, positive strand. The gene is a homolog of a gene that is present three times in the *K. lactis* genome, and which already exists once in another location in the *S. cerevisiae* annotation.

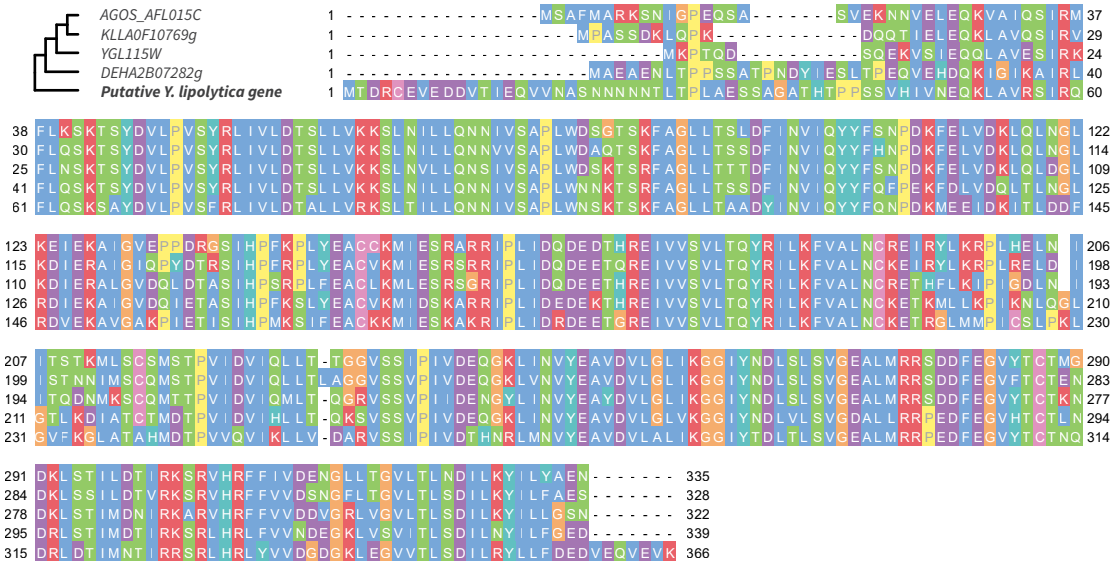


Figure 2.18: A gene missing from the *Y. lipolytica* genome. This gene is located on chromosome Yali0C_3, region 452kbp–454kbp, negative strand. The gene is very similar internally to other genes in the genome. The 3′ end is elongated in this new gene compared with the other genes in the orthogroup. The 5′ end of the orthogroup shows variation in all genes. The *S. cerevisiae* homolog to this gene is SNF4, the activating gamma subunit of the AMP-activated Snf1p kinase complex.

2.6 Comparison of candidate missing genes to RNA-seq data

The HMM hits obtained above already provide significant homology evidence for the presence of missing genes. To further support the evidence for these hits indicating genuine missing genes, the results were compared with publicly available RNA-seq data. RNA-seq is a valuable tool for assessing the presence of transcripts in cellular samples, and is now one of the most commonly used sources of information in locating and constructing gene sequences within a genome (Conesa et al. 2016). Genes that are expressed in a particular sample can be used to provide evidence for both the presence of a gene in the genome and for its intron/exon structure (Boley et al. 2014). As such, RNA-seq is used here to compare the validity of candidate HMM hits compared with the already annotated data.

2.6.1 Accepted hits have higher read coverage than rejected hits

RNA-seq reads were downloaded for each species, with the exception of *E. gossypii* and *D. hansenii*, for which adequate public data were unavailable (see Methods Section 2.7.7). Reads were mapped to each genome and mean read coverage was calculated for candidate hits and compared with read coverage for annotated regions (see Methods Section 2.7.8). For each species, the average read depth ($\bar{d}$) of all bases corresponding to accepted candidate hits was compared to the average read depth of all bases corresponding to rejected candidate hits. As a comparison, $\bar{d}$ was also calculated for both intergenic and genic annotated regions. The results of this analysis are shown in Table 2.11, and illustrated in Figure 2.19.

In all cases, the patterns of average read depth for accepted candidate hits versus rejected candidate hits were in line with the patterns observed for genic versus intergenic annotated regions. In particular, in all cases the $\bar{d}$ value for accepted candidate hits was considerably higher than that for rejected hits, validating the

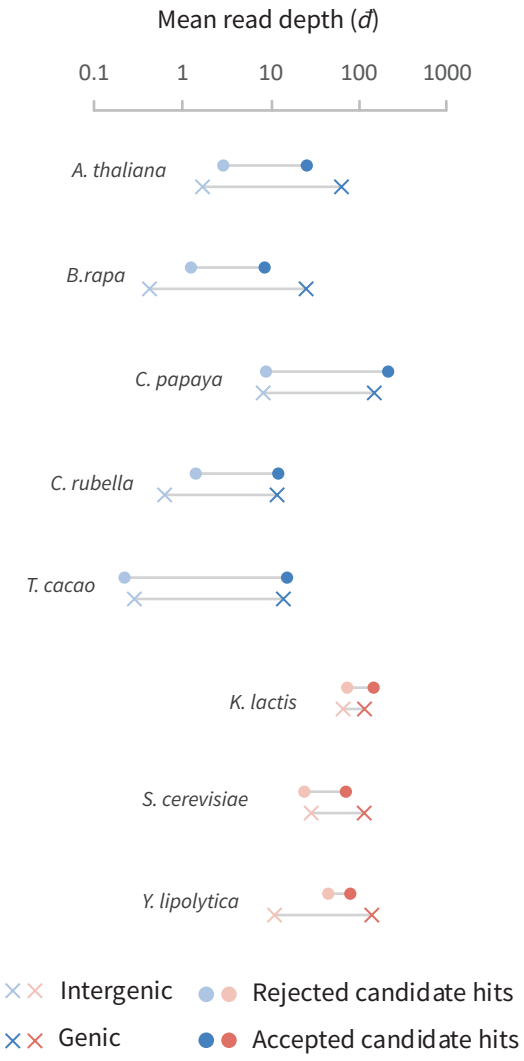


Figure 2.19: Accepted hits have higher read coverage than rejected hits. Crosses represent annotated genic and intergenic data: light crosses are intergenic and dark crosses are genic. Circles represent hits: rejected hits are light, accepted hits are dark. Rejected hits are closer in average read depth to intergenic regions, and accepted hits are closer in average read depth to genic regions.

		Mean coverage ($\bar{d}$)			
	Species	Intergenic	Genic	Rejected candidate hits	Accepted candidate hits
Plants	*A. thaliana*	1.7	64.8	2.9	25.7
	B. rapa	0.4	25.7	1.3	7.9
	C. papaya	8.4	154.7	9.2	225.9
	C. rubella	0.6	11.9	1.4	13.2
	T. cacao	0.3	14.3	0.2	15.6
Fungi	*K. lactis*	66.1	118.1	76.9	153.9
	S. cerevisiae	29.8	117.7	25.1	73.9
	Y. lipolytica	11.3	142.0	46.5	80.1

Table 2.11: Mean coverage ($\bar{d}$) for various regions of the genome: intergenic, genic, rejected candidate hits, and accepted candidate hits. Accepted hits are all those which passed the two filtering steps described in Section 2.4.3. In all cases, mean coverage was higher for accepted hits than rejected hits. Mean coverage for accepted hits was closer to what was observed for genic regions, and mean coverage for rejected hits was closer to what was observed for intergenic regions.

distinction between the two. In some cases, notably *A. thaliana* and *B. rapa*, $\bar{d}$ for accepted candidate hits was lower than for genic hits. This is unsurprising for two reasons: firstly, genes with strong transcriptional profiles are more likely to be annotated through traditional empirical methods; secondly, not all of the accepted candidate hits will end up being genes: some of them may be pseudogenes or other partial genes that are not highly expressed. In any case, the concordance between the coverage profiles of accepted/rejected hits versus genic/intergenic regions adds credence to the validity of the proposed hit set obtained in the previous section.

2.6.2 Accepted hits have higher read coverage than intergenic regions

As a further comparison, the proportion of hits which contained reasonable RNA-seq evidence was compared against simulated background data obtained from each intergenic genome region. For each hit, ten non-overlapping genomic regions of

		Randomly sampled intergenic regions			Accepted candidate hits		
	Species	Count	$\bar{d} > 0$	$\bar{d} > 10$	Count	$\bar{d} > 0$	$\bar{d} > 10$
Plants	*A. thaliana*	24440	1860 (7.6%)	189 (0.8%)	2444	699 (28.6%)	169 (6.9%)
	B. rapa	100400	5391 (5.4%)	561 (0.6%)	10040	3744 (37.3%)	1361 (13.6%)
	C. papaya	36750	5399 (14.7%)	1798 (4.9%)	3675	2977 (81.0%)	2539 (69.1%)
	C. rubella	37420	1929 (5.2%)	213 (0.6%)	3742	1368 (36.6%)	582 (15.6%)
	T. cacao	16270	505 (3.1%)	27 (0.2%)	1627	394 (24.2%)	74 (4.5%)
Fungi	*K. lactis*	420	397 (94.5%)	167 (39.8%)	42	39 (92.9%)	33 (78.6%)
	S. cerevisiae	160	147 (91.9%)	62 (38.8%)	16	15 (93.8%)	7 (43.8%)
	Y. lipolytica	780	524 (67.2%)	89 (11.4%)	78	64 (82.1%)	26 (33.3%)

Table 2.12: Coverage data for hits and for simulated intergenic hits of the same length. In all cases for the plants, strong read evidence is much more common for accepted hits than for the background. In fungi, the distinction is less clear. This is partly due to much of the intergenic regions being taken up by transcribed UTRs, which are not marked in the fungal annotations and so will be counted as "intergenic".

the same length were drawn at random from the portion of the genome labelled as intergenic. For all real and simulated hits, the proportion with any reads and the proportion with reasonable ($\bar{d} > 10$) read evidence were obtained. The results can be seen in Table 2.12 and Figure 2.20. For all plant species, the proportion of accepted hits with any read evidence was much higher than the background. For the fungi the numbers were roughly the same for candidate hits and for the sampled intergenic hits. When considering strong read evidence, however, all species exhibited a much higher proportion of well-supported real hits when compared to the simulated data. Again, for most species the percentage of hits with reasonable read depth was low, tying in with the fact that highly expressed genes are much more likely to have already been annotated. Of note is the proportion of *C. papaya* hits with $\bar{d} > 10$, at 69.1%. This is in line with *C. papaya* coming top of the list for the above assessment of missing genes, indicating that a large number of the hits have very strong experimental and homology evidence.

The clear difference between RNA-seq read data for accepted hits against the

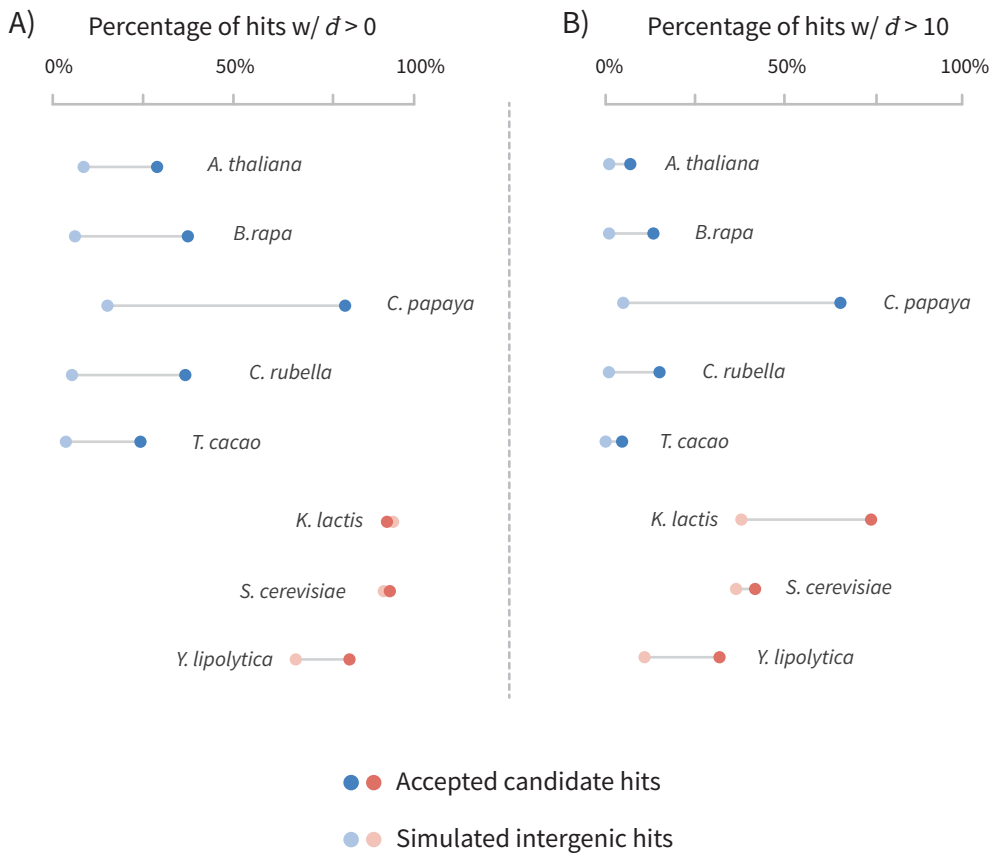


Figure 2.20: Accepted hits have higher coverage than simulated intergenic hits. Coverage distributions for accepted candidate hits versus sampled intergenic hits. A) Percentages of hits with any reads; B) Percentages of hits with $\bar{d} > 10$. In all cases, more accepted candidate hits have read evidence than their simulated intergenic counterparts. This is especially apparent when looking at hits with an average read depth of at least 10.

background demonstrates that many of these candidate genic regions are potentially bona fide genes. The read evidence is however in some cases lower than for most of the already-annotated proteome data, suggesting further considerations must be made in order to truly assert whether or not these findings represent missing genes. Not all accepted hits from above are well supported by RNA-seq data. This could be because the hits come from genes that are not expressed, or are not highly expressed, in the particular sample used, something which is more likely for unannotated genes than for annotated genes, as noted above. In many cases, however, this will be because further evidence needs to be used to assess exactly

whether or not a gene can be deemed to be "real". Despite these uncertainties, the clear differences in read coverage between accepted and rejected candidate hits indicate their worth as potential missing genes.

2.7 Methods

2.7.1 Genome algebra and feature file manipulation

Most genome algebra (intersection, subtraction, merging of genomic regions) was carried out using BedTools version 2.25.0 (Quinlan and Hall 2010), along with various combinations of the standard Unix commands `awk`, `grep`, `sed`, and `sort`, as well as Python version 2.7.12.

2.7.2 Orthogroup calculation

Orthogroups for each complete set of annotated protein-coding genes were calculated using OrthoFinder (Emms and Kelly 2015) version 2.1.2. For plant genes with multiple transcripts, only the primary transcript, as listed in Phytozome (Goodstein et al. 2012), was used for orthogroup inference.

2.7.3 Acquisition of phylogenetic trees

Species trees were calculated using FastTree (Price et al. 2009; Price et al. 2010) version 2.1.9, run with the default parameters on concatenated single-copy ubiquitous orthogroup amino acid FASTA files outputted from OrthoFinder (Emms and Kelly 2015). The roots and outgroups were taken from Phytozome (Goodstein et al. 2012) and Mycocosm (Grigoriev et al. 2014) for plants and fungi respectively. Individual orthogroup gene trees were calculated using FastTree on orthogroup amino acid sequences aligned with MAFFT (see Methods Section 2.7.4).

2.7.4 Alignment of protein and nucleotide sequences

Alignments were carried out on sets of protein amino acid sequences using MAFFT L-INS-i (Katoh and Standley 2013) with the default options. Nucleotide alignments

were obtained by back-translating the amino acid sequences in the alignment according to their underlying coding sequences. Aligning in this way, i.e. by amino acid sequence first, increases the robustness of the nucleotide alignment under synonymous mutations and thus provides higher alignment quality (Wernersson and Pedersen 2003).

2.7.5 Acquisition and implementation of HMMs

For each orthogroup, an HMM was formed from a nucleotide alignment of its constituent sequences (see above). HMMs were formed using Hmmer (Eddy 2011) version 3.1b2, using the command `hmmbuild --informat afa`. Genome databases were constructed using `makehmmerdb --blocksize=10`, and the HMMs were run on these databases using `nhmmer --tformat hmmerfm --dna --cpu 1 --tblout`.

2.7.6 Score distribution estimation

Probability density functions for distributions of adjusted scores were estimated using GAMLSS (Rigby and Stasinopoulos 2005) version 5.0-1 in R version 3.4.1. The density functions fitted were skew-T distributions, specified using `family="ST1"`. The other options specified to GAMLSS were `method=RS()`, `gd.tol=10000000`, `c.cyc=0.001`, and `control=gamlss.control(n.cyc=200)`.

2.7.7 RNA-seq data sources

RNA-seq data were downloaded from the sequence read archive (SRA) (Leinonen et al. 2010), as listed in Table 2.13. Data were downloaded for all species with the exception of two fungal species. There were no public data available for *E. gossypii*. For *D. hansenii*, although public data were available, the reads displayed a very low alignment rate (5.16% of reads compared with an average 92.8% for the other species). It is assumed therefore that these sequencing data are contaminated or are otherwise unsuitable for use.

	Species	SRA ID	Instrument/details
Plants	*A. thaliana*	SRR3932355	Illumina HiSeq 2500, paired end. Wild type Columbia rep1.
	B. rapa	SRR2984945	Illumina HiSeq 2000, paired end. ga-deficient dwarf (gad1-2) +GA rep2.
	C. papaya	SRR3509576	Illumina HiSeq 2500, paired end. SunUp/Sunset cultivars, young hermaphrodite leaf.
	C. rubella	SRR3993756	Illumina HiSeq 2000, paired end. Leaf sample.
	T. cacao	SRR3217315	Illumina HiSeq 2000, paired end. Flower/leaf sample.
Fungi	*K. lactis*	SRR1200528	Illumina Genome Analyzer II, single.
	S. cerevisiae	SRR3396393	Illumina HiSeq 2000, paired end.
	Y. lipolytica	SRR3647200	Illumina HiSeq 2000, paired end.

Table 2.13: Sources of RNA-seq data used to assess hits. All data were downloaded from the Sequence Read Archive (Leinonen et al. 2010). In the case of the plant data, leaf samples were used preferentially where marked.

2.7.8 Alignment and quantification of RNA-seq data

Reads were assessed for quality using FastQC (Andrews 2010), trimmed using Seqtk (Li 2012), and mapped to the genomes using HISAT2 (Kim et al. 2015; Kim et al. 2017) using the default parameters and unpaired or paired-end methods as appropriate. Per-base read depth was calculated for each entire genome using Samtools `mpileup` (Li et al. 2009) with the default parameters. For individual features (genes, coding regions, hits), per-base read depth was obtained by intersecting the global pileup with each individual feature file, using BedTools `intersect` (see Methods Section 2.7.1), and these files were processed using `awk` to obtain average read depth statistics.

2.8 Chapter discussion

In this chapter I introduced two biologically distinct species sets and discussed the properties of their genomes, protein-coding genes, and orthogroups. I then defined and implemented an HMM-based system for assessing genome completeness, using a novel scoring and filtering system. I then proceeded to demonstrate examples of

some valid gene models based around HMM hits, in the context of their orthogroups, and further showed that the HMM hits typically had higher read coverage than intergenic regions and hits which did not pass the filtration criteria applied.

2.8.1 Different species sets display markedly different genome characteristics

The two species sets introduced here, a set of fungal species and a set of plant species, are very different in terms of their size, gene distribution, and typical gene structure. Fungal genomes are small, with highly packed genes and few introns; plant genomes are large, with large intergenic regions and large introns. In addition, the plant species have annotated alternate transcripts for many of their genes, with the notable exception of *C. papaya*. Given that all plant genomes undergo alternative splicing, this is a clue that *C. papaya* may have been afforded less focussed attention for its annotation than the other species.

Unlike the fungi, all of the plant species described here, with the exception of *Arabidopsis thaliana*, have highly unfinished assemblies consisting of many disconnected contigs and scaffolds. The quality of a genome assembly is not necessarily correlated with the quality of its annotation, however genomes with more fragmented assemblies are *a priori* more difficult to annotate. The more complex nature of plant genes compared with fungal genes, coupled with the lesser quality of the assemblies for these species, impacts the success of gene finders on these species. In particular, the task is necessarily going to be harder for the plant genomes, and so one would expect a higher incidence of genome annotation errors.

2.8.2 Incomplete orthogroups can indicate annotation issues

In Section 2.3.3 I described in detail the confounding factors that prevent missing gene estimates from being obtained using absent orthogroup members alone. I noted that many absent orthogroup members can be explained by gene loss and gene gain events, and demonstrated that certain more likely gene loss and gain

events are represented strongly in the distributions of subsets of species found in the orthogroups I calculated. However, I noted that particularly strong missing gene signals could still be observed using this method, and in Section 2.3.4 I considered the number of SCU-1 orthogroups with each species missing. *C. papaya* was missing from a large number of SCU-1 orthogroups, and, given the position of this species in the phylogenetic species tree, this could not easily be explained by gene loss and gain events. The same could not be said for *Y. lipolytica* in fungi, since this is an outgroup with enough distance from the other species to explain these absences through gene loss and gain events. Thus, further evidence was obtained that *C. papaya* is under-annotated, but this method is extremely limited and can only be applied in rare cases.

2.8.3 Filtered HMM searches inform incompleteness estimates

In order to test more directly for the presence of missing genes, I had to actively look for them. Orthogroup amino acid alignments were used to form HMMs, which were used to search genomes for traces of missing genes. Hits to known genes were used to build filtration criteria for these hits, and a further filtration criterion to remove hits to and from repetitive regions was implemented. This resulted in a large number of high-confidence genome hits, representing a large number of orthogroups. Many of these are paralogs to already annotated genes: in particular *B. rapa* had the highest number of accepted hits. Since *B. rapa* has undergone many whole genome duplication events, some of these are likely to be pseudogenes.

To try to provide an estimate of annotation completeness, hits from orthogroups to species not already represented in the given orthogroup were assessed. For three categories of orthogroup, incompleteness scores were found for each genome. In line with previous observations, *C. papaya* had the lowest completeness score in all cases, with its lowest score being in the single-copy category, at 96.1%. The model species

A. thaliana and *S. cerevisiae* had the highest scores in almost all cases. Most fungal genomes had high scores, although some indications of missing genes were observed. *C. rubella* and *B. rapa* had slightly lower scores, at 97.7% and 98.2% completeness respectively considering single-copy orthogroups. Two fungal species, *S. cerevisiae* and *E. gossypii*, received perfect scores. All other scores were generally high, with at least 95% for each score for each species. If these completeness estimates are true, however, this still implies a large number of missing genes. For example, if we extrapolate the completeness score of 96.1% for *C. papaya* to the whole annotated proteome, this would equate to 1,076 unannotated genes. Thus, although the scores are high, the large size of the proteomes in question means that even small improvements of completeness percentage can make a significant difference.

Two of the species here, *A. thaliana* and *C. rubella*, had been assessed previously (Veeckman et al. 2016). In the paper, the authors find that both of these species had among the highest levels of annotation completeness assessed, although *C. rubella* had more missing core gene families than *A. thaliana*. Given that the scores used are different to the ones here, a direct comparison is difficult. However, they can be qualitatively compared. The observations of Section 2.4.5 revealed *A. thaliana* to be very well annotated, in line with what was observed in the Veeckman paper. Similarly, *C. rubella* showed lower levels of annotation than *A. thaliana*, but better annotation levels than the worst-performing *C. papaya*.

2.8.4 RNA-seq evidence supports putative missing genes

To demonstrate the capacity of the filtered hits to find real genes, I manually constructed gene models around several of the hits found, demonstrating high agreement with the other genes in the orthogroup in relation to a multiple sequence alignment. In addition, I demonstrated that, in all cases, hits which passed the filtration criteria had much greater RNA-seq read coverage than both rejected hits and simulated hits to intergenic regions. This further demonstrates the capability

of this system to find *bona fide* genes. In all cases, although the levels were above the background, the number of genes with higher read coverage $\bar{d} > 10$ was low. Given that highly expressed genes are more likely to have been annotated already, this is not unexpected.

2.8.5 Summary

This chapter discussed an approach to assessing genome annotation completeness by searching for missing genes using HMMs. High-quality genome hits from conserved gene families can indicate the prevalence of missing gene type errors, though it is highly likely that they underestimate the extent. In particular, they can offer an indication of the *relative* prevalence of missing gene type errors across a set of species, which can be summarised using a completeness score. As mentioned, however, it is still possible that some of the hits found using the method will be pseudogenes, or will be spurious predictions that have managed to fall through the net. The only way to truly assess the ability of this method to identify missing genes is to build and validate accurate gene models around candidate hits, and confirm the agreement of these gene models with the other genes in their orthogroup.

3 Quantifying gene model errors

3.1 Chapter introduction

This chapter considers the problem of assessing and quantifying the extent of *inaccurate gene models*, again in the specific context of small groups of closely related species, and in the absence of RNA-seq evidence or other experimental data. The introduction to this thesis covers the important points explaining and motivating this chapter. However, as in Chapter 2, the most relevant points are briefly summarised in Section 3.1.1 to enable the reading of this chapter as a stand-alone entity.

3.1.1 Background

Even when a genomic region is correctly labelled as containing a gene, there is still huge scope for error in the precise coordinates and structure of the annotated gene models for that gene. Most intron-exon structures are constructed either manually, from comparison with RNA-seq or EST sequence data, or computationally, by considering intrinsic characteristics such as GC content and polypyrimidine tracts, as well as canonical splice junction markers such as GT splice donors and AG splice acceptors (Yandell and Ence 2012). These signals are often ambiguous, leading to many possibilities for annotation errors, especially for complex gene

models comprising large numbers of exons and introns spanning many thousands of nucleotides.

All analyses that involve genome-derived protein-coding genes can be negatively affected by inaccurate gene models. For example the structure of the gene model can have a considerable impact on inferences such as gene expression levels (Zhao and Zhang 2015), molecular evolution rate analysis (Moretti et al. 2014), ortholog identification (Sjölander et al. 2011) and domain rearrangement analysis (Wu et al. 2012). Furthermore, gene model errors in one genome can be propagated in downstream analyses, compounding the errors both laterally, via cross-species gene annotation, and vertically, by building successive inferences on false information. For example, the human genome is often used as an annotation guide for higher Eukaryotes, and thus errors in the human genome have been repeatedly propagated to other species' genome annotations (Schnoes et al. 2009).

Methods to comprehensively and automatically assess the quality of gene model annotations have not received a great deal of attention. This is in contrast to genome assembly quality metrics, which have been extensively studied (Gurevich et al. 2013; Rahman and Pachter 2013), and genome annotation completeness metrics, which were discussed in Chapter 2 are beginning to gather ground (Veeckman et al. 2016). Similarly to Chapter 2, both extrinsic methods relying on experimental data and non-extrinsic methods relying on homology assessments are valid, but have different use cases depending on the data available. Like in Chapter 2, assessments relying on highly expressed or universally conserved genes have a degree of circularity, and are therefore limited. In this chapter I explore several different metrics for quantifying gene model errors, and use them to assess the species introduced in Chapter 2. In particular, I introduce a phylogenetically local homology-based method for identifying gene model errors via identification of discrepant regions in multiple sequence alignments of orthogroups. I demonstrate the concordance

between this and extrinsic gene model assessment methods, indicating that it can be used as a measure of overall gene model accuracy.

3.1.2 Chapter aims and outline

In this chapter I outline several metrics for assessing the validity of gene models, both by comparison to RNA-seq data and from orthogroup-based, phylogenetically local perspectives. These metrics are designed to be used genome-wide, both to serve as comparative and intrinsic measures for the overall quality of genome annotations, as well as to highlight potential candidate errors for correction. The general aim of the chapter is to demonstrate the validity of the claim that gene model errors can be identified and subsequently corrected by looking for discrepancies in multiple sequence alignments of orthogroups, rather than relying on extrinsic empirical evidence such as RNA-seq or EST sequence data. The extrinsic methods described in the chapter are introduced in order to assess the non-extrinsic method. Briefly, the structure of the chapter is as follows:

- **Section 3.1: Chapter introduction**.
- **Section 3.2: Gene model errors come in several flavours**. In this section I highlight several different categories of gene model error, and provide examples of these errors in the context of RNA-seq evidence and amino acid alignments.
- **Section 3.3: Extrinsic metrics reveal gene model errors**. Here I introduce two extrinsic scoring metrics, which will be used in the remainder of the chapter to assess non-extrinsic methods.
- **Section 3.4: A non-extrinsic metric for gene model assessment**. The main result of the chapter. Here I introduce a non-extrinsic method to highlight potential gene errors by looking for discrepant regions in orthogroup amino acid alignments.

- **Section 3.5: Subcellular localisation predictions indicate start codon accuracy**. Here I discuss a method to assess start codon choice by appealing to the consistency of subcellular localisation predictions, and evaluate this with respect to the non-extrinsic assessment method I introduced earlier in the chapter.
- **Section 3.6: Gene model scores can inform primary transcript choice**. Finally, I evaluate the plant transcripts marked as *primary transcripts* with respect to the assessment metrics introduced in this chapter, and demonstrate that many of the choices of primary transcript are suboptimal with respect to these metrics.
- **Section 3.7: Methods**.
- **Section 3.8: Chapter discussion**.

3.2 Gene model errors come in several flavours

This section describes several categories of gene model errors, based on the nature of the disparity between what is annotated and what is more likely to be the "true" gene model based on aligned RNA-seq evidence and comparison to orthogroup multiple sequence alignements. An example for each error type is given, with both the orthogroup alignment and the gene's RNA-seq profile provided. All figures described examine genes that have been annotated as having only one transcript variant in the genome annotation file, unless explicitly stated. Thus, no evidence is provided in the genome annotation that alternative exon boundaries may be present. Negative-strand genes have been flipped to read left-to-right, matching the accompanying alignments.

3.2.1 Inaccurate exon boundaries

An inaccurate exon boundary occurs when either the splice donor or acceptor site for an intron is marked in the wrong place. This can involve either the shortening or lengthening of exons, and in some cases can affect the frame choice of downstream exons. Figures 3.1 and 3.2 show two examples of this kind of annotation error.

Figure 3.1 shows a *T. cacao gene* (*Thecc1EG000812t1.CGDv1.1*) for which the splice donor site has been annotated downstream of the site that is supported by RNA-seq data. This has resulted in an extended exon. This can be seen clearly in both the spliced aligned read profile for the transcript, and the context of the translated protein within the orthogroup. Below the original gene model is a corrected version, using an earlier splice donor site that is consistent with the RNA-seq read data. The amino acid sequence corresponding to the original gene model causes gaps to be formed in the multiple sequence alignment: the new gene model matches the other sequences in the alignment and requires no gaps.

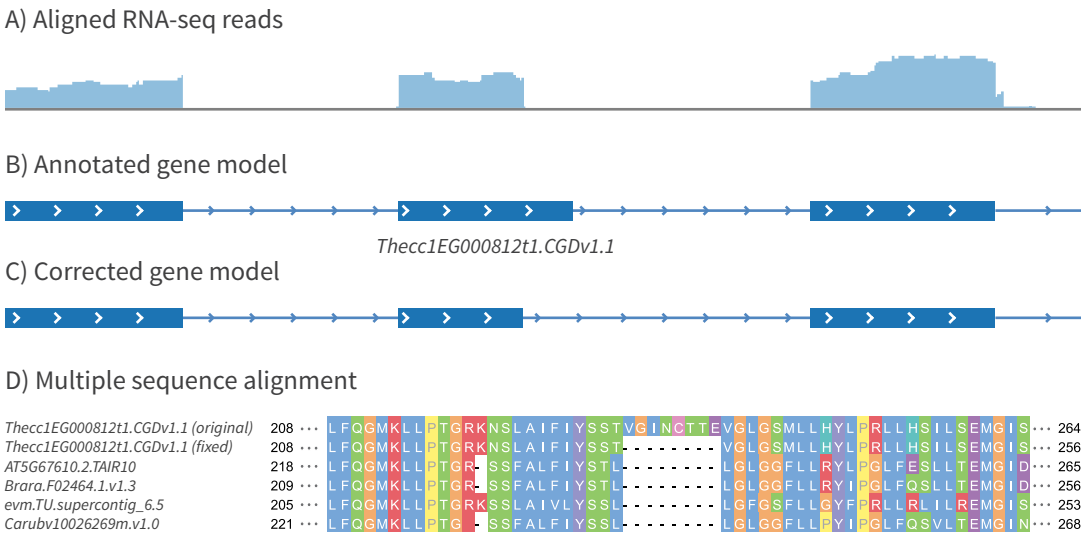


Figure 3.1: Example of extended exon type error in gene *Thecc1EG000812t1.CGDv1.1* from *T. cacao*. A) RNA-seq coverage from spliced-aligned RNA-seq reads; B) The gene model present in the current genome annotation; C) A manually corrected gene model that matches the RNA-seq data; D) A multiple sequence alignment of the corresponding amino acids sequences containing the original annotated gene (original) and the manually corrected gene model (fixed) as well as other genes in the orthogroup.

Figure 3.2 shows a *C. rubella* gene (*Carubv10018017m.v1.0*) for which a splice acceptor site has been annotated downstream of its true location. This site is internal to the "true" exon resulting in a shortened exon lacking several amino acids. Using an earlier splice acceptor site that is consistent with the RNA-seq read data produces an alternative exon that is longer and more consistent with the other sequences in the orthogroup.

3.2.2 Exon omission/inclusion errors

In addition to errors in the specification of exon termini, exons can be omitted completely. Figure 3.3 shows an *A. thaliana* gene (*AT1G22860.1.TAIR10*) for which an exon has been omitted in the gene model annotation, but which can be clearly seen in both the alignment of the orthogroup and the aligned RNA-seq data. Again, an alternative version of the gene that is constructed according to the RNA-seq evidence recovers the consistency of the alignment across all species.

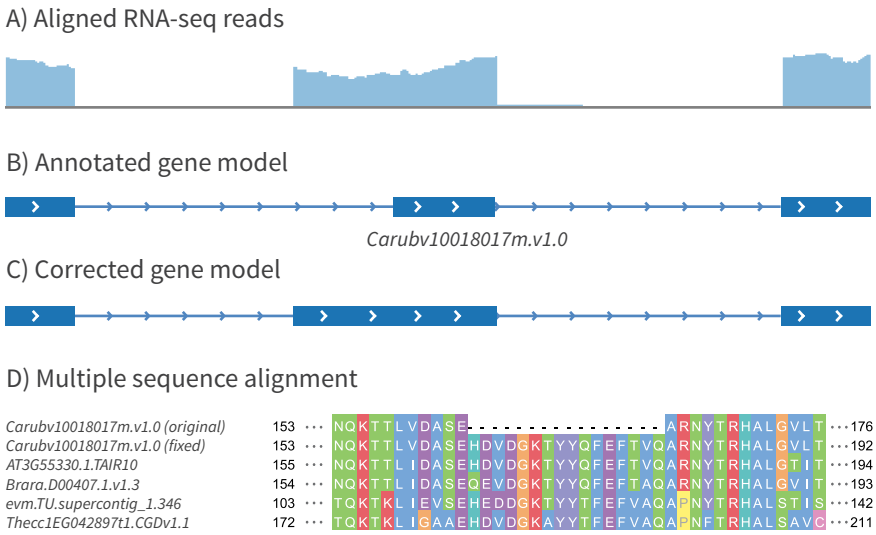


Figure 3.2: Example of a contracted exon type error in gene *Carubv10018017m.v1.0* from *C. rubella*. A) RNA-seq coverage from spliced-aligned RNA-seq reads; B) The gene model present in the current genome annotation; C) A manually corrected gene model that matches the RNA-seq data; D) A multiple sequence alignment of the corresponding amino acids sequences containing the original annotated gene (original) and the manually corrected gene model (fixed) as well as other genes in the orthogroup.

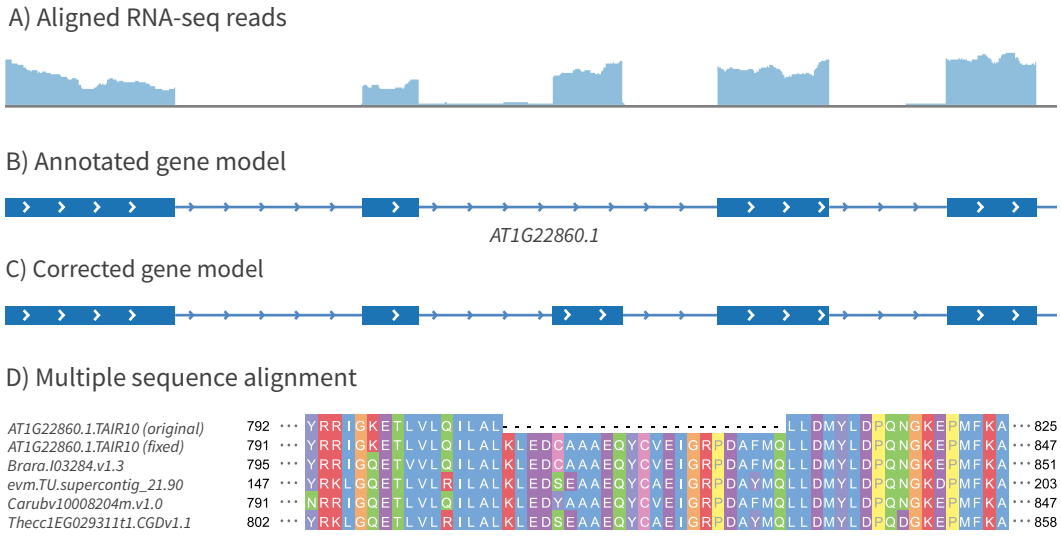


Figure 3.3: Example of an omitted exon. In the annotation for *A. thaliana* gene *AT1G22860.1.TAIR10*, an entire exon appears to be missing when compared with RNA-seq data. A) RNA-seq coverage from spliced-aligned RNA-seq reads; B) The gene model present in the current genome annotation; C) A manually corrected gene model that matches the RNA-seq data; D) A multiple sequence alignment of the corresponding amino acids sequences containing the original annotated gene (original) and the manually corrected gene model (fixed) as well as other genes in the orthogroup.

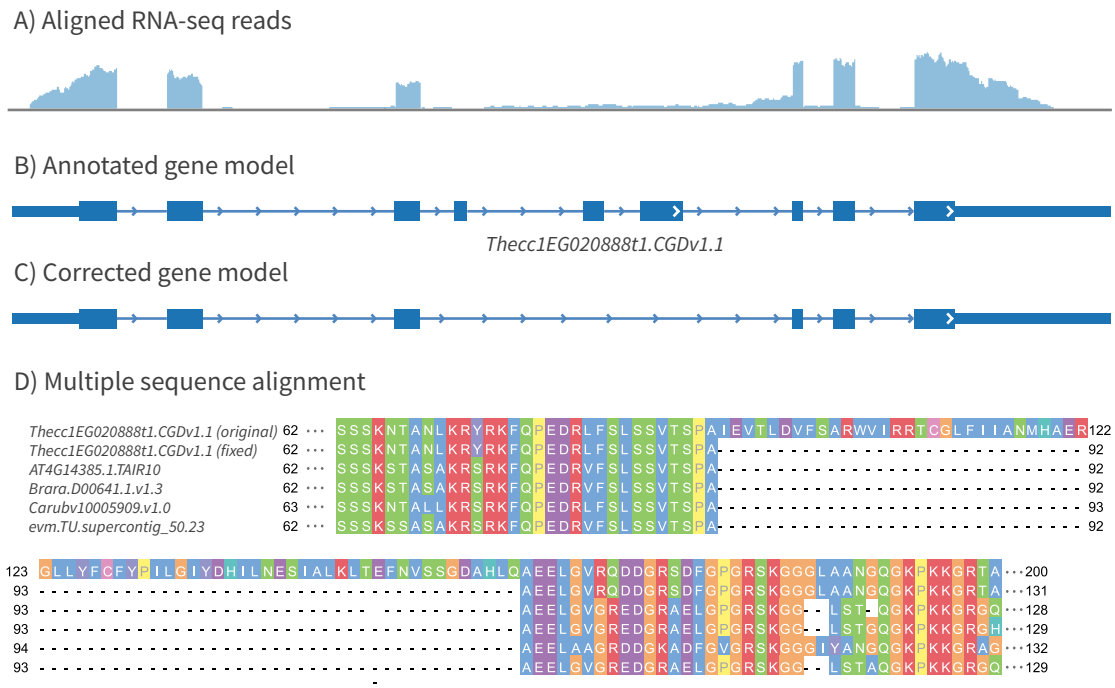


Figure 3.4: Example of extraneous exons in the annotated *T. cacao* gene *Thecc1EG020888t1.CGDv1.1*. A) RNA-seq coverage from spliced-aligned RNA-seq reads; B) The gene model present in the current genome annotation; C) A manually corrected gene model that matches the RNA-seq data; D) A multiple sequence alignment of the corresponding amino acids sequences containing the original annotated gene (original) and the manually corrected gene model (fixed) as well as other genes in the orthogroup.

Figure 3.4 shows a *T. cacao* gene (*Thecc1EG020888t1.CGDv1.1*) for which several exons have been annotated where none appear to be present in the RNA-seq data. Some reads are present in the intron but at a much lower level than the other exonic reads, and none indicate a splice junction. This occurrence of RNA-seq reads in the intron region of the gene is frequently observed and is consistent with detection of nuclear or unprocessed transcripts in which the intron is yet to be spliced. The multiple sequence alignment shows that these additional exons correspond to a stretch of amino acids not represented in the orthologous genes in the other species. Removal of these extraneous exons in line with the RNA-seq evidence recovers the consistency of the alignment.

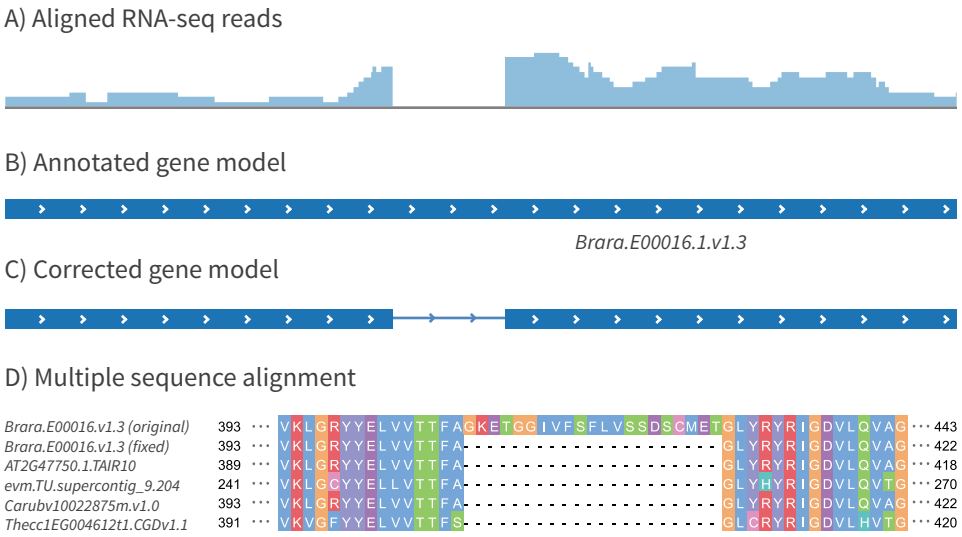


Figure 3.5: Example of a missing intron. Low-coverage RNA-seq data demonstrate the clear presence of an unannotated intron in this *B. rapa* gene, *Brara.E00016.1.v1.3*, however read coverage is low. Fidelity of RNA-seq data can greatly impact its ability to elucidate gene models, a problem which is not present when looking at orthogroups. A) RNA-seq coverage from spliced-aligned RNA-seq reads; B) The gene model present in the current genome annotation; C) A manually corrected gene model that matches the RNA-seq data; D) A multiple sequence alignment of the corresponding amino acids sequences containing the original annotated gene (original) and the manually corrected gene model (fixed) as well as other genes in the orthogroup.

3.2.3 Intron omission/inclusion errors

In addition to errors in the specification of exon boundaries and the spurious omission or inclusion of exons, introns can be omitted or included erroneously. In some cases it is possible to maintain the reading frame of two consecutive exons by also translating the intron and thus two consecutive exons can be joined together by a translatable intron sequence. For example, Figure 3.5 shows a *B. rapa* gene (*Brara.E00016.v1.3*) for which an intron should have been annotated but was not. A clear splice junction can be seen in the RNA-seq data. It is worth noting that in this case the accompanying RNA-seq mapping profile is somewhat "blocky", due to low read coverage in this region: had the read coverage been any lower it is unlikely that RNA-seq-based methods would have been able to identify this missing intron.

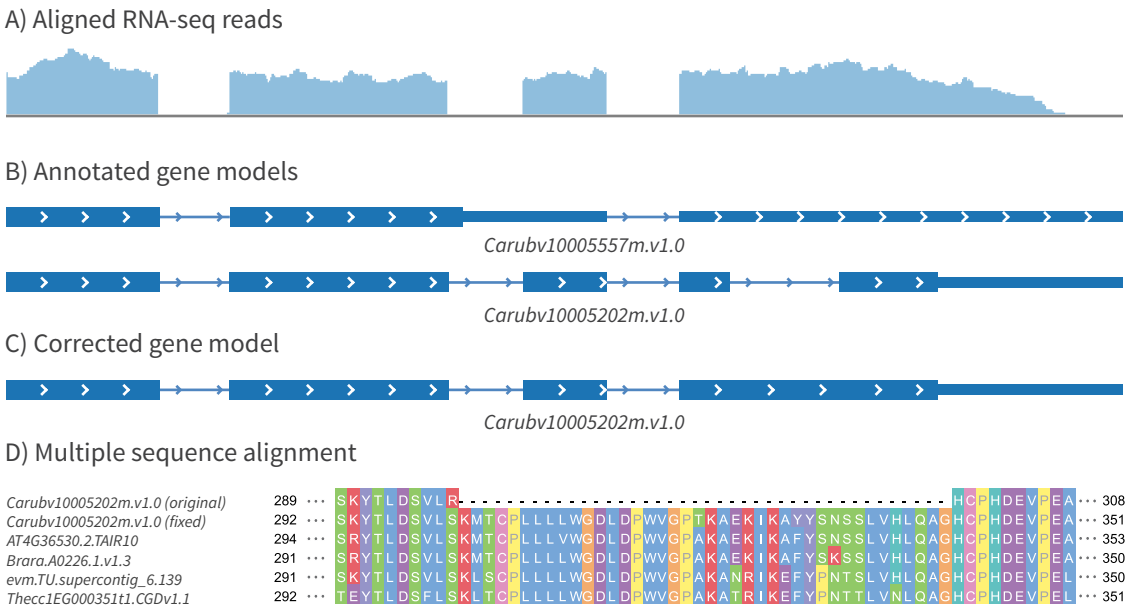


Figure 3.6: Example of an extraneous intron inclusion in *C. rubella* gene *Carubv10005202m.v1.0*. An alternative transcript for this gene is available, although neither the primary nor the alternative transcript matches the RNA-seq data. A) RNA-seq coverage from spliced-aligned RNA-seq reads; B) The gene model present in the current genome annotation; C) A manually corrected gene model that matches the RNA-seq data; D) A multiple sequence alignment of the corresponding amino acids sequences containing the original annotated gene (original) and the manually corrected gene model (fixed) as well as other genes in the orthogroup.

By contrast, inspection of the multiple sequence alignment of orthologs of this gene in other species readily identifies the potential error in the gene model.

Conversely, figure 3.6 demonstrates a case where an intron has been annotated but the data clearly show that no intron is present. This is a *C. rubella* gene for which two transcript variants have been annotated (*Carubv10005202m.v1.0* and *Carubv10005557m.v1.0*), neither of which matches the aligned RNA-seq data. The multiple sequence alignment of the orthogroup shows the primary transcript *Carubv10005202m.v1.0*, and clearly demonstrates that the last two annotated exons should be read through as one long, single exon.

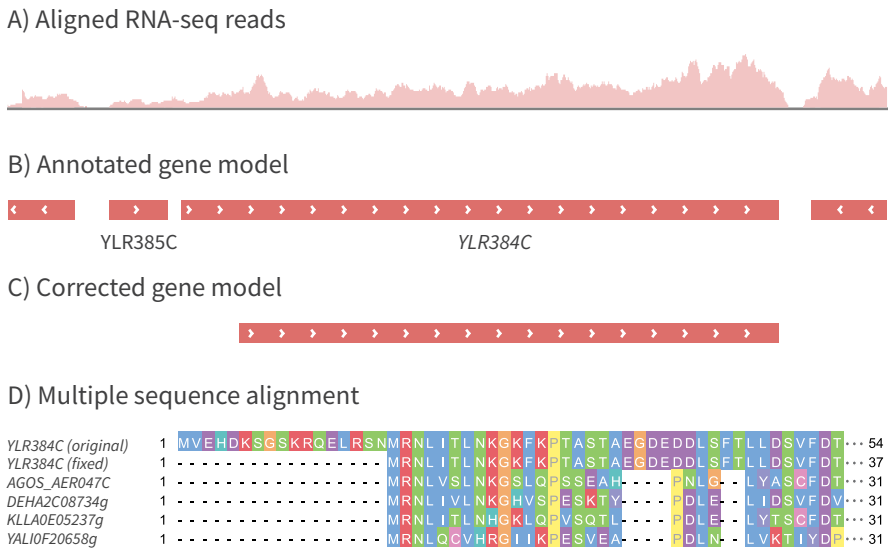


Figure 3.7: Example of an erroneous start codon in *S. cerevisiae* gene *YLR384C*. The location of a start codon is usually impossible to tell precisely from RNA-seq data alone, especially when multiple possible methionine residues exist in close proximity. In contrast, such errors can be clearly elucidated by amino acid alignments. A) RNA-seq coverage from spliced-aligned RNA-seq reads; B) The gene model present in the current genome annotation; C) A manually corrected gene model that matches the multiple sequence alignment; D) A multiple sequence alignment of the corresponding amino acids sequences containing the original annotated gene (original) and the manually corrected gene model (fixed) as well as other genes in the orthogroup.

3.2.4 Inaccurate start codons

Start codon locations can be difficult to determine from RNA-seq evidence alone, due to the presence of 5′ UTRs that are transcribed in addition to the coding sequences. The fact that any methionine codon could be a potential translation start site means that it is very difficult to determine the location of this feature from mapped RNA-seq data. In many cases, however, inaccurate start codons can be observed in multiple sequence alignments, such as for the *S. cerevisiae* gene *YLR384C* shown in Figure 3.7. In this example, several neighbouring single-exon genes can also be seen. In this case, while it is possible that the originally designated start codon is correct, a phylogenetically conserved start codon is also found and thus *a priori* one would expect the conserved one to be more likely to be correct.

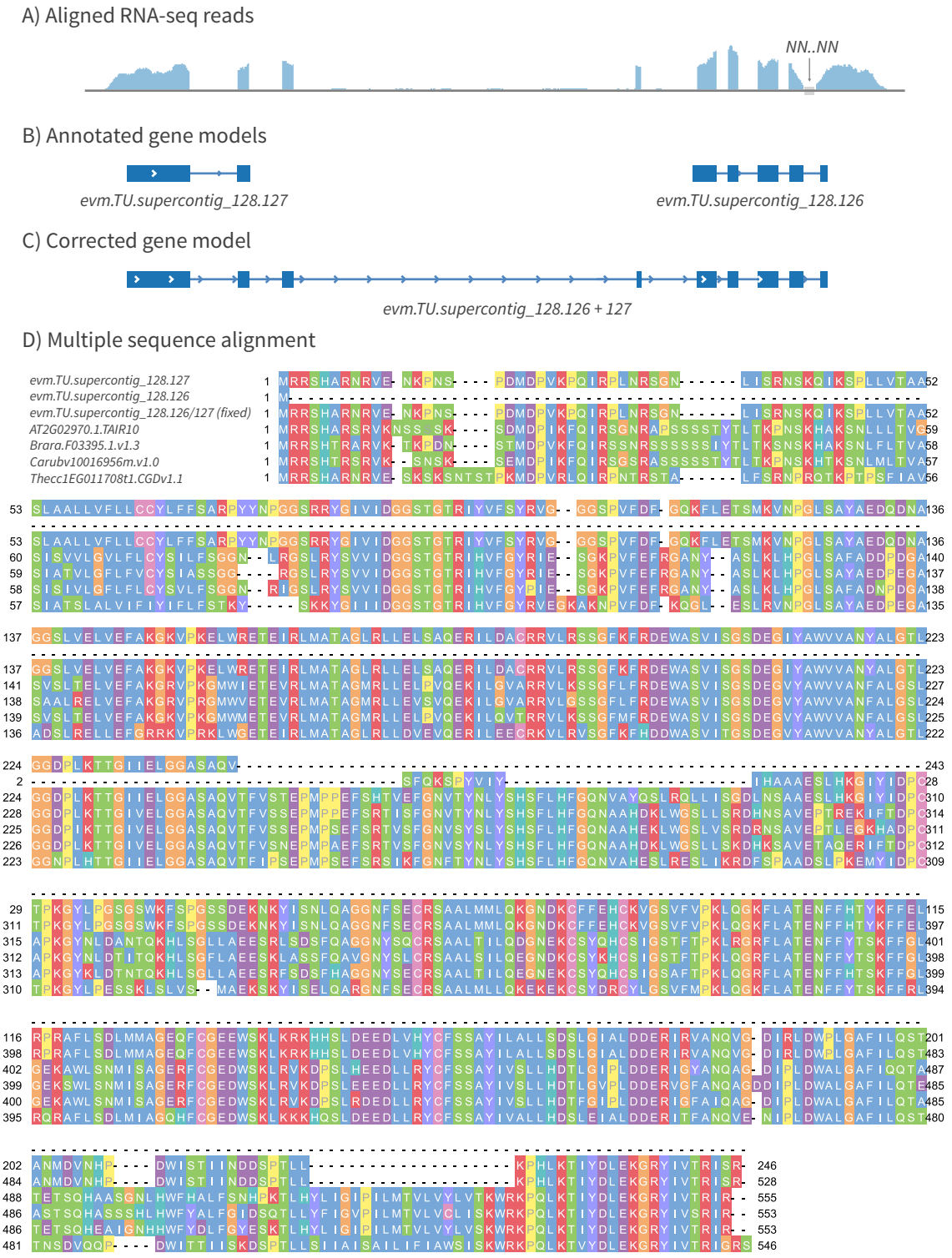


Figure 3.8: Example of a split gene. These are two *C. papaya* genes, *evm.TU.supercontig_128.126* and *evm.TU.supercontig_128.127*, that should have been annotated as a one. A) RNA-seq coverage from aligned RNA-seq reads; B) Currently annotated gene model; C) Manually corrected gene model that... *(continued on page 89)*

Figure 3.8: *(continued from page 88)...* matches the RNA-seq data; D) Multiple sequence alignment of the corresponding amino acids sequences containing the original annotated gene (original) and the manually corrected gene model (fixed) as well as other genes in the orthogroup. The separate gene models are not supported by the RNA seq data; Note that, due to a small patch of "N" bases in the genome assembly, a small portion of the 3′ end of the gene could not be constructed completely. This does not affect the capacity for the two separate gene parts to be joined

---

3.2.5 Split and fused genes

Sometimes a single gene can be annotated as two separate genes. Conversely, two separate genes can be fused into a single gene. Figure 3.8 shows two annotated *C. papaya* genes (*evm.TU.supercontig_128.127* and *evm.TU.supercontig_128.126*), which the RNA-seq data suggests should have been annotated as a single gene. This can happen when an intron is unusually long compared with other genes observed in the genome, causing gene finders to identify genic regions but to fail to join them. The multiple sequence alignment shows both genes, as well as a manually constructed gene model containing both. In the multiple sequence alignment, each of the separately annotated gene models are consistent for around half of the alignment length, the other half being gaps. The joined-up version of the gene is consistent with the rest of the orthogroup.

Figure 3.9 shows an example of a fused gene (*evm.TU.supercontig_3.62*) in *C. papaya*. The depth of the coverage of the RNA-seq evidence for the two halves of the annotated genes are dramatically different, which indicates that the two halves of the genes are actually separate; furthermore there is no evidence of any splice junction in between them. This type of error can be seen in the multiple sequence alignment as a large region of the gene in question which is not represented in any of the other sequences in the orthogroup and which has resulted in a gappy erroneous alignment of the genes. The figure also shows an amended gene model based on the RNA-seq evidence.

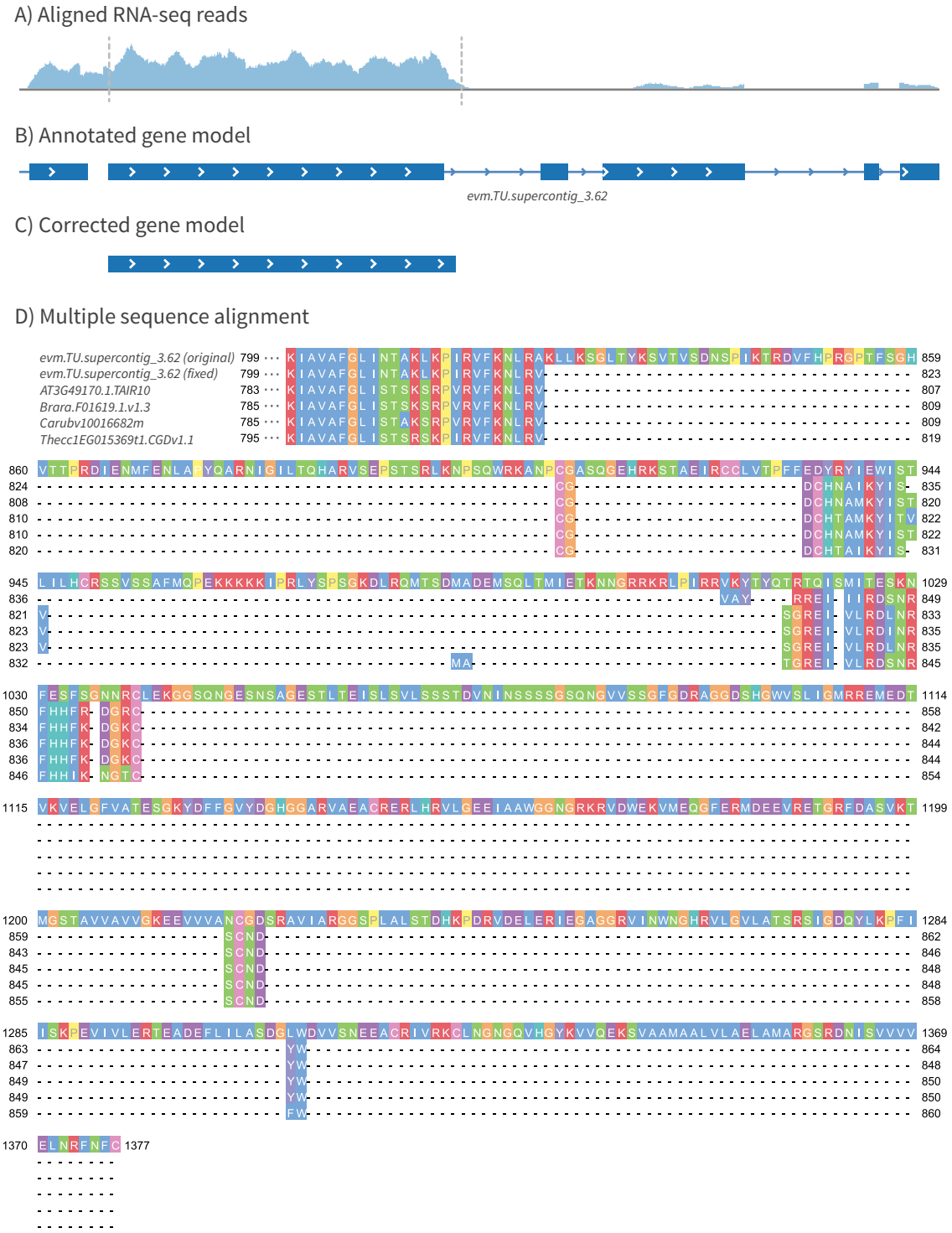


Figure 3.9: Example of fused genes. The gene *evm.TU.supercontig_3.62* in *C. papaya* has been fused with a neighbouring gene. There are no RNA-seq splice junctions joining the two halves of the annotated gene, and the alignment... *(continued on page 91)*

Figure 3.9: *(continued from page 90)...* shows large regions of disagreement with the sequence in question. A) RNA-seq coverage from spliced-aligned RNA-seq reads; B) The gene model present in the current genome annotation; C) A manually corrected gene model that matches the RNA-seq data; D) A multiple sequence alignment of the corresponding amino acids sequences containing the original annotated gene (original) and the manually corrected gene model (fixed) as well as other genes in the orthogroup.

3.2.6 Complex errors

Sometimes several of the above errors can be combined to produce more complex error types. This can occur in particular when inaccurate features and feature boundaries cause frameshifts to the reading frame of subsequent exons. Figure 3.10 is an example from *A. thaliana* (*AT5G47380.1.TAIR10*). In this gene annotation, what should have been a terminal exon has been truncated, and an additional exon has been placed after it. No intron is present in the RNA-seq data before the final annotated exon. The "fixed" version of the transcript has had the C-terminal exon removed, and the new final exon extended slightly to reach a stop codon. This is a much better match for the rest of the genes in the orthogroup. Note also the high levels of mapped reads within the region occupied by the annotated intron. This is due to an alternate splice form being present in the transcriptome sample, as well as either a second alternate splice form involving complete omission of the intron, transcribed non-coding RNA in the intron, or pre-spliced transcripts being included in the sample. This gene is therefore quite challenging to reconstruct from RNA-seq evidence alone due to significant retention of intron sequence.

3.2.7 Accurate gene model annotation is not always possible

Sometimes the creation of an accurate gene model is hindered by flaws inherent to the genome assembly containing it. This is important to note when considering the overall gene model accuracy of a set of gene model annotations: if it is concluded that a genome annotation needs improving, it may not always be possible to do so until the underlying assembly is updated.

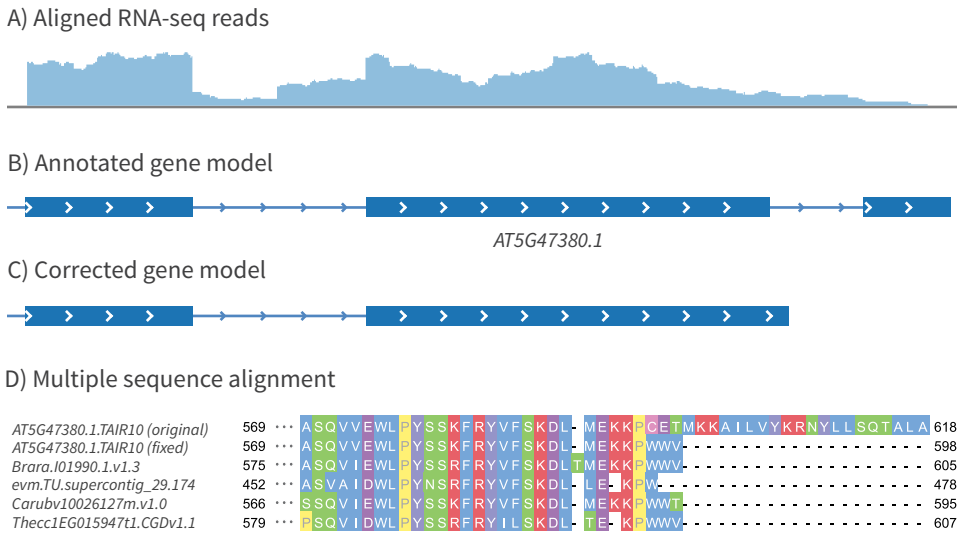


Figure 3.10: Example of a complex gene model error. Sometimes error types are not quite as clear-cut as the categories described in this section. This example from *A. thaliana* shows an added C-terminal exon, accommodated by a contracted previous exon. A) RNA-seq coverage from spliced-aligned RNA-seq reads; B) The gene model present in the current genome annotation; C) A manually corrected gene model that matches the RNA-seq data; D) A multiple sequence alignment of the corresponding amino acids sequences containing the original annotated gene (original) and the manually corrected gene model (fixed) as well as other genes in the orthogroup.

Figure 3.11 demonstrates one particular type of assembly flaw. The figure shows two genes in *C. papaya*, *evm.supercontig_9.204* and *evm.supercontig_9.205*, which should have been annotated as a single gene. In this case, the split was caused by a large run of "N" nucleotides in the gap between the two gene fragments. Such "N" gaps typically represent regions where two contigs have been scaffolded together and many gene prediction algorithms do not predict genes that span "N" gaps, thus preventing correct gene model construction. As the requisite DNA sequence was absent from the assembly it was not possible to construct a "fixed" version of the gene model using RNA-seq data.

Similarly, if a gene is found towards the very end of a scaffold, gene models may be truncated, and it may not be possible to recover them: even if the other half of such a broken gene were present at the end of another scaffold, genes cannot be annotated over scaffold boundaries. Finally, sometimes mapped RNA-seq

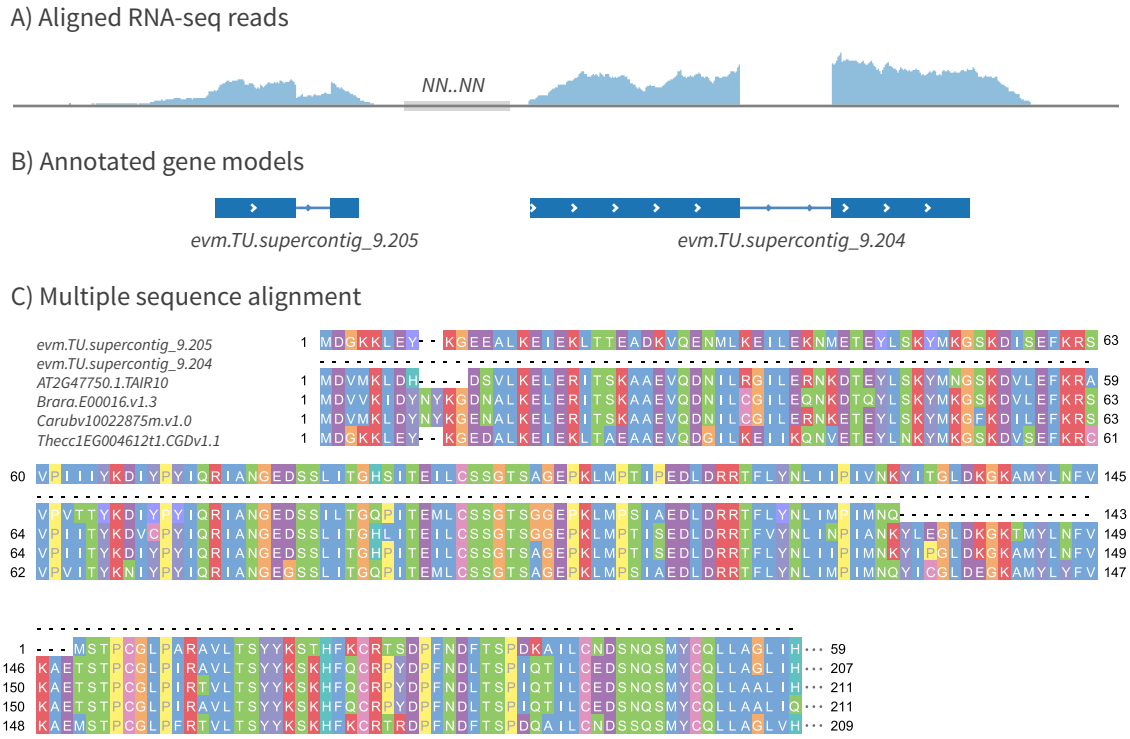


Figure 3.11: Assembly errors can cause annotation errors. This example shows two annotated genes that in reality should be one in *C. papaya*: *evm.TU.supercontig_9.205* and *evm.TU.supercontig_9.204*. This error is slightly different, as it is caused by an issue in the underlying genome sequence. A large region of contiguous "N" bases in between the two "genes" has prevented both accurate gene construction and RNA-seq mapping. The large region missing from the start of *evm.TU.supercontig_9.204* in the orthogroup alignment would be recovered if the gene was re-fused with the full *evm.TU.supercontig_9.205* sequence and the inaccessible intermediate portion. A) RNA-seq coverage from spliced-aligned RNA-seq reads; B) The gene model present in the current genome annotation; C) A manually corrected gene model that matches the RNA-seq data; D) A multiple sequence alignment of the corresponding amino acids sequences containing the original annotated gene (original) and the manually corrected gene model (fixed) as well as other genes in the orthogroup.

data indicate small insertions or deletions (indels) compared with the genome assembly. This is usually nothing more than the result of a particular allele choice for a genomic region, although sometimes it can be the result of a mistake in the genome assembly process. In either case, ordinary gene annotations can sometimes be tripped up by these, meaning sensible, in-frame gene models cannot be predicted. Thus, not all gene model errors are true gene model errors: some are artefacts of assembly inaccuracies.

3.3 Extrinsic metrics reveal gene model errors

The accuracy of a single gene model can only truly be tested by extrinsic means. The most concrete approach for obtaining the gene model for a single transcript would be to collect the mature protein and obtain its amino acid sequence directly using a protein sequencing method such as Edman degradation (Edman 1950) or mass spectrometry (Burgess et al. 2010; Barrera and Robinson 2011; Cox and Mann 2011). The protein sequence can then be aligned back to the genome sequence to construct the gene model. While useful for individual studies, such methods are time-consuming, expensive, and not readily scalable to genome-wide applications.

Incorporation of experimental data, such as RNA-seq, is the most common approach for assisting in determining gene models. However, RNA-seq data are not always available for an organism. Even when it is available, it can be extremely difficult to get single or aggregated RNA-seq data sets in which all genes that are present in the proteome are expressed (García-Ortega and Martínez 2015). This is particularly true for multicellular organisms where proteins may be expressed only in certain cell types, under certain conditions, or during a particular stage of the life cycle of the organism. Furthermore, genome assembly errors (such as the "N" gaps described above) and read mapping ambiguities can make accurate representation of some genes impossible, especially for repeat-rich regions, or regions containing retrotransposons.

In the absence of adequate extrinsic data, sequence homology can be used to assess gene model validity. Section 3.4 of this chapter will aim to demonstrate how assessments of orthogroup alignment consistency can be used as a proxy for overall quality of a set of gene models. However, discrepancies in orthogroup alignments, while often indicative of errors, do not provide conclusive evidence for them. In order to assess any non-extrinsic methods of gene model error analysis, it is necessary to first construct extrinsic metrics against which to compare them.

This section describes two such assessment metrics for judging the quality of a gene model compared to RNA-seq data, attempting to formalise and automate the notion of an annotated gene model "matching" its aligned read profile.

3.3.1 Splice junction recall identifies erroneous splice junctions

In keeping with the most common method of manually determining gene models, the first accuracy metric introduced here involves comparing annotated splice sites with RNA-seq mapping junctions. The spliced-read mapper HISAT2, which was used here to align RNA-seq read data, makes judgements about whether a single read should be split over multiple regions. Reads that are split contain *mapping junctions*. Biologically, these mapping junctions correspond to exon-exon splice junctions. *A priori*, one can expect correctly annotated splice junctions to correspond to RNA-seq mapping junctions more frequently than incorrectly annotated splice junctions.

The *splice junction recall score* (jR) measures the agreement between annotated splice junctions and RNA-seq mapping junctions. An illustration of how the jR score is obtained can be seen in Figure 3.12. For a given gene G and a genomic region R spanning the ends of the gene, define the set of annotated splice junctions in G to be J_G, and the set of RNA-seq mapping junctions present in the data mapped to R by J_R. Define the splice junction recall by

$$jR\left(J_G, J_R\right) = \begin{cases} \frac{|J_G \cap J_R|}{|J_G|} & \text{if } J_G \neq \emptyset \\ 1 & \text{if } J_G = \emptyset \end{cases}$$

This provides a measure of how many splice junctions are represented in the RNA-seq data. A gene with no introns marked will receive a score of 1. This is a one-way score: it only assesses how many of the junctions present in a given gene model are also found in the RNA-seq data that map to the region containing it, and not vice versa. This is for two reasons: firstly it is possible that alternative transcripts exist for a given gene, causing multiple possible splice junctions to

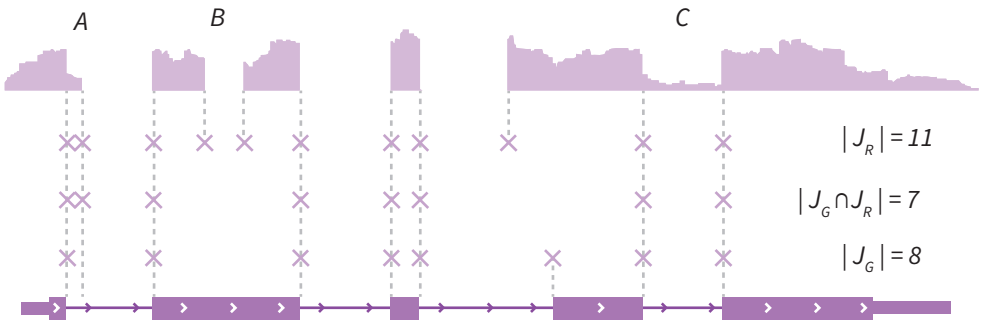


Figure 3.12: Illustration of splice junction recall (jR) score. In this case the jR score is $jR(J_G, J_R) = \frac{7}{8} = 0.875$. The top shows aligned RNA-seq reads, the bottom shows the annotated gene model. A) Multiple possible alternate transcripts are indicated by the read profile, with two possible splice junctions at the end of the first exon of the gene; B) Unannotated introns are not detected by the jR score; C) Annotated and read-spliced introns can still contain reads.

exist; secondly, both non-coding RNA and coding sequences for other genes can inhabit the same genomic neighbourhood of any given gene, both of which can contain additional splice junctions that are independent of the gene model under investigation. It would be unfair to detract from a gene model's score on the basis of other transcripts present in its neighbourhood.

The one-way nature of the jR score means that it will fail to spot some particular types of error, namely unannotated introns and exons: this is assessed with the coverage score, defined in Section 3.3.3. A properly annotated gene should have a jR score of 1, provided the data informing the score is of a sufficiently high resolution. A score of 1 would mean that every splice junction in the gene model has supporting evidence in the RNA-seq data that map to this gene.

3.3.2 Clear differences in splice junction recall scores between species

The genes annotated in Species Sets P and F were assessed according to their splice junction recall score. Mapped RNA-seq data were obtained as described in Chapter 2, Section 2.7.7, and splice junctions were obtained from the mapped reads

using RegTools (Griffith and Griffith 2018) (see Methods Section 3.7.1). Splice junction recall was calculated for all sufficiently expressed genes. Note that, for very lowly expressed genes, read data can sometimes fail to indicate splice junctions even when they may be biologically present. To account for this, assessments were restricted to genes with a basic level of expression (BLE): that is, an average read depth ($\bar{d}$) of 10 reads or more across the annotated coding regions. In addition, assessments were stratified into genes with low (LLE), medium (MLE), and high (HLE) levels of expression, defined to be genes with $10 < \bar{d} \leq 50$, $50 < \bar{d} \leq 100$, and $\bar{d} > 100$ respectively. Note that highly expressed genes are less likely to have imperfect *jR* scores: this is because these highly expressed genes are more likely to have been found in EST libraries or other resources that may have been used to help guide the algorithms used to annotate them.

In addition to distinguishing genes based on their average expression levels, a distinction is made between genes which have an alternative transcript marked and genes which do not. For the plant data sets, assessments were made for all transcripts, primary transcripts (as marked in Phytozome (Goodstein et al. 2012)) and single-transcript genes. Genes with alternative transcripts may be forgiven for having imperfect *jR* scores if an alternative transcript is present in the read profile. Thus, although applying the junction score analysis to all genes can reveal inadequacies in the primary transcript choice, applying to only genes with no alternative transcripts marked could be considered a better measure of the true fidelity of an annotated set of gene models.

Figure 3.13 and Table 3.1 show the percentages of transcripts for each species with imperfect *jR* scores, i.e. $jR < 1$, for various levels of expression. All species have some genes with imperfect *jR* scores. For the plant species, *T. cacao* and *C. papaya* appear to have the by far the highest proportion of inaccurate gene models from a splice junction perspective. For the fungi, the vast majority of annotated genes only have a single exon (see Chapter 2, Table 2.4) and thus are not amenable

		% of transcripts with imperfect *jR* scores								
		All genes			Primary transcripts			Single-transcript genes		
	Species	Low	Med.	High	Low	Med.	High	Low	Med.	High
Plants	*A. thaliana*	12.6%	10.3%	10.1%	7.8%	7.5%	6.0%	6.4%	2.7%	1.0%
	B. rapa	8.3%	4.9%	4.2%	6.8%	5.0%	4.3%	7.2%	4.4%	3.8%
	C. papaya	37.6%	42.1%	36.4%	37.3%	42.1%	36.4%	37.7%	42.0%	36.4%
	C. rubella	10.7%	9.3%	7.4%	9.4%	8.4%	6.9%	9.8%	8.1%	6.5%
	T. cacao	29.5%	23.1%	19.2%	21.7%	19.3%	16.4%	19.4%	13.9%	12.2%
Fungi	*K. lactis*	0.3%	0.4%	0.3%						
	S. cerevisiae	0.7%	0.4%	0.1%						
	Y. lipolytica	2.0%	2.7%	1.5%						

Table 3.1: Assessments are stratified into genes with low (LLE), medium (MLE), and high (HLE) levels of expression, defined to be genes with $10 < \bar{d} \leq 50$, $50 < \bar{d} \leq 100$, and $\bar{d} > 100$ respectively. All genes have some imperfect gene models. For the plants, *T. cacao* and *C. papaya* genes are most likely to be imperfect. For genes without alternate transcripts marked, *A. thaliana* has the fewest imperfect gene models, as expected. Fungi have very few genes that are imperfect from a splice junction perspective, as most of them contain only a single exon. Splice junction recall scores for all assessable genes can be found in the Supplementary Material at `https://doi.org/10.5281/zenodo.1198798`, directory `ch3/1-juncScores`.

to splice junction analysis. This is reflected in the very small numbers of imperfect *jR* scores, and the data are so sparse that it is difficult to make concrete assertions. Of the three fungal species analysed, however, *Y. lipolytica* appears to have the greatest percentage of inaccurate gene models, and thus could be considered the least-well annotated genome of this set from a splice junction perspective.

The observations for *C. papaya* are directly in line with what was observed in Chapter 2, indicating that the *C. papaya* genome is not only under-annotated but that its gene models are also sub-par. In contrast to Chapter 2, *T. cacao* also shows strong signs of poor gene models. It was observed that *T. cacao* has large numbers of genes and very large introns, often containing other genes, and so it is likely that this genome has been *over-annotated*. As expected, genes from *A. thaliana* and *S. cerevisiae*, the well annotated model species, have the lowest incidence of

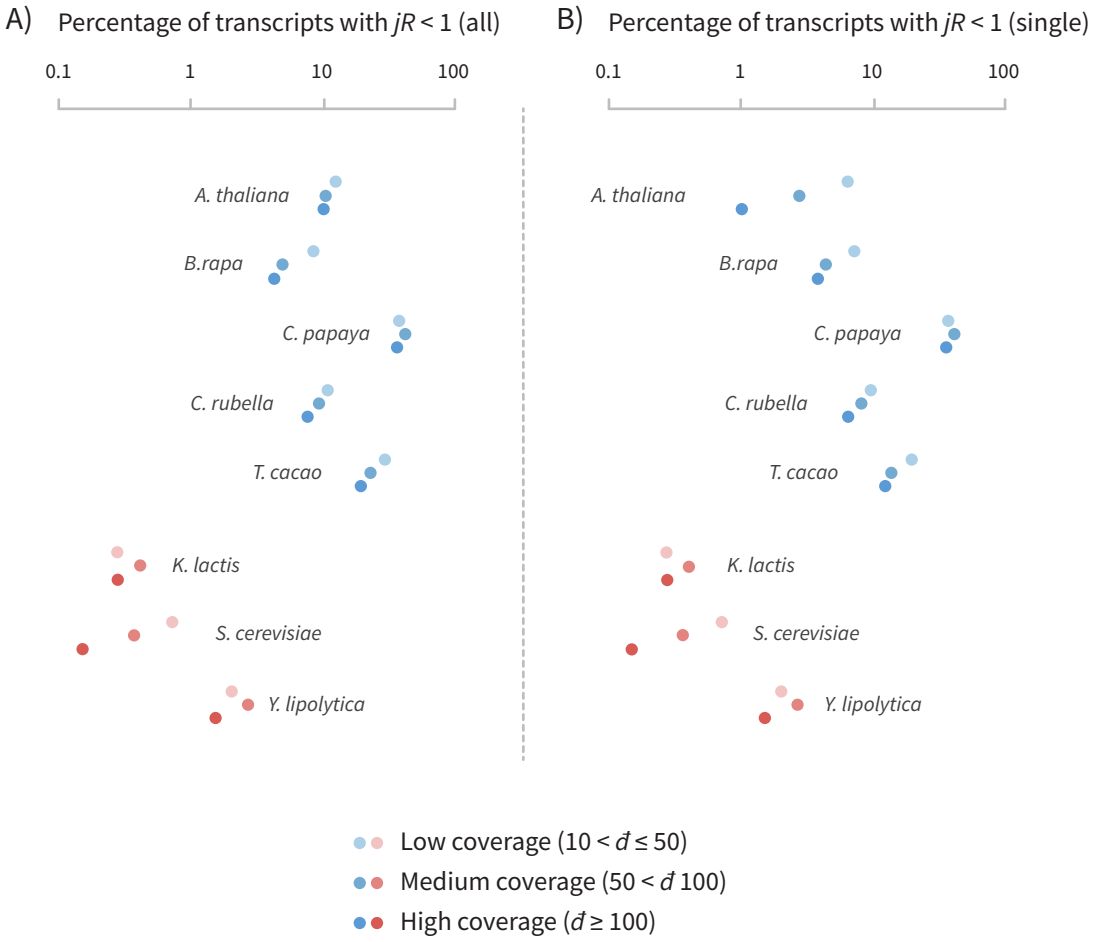


Figure 3.13: Distributions of imperfect splice junction recall scores. Graphical representation of Table 3.1. *C. papaya*, *T. cacao*, and *Y. lipolytica* stand out as the species with the highest number of imperfect *jR* scores for their respective species sets. Note that in most cases, genes with higher expression are less likely to have imperfect *jR* scores. This is because highly expressed genes are more likely to have had their RNA-seq profiles leveraged for genome annotation, and thus are more likely to be correct.

imperfect *jR* scores. For all plant species except *C. papaya*, which has very few alternate transcripts marked, single-transcript genes typically exhibit more perfect *jR* scores than those with multiple transcripts annotated. Finally, in nearly all cases, higher-expressed genes typically have a much lower incidence of imperfect *jR* scores than the lower-expressed genes. As noted above, this is likely because highly expressed genes are a lot more likely to have been picked up by genome annotation pipelines that have been supported by RNA-seq or EST data.

3.3.3 Coverage score identifies missing introns and exons

The splice junction recall score described above is useful for assessing the presence of annotated splice sites in a genome by comparing them to mapped RNA-seq data. As noted, however, it is unable to pick up errors such as unannotated introns and exons and misplaced start codons. Furthermore, "almost" correct splice junctions are treated identically to absent splice junctions. A transcript could have several of its splice junctions annotated only a few bases away from the true splice site, indicating a transcript that would be biologically very close to the true transcript, but that would perform poorly in *jR*-score assessments. An additional score is required to tackle these cases.

Another way to assess the validity of a gene model using empirical RNA-seq data is to consider the *read coverage* of the predicted exons for that gene model. Exons are retained in RNA transcripts and thus should have higher read coverage than introns. This is not completely clear-cut: alternative transcripts and transcription of non-coding sequences can mean that the introns of a given transcript may contain independent transcribed regions from another. Generally, however, one should expect the best gene model to be the one that is best represented in terms of both exon coverage and intron non-coverage.

The *coverage score* is a measure of how well the exonic and intronic bases indicated by a given gene model are matched by RNA-seq data. A graphical representation of coverage score calculation can be seen in Figure 3.14. Given that gene expression levels can vary considerably and irregularly across the length of a transcript (Hansen et al. 2010; Jiang and Salzman 2015; Conesa et al. 2016; Wang et al. 2016), care must be taken to ensure the expression profile for a gene region is properly interpreted. For example, sample preparation methods can bias coverage towards the centre and 3′ ends of the transcript; furthermore, jagged read profiles and transcription of antisense regions and other intronic ncRNAs can cause

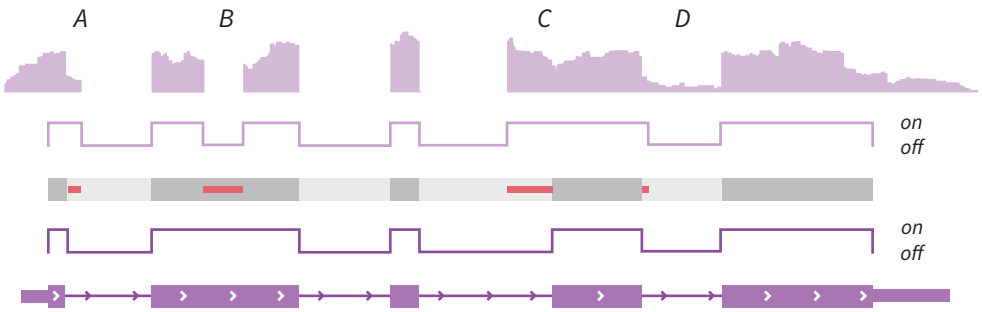


Figure 3.14: Illustration of coverage score. From top to bottom: mapped read depth profile, binary representation of read profile, annotated gene with discrepancies (marked in red), annotated gene binary representation, annotated gene model. A) "Off" regions are defined to have one fifth or lower the coverage of "on" regions: as such, inaccuracies in deciding which regions are on or off can occur; B) A missed intron causes a discrepancy; C) An inaccurate splice site is a genuine discrepancy; D) Another slight discrepancy caused by transcribed intronic regions.

expression profiles to be highly non-binary (Pelechano and Steinmetz 2013). To mitigate this, a rolling threshold approach is used. For a gene region R, and a genomic coordinate $x \in R$, the expression characteristic χ is defined as:

$$\chi(x) = \min\left(\max\left(\{\rho(y) : y \in R, y \leq x\}\right), \max\left(\{\rho(y) : y \in R, y > x\}\right)\right),$$

where $\rho(y)$ is the read count at genomic coordinate y. Bases in the gene region to which the RNA-seq data have been mapped are categorised based on whether they are likely to correspond to exonic or non-exonic regions: a base x is considered to be *on* (i.e. likely included in the mature mRNA) if $\rho(x) > \frac{\chi(x)}{5}$, and *off* (i.e. likely not included in the mature mRNA) if $\rho(x) \leq \frac{\chi(x)}{5}$. The coverage score for a gene model $G = \{G_1, \ldots, G_n\}$, where the G_i are alternately exons and introns, is defined to be:

$$C(G) = \frac{1}{n}\left(\sum_{G_i \text{ exonic}} \frac{|\{x \in G_i : x \text{ on}\}|}{|G_i|} + \sum_{G_j \text{ intronic}} \frac{|\{x \in G_j : x \text{ off}\}|}{|G_j|}\right)$$

that is, the average length-adjusted coverage score for each individual feature in the gene model.

Coverage scores were calculated for all single-transcript genes with basic levels of expression ($\bar{d} > 10$). Table 3.2 shows for each species the number of single-transcript genes which had imperfect scores (i.e. $C < 1$), and scores below various thresholds. The various thresholds are essentially arbitrary, but serve two purposes: to highlight seriously erroneous gene models (which would have very low coverage scores) and to forgive natural variations in RNA-seq coverage (which would reduce the coverage score and thus could cause many correct gene models to have scores slightly less than 1). In all species, relatively few transcripts had perfect scores: this is the combined effect of unannotated alternate transcripts / transcribed non-coding RNA and uncertainties in assessing on versus off regions in the RNA-seq data. The results become clearer when stratifying genes based on their scores. When considering scores of 0.95 or lower, *C. papaya* and *T. cacao* stand out once again as the plant genomes with the largest number of poor-scoring annotations. This distinction remains for thresholds of 0.75 and 0.5, although the data are significantly reduced by this stage. The fungal genomes show very little consistent difference in their scores. This is again because splicing is rare in fungal genomes, and so imperfections are much more likely to come from inconsistent read profiles within exons.

3.3.4 Coverage and splice junction recall are correlated

Figure 3.15 shows a scatter plot comparing the data for both the coverage score and the junction score. Shown are the percentages of single-transcript genes with basic expression levels ($\bar{d} > 10$) that have an imperfect junction score versus the percentage of single-transcript genes which have coverage score less than 0.95.

For fungal species, all percentages are fairly low, due to the fact that both scores are designed to pick up splicing discrepancies, and splicing is rare in fungal genes. Thus many of the imperfect scores for the fungi will be due to RNA-seq inconsistencies and mapping errors, and no correlation was observed. For the plants, *A. thaliana*, *B. rapa* and *C. rubella* all have fairly low percentages of low-scoring

	Species	% $C < 1$	% $C < 0.95$	% $C < 0.75$	% $C < 0.5$
Plants	*A. thaliana*	25.9%	5.9%	1.6%	0.1%
	B. rapa	37.9%	8.1%	1.4%	0.1%
	C. papaya	51.1%	34.0%	11.6%	0.7%
	C. rubella	41.2%	10.8%	1.5%	0.1%
	T. cacao	39.0%	20.1%	6.3%	0.6%
Fungi	*K. lactis*	50.4%	9.7%	0.8%	0.1%
	S. cerevisiae	15.0%	5.2%	2.5%	0.7%
	Y. lipolytica	24.7%	7.1%	2.3%	0.4%

Table 3.2: Percentages of single-transcript genes for each species passing various thresholds of coverage score (C). In each species, few genes have few perfect scores ($C = 1$), although *C. papaya* has the largest percentage and the two model species have the smallest percentages. The stratification is clearer for $C < 0.95$. This reduces the effects of insubstantial or uneven RNA-seq data, and of small amounts of intronic non-coding regions being transcribe. In particular, *C. papaya* and *T. cacao* stand out. For lower thresholds, the data become greatly reduced. Coverage scores for all assessable genes can be found in the Supplementary Material at `https://doi.org/10.5281/zenodo.1198798`, directory `ch3/2-covScores`.

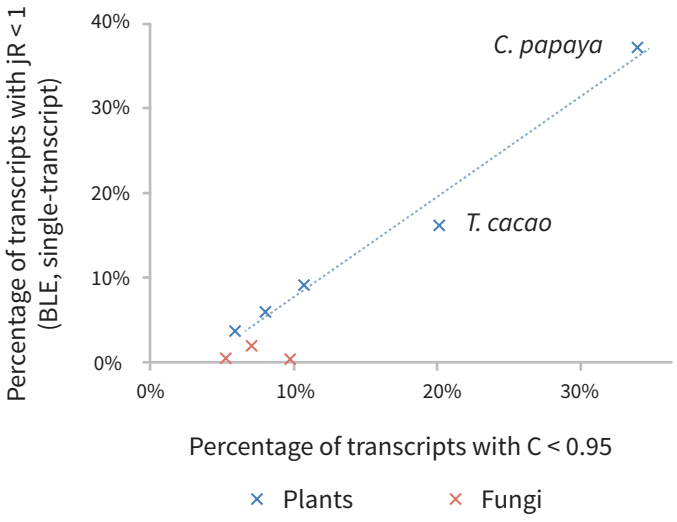


Figure 3.15: Coverage and splice junction recall scores are correlated. Shown is the percentage of single-transcript genes with imperfect coverage scores, against the percentage or single-transcript genes with imperfect splice junction recall scores. The plant data are correlated ($R^2 = 0.980$, $p = 0.0066$); the fungal scores are not. This is due to low incidence of introns in fungi, and so most of the poor scores will have been contributed by RNA-seq inadequacies. Shown is a trend line for the plant data set.

genes by both measures. *C. papaya* and *T cacao* have high percentages of low-scoring genes by both measures. For the plants, the two scores are correlated with an R^2 value of 0.980 ($p = 0.0066$). The correlation observed here suggests that both scores reveal underlying inadequacies in the genomes assessed. The scores are not identical, and thus can reveal different aspects about the accuracy of the gene models.

3.4 A non-extrinsic metric for gene model assessment

The two scores discussed so far make use of extrinsic data to infer levels of gene model accuracy. As observed, these scores do not work well for species in which splicing is rare, as is the case for members of Species Set F. In addition, the scores are affected by the fidelity of the RNA-seq data used, with genes with higher read coverage typically exhibiting better splice junction recall and coverage scores than those with a lower mean coverage. This section describes the main method of this chapter, an orthogroup-based non-extrinsic approach for assessing gene models from the perspectives of alignment discrepancies, which aims to avoid some of the problems inherent to extrinsic data. These are compared to the above extrinsic scores to support the claim that alignment discrepancies can be used to identify inaccurate gene models.

3.4.1 A scoring system for amino acid alignments

The examples of various gene model error types discussed in Section 3.2 compared two ways of visually identifying gene model errors: comparison with RNA-seq data and observation of discrepancies in orthogroup amino acid alignments. In the latter case, discrepancies can most easily be identified when the amino acid alignment is sufficiently coherent, i.e. containing few gaps and exhibiting a high degree of intra-orthogroup agreement. In order to automate an analysis of orthogroup alignment discrepancies and restrict them to easy-to-assess regions, it is necessary

to formalise the notion of a well-formed multiple sequence alignment. This will be done using the *alignment score*. This section will define and discuss briefly some of the characteristics of this score.

There is no standard scoring system for amino acid multiple sequence alignments, although pairwise alignment scoring systems based on amino acid substitution have been well studied. The PAM (Dayhoff et al. 1978) and Blosum (Henikoff and Henikoff 1992) systems for pairwise alignment scoring both provide methods for comparing protein sequences based on their amino acid content. Extensive comparisons (Henikoff and Henikoff 1993; Vogt et al. 1995) suggest that the Blosum model is the most accurate and more generally applicable to full-length amino acid sequences (Pearson 2013). As such, it is typically used as the standard today, for example in the ubiquitously used BLAST algorithm (Altschul et al. 1990), and is generally considered to be the canonical amino acid comparison matrix. In particular, the Blosum62 matrix is used as the basis for the MAFFT multiple sequence alignment algorithm used here. Blosum62 will therefore be the substitution matrix used in the following analyses.

For a given multiple sequence alignment, define the *alignment score* α to be the sum of columnwise internal Blosum62 scores. More precisely, an amino acid alignment of length l comprising k species can be considered as an ordered sequence $A = (C_n)_{n=1}^{n=l}$ of columns $C_n = (c_1^n, \cdots, c_k^n)$. The *column score* γ for a column C_n is defined as the average pairwise Blosum62 score for amino acids in that column:

$$\gamma\left(C_n\right) = \frac{\sum_{1 \leq i < j \leq k} \mathrm{Blos}\left(c_i^n, c_j^n\right)}{k}$$

$$\alpha\left(A\right) \sum_{n=1}^{l} \gamma\left(C_n\right)$$

Gaps are scored -1 against any amino acid, and gap-gap pairs are scored 0. The adjusted alignment score $\bar{\alpha}$ is defined as $\bar{\alpha} = \frac{\alpha}{l}$, where l is the alignment length. The α score is the same as the sum-of-pairs score optimised by the multiple

sequence alignment tool MUSCLE (Edgar 2004), and is similar to the weighted sum-of-pairs score used by MAFFT (Katoh and Standley 2013). A basic Python implementation of the alignment score can be found in the Supplementary Material at `https://doi.org/10.5281/zenodo.1198798`, directory `ch3/3-aliScore`.

3.4.2 Poor alignment scores reflect low sequence similarity

To appreciate the meaning of the alignment scores, it is helpful to consider directly how they behave under alignment imperfections. For a column in a *perfect* alignment, that is, one in which all sequences are exactly equal, the highest and lowest possible column scores under Blosum62 are 11.0 and 4.0 respectively. These correspond to the highest and lowest self-self scores in the Blosum62 matrix. The expected score for a perfect alignment A is

$$\mathbb{E}\left(\bar{\alpha}\left(A\right)\right) = \sum_{\alpha \in \Omega} \eta_a \mathrm{Blos}\left(a, a\right)$$

where Ω is the set of amino acids ($|\Omega| = 20$), and η_a is the relative fractional abundance of an amino acid a, with $\sum_{\in A} \eta_a = 1$. The mean expected perfect scores across the plant and fungal species are 5.21 (s.d. 0.0051) and 5.19 (s.d. 0.0086) respectively. Figure 3.16 shows the expected score for simulated perfect alignments of various lengths, with amino acid distributions based on the average distribution for the plant and fungal data set combined. Although the expected perfect score is not dependent on the length of the alignment, the score clearly stabilises for longer sequences.

To demonstrate how the alignment score can be used as an indicator of sequence disagreement, simulated orthogroup alignments of randomly mutated sequences were produced. Starting with a random amino acid sequence of length 1000, ten variants were produced, with each residue "mutated" with probability μ. If a residue is chosen for mutation, it is either changed to a different amino acid (with probability 0.8) drawn from the same distribution as that observed for the

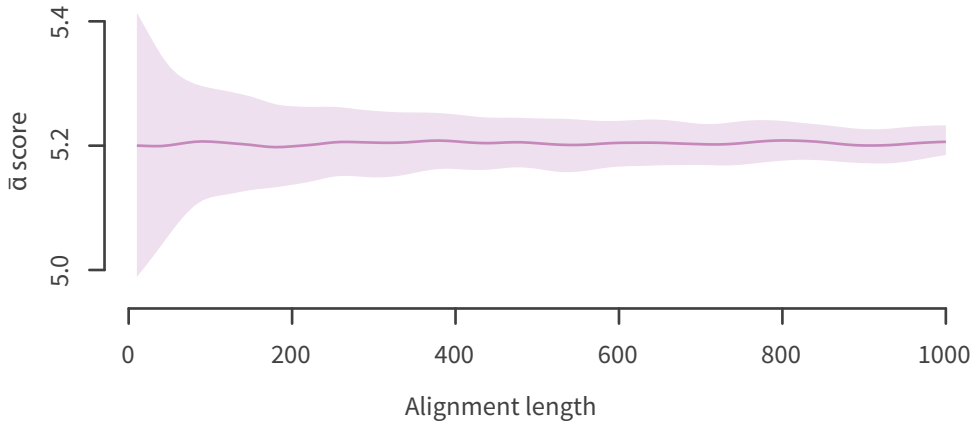


Figure 3.16: Alignment scores of simulated perfect aligments. Alignment scores approach the expected as the alignment length increases. For each alignment length, from 10 to 1,000, with a step size of 10, 1,000 repetitions were performed. The region between the upper and lower quartiles is indicated by the shaded band.

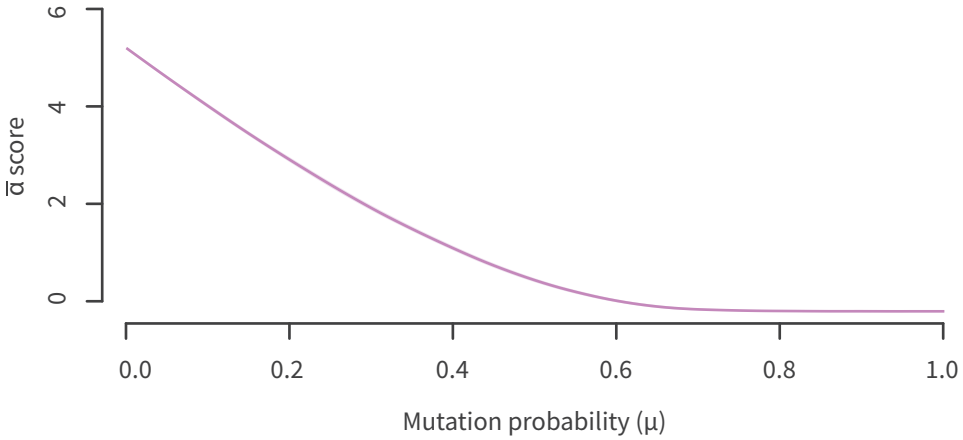


Figure 3.17: Alignment score rapidly decreases for broken alignments. Amino acid alignment score decreases rapidly as "mutation" probability μ increases. For each mutation probability, from 0 to 1 with a step size of 0.01, 1,000 repetitions were performed. The region between the upper and lower quartiles is indicated by a shaded band, however the uncertainty is so small that it is almost impossible to see here.

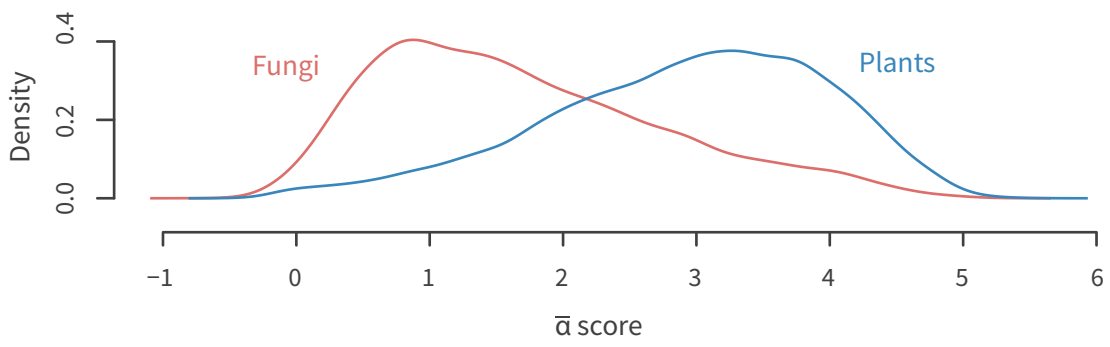


Figure 3.18: Fungal orthogroups have lower alignment scores. Shown are the adjusted alignment scores for all orthogroups from plants and fungi. Adjusted alignment scores for fungal orthogroups are generally lower than those for plants: this is a consequence of increased evolutionary distance within the fungal species set compared with the plants.

annotated fungal and plant sequences, or changed to a gap (probability 0.2). These sequences are then re-aligned and the $\bar{\alpha}$ score is calculated. This simulates a set of ten genes drawn from a single ancestral gene, each having undergone uniform random mutation. The distribution of scores for various μ is shown in Figure 3.17. The figure clearly shows that the alignment score rapidly decreases the more a set of sequence deviates internally. Low alignment score is therefore indicative of poor-quality amino acid alignments.

3.4.3 Fungal orthogroups have lower alignment scores

The overall distributions of $\bar{\alpha}$ scores for plant and fungal orthogroups in Species Sets P and F can be seen in Figures 3.18 and 3.19. Figure 3.18 shows the density distributions of adjusted scores from all plant and fungal orthogroups. Plant orthogroups seem to have generally much better alignments than the fungal orthogroups. This indicates that the sequences in the fungal data set are generally more different from one another than in the plant data set: this is simply a function of differences in mutation rate, generation time and divergence time for the species sets in question. Moreover, fungal genomes are known to be somewhat "plastic" and are prone to frequent mutations and rearrangements (McTaggart et al. 2016).

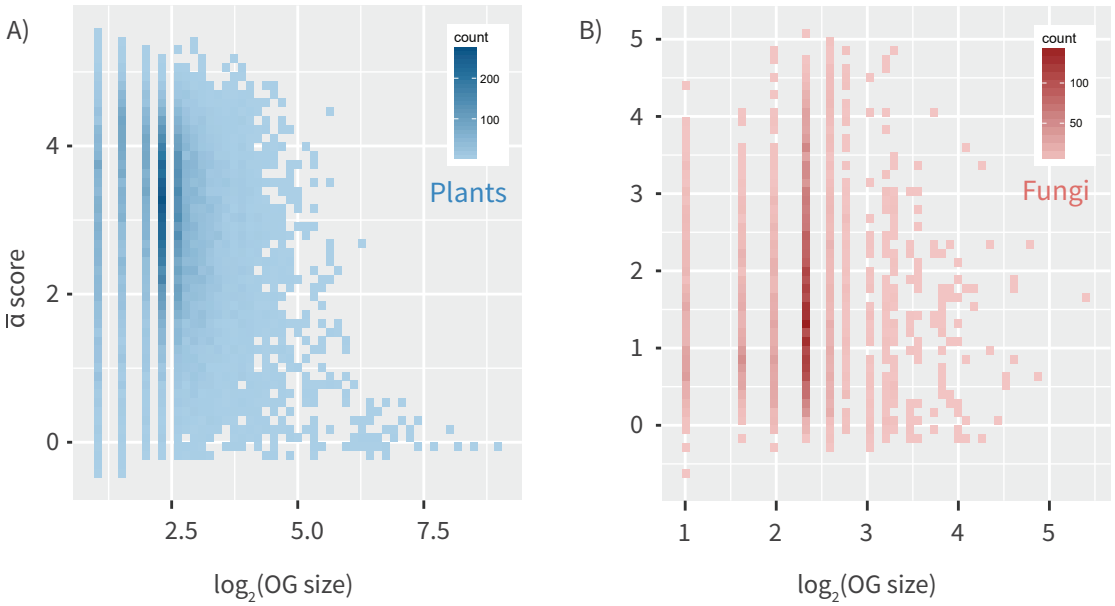


Figure 3.19: Larger orthogroups have lower alignment scores. Shown are density distributions of orthogroup entropy scores for plants and fungi: A) Plants; B) fungi. Darker patches indicate more orthogroups with the indicated α score and orthogroup size. Most orthogroups are of size 5 and have a score between 3 and 4 for plants, and between 1 and 2 for fungi. Larger plant orthogroups tend to have very poor scores. This is not an inherent consequences of orthogroup size: rather, large orthogroups tend to include highly duplicated gene paralogs, allowing considerable intra-orthogroup variation.

Figure 3.19 compares orthogroup length to the alignment score. Notably, very large orthogroups for plants have very low alignment scores. This is because large orthogroups necessarily contain large numbers of paralogous genes within a single species, which usually exhibit considerably more variation than homologous genes between species.

On its own, the alignment score described above does not provide a great deal of information about the validity of the sequences in an alignment. It may, however, be leveraged in three ways: firstly, it can be used to partition alignment sets or sets of aligned regions into high-fidelity and low-fidelity groups, for high-confidence analysis of subregions. Secondly, isolated sections of an alignment whose sub-alignment scores disagree strongly with the rest of the alignment can be considered potential candidates for gene model errors. Finally, given possible sequence variants and

the resulting multiple choice in orthogroups, alignments can be compared. The most coherent alignment will be the one with the highest α score, and the genes contained therein could be considered the optimal choices.

3.4.4 Pools, bridges, and discrepant genes

The alignment score described in Section 3.4.1 can be used to identify well-aligned regions of orthogroup alignments, and consequently "obvious" adjacent regions exhibiting suspicious strings of gaps. As was seen in Section 3.2, such regions can correspond to gene model errors. This section provides an automated method to search for the kinds of gappy regions that can suggest gene model errors, by considering only well-aligned regions of amino acid orthogroup alignments.

For any given amino acid alignment, define *pools* and *bridges* as follows. A *pool* is five or more contiguous positions of the alignment for which one sequence has only gaps, and all other sequences are gap-free in that region (Figure 3.20A). Conversely, a *bridge* is five or more contiguous positions of an alignment for which one sequence is gap-free, and all other sequences have gaps (Figure 3.20B). The length of five amino acids here is arbitrary, though given that many gaps of size 1–3 are genuine small insertion or deletion events, it was considered a sensible minimum threshold.

For both pools and bridges it is additionally required that the ten amino acids flanking the region in question are completely gap-free and have an adjusted alignment score of 4 or more (this is the minimum possible perfect score for any alignment and is lower than the expected adjusted score). Again this choice is arbitrary but was considered a sensible threshold. If a flanking region of length ten cannot be constructed due to the pool or bridge being near the beginning or end of the alignment, the flanking region is taken as the region from the relevant end of the pool/bridge until the relevant edge of the alignment (Figure 3.20C). If the pool/bridge is at the very end or very beginning of the alignment, such that no flanking region whatsoever can be constructed on one or both sides of the

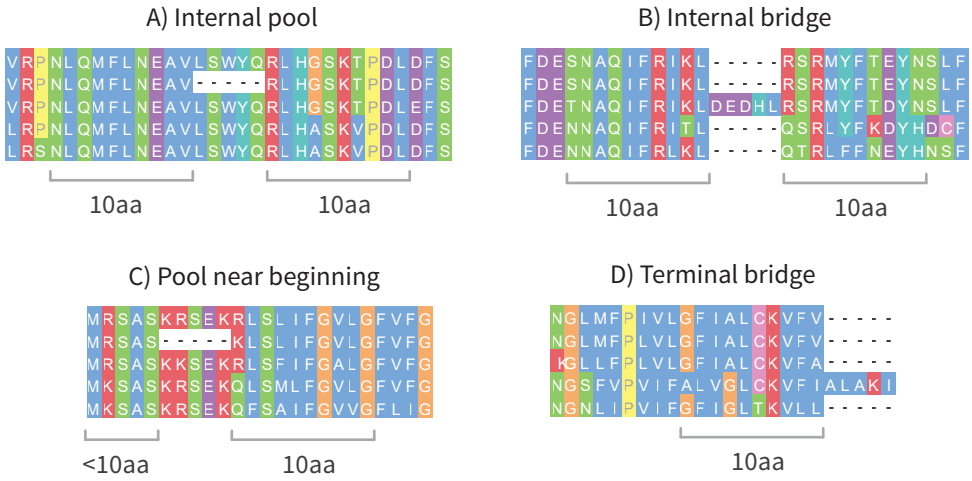


Figure 3.20: Illustration of pools and bridges. A) Pool within an alignment flanked by 10aa high-quality region; B) Bridge within an alignment flanked by 10aa high-quality region; C) bridge near beginning of alignment is judged on the flanking material available; D) Bridge at the end of an alignment doesn't require a right-hand flanking region.

region, this constraint is not implemented (Figure 3.20D). A feature that is either a pool or a bridge will be referred to as a *discrepant region*, and a gene containing a discrepant region will be referred to as a *discrepant gene*.

3.4.5 Many orthogroups contain clear discrepancies

All primary plant and fungal orthogroups (i.e. formed from primary transcripts) containing three or more genes were searched for discrepant regions. In total, 50 such regions were found in fungal orthogroups; 2,454 were found in plant orthogroups. In SCU orthogroups, 37 discrepant regions were found in fungal orthogroups, and 592 were found in plant orthogroups. Breakdowns of the number of regions for each species can be seen in Figure 3.21 for all orthogroups containing three or more sequences, and in Figure 3.22 for SCU orthogroups.

In both SCU and all-orthogroup cases, the same sets of species exhibit the highest degree of potential error as previously observed: *C. papaya* and *T. cacao* for the plants and *Y. lipolytica* for the fungi (Figure 3.21). Similarly *A. thaliana* and *S. cerevisiae* show the least potential error. The pattern is clearer for SCU orthogroups:

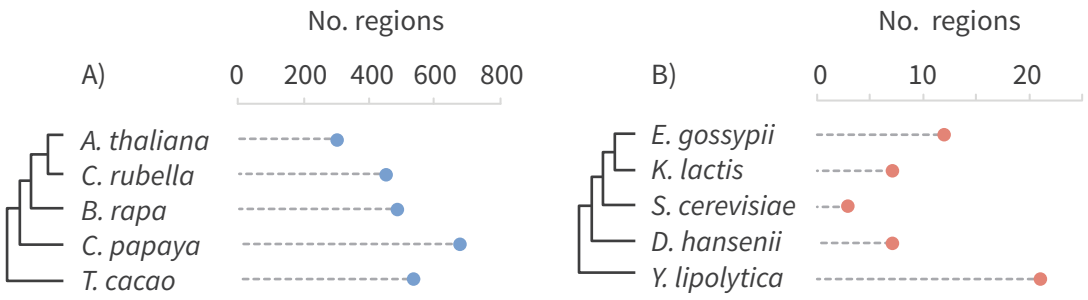


Figure 3.21: Some species have more discrepancies than others (all orthogroups). This figure shows the number of discrepant regions for genes across all orthogroups of the two species sets. A) Distribution for plants; B) Distribution for fungi. Although *C. papaya* and *T. cacao* lead the way for the plant data set, the difference between them and the others is small. For the fungal data set, *Y. lipolytica* exhibits the largest number of discrepant sequences. In both cases, the model species have fewer discrepant regions than the other species.

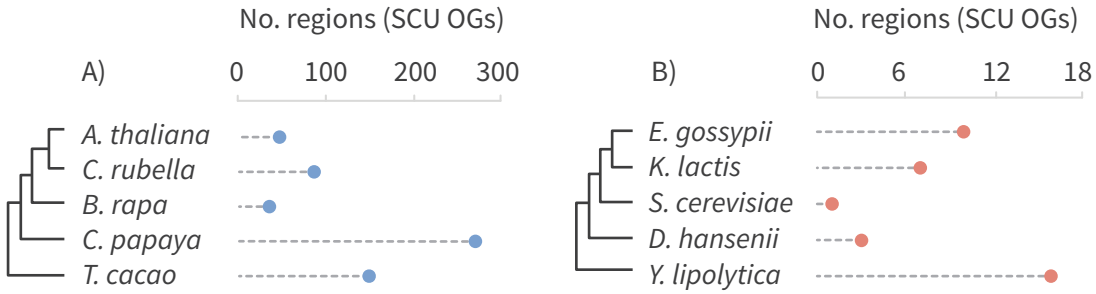


Figure 3.22: Some species have more discrepancies than others (SCU orthogroups). This figure shows the number of discrepant regions for genes in SCU orthogroups only. A) Distribution for plants; B) Distribution for fungi. In this case, the lead for *C. papaya* and *T. cacao* in the plants is much clearer, while the fungal distribution remains largely the same. Although *C. papaya* and *T. cacao* lead the way for the plant data set, the difference between them and the others is small. For the fungal data set, *Y. lipolytica* exhibits the largest number of discrepant sequences, as expected. In both cases, the model species have fewer discrepant regions than the other species, with the exception of *B. rapa*, which also has few discrepant regions.

genes with many paralogs are much more likely to have genuine differences than genes that have remained single-copy in all species (Figure 3.22). The agreement between these assessments and the extrinsic assessments considered earlier suggests that this approach is indeed picking up signals of erroneous gene models.

Discrepant orthogroups can be used to form a rough accuracy score for annotated gene model sets. Define the *discrepancy score* by dividing the number of discrepant

genes in SCU orthogroups for a species by the total number of SCU orthogroups for that species set. Discrepancy scores for each species are shown in Table 3.3. For the fungi, all scores are very low. This is for two reasons. Firstly, since fungal genes are simpler, the scope for error is lower. Secondly, fungal orthogroups are typically less well aligned than plant genomes, and the strict alignment score criteria are likely to cause many genuine discrepant regions to be missed. The plant scores are higher, with the highest discrepancy score obtained by *C. papaya* at 7.0%, but still fairly low compared, for example, with the estimates produced in Section 3.3. Due to the strict criteria imposed, the numbers obtained here are likely to provide a considerable underestimate for the number of gene model errors present, but should give an indication of the *distribution* of the errors across the species. Nevertheless, this basic analysis of discrepant regions reveals patterns of potential gene model errors that are consistent with previous analyses. In particular, the strict nature of the discrepancy criteria means that these errors are very likely to be real.

3.4.6 Discrepant plant genes have poor splice junction recall

Due to the possible presence of genuine evolutionary sequence changes, discrepant genes do not necessarily indicate an inaccurate gene model. To test the correlation between discrepant genes and gene model inaccuracies, the results were compared to the extrinsic scores defined earlier. As before, a distinction is made between SCU orthogroups and all orthogroups, with the assumption that SCU orthogroups should generally be cleaner and more likely to provide biologically relevant results.

The first two analyses here, considering splice junction recall score (jR) and coverage (C) scores as defined in section 3.3, are restricted to discrepant genes for which the discrepant region is not initial, i.e. the disagreeing sequence segment did not appear at the very beginning of the alignment. A discrepancy at the beginning of an alignment usually indicates an inaccurate start codon, which cannot be detected by either the jR or C scores. Such a gene will be referred to as a *non-initial*

	Species	No. discrepant genes in SCU OGs	Score
Plants	*A. thaliana*	48	1.2%
	B. rapa	37	1.0%
	C. papaya	272	7.0%
	C. rubella	88	2.3%
	T. cacao	147	3.8%
Fungi	*E. gossypii*	10	0.4%
	D. hansenii	3	0.1%
	K. lactis	7	0.3%
	S. cerevisiae	1	0.0%
	Y. lipolytica	16	0.6%

Table 3.3: Discrepancy scores for all species. These scores are obtained by finding the number of discrepant genes for each species in SCU orthogroups, and dividing by the total number of orthogroups. Given the strict criteria imposed, these are likely to be underestimates of incompleteness, but can flag up relative levels between species. In particular, for fungi, alignments are often of a poorer quality, and so the high-quality regions required for discrepancy identification will be rarer. FASTA format amino acid alignments for all orthogroups containing discrepant genes can be found in the Supplementary Material at `https://doi.org/10.5281/zenodo.1198798`, directory `ch3/4-DOGs`.

discrepant gene (niDG), and all other genes (including *initial discrepant genes*, or iDGs), will be referred to as the *background*. Note that the distinction is not entirely clear cut, as a gene can contain multiple discrepant regions, so the background can still contain discrepant genes. Furthermore, inaccurate start codons do not always manifest at the very beginning of an alignment, due to ambiguities in the alignment algorithm used, and so some niDGs may still represent start codon choice errors.

Table 3.4 and Figure 3.23 show the mean splice junction recall (jR) scores for assessable (i.e. having basic level of expression, $\bar{d} > 10$) niDGs versus the background. For the plants, jR scores for discrepant genes are clearly lower, for both all orthogroups and for SCU orthogroups. For the fungi, the scores are all perfect for niDGs, and nearly all perfect for the background. There were only 19

and 13 niDGs for all orthogroups and SCU orthogroups respectively in the fungi, of which 13 and 10 respectively were assessable. All of these happened to have perfect scores – thus the resolution here is fairly low and not representative. This reinforces the fact that fungi have few splice junctions, and this is reflected in the low incidence of niDGs and typically high junction scores. Discrepant genes in the fungi here are likely to be either missing intron annotations, which are not detected by the *jR* score, or sequence insertion events, meaning the gene is discrepant but does not correspond to a gene model error.

Table 3.5 and Figure 3.24 show the percentages of assessable niDGs which have imperfect *jR* scores, compared with the background. The distinction here is profound for plants, with 72.4% of niDGs from all orthogroups having imperfect junction recall scores compared with the background of 18.0%. This is even more stark for SCU orthogroups, for which 81.5% of all niDGS have imperfect junction recall scores, compared with 18.4% of the background. Once again, for fungi there is very little information to go on, with essentially all genes in all cases having perfect *jR* scores. Thus, considering all species, at least 72% of discrepant genes contain gene model splice junction errors.

3.4.7 Discrepant plant genes have suboptimal coverage

As noted, splice junction recall scores are unable to detect missing introns and exons, something that the coverage score is able to do. Table 3.6 and Figure 3.25 show the mean coverage scores for assessable niDGs versus the background, for all orthogroups and SCU orthogroups. In all cases, the mean coverage scores were lower for niDGs than the background. This is in line with what was seen for the splice junction recall score in plants. For fungi, this suggests that some of the discrepancies are due to missing introns, which were not picked up by the splice junction recall score. The differences here were small: many discrepancies will be minor and will only have a small effect on the resultant score. Furthermore, many

	Mean splice junction recall scores			
	All orthogroups		SCU orthogroups	
Species set	niDGs	Background	niDGs	Background
Plants	0.85	0.94	0.85	0.94
Fungi	1.0	0.99	1.0	0.99

Table 3.4: Mean splice junction recall scores for non-initial discrepant genes (niDGs). The scores are the same for genes from single copy ubiquitous (SCU) orthogroups and the others. For plants, the mean splice junction recall score is markedly less than the background. For fungi, it is slighlty higher than the background, however the data are so sparse that this result is inconclusive.

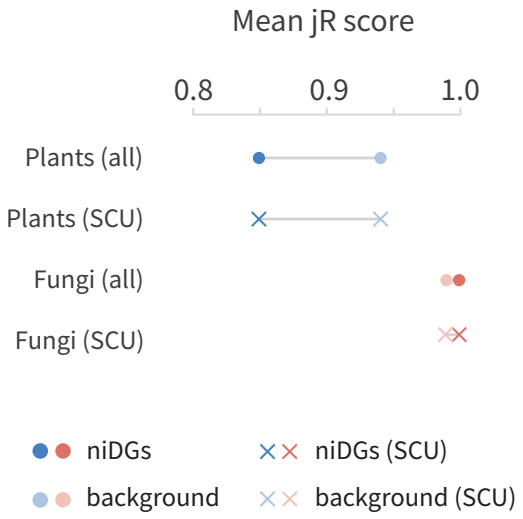


Figure 3.23: Mean splice junction recall score is lower for discrepant genes. Shown are mean splice junction recall scores for non-initial discrepant genes (niDGs), a visual representation of Table 3.4. Splice junction recall scores are typically higher for plants; for the fungi the results are inconclusive due to sparse data. Many of the niDGs for fungi will be missing introns, which will not affect the *jR* score.

genes in the *background* category will have gene model errors, and thus will receive lower scores, but won't have been flagged up as niDGs, further narrowing the scores.

Table 3.7 and Figure 3.26 show the percentages of assessable niDGs which have imperfect coverage versus the background. Again, in all cases a higher percentage of niDGs have imperfect scores compared with the background: nearly 100% for niDGs from SCU orthogroups. As noted previously, a large number of genes will have imperfect *C* scores even if their gene models are correct, due to inconsistencies

	% gene models with imperfect *jR* scores			
	All orthogroups		SCU orthogroups	
Species set	niDGs	Background	niDGs	Background
Plants	72.4%	18.0%	81.5%	18.4%
Fungi	0.0%	0.9%	0.0%	0.9%

Table 3.5: Percentages of discrepant gene models with imperfect splice junction recall scores. Far more discrepant genes have imperfect *jR* scores compared with the background for plants. For fungi, essentially all genes have perfect scores: again this is a combination of sparse data and fungal genes having few introns.

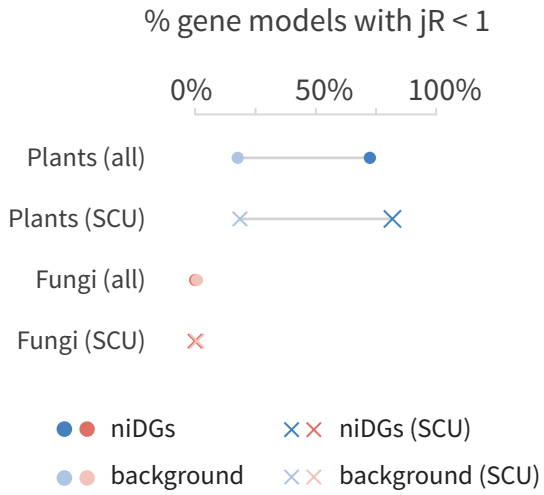


Figure 3.24: Discrepant gene models commonly have imperfect splice junction recall. Shown are percentages of gene models with imperfect splice junction recall scores, a visual representation of Table 3.5. For plants, most niDGs have imperfect *jR* scores, compared with very low percentages for the background. For fungi, the data are sparse and once again inconclusive.

in the RNA-seq read profiles, however the distinction here indicates that many more real coverage problems are occurring for niDGs than for the background. Thus, at least 89% of genes identified as discrepant contain potential gene model errors assessable by RNA-seq coverage.

Given that some of the imperfect coverage scores observed above may be due to ambiguous RNA-seq mapping profiles, it is instructive to consider a lower threshold of coverage score than "perfect". Table 3.8 and Figure 3.27 show the percentages of

	Mean coverage scores			
	All orthogroups		SCU orthogroups	
Species set	niDGs	Background	niDGs	Background
Plants	0.932	0.966	0.938	0.966
Fungi	0.980	0.984	0.975	0.984

Table 3.6: Mean coverage scores for non-initial discrepant genes (niDGs) versus the background. In all cases, the mean coverage score was lower for niDGs than the background, however the difference is small. This is because many of the discrepancies will only be small, and thus will not have a profound effect on the score. Furthermore, some of the background will also contain gene model errors, which will reduce the score.

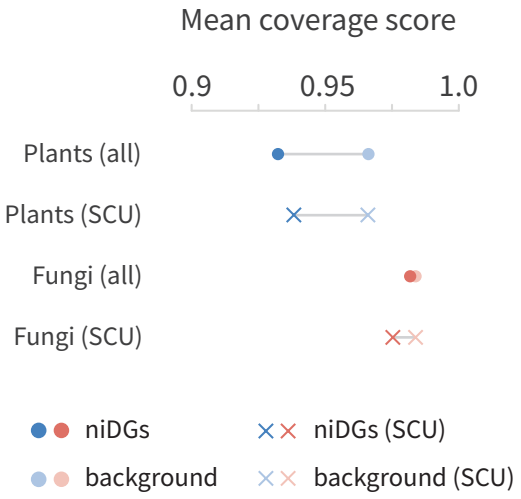


Figure 3.25: Mean coverage score is lower for discrepant genes. Shown are mean coverage scores for non-initial discrepant genes (niDGs) versus the background. Graphic representation of data in Table 3.6. In all cases, the mean coverage score was lower for niDGs than the background.

assessable niDGs with C scores lower than 0.95, versus the background. Again, in all cases, more niDGs have scores < 0.95 compared with the background. However, in all cases the number of gene models with $C < 0.95$ is low compared with $C < 1$. This suggests that many genuine inaccurate gene models may have been discarded when reducing the threshold, i.e. that some of the gene model errors in the niDGs are not sufficient to reduce the coverage score to below 0.95. Nevertheless, in all cases niDGS are at least twice as likely to have low coverage scores than the background.

	% gene models with $C < 1$			
	All orthogroups		SCU orthogroups	
Species set	niDGs	Background	niDGs	Background
Plants	89.3%	46.1%	95.6%	46.4%
Fungi	46.2%	28.6%	50%	28.6%

Table 3.7: Percentages of gene models with imperfect coverage scores. In all cases, a greater percentage of niDGs had imperfect coverage scores than the background. This is more profound for the plant species, as many of the niDGs will correspond to start codon errors that happen to have aligned

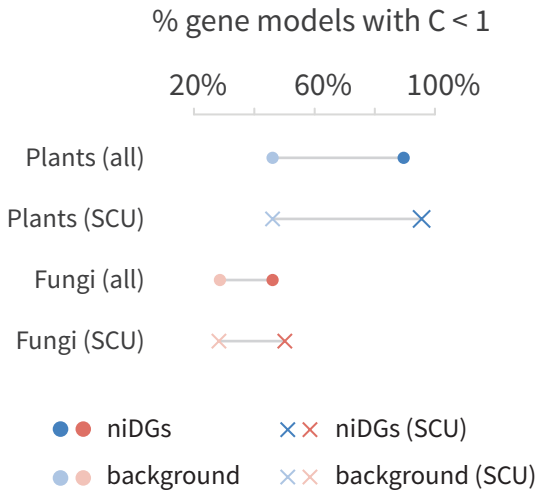


Figure 3.26: Imperfect coverage scores are common for discrepant genes. Shown are percentages of gene models with imperfect coverage scores. Graphical representation of data in Table 3.7. In all cases, a greater percentage of niDGs had imperfect coverage scores than the background.

In summary, discrepancies in orthogroup alignments show a strong capacity to identify gene model errors that can also be identified by extrinsic metrics, as demonstrated by the agreement between the two approaches. For the plant data, the signals are very strong, but for the fungal data, the data are sparse and so less clear signals were observed. This was however also the case for the extrinsic data. Thus, discrepant genes and the discrepancy score can take the place of extrinsic metrics in the absence of experimental data.

	% gene models with $C < 0.95$			
	All orthogroups		SCU orthogroups	
Species set	niDGs	Background	niDGs	Background
Plants	43.4%	17.1%	42.9%	17.3%
Fungi	15.4%%	7.2%	20.0%	7.2%

Table 3.8: Percentages of gene models with $C < 0.95$. Here, the distinction between niDGs and the background is stronger than for $C < 1$. Around 40% of niDGs have a coverage score less than 0.95. Very few fungal genes have low coverage.

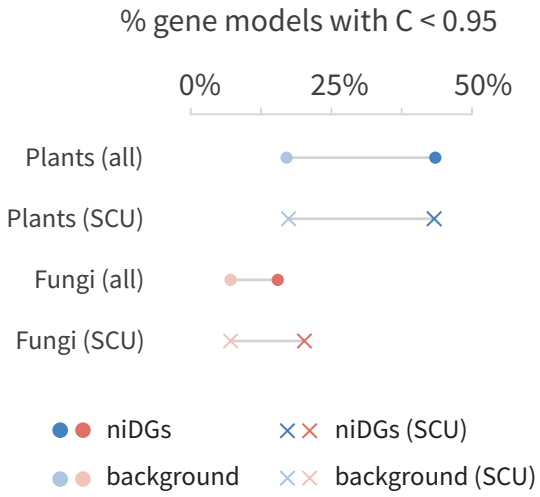


Figure 3.27: Low coverage scores are common for discrepant genes. Shown are the percentages of gene models with $C < 0.95$, a graphical depiction of data in Table 3.8. There is a clear distinction here in all cases beterrn niDGs and the background, with many more niDGs exhibiting lower coverage scores.

3.5 Subcellular localisation predictions indicate start codon accuracy

Notably, neither of the extrinsic assessments described in Section 3.3 – the coverage score nor the splice junction recall score – was able to identify inaccurate translation start sites, due to the ambiguity of optimal methionine choices in transcribed 5′ UTRs. Fortunately, the level to which start codons are inaccurate can be elucidated by orthogroup analysis.

Most genes are highly consistent in the organelle or cellular location in which

their mature translated proteins end up (Parton et al. 2014). The principle method of protein localisation is through protein-protein interactions that help guide nascent proteins to their correct destinations within a cell. These protein interactions typically make use of short motifs, several of which are present at the 5′ end of the protein sequence. Many bioinformatic tools have been produced that analyse the 5′ ends of protein sequences to predict their subcellular localisations. Inaccurately annotated start codons affect the success of these tools in obtaining accurate localisation predictions.

A priori, it is parsimonious to assume that orthologous genes in closely related species are targeted to the same subcellular location. While there will inevitably be exceptions to this assumption through genuine gain or loss of targeting motifs, these exceptions should be comparatively rare. An example of an inaccurate gene model choice causing inaccurate protein localisation prediction was shown in Chapter 1, Figure 1.6. Assessing the consistency of subcellular localisation predictions across the set of genes in an orthogroup can provide clues as to the validity of start codon choices for gene models within an orthogroup, and flag up start codon annotation errors.

3.5.1 Extrinsic analyses agree with localisation inconsistencies

To assess the extent to which particular members of the species sets under consideration exhibited discordant predicted subcellular localization, a highly restricted data set was analysed, consisting only of SCU orthogroups. Subcellular localisation predictions for both the plant and fungal datasets were determined using TargetP (Emanuelsson et al. 2000), as described in Methods Section 3.7.2. SCU orthogroups with uniform localisation predictions except for in one species were considered. Figure 3.28 shows the number of times each species was the odd one out for these datasets.

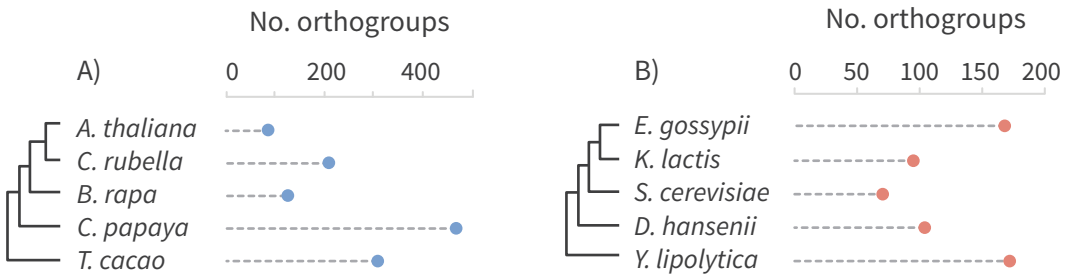


Figure 3.28: Localisation inconsistencies suggest poor start site annotation. Shown are the numbers of localisation "odd-one-out" SCU orthogroups for each species: that is, orthogroups that would have perfect subcellular localisation consistency were it not for that species. This shows, for each species, the number of SCU orthogroups with imperfect subcellular localisation consistency, such that that species is the *only one* causing disagreement. A) In plants, *C. papaya* and *T. cacao* are once again the species providing the least consistency. B) In fungi, this start codon analysis reveals more information than the coverage and splice junction assessments above, identifying *E. gossypii* and *Y. lipolytica* as the potential worst candidates.

As always, this assessment is likely to be tied up to some degree with the phylogenetic structure of each species group: one might expect that outgroups would be more likely to be the "odd one out" than ingroups. That said, clear signals for both *C. papaya* and *T.cacao* in the plant data set can be seen: in particular *C. papaya* is the odd one out more often than *T. cacao* even though it is not the most distant outgroup. This lines up with observations above that *C. papaya* and *T. cacao* are likely to be the least well annotated of the species here. For fungi, this 5′ analysis reveals a lot more than the splice junction and coverage scores above. In particular, it is effective for the species without RNA-seq data: *E. gossypii* and *D. hansenii*. Here, *E. gossypii* and *Y. lipolytica* are the worst performing, again in line with what was seen using the RNA-seq based assessments above for *Y. lipolytica*.

3.5.2 Initial discrepancies correlate with localisation inconsistencies

The consistency of localisation predictions for genes in an orthogroup was also considered relative to the presence of discrepant genes. The localisation consistency was formally defined as an entropy score across the predictions for each gene in

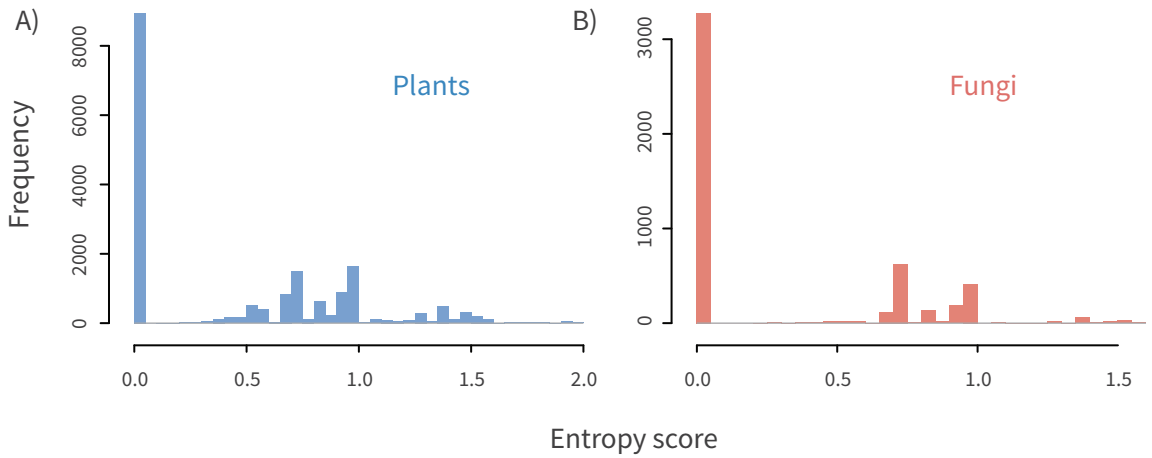


Figure 3.29: Most orthogroups have consistent localisation. Shown are entropy scores for primary orthogroups for plants and fungi. A) Entropy scores for all 18,330 plant orthogroups; B) Entropy scores for all 4,994 fungal orthogroups. Only orthogroups consisting of primary transcripts were assessed. An entropy score of zero means that every gene is predicted to localise to the same subcellular compartment. Entropy scores greater than zero quantitatively indicate the level of conflict. The majority of scores in both cases are perfect (zero entropy). Subcellular localisation predictions and entropy scores for all orthogroups can be found in the Supplementary Material (`https://doi.org/10.5281/zenodo.1198798`, directory `ch3/5-subcell`).

an orthogroup (see Methods Section 3.7.3). The best possible entropy score is 0, indicating total agreement across all sequences within the orthogroup, and the entropy score increases as agreement decreases. As an example, entropy scores across all primary orthogroups for plants and fungi are shown in Figure 3.29.

Define a *discrepant orthogroup*, or DOG, to be an orthogroup containing a discrepant gene. Unlike splice sites, assessments relating to subcellular localisation accuracy depend on features at the start of the alignment. Thus, in this case assessments are restricted to initial discrepant genes. For these purposes, *initial discrepant orthogroups*, or iDOGs, will be considered: these are defined as orthogroups that contain initial discrepant genes.

Table 3.9 and Figure 3.30 show the mean entropy scores for iDOGs and all other orthogroups that are not iDOGs (the *background*: this includes non-initial DOGs as well as orthogroups that do not contain any discrepant genes).

	Mean entropy scores			
	All orthogroups		SCU orthogroups	
Species set	iDOGs	Background	iDOGs	Background
Plants	0.507	0.461	0.569	0.519
Fungi	0.357	0.293	0.338	0.283

Table 3.9: Mean entropy scores for initial discrepant orthogroups (iDOGs) vs background. Here, iDOGs are initial discrepant orthogroups, that is orthogroups with a discrepant region at the beginning of their multiple sequence alignment. In all cases, entropy scores are worse for iDOGs than for the background. The pattern is similar for all orthogroups and SCU orthogroups, though the entropy scores are generally higher (worse) for SCU orthogroups. This is because differences in subcellular localisation are amplified for smaller groups in the entropy score.

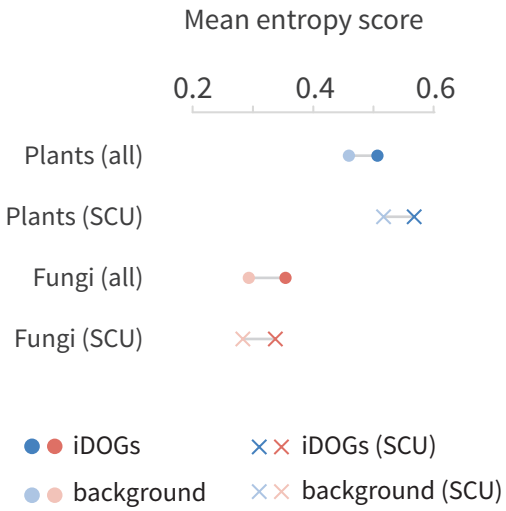


Figure 3.30: Mean entropy score is lower for discrepant orthogroups. Shown are mean entropy scores for initial discrepant orthogroups (iDOGs) vs background. Graphical representation of the data in Table 3.9. In all cases, the entropy is higher (worse) for iDOGs than for the background.

	% orthogroups with nonzero entropy			
	All orthogroups		SCU orthogroups	
Species set	iDOGs	Background	iDOGs	Background
Plants	62.0%	50.9%	61.9%	55.5%
Fungi	45.2%	34.3%	62.0%	33.5%

Table 3.10: Percentage of iDOGs with nonzero – i.e. imperfect – entropy scores. In all cases, large proportions of orthogroups have nonzero entropy scores. Slightly more iDOGs have imperfect entropy, although the distinction id relatively small. This suggests that in many cases, the discrepancy at the beginning of a gene is not sufficient to inhibit the accurate prediction of subcellular localisation.

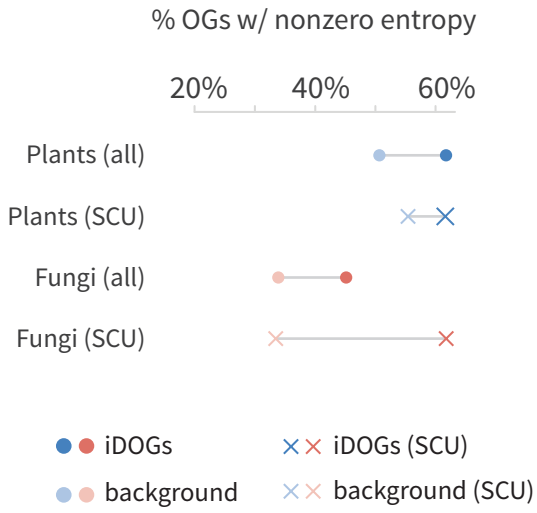


Figure 3.31: Inconsistent localisation is more common for discrepant orthogroups. Shown are percentages of iDOGs with nonzero entropy scores. Graphical representation of the data in Table 3.10. In all cases, a greater percentage of iDOGs disagree in their subcellular localisation predictions than one would expect for the background.

As before, all orthogroups here contain 3 or more sequences. In all cases, the orthogroups containing pools and bridges at the beginnings of their alignments show a higher entropy score than the background, indicating more disagreement between sequences in terms of their localisation prediction when an initial discrepant region is present. This is not surprising, but provides a concrete indication of the validity of using discrepant regions to identify biologically relevant gene model errors at the starts of protein sequences.

Finally, Table 3.10 and Figure 3.31 show the percentages of orthogroups in each category that had nonzero entropy scores. In all cases, discrepant orthogroups were more likely to have nonzero entropy than the background. It should be noted that not all start codon errors will inhibit subcellular localisation prediction, especially if the annotated start codon is near to the true start site or if no targeting peptide is present in the protein: thus the actual level of inaccurate start codons represented by iDOGs is likely to be higher than what is suggested by these data.

3.6 Gene model scores can inform primary transcript choice

To finish this chapter, annotations containing multiple transcripts were assessed for the optimal transcript choice under the metrics described above. For organisms in which splicing occurs, most well-developed genome annotations contain several alternative gene models for any given gene. For each gene, there will sometimes be one that is marked as the *primary transcript*. This is almost always the gene model that is used in any cross-species analyses: given that different alternative gene models produce different proteins, the results of any cross-species analysis are dependent on the variant of the gene model that is selected. The rationale for the choice of primary transcript is not stated in any genome annotation and typically is not stated in comparative genomic studies.

Comparing primary and alternative transcripts is useful for two reasons. Firstly, it helps to identify cases where the primary transcript marked does not appear to be the optimal choice, by analysing the transcripts under the scores defined above. Secondly, it provides a ready-made set of transcript variants on which comparative measures of alignment and gene model consistency can be assessed.

3.6.1 Alternate transcript choices can improve alignment scores

For the plant species set, primary transcripts were compared with annotated alternative transcripts. For each orthogroup, an alternative orthogroup was formed using a choice of alternate transcripts for each gene. To keep calculations in the realms of feasibility, orthogroups with 1,000 or more combinatorial transcript variant combinations were not considered. Out of 18,330 orthogroups considered, only 35 had 1,000 or more combinations and were left out of this analysis. The broadest scope for transcript choice was OG0000012 with 42,467,328 combinations. After disregarding orthogroups for which no alternative transcripts were available, 9,579 orthogroups remained with transcript variants.

Of the 9,579 plant orthogroups with alternate forms, 35.1% had variants that had a better full alignment score α; 39.4% had a variant that had a better adjusted alignment score $\bar{\alpha}$. This shows two things: firstly, some of the primary transcript choices may not be those that are most conserved between species. Secondly, given that improvements can be made by replacing primary transcripts with alternate transcripts, there may well be other, unannotated transcript variants that improve the alignment score for a given orthogroup, especially in species for which very few alternatives are annotated, such as *C. papaya*.

3.6.2 Alternate transcript choices can improve localisation consistency

In addition to the alignment score, orthogroups were assessed for the consistency of their subcellular localisation consistency, as defined earlier via the entropy score. Of the 9,579 orthogroups that had an alternate gene models, the entropy could be improved for 1,596 (16.7%) by changing the choice of model for one or more of the constituent genes. Moreover, 694 orthogroups (7.2%) whose primary transcripts gave an imperfect entropy score had an alternative that resulted in a

perfect score for the orthogroup. From the perspective of correctly placed start codons, most of these alternative choices can be considered "more correct" than their listed primary transcripts.

3.6.3 Many primary transcripts have suboptimal coverage and splice junction recall

Finally, coverage and splice junction scores as defined in Section 3.3 were compared for primary and non-primary transcripts for plant genes. Table 3.11 shows the percentage of transcripts that are better represented in terms of their splice junction recall score by RNA-seq profiles for genes of varying average read depth. As was noted in Table 2.3 in Chapter 2, the number of alternate transcripts marked varied considerably between species, and many gene finders only report a single transcript. In particular, *C. papaya* has very few alternate transcripts marked – only 34 genes with alternate transcripts. Thus no conclusions can be drawn for *C. papaya*. For the other species, all had some suboptimal primary transcript choices from a splice junction recall perspective. *T. cacao* in particular showed the greatest scope for improvement. In total, 19,4%, 16.8%, and 16.9% of low, medium, and highly expressed genes respectively, had suboptimal primary transcript choices from a splice junction perspective.

Table 3.12 shows the percentages of genes with alternative transcripts for which the primary transcript is better or worse than an alternative from an RNA-seq coverage perspective. In many cases for all species, slightly better transcripts are available, and in total 32% of marked primary transcripts were suboptimal. Across all genes, *T. cacao* showed the most scope for improvement. Looking at transcripts that differ considerably (a score difference of 0.1 or more), *C. papaya* and *T. cacao* once again come off the worst.

		% multi-transcript genes with lower *jR* scores than an alternative		
	Species	Low	Medium	High
Plants	*A. thaliana*	17.0%	15.4%	14.4%
	B. rapa	13.6%	10.7%	17.0%
	C. papaya	0.0%	N/A	20.0%
	C. rubella	8.0%	7.8%	11.9%
	T. cacao	25.2%	20.6%	21.3%
	All species	19.4%	16.8%	16.9%

Table 3.11: Percentages of suboptimal primary transcript choices from a junction perspective, at various levels of expression. Here the data for *C. papaya* are sparse: there are only 34 genes with alternative transcripts marked, and there were only 14 with a basic level of expression, 2 of which had low expression and 12 of which had high expression. Thus, no conclusions can be drawn in the *C. papaya* case. For all other species, there were sufficient data. All of the species have some primary transcript choices that are suboptimal. Across all levels of expression, *T. cacao* shows the greatest scope for improving primary transcripts from a splice junction recall score perspective.

		% of multi-transcript genes (suboptimal coverage scores)		
	Species	Much better transcript available (scores differ by > 0.1)	Better transcript available	Primary transcript is best transcipt
Plants	*A. thaliana*	4.9%	28.1%	71.8%
	B. rapa	3.8%	27.2%	72.7%
	C. papaya	7.1%	21.4%	78.5%
	C. rubella	3.8%	18.6%	81.3%
	T. cacao	6.6%	38.9%	61.0%
	All species	5.4%	32.0%	68.0%

Table 3.12: Percentages of primary transcripts that are suboptimal choices with respect to coverage score. All species have some suboptimal genes. *T. cacao* has the largest proportion of suboptimal primary transcript choices. The data for *C. papaya* here are very sparse. When restricting to more highly-differing transcript choices (score difference of 0.1 or more), here *C. papaya* and *T. cacao* prevail once again as the species with the greatest room for improvement, although in all cases few genes exhibit this characteristic.

3.7 Methods

3.7.1 Obtaining mapping junctions

Mapping junctions were extracted from BAM files using RegTools, with a minimum intron size of 4, using the command `regtools junctions extract -i 4`. Junctions were stripped of flanking regions using `awk`, and the minimum read evidence for a junction was set at 10 reads.

3.7.2 Subcellular localisation prediction

Subcellular localisation predictions for amino acid sequences in both the plant and fungal datasets were determined using TargetP (Emanuelsson et al. 2000). TargetP was run with the default options for the fungal sequences. For the plant data set, TargetP was run with the `-P` option to predict chloroplast targeting sequences.

3.7.3 Calculation of entropy score

The localisation consistency $H(O)$ for an orthogroup $O = \{g_1, \ldots, g_k\}$ is defined as an entropy score across the categories observed for each gene:

$$H(O) = -\frac{1}{|O|} \sum_{C \in \mathcal{C}(O)} \frac{|C|}{|O|} \cdot \log\left(\frac{|C|}{|O|}\right)$$

where $\mathcal{C}(O) = \{C_1, \ldots, C_n\}$ is the partition of genes in O into their localisation categories, $C_i \cap C_j = \varnothing \ \forall i, j$, and $\bigcup_{i=1}^{n} C_i = O$.

3.8 Chapter discussion

In this chapter I introduced several extrinsic and non-extrinsic measures for assessing the accuracy of gene models and used them to assess publicly available annotations for the species sets described in Chapter 2. In particular, I developed a non-extrinsic, orthogroup-based approach for identifying gene model errors, and assessed this in relation to other extrinsic and non-extrinsic assessment methods. Finally, I assessed publicly marked primary transcripts and demonstrated that many of them are the suboptimal choice relative to the metrics described here.

3.8.1 Coverage and splice junction assessments are correlated

In Section 3.3 I described two extrinsic metrics for assessing gene model accuracy, the *splice junction recall score* and the *coverage score*. Each has benefits and drawbacks. The splice junction recall score measures the extent to which splice junctions are reflected in mapped read junctions in RNA-seq data. It cannot detect missing features, such as missing exons or introns. The coverage score is able to detect all absent or erroneously present introns and exons, as well as inaccurate exon boundaries, however it can suffer if mapped RNA-seq data are noisy, in particular when alternate transcripts are present. Neither score can detect inaccurate start codons, although genes that are particularly over-extended at the $5'$ end can sometimes receive lower coverage scores. For fungi, both scores were nearly always perfect in all species, reflecting the low incidence of splicing in fungal gene models. For plants, a strong correlation between the two scores was observed ($R^2 = 0.983$, $P = 0.990$). The scores reveal different aspects of gene model errors, and thus this correlation supports their ability to reveal underlying poor quality gene annotations.

3.8.2 *C. papaya* and *T. cacao* show strong error signals

Two plant species, *C. papaya* and *T. cacao*, showed particularly strong signals for missing gene errors for both extrinsic scores considered here. *C. papaya* and *T. cacao* had 34.0% and 20.0% of genes with reasonable levels of expression ($\bar{d} > 10$) and low coverage scores ($C < 0.95$), and 36.4% and 12.2% respectively of genes with high average read depth ($\bar{d} > 100$, very strong evidence) had imperfect splice junction recall scores – the anntotated splice junctions were not supported by the data. This result indicates that these two annotations are in particular need of care and attention, however the other species are not out of the woods: *B. rapa* and *C. rubella* for example had many imperfect splice junctions (3.8% and 6.5% respectively for highly expressed genes), and low coverage scores (8.1% and 10.8% respectively). The two model species tended to have high, but still not perfect, scores. Thus, although *T. cacao* and *C. papaya* show the strongest signals for inaccurate gene models, they are not the only species here for which gene model inadequacies are indicated.

3.8.3 Discrepant genes are evidence for gene model errors

For two of the fungal species used here (*E. gossypii* and *D. hansenii*), adequate RNA-seq data sets were not publicly available. This, in addition to RNA-seq based methods being powerless to detect inaccurate start sites, motivates orthogroup-based, non-extrinsic methods for assessing overall gene model accuracy. In Section 3.4, I detailed an orthogroup-based assessment system based on discrepant sequences for a particular subset of well-flanked regions. This strict subsetting means the error identification system is not applicable to all genes, but can be used to inform an overall relative measure of the accuracy of a set of gene models. In particular, the distribution of discrepancy scores for the various species was broadly in line with what was observed using the extrinsic scores, with *C. papaya* and *T. cacao* both scoring the worst at 7.0% and 3.8% respectively of genes in SCU

orthogroups being discrepant. I showed that discrepant sequences flagged using this method typically had far lower coverage and splice junction recall scores compared with the background, however for fungal genes, in which splicing is rare, no strong signal was seen. Thus, for complex gene models, the method demonstrates a strong capacity for identifying errors in individual genes and for indicating global annotation deficiencies.

3.8.4 Non-extrinsic methods reveal otherwise hidden errors

The method of discrepant orthogroups is particularly useful in highlighting errors that cannot be assessed using extrinsic means. This applies to rarely or very lowly expressed genes, for which data is typically unavailable, and to start codon errors, which cannot be assessed using RNA-seq given the presence of transcribed UTRs upstream of the translation start site. In Section 3.5 I demonstrated that discrepancies at the beginnings of alignments correlate with inconsistent subcellular localisation predictions within orthogroups. Thus orthogroup-based methods for assessing gene model accuracy can reveal aspects of gene model accuracy that elude the more widely used RNA-seq-based methods.

3.8.5 Primary transcript choices are frequently suboptimal

Finally, I considered the set of primary transcripts annotated for the plant data set. These transcripts are intended to be representative gene models for each gene, and are typically the set of gene models used for genome-wide and cross-species analyses. I demonstrated that, by all of the metrics described, many transcripts marked as primary did not actually represent the optimal choice of gene model. In total, 35.1% of orthogroups assessed could have their alignment scores improved by substituting one of the primary transcripts for an alternative. I found that 7.9% of orthogroups with inconsistent subcellular localisation predictions could be made consistent by substituting a primary for an alternate transcript, 16.9% of

highly expressed alternate transcripts had better *jR* scores, and 32.0% had a better coverage score than the marked primary version. This has significant downstream implications for any analysis using these primary transcripts. Thus, the assessment methods used in this chapter could be used not only to improve gene models, but to add more systematic heuristics to the identification of primary transcripts.

3.8.6 Summary

The accuracy measures described in this chapter suggest that many gene models do not appear to be optimal when assessed in relation to extrinsic data and when compared to homologs from other species. Orthogroup-based methods support, and are supported by, the extrinsic methods, but not all discrepancies observed will indicate the actual biological presence of an alternative, more consistent gene model. In many cases, genuine biological differences will have caused exons to shorten or lengthen, introns or exons to be lost or introduced, or start codons to shift. To strengthen the capacity of the orthogroup-based method to assess the extent of gene model errors, attempts must be made to specifically reconstruct in a discrepant gene the concordant features in its homologs. Nevertheless, the methods described here demonstrate a high capacity to identify the prevalence of gene model errors in public annotations.

4

OrthoFiller: Utilising data from multiple species to improve the completeness of genome annotations

The content of this chapter was accepted for publication in BMC Genomics in May 2017 (Dunne and Kelly 2017). A copy of the published version is included for reference as Appendix A.

4.1 Chapter introduction

4.1.1 Background

In Chapter 2 I discussed the problem of missing genes – genes which are present in an organism's DNA sequence but which lack annotations for a given genome assembly. In particular, I discussed methods for quantifying the extent of missing genes in the context of a specific set of species, by using hidden Markov models and a novel, evidence-based probabilistic filtration method to identify the genomic coordinates of putative missing genes.

Due to the possible presence of pseudogenes, retrotransposons, repetitive sequences, and common domains in a genome sequence, even high-quality hits

do not provide complete confidence regarding the genuine presence of a given putative gene. Although they provide strong evidence, the hits obtained using the method in Chapter 2 could still fail to yield valid gene models. Even then, any gene model built around them must display a certain level of similarity to the gene family used to locate them before being deemed a *bona fide* member of the target orthogroup. In order to truly quantify the extent of missing genes from high quality hits, further efforts must be made to ensure gene models resembling the source orthogroup can be constructed around them. In doing so, not only do we get a more accurate method for determining the extent of missing genes, but we get a framework in which to fix such errors.

In this chapter I present OrthoFiller, a self-contained tool designed to address the problem of finding unannotated but present evolutionarily conserved genes and updating and thus improving a species' genome annotation. OrthoFiller aims to simultaneously leverage data from multiple species to mutually improve the genome annotations of all species in a given input set. OrthoFiller extends the work of Chapter 2 by building gene models around high-quality hits and verifying them in the context of their orthogroups. OrthoFiller is run as a python script, and is available to download under the GPLv3 license at `https://github.com/mpdunne/orthofiller`, where instructions for running the program can also be found.

4.1.2 Chapter aims and outline

In this chapter I present the OrthoFiller algorithm and evaluate it under several criteria. By assessing the recovery of genes from artificially depleted plant and fungal gene sets, I demonstrate the accuracy and utility of OrthoFiller for finding missing genes and improving genome annotations. Furthermore, I show that applying OrthoFiller to existing "complete" genome annotations can identify and annotate substantial numbers of overlooked genes in these two sets of species, yielding significant improvements in the completeness of their genome annotations.

The content of this chapter was accepted for publication in BMC genomics in May 2017, and the original published version is included for reference as Appendix A. The chapter is laid out in roughly the same structure as the original published version (Results, Methods, Discussion), and the content is the same with three exceptions. Firstly, material already covered in this thesis is not repeated here, and where applicable I refer the reader back to the relevant section in an earlier chapter. Secondly, the RNA-seq data sources used were changed to match the data sources introduced earlier in this thesis. Thirdly, since publication I identified some bugs in the code relating to running parallel HMM searches on multiple genomes, as well as relating to the proper reporting of novel genes. These issues are discussed in more detail in Section 4.2.9, and the novel found gene set has been updated accordingly. These changes did not affect the other results of the Chapter.

4.2 Results

4.2.1 Algorithm overview and evaluation criteria

OrthoFiller aims to find genes that are present in a species' genome sequence, but which have no predicted gene model in the genome annotation for that species. It takes a probabilistic, orthology-based approach to gene identification, leveraging information from multiple species simultaneously to improve the completeness of the genome annotations for all species under consideration. OrthoFiller is not designed for *ab initio* gene prediction and requires that each genome under consideration possesses a basic level of annotation, taken to be at least 100 annotated genes. This basic level of annotation is required for training the gene finder Augustus, and the more input genes are used the more accurate the training will be. The genomes should ideally be from a set of related species from the same taxonomic group (genus, family, order or class). The algorithm uses three novel scoring and filtration metrics (including the hit quality filter introduced in Chapter 2) to filter gene

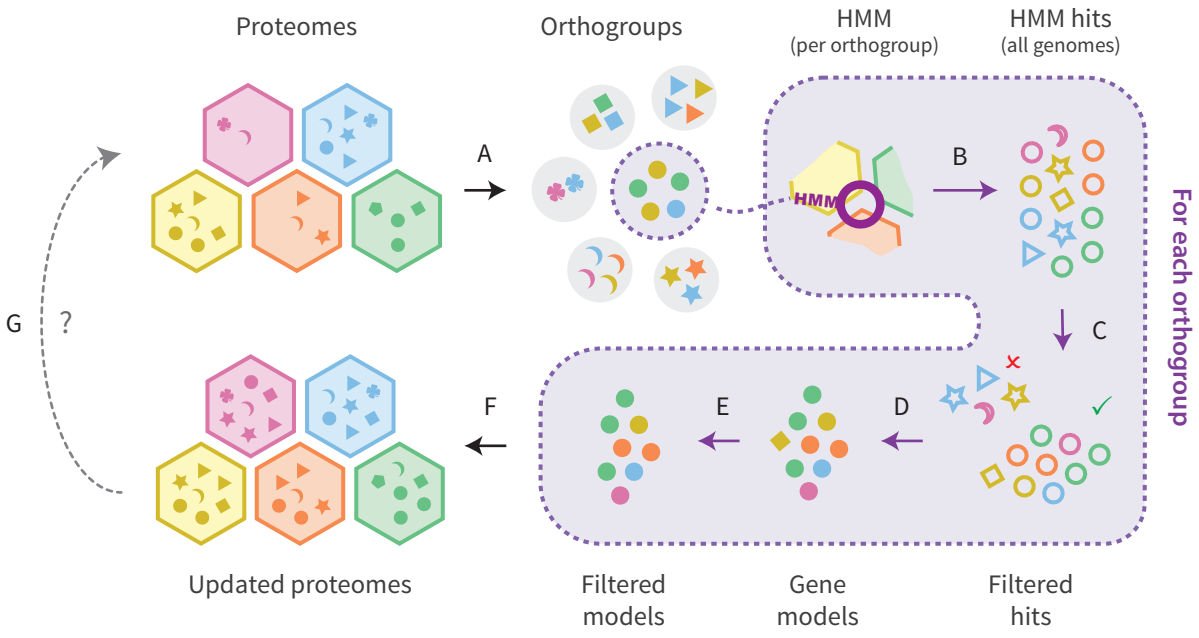


Figure 4.1: Workflow diagram for the OrthoFiller algorithm. A) Annotated proteomes are subdivided into orthogroups using OrthoFinder. B) Protein sequences in each orthogroup are subjected to multiple sequence alignment, back-translated to DNA and used to create hidden Markov models (HMMs). These HMMs used to search each genome in the set. C) The hits are evaluated and low-quality hits are filtered out. D) Gene models are constructed around each retained hit using Augustus. E) The new gene models are compared to the hints that were used to generate them, and filtered to remove those which bear insufficient similarity to the hints. F) The filtered genes are clustered into orthogroups and genes that are successfully placed into the orthogroup that was used to identify them are retained. G) The process may be run once, or iteratively until no further genes are found.

models constructed around high-quality HMM hits, resulting in high-confidence annotations of missing orthogroup members.

An overview of the OrthoFiller algorithm can be seen in Figure 4.1. The input for OrthoFiller is a set of gene coordinates in GTF format and a set of corresponding genome nucleotide FASTA sequences. Protein sequences are extracted from the genome FASTA files using the coordinates in the GTF files and a user-selected translation table. OrthoFiller then uses OrthoFinder (Emms and Kelly 2015) to cluster the genes from all species into orthogroups. The protein sequences of each orthogroup are aligned and the source nucleotide sequences for these proteins are threaded back through the protein multiple sequence alignment to create alignments

of the nucleotide sequences of each orthogroup. Each nucleotide alignment is used to build a hidden Markov model (HMM) that is used to search the entire genome sequence of each species under consideration. The bit scores for hits returned by these HMMs are used to learn the score distributions of true positive and false positive hits, and these are used to filter HMM hits that do not overlap with an existing annotated gene. The hits that pass these filtering criteria are carried forward for further assessment. This part of the OrthoFiller algorithm is essentially the same as the procedure detailed in Chapter 2, Section 2.4, and the details are briefly outlined again in Methods Section 4.3.1.

Each hit that survives this filtration is considered to be a potential genic region, or *hint*. The Augustus gene finder (Stanke and Morgenstern 2005) is used to build gene models around the HMM hints (see Methods Section 4.3.2. Gene models constructed by Augustus are subjected to two successive rounds of assessment and filtration. Firstly, the predicted gene models are compared against the hints that were used to inform them using the *hint F-score* (see Methods Section 4.3.3). If the gene model and its source hint are not sufficiently similar, the gene model is considered to be unrelated to the hint, and thus to the orthogroup used to inform its prediction. Secondly, the newly predicted genes that satisfy the first criterion are subjected to orthogroup inference using the full set of existing and newly predicted genes. Those newly predicted genes that are clustered into an orthogroup whose HMM was used to predict them are then accepted as *bona fide* genes and added to the genome annotation (Methods Section 4.3.4). Thus, genes predicted by OrthoFiller satisfy stringent orthology-based criteria for asserting their existence.

To demonstrate the utility of OrthoFiller on real data it was applied independently to the two species sets introduced in Chapter 2, Section 2.2.1. OrthoFiller was assessed using these datasets in two ways: first via simulating an incomplete genome annotation by randomly removing entries from the genome annotation of the representative best-annotated model species from each set, and assessing

the accuracy of OrthoFiller in recovering the removed genes; second by applying OrthoFiller to the complete datasets and validating the novel detected genes through analysis of publicly available RNA-seq data.

Two measures were used to assess the quality of recovered genes: the protein F-score (pF) and the orthogroup F-score (oF), both defined in Methods Section 4.3.6. These scores were calculated for all genes "recovered" by OrthoFiller (i.e. OrthoFiller results that overlapped one of the removed genes), by comparing the recovered gene with the removed gene and assuming that the original removed gene model was correct. Genes that are unique to the test species that lack homologs in other species were not analysed in this assessment, as OrthoFiller was designed to find evolutionarily conserved genes. As there were no publicly available comparable methods that perform the same task as OrthoFiller, the method was assessed in comparison to performing the analysis without conducting the OrthoFiller evaluation and filtration steps. i.e. using unfiltered HMM hits from the orthogroups as hints for the *de novo* gene finder Augustus, and accepting all identified gene models that did not overlap an existing gene.

4.2.2 Evaluation of OrthoFiller on *A. thaliana* after removal of 10% of gene annotations

OrthoFiller was applied to artificially depleted sets of *A. thaliana*, with 10% of annotations removed, repeated five times with a distinct subset of 2,410 genes removed on each repetition (see Methods Section 4.3.5). OrthoFiller was run on each of these sets along with the full gene sets from the other four species in Species Set P, but only predicting genes for *A. thaliana*. Results were compared to results obtained by running the Augustus gene finder, trained with the input gene set, and omitting the three OrthoFiller filtration steps: that is, using unfiltered HMM hits as hints, and accepting any resulting gene that did not overlap a gene that was already in the annotation. The results of running OrthoFiller and the *de novo*

	10% annotations removed			
	OrthoFiller		*de novo*	
No. genes removed	2410		2410	
Total genes found	1233	(37.5)	13184	(426.7)
Found genes which overlap removed genes	1107	(37.5)	4918	(130.4)
Total recovered genes	1036	(31.72)	2269	(16.4)
Precision	0.90	(0.03)	0.37	(<0.01)
Recall	0.43	(0.01)	0.94	(<0.01)
Number of split genes	67	(5.4)	1214	(23.6)
Mean pF score of found genes	0.75	(0.01)	0.45	(< 0.01)
Mean oF score of found genes	0.94	(< 0.01)	0.51	(< 0.01)
High-quality found genes ($pF \geq 0.95$)	432	(29.3)	641	(27.4)
Lower-quality found genes ($pF < 0.95$)	675	(32.1)	4277	(147.8)
Mean pF score of lower-quality genes	0.61	(0.02)	0.31	(< 0.01)
% of lower-quality genes with $oF < 0.95$	34.6	(1.9)	81.1	(0.9)

Table 4.1: Results for recovery of *A. thaliana* genes in the case of 10% annotation removal, OrthoFiller and *de novo*. Numbers shown are mean values over 5 iterations; bracketed numbers are the standard deviations.

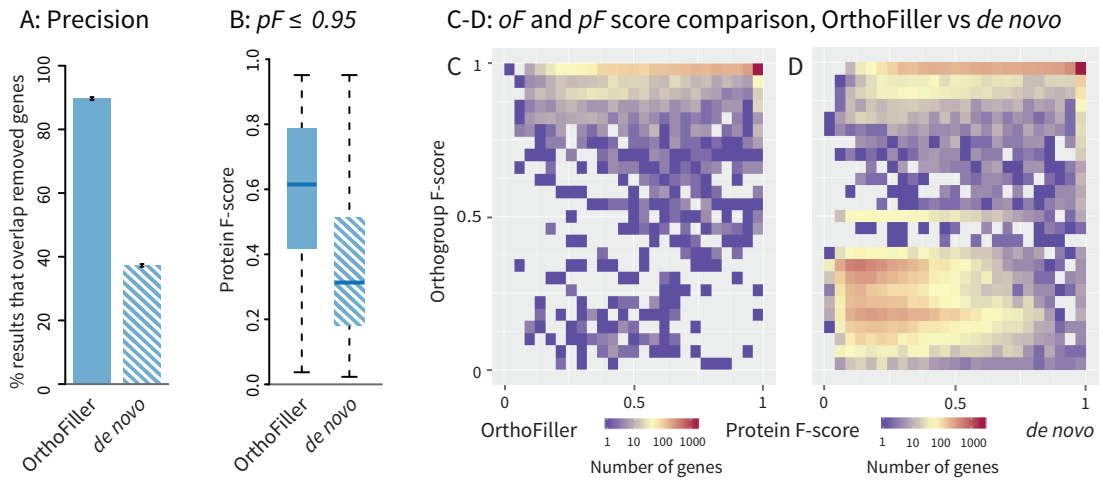


Figure 4.2: OrthoFiller vs *de novo* recovery statistics, plants, 10% depletion. OrthoFiller successfully avoids a large number of poor-quality gene recoveries. A) Using OrthoFiller 89.9% of found genes overlapped removed gene annotations, compared with 37.3% without filtration; B) A boxplot of protein F-scores for genes predicted using OrthoFiller, or in the absence of OrthoFiller filtration, that had a protein F-score of ≤ 0.95. Scores were typically lower for the *de novo* runs. C) Density plot showing the protein and orthogroup F-scores for all recovered genes using OrthoFiller. D) Density plot showing the protein and orthogroup F-scores for all recovered genes in the absence of OrthoFiller filtration.

prediction on these sets are shown in Table 4.1 and Figure 4.2. Here, and for the rest of this chapter, *precision* will be defined as the fraction of found genes which match genes that were removed from the annotation, and *recall* will be defined as the fraction of removed genes that were recovered.

After application of OrthoFiller to the 10% depleted *A. thaliana* data sets, an average of 1,233 genes were output, 89.8% of which overlapped removed genes. Of these, 39.0% were recovered to very high accuracy ($pF \geq 0.95$), and the mean protein score of the lower-quality genes was 0.61. In total, 95% of these lower-quality genes were still placed in the correct orthogroup after orthogroup inference, indicating that the lower protein scores were typically a product of inaccurate gene models rather than the gene itself being incorrect, and orthogroup inference remained reliable.

In the absence of OrthoFiller filtration, an average of 13,184 genes were discovered, vastly more than the 2,410 genes removed, indicating that a large number of these were erroneous: only 37.3% of these genes overlapped removed genes. Of those that overlapped removed genes, 24.7% were split into multiple parts, and only 13.0% were high quality in terms of their protein F-score. The average protein F-score of the lower-quality genes was 0.31, indicating a large number of wrong genes, and 81.1% of these lower-quality genes also had a low (≤ 0.95) orthogroup F-score. Thus in the absence of OrthoFiller filtration, huge numbers of unreliable genes were found.

Figure 4.2 illustrates the results for *A. thaliana* gene recovery with and without OrthoFiller filtration. Figure 4.2A shows the precision, i.e. the percentages of found genes that overlap removed genes, and highlights the disparity between the precision for OrthoFiller vs. *de novo* prediction. Since *A. thaliana* can be assumed to be mostly complete, the many predicted genes reported by the *de novo* method are likely to be spurious. Figure 4.2B shows the distributions of pF scores for recovered genes of lower quality ($pF \leq 0.95$). The lower quality genes are, on

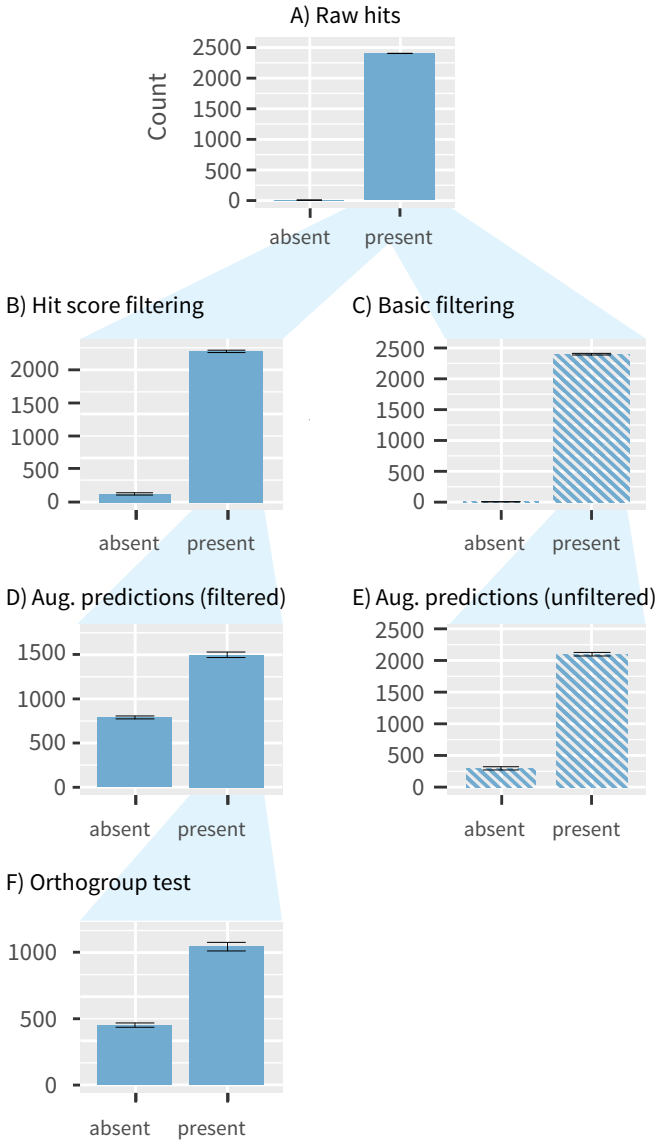


Figure 4.3: Gene recoveries at each OrthoFiller stage, plants, 10% depletion. A) The number of deleted genes that obtained hits from one or more orthogroup HMMs. B) The number of deleted genes that had hits after OrthoFiller hint filtration. C) No hint filtration. D) The number of deleted genes for which a gene prediction was made using Augustus that satisfied OrthoFiller filtration tests. E) The number of deleted genes for which a gene prediction was made using Augustus in the absence of OrthoFiller filtration. F) The number of newly predicted genes that were retained or discarded based on the orthogroup assignment filter step in OrthoFiller.

average, better in the case of OrthoFiller filtration than *de novo*. Figures 4.2C and D further show the distinction between OrthoFiller and *de novo* predictions. Using OrthoFiller, 37.1% of genes had both a very high (≥ 0.95) protein and orthogroup F-score. In the *de novo* case, only 11.4% had both scores very high. Furthermore, in the *de novo* case, 52.4% of recovered genes had both scores very low (< 0.5), compared with 1.4% of genes found using OrthoFiller. Thus in this case using OrthoFiller considerably reduces the proportion of found genes which are erroneous. Although many genes with lower pF and oF scores still make it through OrthoFiller, there is clearly a high abundance of accepted genes in the lower left-hand corner of the F-score comparison graph for *de novo* predicted genes, indicating poor protein and poor orthogroup resemblance between the predicted gene and the gene it has supposedly recovered.

Figure 4.3 shows the progression of recovered genes from 10% depleted *A. thaliana* data sets. All removed genes were hit by some HMMs. Relatively few of these were lost under hit filtering (note that this is only considering hits to removed genes, and considers entire genes being discarded rather than the hits themselves: many spurious hits to other regions of the genome will have been removed at this stage). Most recovered genes are lost at the hint filter stage and the orthogroup test stage. This suggests that inaccurate gene models may be responsible for the incomplete gene annotation recovery.

4.2.3 Evaluation of OrthoFiller on *A. thaliana* after removal of 90% of gene annotations

Table 4.2 shows the results of OrthoFiller upon application to artificially depleted sets of *A. thaliana* gene annotations, this time with 90% of annotations removed. For each iteration, 21,683 genes were removed. OrthoFiller and the *de novo* reduced version were run on these sets in the same manner as for the 10% depleted set.

	90% annotations removed			
	OrthoFiller		*de novo*	
No. genes removed	21683		21683	
Total genes found	11480	(96.2)	42504	(223.5)
Found genes which overlap removed genes	11344	(89.0)	35609	(149.5)
Total recovered genes	10380	(59.7)	20431	(33.0)
Precision	0.99	(<0.01)	0.84	(<0.01)
Recall	0.48	(<0.01)	0.94	(<0.01)
Number of split genes	945	(34.4)	7451	(37.7)
Mean pF score of found genes	0.70	(< 0.01)	0.55	(< 0.01)
Mean oF score of found genes	0.89	(< 0.01)	0.64	(< 0.01)
High-quality found genes ($pF \geq 0.95$)	3420	(21.3)	9080	(99.5)
Lower-quality found genes ($pF < 0.95$)	7924	(99.5)	26530	(201.3)
Mean pF score of lower-quality genes	0.57	(< 0.01)	0.33	(< 0.01)
% of lower-quality genes with $oF < 0.95$	43.4	(0.6)	73.9	(0.1)

Table 4.2: Results for recovery of *A. thaliana* genes in the case of 90% annotation removal, OrthoFiller and *de novo*. Numbers shown are mean values over 5 iterations; bracketed numbers are the standard deviations.

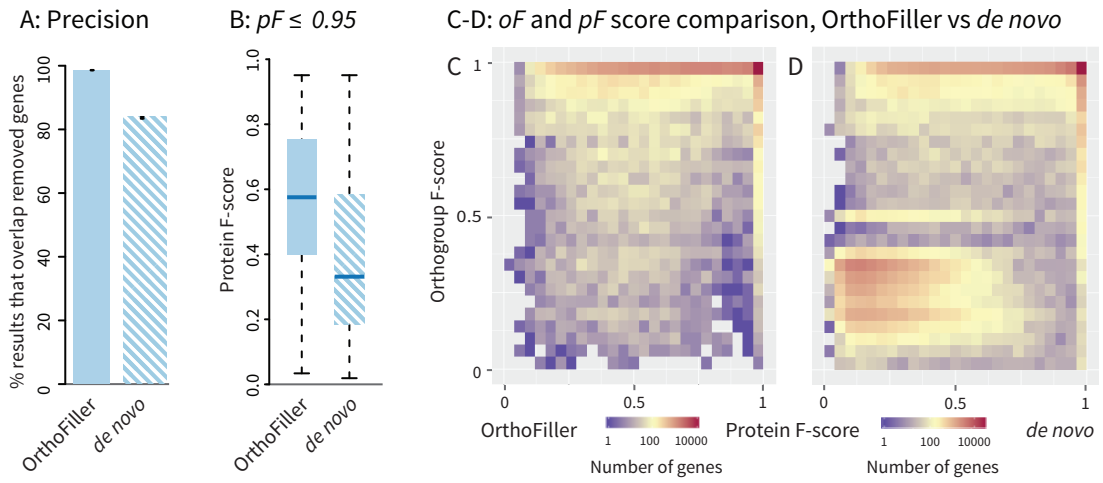


Figure 4.4: OrthoFiller vs *de novo* recovery statistics, plants, 90% depletion. OrthoFiller filters succeed in considerably reducing the proportion of low-quality recovered genes. A) Using OrthoFiller 98.8% of found genes overlapped removed gene annotations, compared with 83.7% without filtration; B) A boxplot of protein F-scores for genes predicted using OrthoFiller, or in the absence of OrthoFiller filtration, that had a protein F-score of ≤ 0.95. C) Density plot showing the protein and orthogroup F-scores for all recovered genes using OrthoFiller. D) Density plot showing the protein and orthogroup F-scores for all recovered genes in the absence of OrthoFiller filtration.

The results from applying OrthoFiller to 90% depleted *A. thaliana* data sets, while still indicating considerable improvement compared with *de novo* prediction, were less clear than in the 10% case in terms of the quality of recovered genes. Using OrthoFiller, 11,480 genes were discovered, of which 98.8% overlapped genes that had been removed. Of these, 30.1% had a very high protein F-score, and the average protein F-score for those of lower quality remained relatively high at 0.57. In total 56.5% of genes with lower protein F-scores ended up in the right orthogroup after orthogroup inference.

In the absence of OrthoFiller filtration, again vastly more genes were found than removed, with an average of 42,504 genes found. Of these, 83.8% overlapped removed genes, however many (20%) were split into multiple parts. Only 25% of the overlapping genes were recovered to high accuracy ($pF \geq 0.95$), and the average pF score for the lower-quality genes was low at 0.33, indicating that a large number of these genes bore no resemblance to the removed gene they overlapped. In addition, 73.9% of lower-quality genes had low-quality orthogroup reclustering, indicating that the proteins were sufficiently mis-predicted that they disrupted orthogroup inference. This shows that, although slightly more genes were recovered in the *de novo* case, considerably more noise and erroneous predictions are produced.

Figure 4.4 shows quality statistics for OrthoFiller and *de novo* gene recovery on the 90% depleted *A. thaliana* data sets. Figures 4.4 C-D show the distribution of orthogroup F-scores versus protein F-scores for recovery in the 90% removal case for *A. thaliana*. Using OrthoFiller, 24.6% of genes were recovered to a dually high quality, compared with 20.3% in the *de novo* case. Furthermore, 38.1% of genes recovered without OrthoFiller were of dually low quality ($pF, oF < 0.5$), compared with only 3.4% of genes recovered with OrthoFiller. Thus, the performance characteristics of OrthoFiller on the 10% and 90% plant datasets are broadly consistent (e.g. 42.9% and 47.9% recovery respectively, 39.0% and 30.1%

high-accuracy recoveries respectively), and both contain a considerably smaller proportion of clearly erroneous genes than would be found without filtering.

Figure 4.5 shows the progression of genes recovered from 90% depleted *A. thaliana* that were lost due to filtration steps. Again, all removed genes were hit by some HMMs. Relatively few of these were lost under hit filtering. Most recovered genes are lost at the hint filter stage here. In particular, many of the hits could not be converted into adequate gene models even in the absence of filtering, and thus were lost at the Augustus prediction stage in the *de novo* case.

4.2.4 Evaluation of OrthoFiller on *S. cerevisiae* after removal of 10% of gene annotations

Similarly to the above, OrthoFiller was run on artificially depleted subsets of the *S. cerevisiae* gene set along with the other species from Species Set F. Table 4.3 shows the results of OrthoFiller upon application to artificially 10% depleted sets of *S. cerevisiae* gene annotations (see Methods Section 4.3.5). For the fungi, 10 repetitions were performed, and the gene sets removed each time were disjoint. For each of the 10 iterations, 513 genes were removed. OrthoFiller results were compared to results obtained as above by running the Augustus gene finder, trained with the input gene set, using unfiltered HMM hits as hints, and accepting any resulting gene that did not overlap a gene that was already in the annotation.

After running OrthoFiller on the 10% depleted *S. cerevisiae* data set, an average of 160 genes were predicted in the genome of *S. cerevisiae* that were not present in the submitted genome annotation file. Of these, 98.9% overlapped genes that were deleted from the original annotation. 96.1% of genes were recovered to high accuracy ($pF \geq 0.95$). The mean protein F-score of the remaining genes of lower accuracy ($pF < 0.95$) was still high at 0.89 (Figure 4.6B). All of the genes that had lower gene model accuracy were placed in exactly the same orthogroup as expected when the sequences were subjected to orthogroup inference. Thus, although typically around

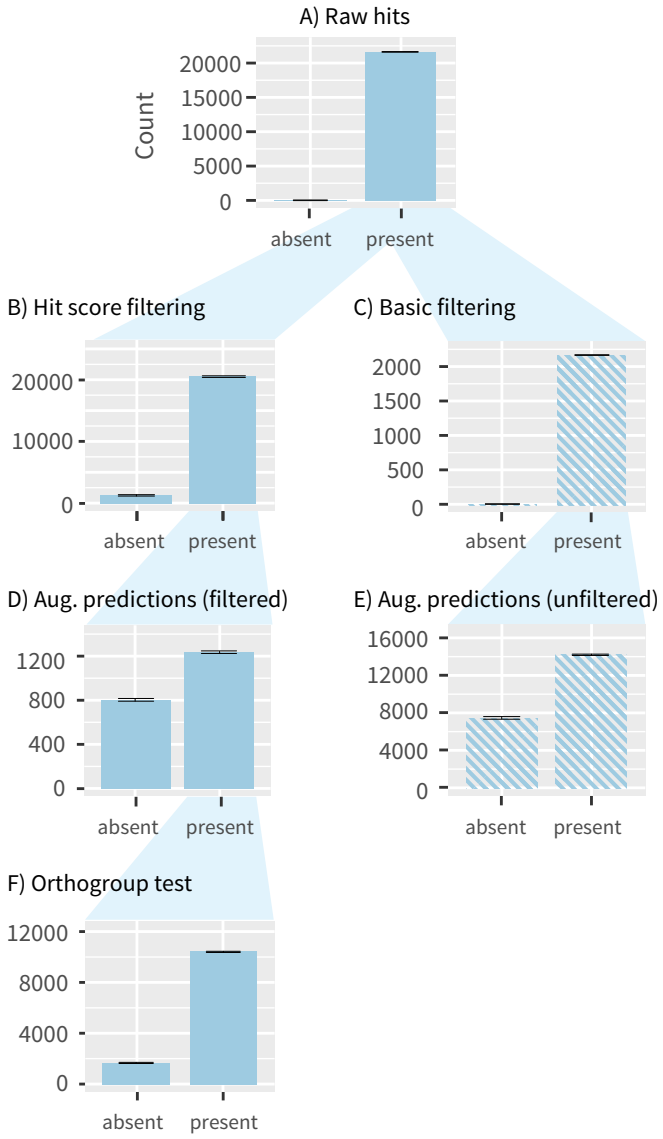


Figure 4.5: Gene recoveries at each OrthoFiller stage, plants, 90% depletion. A) The number of deleted genes that obtained hits from one or more orthogroup HMMs. B) The number of deleted genes that had hits after OrthoFiller hint filtration. C) No hint filtration. D) The number of deleted genes for which a gene prediction was made using Augustus that satisfied OrthoFiller filtration tests. E) The number of deleted genes for which a gene prediction was made using Augustus in the absence of OrthoFiller filtration. F) The number of newly predicted genes that were retained or discarded based on the orthogroup assignment filter step in OrthoFiller.

6 of the gene models differed from the original reference gene model, this difference was not sufficient to disrupt downstream identification of orthologous genes.

To provide a comparison, in the absence of the OrthoFiller evaluation steps an average of 582 genes were identified, of which only 79.2% overlapped with genes that were deleted from the original annotation and an average of 120 genes were not predicted as genes in the original *S. cerevisiae* genome (Figure 4.6A), indicating many of these were erroneous. In addition, on average 11 genes were split upon recovery. Of the genes that overlapped with genes present in *S. cerevisiae* genome annotation 89.3% were genes recovered with high accuracy ($pF \geq 0.95$) and the mean protein F-score of those recovered to a lower accuracy was 0.57 (Figure 4.6B), considerably lower than with OrthoFiller. Of the lower-quality genes, an average of 42.1% had an orthogroup F-score less than or equal to 0.95, compared with 4.0% for OrthoFiller. Moreover, 39.4% of the lower-quality genes were sufficiently mis-predicted that they failed to be placed in an orthogroup, or were placed in an orthogroup that shared no members with the orthogroup that contained the original gene. Thus in the absence of OrthoFiller filtration, more genes were recovered but an average of 11 genes were fragmented, 39.4% of the found genes bore insufficient similarity to the reference gene to facilitate orthogroup inference, and 42.1% were sufficiently mis-predicted that the results of orthogroup inference was altered.

Figure 4.6C-D shows the distribution of orthogroup F-scores versus protein F-scores obtained following application of OrthoFiller to this test dataset. In these panels, results from all 10 runs have been pooled together. The majority of genes recovered with OrthoFiller had both high protein and orthogroup F-scores (Figure 4.6C): 94.4% had both F-scores ≥ 0.95. This indicates that the majority of predicted genes are identical (or nearly identical) to the original removed gene and that when subjected to orthogroup inference they were clustered in the correct orthogroup. Imperfect protein F-scores can be explained by discrepancies in intron/exon and start/stop codon choices between the removed and recovered gene models. Imperfect

	10% annotations removed			
	OrthoFiller		*de novo*	
No. genes removed	513		513	
Total genes found	160	(11.4)	582	(21.2)
Found genes which overlap removed genes	158	(11.2)	462	(11.9)
Total recovered genes	158	(11.1)	447	(9.6)
Precision	0.99	(0.07)	0.79	(0.02)
Recall	0.30	(0.02)	0.87	(0.02)
Number of split genes	0	(0.3)	11	(4.5)
Mean pF score of found genes	0.99	(< 0.01)	0.95	(0.01)
Mean oF score of found genes	0.99	(< 0.01)	0.96	(0.01)
High-quality found genes ($pF \geq 0.95$)	152	(9.1)	412	(9.4)
Lower-quality found genes ($pF < 0.95$)	6	(3.2)	49	(12.2)
Mean pF score of lower-quality genes	0.89	(0.04)	0.57	(0.07)
% of lower-quality genes with $oF < 0.95$	4.0	(8.1)	42.1	(9.3)

Table 4.3: Results for recovery of *S. cerevisiae* genes in the case of 10% annotation removal, OrthoFiller and *de novo*. Numbers shown are mean values over 10 iterations; bracketed numbers are the standard deviations.

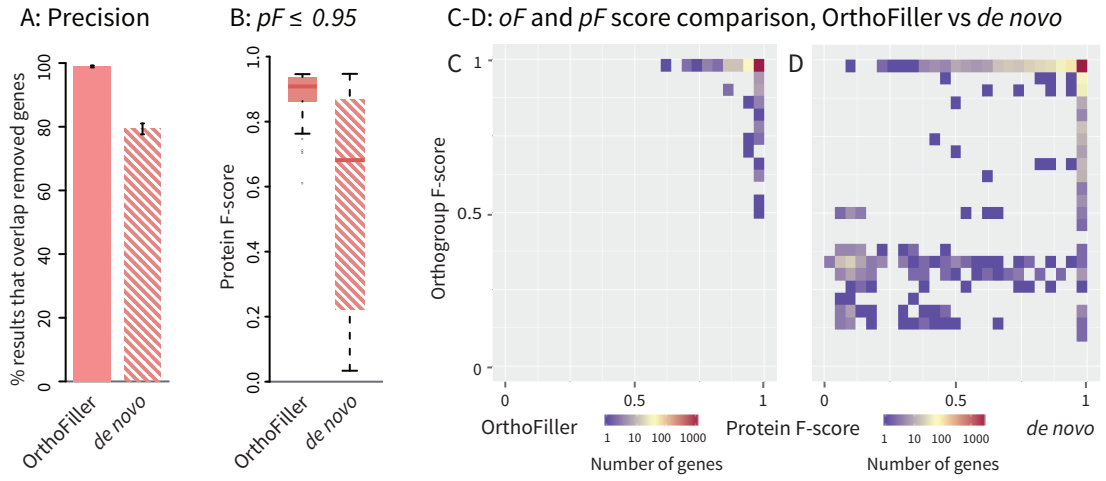


Figure 4.6: OrthoFiller vs *de novo* recovery statistics, fungi, 10% depletion. A) Using OrthoFiller an average of 158 of 160 found genes (99.0%) had genomic locations matching any of the 513 deleted genes. In the absence of OrthoFiller filtration an average of 447 of 582 (79.2%) of the found genes matched any of the deleted genes. B) A boxplot of protein F-scores for genes predicted using OrthoFiller, or in the absence of OrthoFiller filtration, that had a protein F-score of ≤ 0.95. C) Density plot showing the protein and orthogroup F-scores for all recovered genes using OrthoFiller. D) Density plot showing the protein and orthogroup F-scores for all recovered genes in the absence of OrthoFiller filtration.

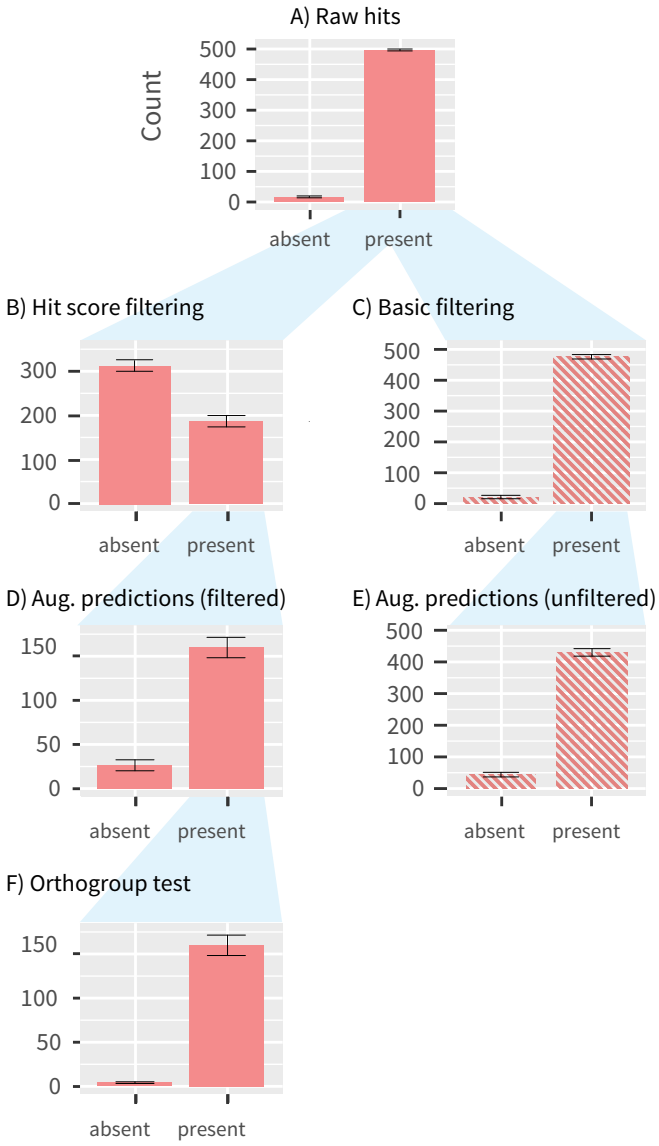


Figure 4.7: Gene recoveries at each OrthoFiller stage, fungi, 10% depletion. A) The number of deleted genes that obtained hits from one or more orthogroup HMMs. B) The number of deleted genes that had hits after OrthoFiller hint filtration. C) No hint filtration. D) The number of deleted genes for which a gene prediction was made using Augustus that satisfied OrthoFiller filtration tests. E) The number of deleted genes for which a gene prediction was made using Augustus in the absence of OrthoFiller filtration. F) The number of newly predicted genes that were retained or discarded based on the orthogroup assignment filter step in OrthoFiller.

orthogroup F-scores were due to fluctuations in orthogroup membership. Figure 4.6D shows the results in the absence of OrthoFiller processing. In this case, 85.3% were of dually high quality. In particular, 3.2% of predicted genes had both a low (<0.5) protein and orthogroup F-score, indicating those predicted genes were sufficiently incorrect to cause errors in orthologous gene identification. Thus, although OrthoFiller does not recover all deleted genes (30.8% of removed genes), application of OrthoFiller resulted in the recovery of high-quality gene annotations that contain few (in this example there are none) incorrectly predicted genes.

Figure 4.7 shows the progression of recovered genes that were lost during each OrthoFiller filtration step. In contrast to the plants, here most genes were discarded due to low-quality hits. Most hits that made it through the hit filtration step went on to produce high-quality gene models.

4.2.5 Evaluation of OrthoFiller on *S. cerevisiae* after removal of 90% of gene annotations

Figure 4.8 shows the performance statistics for OrthoFiller using a version of *S. cerevisiae* genome where 90% of gene annotations were removed (4,615 annotations). This represents an extreme case where a genome has minimal annotation. Again, the experiment was run 10 times, this time with disjoint subsets of 10% of genes left remaining. Here, application of OrthoFiller resulted in the identification of on average 1,472 genes that overlapped any of the removed genes, and typically one or no genes were found that did not match one of the removed genes. Of the found genes, 95.2% were recovered with a protein F-score of 0.95 or greater. Of the genes with lower protein F-scores (Figure 4.8B), only 4.7% had an orthogroup F-score < 0.95. As before, although these gene models differed from the original reference gene model, this difference was not sufficient to disrupt downstream identification of orthologous genes.

	90% annotations removed			
	OrthoFiller		*de novo*	
No. genes removed	4615		4615	
Total genes found	1473	(47.5)	4270	(28.2)
Found genes which overlap removed genes	1472	(47.4)	4163	(18.3)
Total recovered genes	1471	(47.3)	4074	(22.4)
Precision	1.00	(0.03)	0.97	(<0.01)
Recall	0.32	(0.01)	0.88	(<0.01)
Number of split genes	1	(0.7)	89	(8.49)
Mean pF score of found genes	0.99	(< 0.01)	0.96	(< 0.01)
Mean oF score of found genes	0.99	(< 0.01)	0.96	(< 0.01)
High-quality found genes ($pF \geq 0.95$)	1402	(44.76)	3766	(33.6)
Lower-quality found genes ($pF < 0.95$)	71	(5.6)	397	(23.1)
Mean pF score of lower-quality genes	0.88	(< 0.01)	0.63	(0.01)
% of lower-quality genes with $oF < 0.95$	4.8	(1.4)	38.7	(1.1)

Table 4.4: Results for recovery of *S. cerevisiae* genes in the case of 90% annotation removal, OrthoFiller and *de novo*. Numbers shown are mean values over 10 iterations; bracketed numbers are the standard deviations.

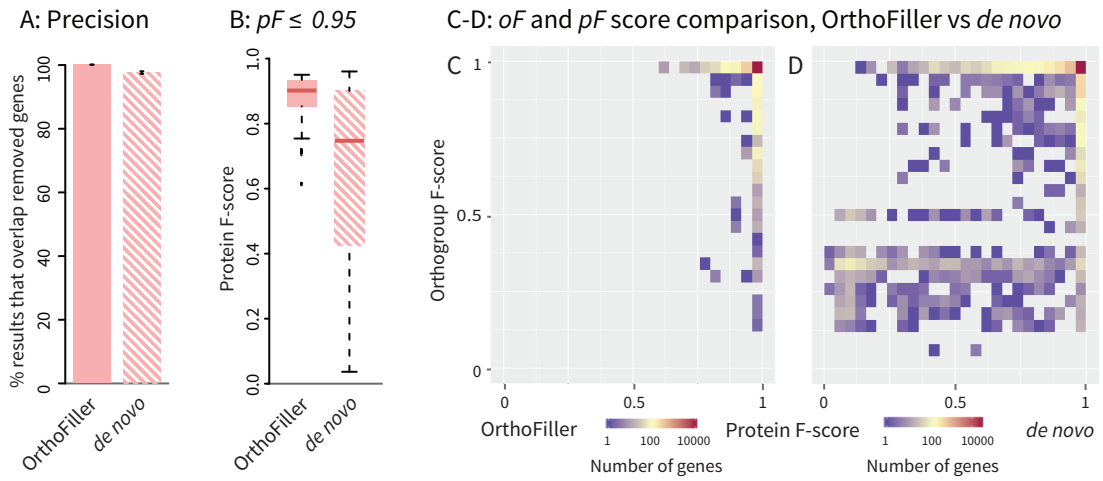


Figure 4.8: OrthoFiller vs *de novo* recovery statistics, fungi, 90% depletion. Far more low-quailty genes are produced in the absence of OrthoFiller filtration. A) Using OrthoFiller an average of 1,472 of 1,473 found genes (> 99.9%) matched any of the 4,615 deleted genes. In the absence of OrthoFiller filtration an average of 4,163 of 4267 (97.5%) of the found genes matched any of the deleted genes. B) A boxplot of protein F-scores for genes predicted using OrthoFiller, or in the absence of OrthoFiller filtration, that had a protein F-score of ≤ 0.95. C) Density plot showing the protein and orthogroup F-scores for all recovered genes using OrthoFiller. D) Density plot showing the protein and orthogroup F-scores for all recovered genes in the absence of OrthoFiller filtration.

In the absence of OrthoFiller filtration, an average of 42,670 genes were found, of which 97.5% overlapped the removed genes. An average of 89 genes were split. On average, 90.4% of the found genes had a protein F-score ≥ 0.95. Of the genes with lower protein F-scores, 39.4% had an orthogroup F-score lower than 0.95, and 32.3% were sufficiently mis-predicted that they failed to be placed in any orthogroup at all, or in an orthogroup completely different to the one that was used to find them. 2.0% of *de novo* predicted genes were of dually low quality ($pF, oF < 0.5$).

Figure 4.8 parts C-D show the distribution of orthogroup F-scores versus protein F-scores for recovery in the 90% removal case. Figure 4.8C shows that most genes were recovered well, with 89.8% genes predicted correctly and placed in the correct orthogroup when subject to orthogroup inference ($pF \geq 0.95$, orthogroup F-score $\geq$ 0.95). Interestingly, there are many genes that are predicted correctly but are placed into a slightly different orthogroup to what was expected. This is due to changes in orthogroup membership caused by the many still-missing genes. Thus, although the input datasets are dramatically different the performance characteristics of OrthoFiller on the 10% and 90% datasets are broadly consistent (e.g. 30.8% and 31.9% recovery respectively, 94.1% and 89.8% high-accuracy recoveries respectively).

Finally, Figure 4.9 shows the quantities of recovered genes that are lost at each filtration stage for the 90% depleted *S. cerevisiae* data set. Again, most genes are lost at the hit filtration stage in OrthoFiller.

4.2.6 OrthoFiller mainly predicts simple gene models

Table 4.5 shows statistics for the genes found using OrthoFiller compared with the *de novo* predicted genes found using Augustus. The fungal genes generated by both methods broadly resemble the gene models from the original annotation, which is unsurprising given that the fungal genes were mostly fairly simple (mean 1.06 and 1.07 exons from OrthoFiller and *de novo* runs respectively, compared with 1.08 in the annotation). The plant genes obtained using both methods typically

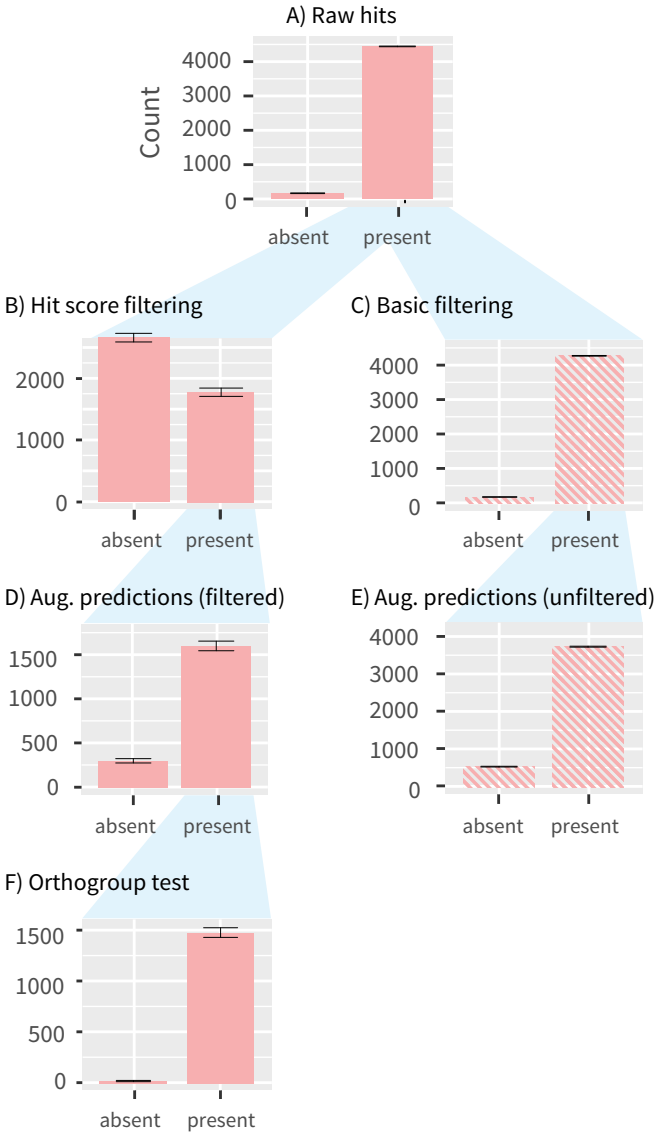


Figure 4.9: Gene recoveries at each OrthoFiller stage, fungi, 90% depletion. A) The number of deleted genes that obtained hits from one or more orthogroup HMMs. B) The number of deleted genes that had hits after OrthoFiller hint filtration. C) No hint filtration. D) The number of deleted genes for which a gene prediction was made using Augustus that satisfied OrthoFiller filtration tests. E) The number of deleted genes for which a gene prediction was made using Augustus in the absence of OrthoFiller filtration. F) The number of newly predicted genes that were retained or discarded based on the orthogroup assignment filter step in OrthoFiller.

			Mean (all found genes)							
			No. exons		Exon size		Intron size		Transcript size	
Plants	10% depleted	OrthoFiller	1.50	(0.90)	576	(597)	178	(191)	862	(634)
		de novo	2.50	(2.10)	266	(359)	203	(242)	662	(727)
	90% depleted	OrthoFiller	1.59	(0.98)	517	(555)	172	(201)	823	(604)
		de novo	2.57	(2.48)	246	(326)	179	(218)	633	(655)
Fungi	10% depleted	OrthoFiller	1.06	(0.25)	1461	(1146)	233	(217)	1546	(1131)
		de novo	1.17	(0.42)	1401	(1221)	195	(206)	1640	(1361)
	90% depleted	OrthoFiller	1.05	(0.22)	1386	(1050)	250	(213)	1453	(1036)
		de novo	1.08	(0.28)	1405	(1122)	220	(198)	1523	(1135)

Table 4.5: Gene model statistics for genes found using OrthoFiller versus *de novo*, for each of the various cases. Values shown are means across all genes from all repetitions, with standard deviations in brackets. Genes recovered by OrthoFiller tend to be simpler, with fewer, larger exons. In addition, genes found using OrthoFiller tend to be slightly longer.

displayed a much lower exon count than the genes in the original annotation (mean 1.55 and 2.54 exons from OrthoFiller and *de novo* runs respectively, compared with 5.17 exons in the annotation). The exon counts were particularly low in the case of OrthoFiller. This is likely because complex genes are less easy to predict accurately, thus the predicted genes in these loci are less likely to match their informant hints and more likely to have been removed at an early stage in the algorithm by the OrthoFiller filtration steps.

4.2.7 OrthoFiller behaves consistently under addition and removal of species

In order to test the behaviour of OrthoFiller under different choices of species quantities and relatedness, the algorithm was run on expanded and reduced versions of species sets P and F. Three additional species were used for each of the fungal and species data set. The additional included species are listed in Table 4.6 as species sets P+ and F+ for plants and fungi respectively. In both cases, the additional

	Species Name	Source	Strain	Taxonomy ID	References
Plants +	*Brassica oleraceae capitata*	JGI	v1.0	3716	[4]
	Citrus clementina	JGI	V1.0	85861	[2,3]
	Populus trichocarpa	JGI	v3.1	3694	[2,4]
Fungi +	*Aspergillus nidulans*	JGI	FGSC A4	227321	[5]
	Schizosaccharomyces pombe	JGI	972h-	4896	[6,7]
	Zygosaccharomyces rouxii	JGI	CBS732	559307	[8]

Table 4.6: Additional species sets used for algorithm assessment, with sources, strain/cultivar information, and references. All data sets were downloaded from the Joint Genome Institute (JGI) website, specifically Phytozome (Goodstein et al. 2012) and Mycocosm (Grigoriev et al. 2014) for plants and fungi respectively. References: [1]Goodstein et al. 2012, [2]Liu et al. 2014, [3]Wu et al. 2014, [4]Tuskan et al. 2006, [5]Galagan et al. 2005, [6]Wood et al. 2002, [7]Wood et al. 2012, [8]Génolevures Consortium 2009

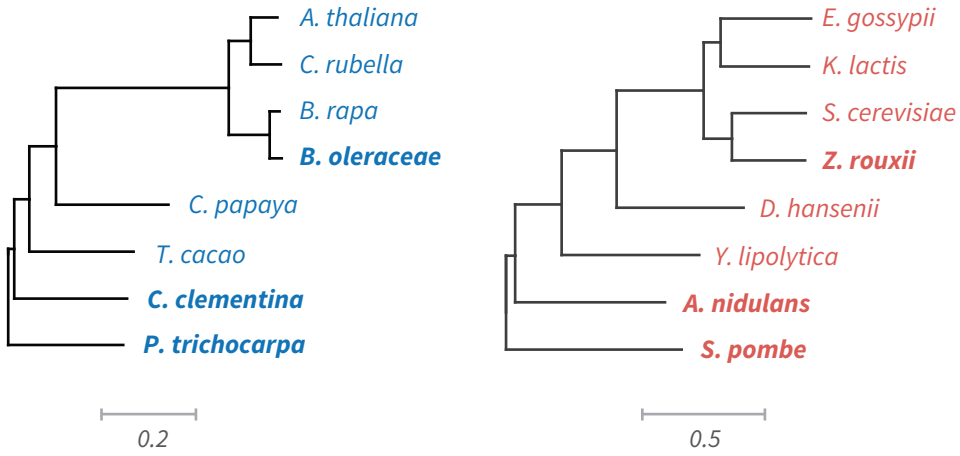


Figure 4.10: Phylogenetic trees for additional species sets. For each species set, an ingroup, an outgroup, and a distant outgroup was chosen.

species were chosen as follows: one species internal to the phylogenetic tree of the existing species set (*B. oleraceae*, *Z. rouxii*), one species external to the phylogenetic tree (*C. clementina*, *A.nidulans*), and one species that was more distantly related (*P. trichocarpa*, *S. pombe*). Phylogenetic trees showing the evolutionary context of the additional species can be seen in Figure 4.10 (see Chapter 2, Methods Section 2.7.3 for details on tree calculation).

For both the plant and fungal species, the behaviour of OrthoFiller was highly

consistent across the species sets, with the exception of the case when only two species are included, as can be seen in Figure 4.11. It is therefore recommended that users include at least three well-related species in an OrthoFiller analysis. Similarly for the fungal species (Figure 4.12), the results were consistent across runs with 3 or more species. Note that for each species set used here the complement of "discoverable" genes was variable due to differences in orthogroup membership, and different random subsets were used for each OrthoFiller run. It is therefore not possible to directly compare the recovery statistics of individual genes across the different expanded and contracted species sets.

4.2.8 OrthoFiller recovers mostly verified fungal genes

The genes recovered from *S. cerevisiae* when applying OrthoFiller after 10% and 90% removal were compared with the dubious/verified classifications provided by the SGD (Dwight et al. 2004), as shown in Table 4.7, averaged over 10 runs with different disjoint subsets removed (disjoint subsets remaining in the 90% case). The SGD classifications mark 5,103 genes as verified, 784 as dubious, and 806 as uncharacterised. OrthoFiller recovered no dubious genes, but a small proportion of recovered genes were uncharacterised (1.83% of genes in the 10% case and 2.26% in the 90% case). Without filtering, a small amount of recovered genes were classified as dubious (0.022% and 0.041% for 10% and 90% depleted gene sets respectively), and considerably more genes were uncharacterised (6.70% and 6.83% for 10% and 90% depleted gene sets respectively). Thus genes predicted with OrthoFiller are more reliable according to the SGD than those predicted without filtering.

4.2.9 OrthoFiller detects hundreds of conserved genes not present in the reference genome annotations

In addition to testing the ability of OrthoFiller to recover already predicted genes from simulated depleted datasets, the algorithm was applied to both of the sets of original annotations from Species Sets P and F, to assess the potential for

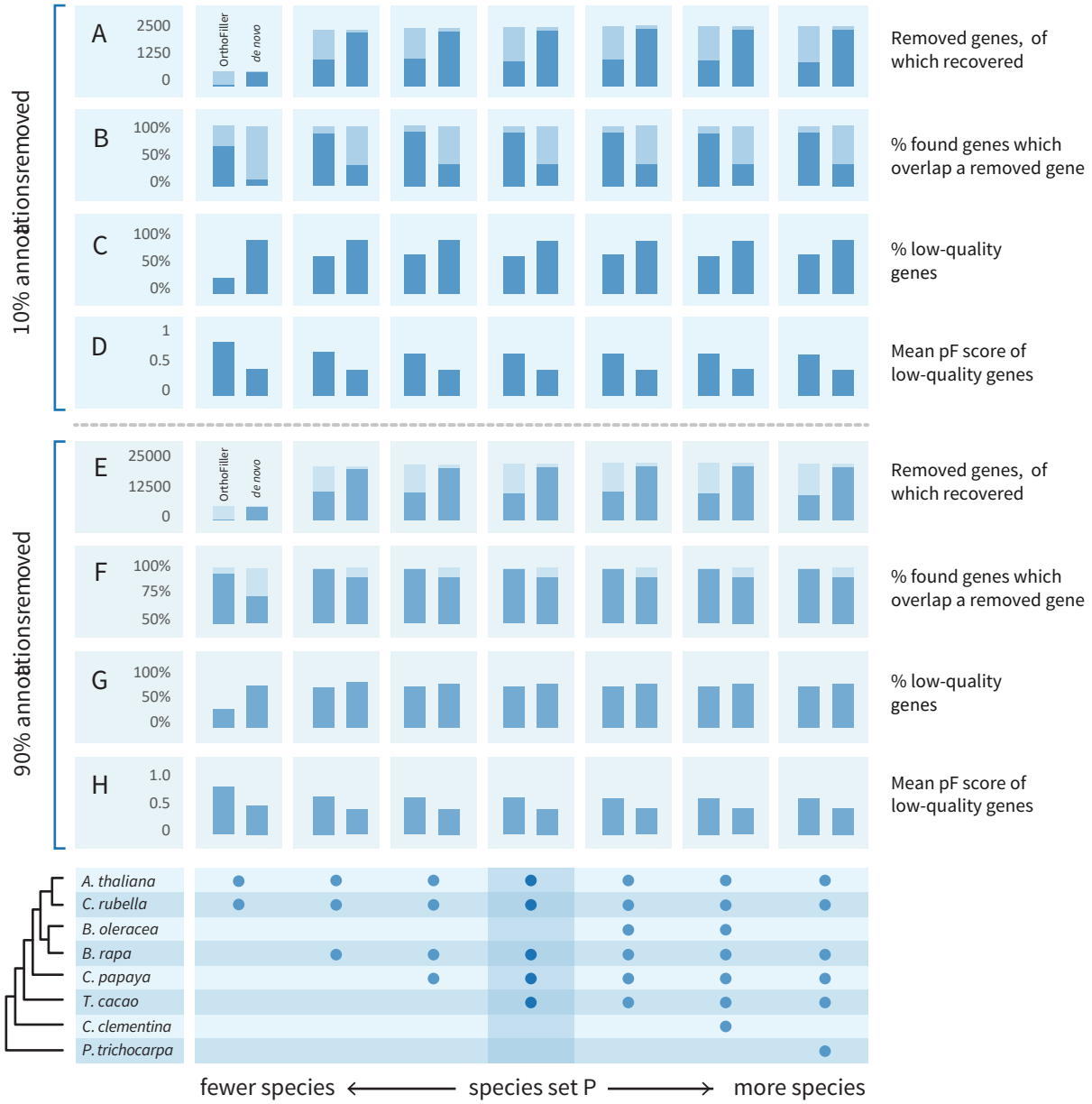


Figure 4.11: OrthoFiller behaves consistently for expanded plant species set. Species sets are indicated by bullets on the bottom panel; paired bars are OrthoFiller results (left) followed by *de novo* results (right). Behaviour of OrthoFiller is consistent across different numbers and relatedness of species, with the exception of when only two species are used. A, E) Number of genes removed from *A. thaliana*, and the proportion recovered. B, F) Proportion of genes found which overlap removed genes is consistently higher for OrthoFiller than for its *de novo* counterpart. C, G) Percentage of recovered genes of low quality, i.e. with $pF < 0.95$, is consistently lower in OrthoFiller than the *de novo* version. D, H) Mean pF score of lower-quality genes is consistently higher for OrthoFiller than for the *de novo* version.



Figure 4.12: OrthoFiller behaves consistently for expanded fungal species set. Species sets are indicated by bullets on the bottom panel; paired bars are OrthoFiller results (left) followed by *de novo* results (right). Behaviour of OrthoFiller is consistent across different numbers and relatedness of species, with the exception of when only two species are used. A, E) Number of genes removed from *S. cerevisiae*, and the proportion recovered. B, F) Proportion of genes found which overlap removed genes is consistently higher for OrthoFiller than for its *de novo* counterpart. C, G) Percentage of recovered genes of low quality, i.e. with $pF < 0.95$, is consistently lower in OrthoFiller than the *de novo* version. D, H) Mean pF score of lower-quality genes is consistently higher for OrthoFiller than for the *de novo* version.

	10% annotations removed		90% annotations removed	
Classification	OrthoFiller	*de novo*	OrthoFiller	*de novo*
Dubious (mean)	0 %	0.022 %	0 %	0.041 %
Verified (mean)	98.17 %	93.28 %	97.74 %	93.13 %
Uncharacterised (mean)	1.83 %	6.70 %	2.26 %	6.83 %

Table 4.7: SGD reliability classifications for recovered genes. OrthoFiller recovers a higher percentage of "verified" genes than the *de novo* method without filtering, for both levels of depletion.

novel genes to be discovered. In total, 32 novel genes were found across the fungal species evaluated, and 861 novel genes were found across the plant species. The number of genes found for each species in each set is listed in Table 4.8, along with some gene model statistics for the new genes. As with what was previously observed, these genes tend to be simpler than the genes that were already annotated for the species in question. GTF coordinates and FASTA amino acid sequences are available to download for all novel genes from `https://doi.org/10.5281/zenodo.1198798`, directory `ch4/1-newGenes`.

Note that the numbers here are different from those in the published version of this chapter as seen in Appendix A. This is for three reasons. Firstly, when running multiple HMM searches on multiple genomes in parallel, Hmmer (version 3.1b2) suffered frequent memory leaks and crashes. This had prevented several of the HMM searches from completing correctly, and OrthoFiller was updated to split up the HMM searches in a more memory-efficient manner. This notably caused an increase in the number of reported genes in *B. rapa*. This did not affect the other results of this chapter, for which searches were only carried out on a single genome. Secondly, some of the genes reported as novel in the published version of this work were in fact already included in the input gene set. These were genes built from hits that had not overlapped the gene in question, usually due to matching an exon that was not annotated in the original transcript due to

	Species	Gene count		Mean (new genes, sizes in bp)							
		Annotated	New	No. exons		Exon size		Intron size		Transcript size	
Plants	*A. thaliana*	27416	136	1.30	(0.56)	492	(523)	250	(211)	640	(570)
	B. rapa	40492	172	1.48	(0.75)	458	(499)	520	(1506)	676	(549)
	C. papaya	27695	228	1.36	(0.69)	451	(458)	1235	(4133)	616	(486)
	C. rubella	26521	249	1.41	(0.83)	491	(493)	294	(413)	692	(597)
	T. cacao	29452	76	1.28	(0.53)	461	(422)	666	(1331)	589	(442)
Fungi	*E. gossypii*	4768	2	1.50	(0.50)	441	(152)	134	(0)	661	(149)
	D. hansenii	6272	13	1.23	(0.58)	1404	(1270)	277	(86)	1728	(1275)
	K. lactis	5076	7	1.29	(0.45)	1638	(1283)	99	(25)	2106	(1601)
	S. cerevisiae	6600	1	1.00	(0)	330	(0)	N/A	(N/A)	330	(0)
	Y. lipolytica	6447	9	1.44	(0.50)	1114	(1814)	185	(135)	1609	(1995)

Table 4.8: Counts of found genes for each of Species Sets P and F. In plants, 863 new genes were found; in fungi, 32 new genes were found. For plants, genes were typically smaller and simpler than was observed for the existing annotated genes. Statistics shown are mean values, with standard deviations in brackets.

gene model error or variation. These hits were subsequently built into genes that overlapped a gene in the input set, but were not explicitly removed. Updating this part of the algorithm notably caused a reduction in the number of new genes found in *C. papaya*. Again, this did not affect the other results of this chapter, as the other analyses explicitly accounted for genes in the input set. Thirdly, the paper reported two new genes in *S. cerevisiae*, and here only one is marked. The other *S. cerevisiae* gene reported in the paper was not included in the *input* gene set, and so OrthoFiller was correct to find it. However, the input set was created from a GFF3 file downloaded from the SGD (Dwight et al. 2004), using all entries marked as `gene`. The found gene reported in the paper was marked in the GFF3 file as a `transposable_element_gene`, namely *YGR109W-A* and so was not included in the input set. Given that this gene was already known, it was not included here. Again, this did not affect the other results of the chapter, as in all other species no distinction between gene types was marked in the downloaded feature files.

	Species	SRA ID	Genes in original annotation		Novel genes	
			Total	W/ reads	Total	W/ reads
Plants	*A. thaliana*	SRR3932355	27416	24514 (89.4%)	136	64 (47.1%)
	B. rapa	SRR2984945	40492	34709 (85.7%)	172	132 (76.7%)
	C. papaya	SRR3509576	27751	23973 (86.4%)	228	152 (66.7%)
	C. rubella	SRR3993756	26521	19244 (72.6%)	249	130 (52.2%)
	T. cacao	SRR3217315	29452	24155 (82.0%)	76	33 (43.4%)
Fungi	*K. lactis*	SRR1200528	5076	5075 (100.0%)	7	7 (100%)
	S. cerevisiae	SRR3396393	6572	6553 (99.7%)	1	1 (100%)
	Y. lipolytica	SRR3647200	6447	6434 (99.8%)	9	8 (88.9%)

Table 4.9: Coverage statistics for novel genes in plant and fungal species. Genes with reads are classed as those with any reads on CDS regions of any transcript variant of each gene.

4.2.10 Novel genes are supported by RNA-seq data

To be detected as a novel gene OrthoFiller requires genes to pass rigorous sequence similarity tests to genes in other species (including empirical evaluation of sequence similarity scores to distinguish real from spurious hits), which in itself provides evidence for the existence of predicted genes through homology. To provide additional evidence for the existence of the novel predicted genes they were subjected to analysis using publicly available RNA-seq data from the Sequence Read Archive (SRA – Leinonen et al. 2010). RNA-seq data sets were obtained and aligned as described in Chapter 2, Methods Sections 2.7.7 and 2.7.8.

Table 4.9 shows the percentages of the novel genes that had RNA-seq evidence. For most genomes, most genes predicted by OrthoFiller are highly supported by RNA-seq evidence, with almost all of the novel fungal genes exhibiting RNA-seq evidence, and 55.2% of novel plant genes being supported. Given that the plant RNA-seq datasets come from single tissue samples under a single condition it is not expected that all genes will be detected in these samples. For example, not all of the original annotated genes from the source datasets had reads. It should also

be noted that genes that are represented in RNA-seq reads are more likely to have been annotated already, given that many genome annotation pipelines rely on such data to perform their analyses (Yandell and Ence 2012). Figures 4.13–4.20 show examples of RNA-seq read coverage for newly predicted genes from each species, alongside gene trees calculated for the newly updated orthogroup (see Chapter 2, Methods Section 2.7.3 for details on tree calculation).

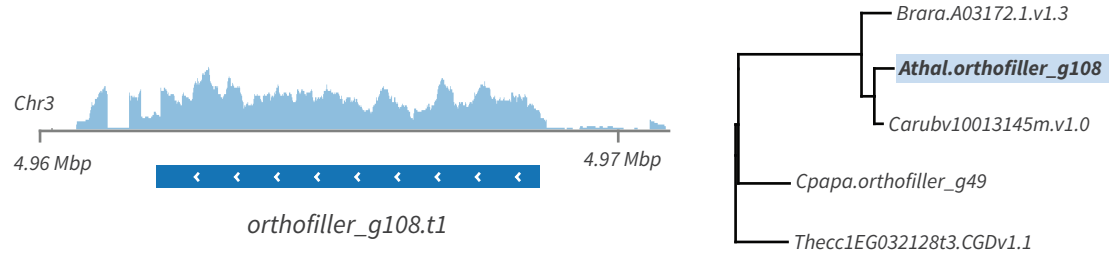


Figure 4.13: A missing gene recovered in *A. thaliana*. This gene, along with another recovered gene in *C. papaya*, completes a single-copy ubiquitous orthogroup.

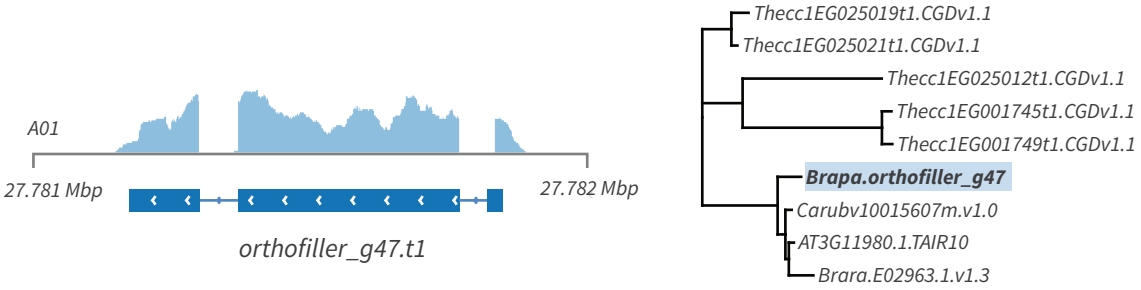


Figure 4.14: A missing gene recovered in *B. rapa*. This gene is present in all other species, but had no homolog annotated in *B. rapa*.

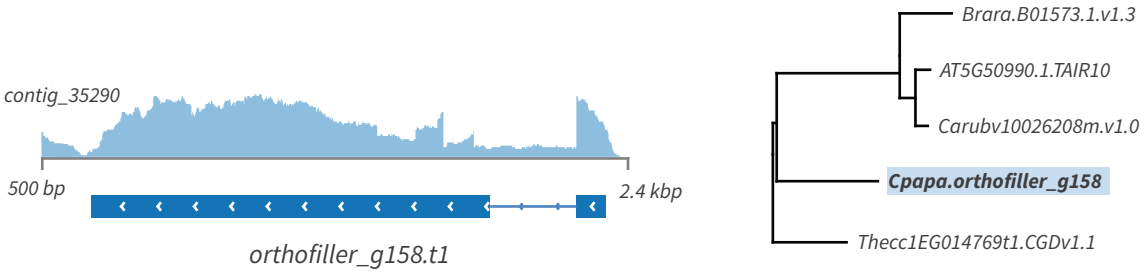


Figure 4.15: A missing gene recovered in *C. papaya*. This gene is single-copy in all other species, but had no homolog annotated in *C. papaya*. The RNA-seq read mapping profile suggests that this gene has multiple alternate splice forms.

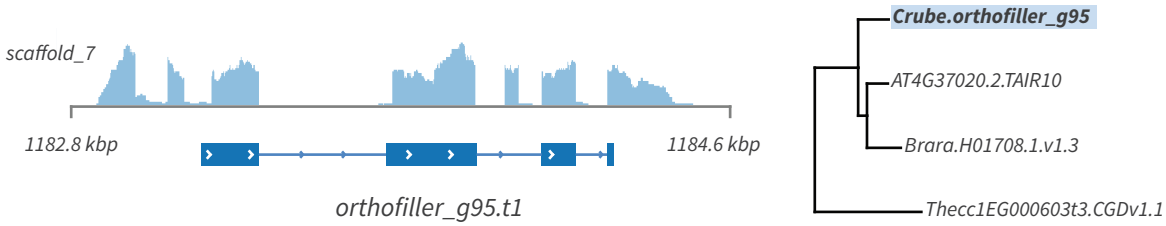


Figure 4.16: A missing gene recovered in *C. rubella*. Here several exons are missing from the recovered gene model. This gene was already annotated in all species except *C. papaya* and *C. rubella*

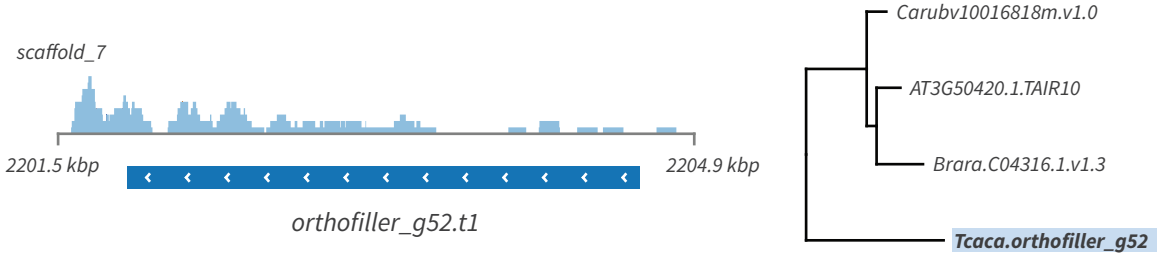


Figure 4.17: A missing gene recovered in *T. cacao*. This gene has fairly low coverage.

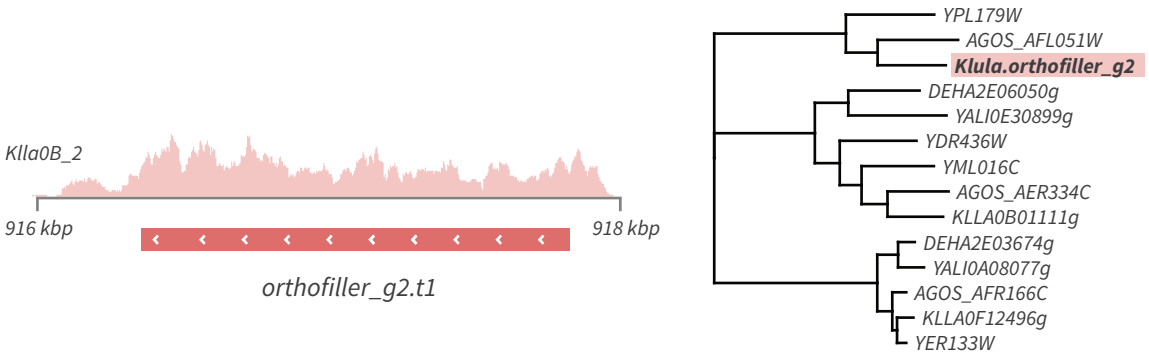


Figure 4.18: A missing gene recovered in *K. lactis*. This gene appears to be multi-copy in all other species, and already has a paralog annotated in *K. lactis*.

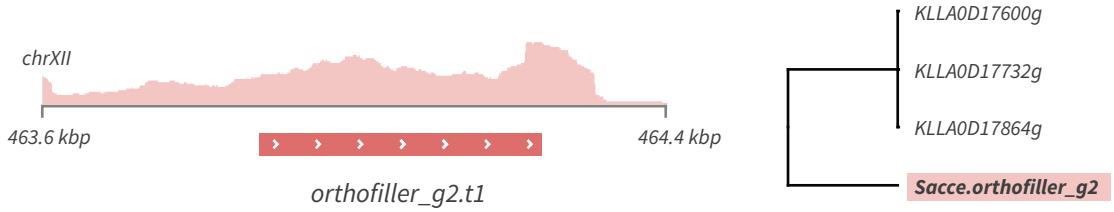


Figure 4.19: A missing gene recovered in *S. cerevisiae*. This gene has three identical homologs in *K. lactis*. Note that this is one of the genes identified in Chapter 2, but on orthogroup reclustering the new gene has displaced the other *S. cerevisiae* gene (*YLR154W-C*) from the orthogroup. This is an issue with OrthoFinder: the two genes are identical except for 16 extra amino acids at the 5′ end of *YLR154W-C* – probably an erroneous translation start site. All the other genes in this orthogroup are 110 amino acids long, and the new gene here matches the *K. lactis* genes more closely than *YLR154W-C*. It is likely that the small size of these genes inhibited proper orthogroup identification.

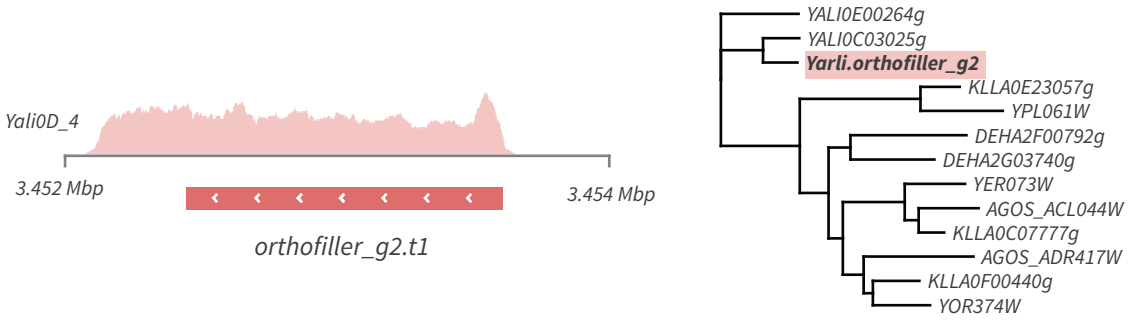


Figure 4.20: A missing gene recovered in *Y. lipolytica*. The gene is a paralog of two already-annotated *Y. lipolytica* genes.

4.3 Methods

4.3.1 Acquisition and evaluation of HMM hits

The first step in the OrthoFiller algorithm is almost identical to the HMM hit filtration system described in Chapter 2, Section 2.4.3. OrthoFiller begins by extracting and translating protein-coding genes according to genomic coordinates specified in inputted GTF gene annotation files. These sets of protein amino acid sequences are then clustered into orthogroups using OrthoFinder (Emms and Kelly 2015), and the resulting subsets are aligned using MAFFT L-INS-i (Katoh and Standley 2013), and back-translated into aligned nucleotide sequences, as described in Chapter 2, Methods Section 2.7.4. HMMs are then formed from the resulting orthogroups, and run across the genomes using Hmmer (Eddy 2011). Distributions for on-target and off-target hits are estimated using GAMLSS in R (Rigby and Stasinopoulos 2005) by fitting skew-T distributions, which are used to generate probabilistic filters that aim to distinguish between relevant and spurious hits. These filters are produced exactly as described in Chapter 2, Section 2.4.3. The additional filter described in Chapter 2, which aimed to dispose of hits from highly repetitive regions, was not used here, as other subsequent filters were used instead. The resulting set of hits is brought forward for further analysis as part of the OrthoFiller algorithm.

4.3.2 Acquisition of putative gene models

Once a set of high-quailty HMM hints has been obtained, the Augustus gene finder (Stanke and Morgenstern 2005) is used to predict genes around those hints. Prior to this, gene model characteristics from already-annotated genes in each species are used to train Augustus, using the included `autoAugTrain.pl` package. Augustus uses the hints as clues to the locations of genes, and preferentially builds genes around them. Each hint is specified to Augustus as an `exonpart`. Aside from the

hint specification, Augustus is run with the default parameters. Once Augustus has been run, any results that do not overlap a hint are discarded.

4.3.3 Evaluation of putative gene models: the hint filter

For any given genic region, it is usually possible to construct some kind of gene model in or near that region. As such, it is plausible that Augustus could create genes in the specified regions of interest that are transcriptionally and translationally valid, but which fail to bear any resemblance to the hint set used to construct them. For this reason, a further filtering step is required to minimise the propagation of such errors. This step will be referred to as *hint filtering*.

The *hint F-score* is a way of assessing how well the hints provided by HMM hits correspond to a given associated gene, be it an already annotated gene or an Augustus-constructed gene model. For a gene G and a set of hits $h_1^O ... h_n^O$, where each hint is derived from the same HMM defined by an orthogroup O, define the hint region for O on G as

$$H_G^O = \bigcup \left\{ h_i^O : h_i^O \text{ overlaps } G \right\}$$

that is, the union of all hits from O which overlap G. For a gene G and a hint region $H = H_G^O$, define the *hint F-score* as

$$hF(H, G) = \frac{2 \cdot hP(H, G) \cdot hR(H, G)}{hP(H, G) + hR(H, G)}$$

where

$$hP(H, G) = \frac{|H \cap G|}{|H|}; \qquad hR(H, G) = \frac{|H \cap G|}{|G|}.$$

As with the hit filter, insights into the construction of a hint filter can be obtained by comparing the results of *on-target* and *off-target* hits. As before, an *on-target* hit is an HMM hit to the genome which overlaps a gene from the orthogroup used to obtain it. Similarly, an *off-target* hit is an HMM hit to the genome which overlaps a gene from an orthogroup other than the one used to

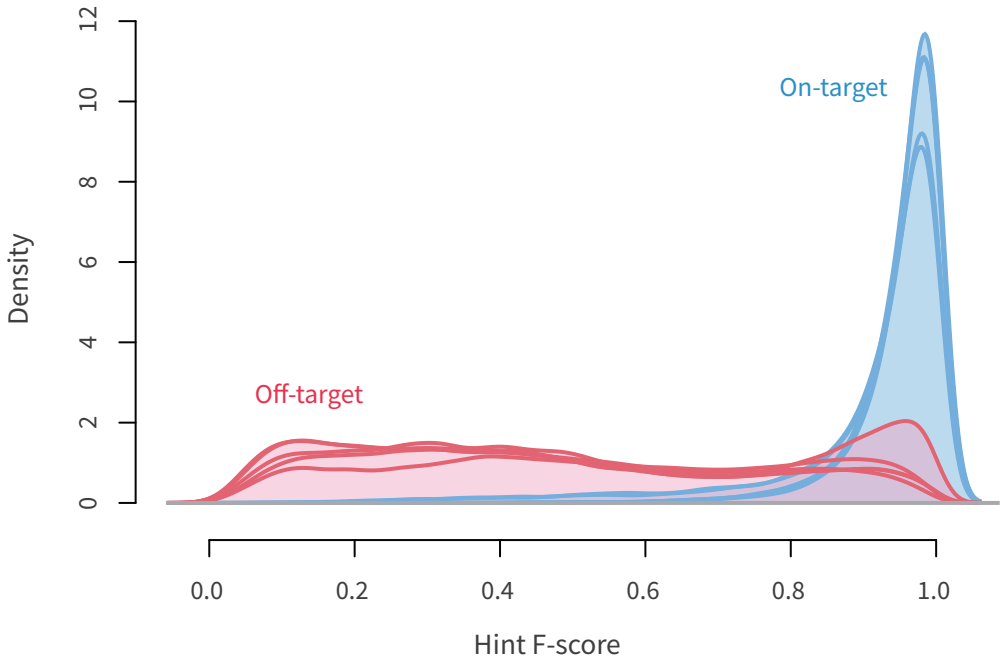


Figure 4.21: Hint F-score distributions for hits to plant genes: kernel density plots of hint F-scores for on-target and off-target hits from plant orthogroups. The on-target hits display a clear peak close to 1.0. The off-target hits have much more varied scores, in particular many off-target hits have scores that match those for on-target hits. A particular peak can be seen for *B. rapa*: this corresponds to large-scale duplications and rearrangements in this genome, in particular paralogs that have not all been clustered into the same orthogroup due to mutations.

obtain it, or which has not been clustered into an orthogroup. In the hit filtration step, this distinction was used to analyse characteristics of the adjusted bit scores for HMM hits. Here, the two hit classes will be used to assess the similarity of the HMM hits to the genes they putatively represent.

Figures 4.21 and 4.22 show the distributions of hint F-scores for the two data sets, for both off-target and on-target hits. Each line represents the data for one species. For both plants and fungi, on-target hits usually bear strong resemblance to the genes that they overlap, as evidenced by the clear peak close to 1.0 from all off-target hits (mean values 0.936 and 0.985 for plants and fungi respectively).

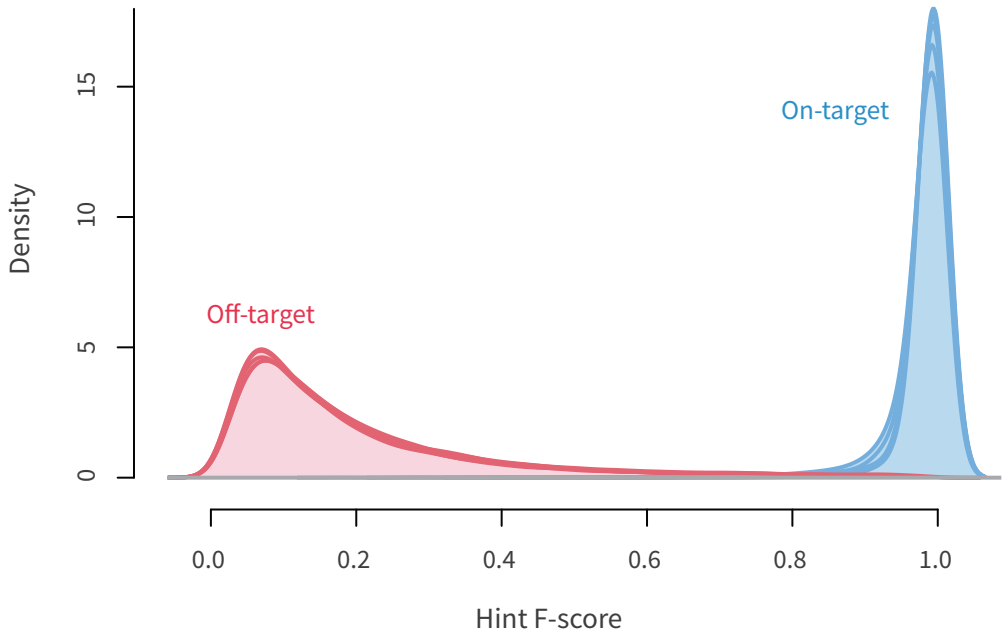


Figure 4.22: Hint F-score distributions for hits to fungal genes: kernel density plots of hint F-scores for on-target and off-target hits from fungal orthogroups. The on-target hits display a clear peak close to 1.0. The off-target hits have typically very low scores..

For both species sets, off-target hits typically bear less resemblance to the genes that they overlap (mean values 0.453 and 0.203 for plants and fungi respectively). This is very clear for the fungi; for the plants, many off-target hits also show high hint F-scores. This is due to higher levels of paralogous genes in plants, which may not always be clustered into the same orthogroup. The highest peak for the off-target hits is for *B. rapa*, which has undergone many whole genome duplication events and thus can be expected to have high levels of paralogous genes. By considering these data, a fixed cut-off value of 0.8 for the hint F-score was determined: any gene models produced by Augustus whose hints do not yield a hint F-score of 0.8 or more are discarded.

4.3.4 Orthogroup membership evaluation

Once putative gene models have been filtered, the original annotated gene sets are reclustered into orthogroups together with all the candidate novel genes that

passed the filtration steps above, again using OrthoFinder. The orthogroup of each newly predicted gene is compared with the orthogroup(s) whose HMMs were used to predict that gene, by assessing the overlap between the original and the new orthogroups.

Any given candidate new gene could have had multiple orthogroups informing its prediction. Furthemore, addition of new genes can cause fluctuations in orthogroup membership between the old and the new proteomes. It is therefore only required that the new orthogroup into which the gene is clustered has non-zero overlap with at least one of the orthogroups used to predict it, and genes which do not fulfil this criterion are discarded. Any gene models that survive this step are returned as OrthoFiller results.

4.3.5 Obtaining simulated data sets

Annotated gene sets for *A. thaliana* and *S. cerevisiae* were depleted artificially by 10% and 90%, i.e. by removing 10% and 90% of all "discoverable" gene annotations. Here, *discoverable* genes are defined to be genes that are contained in an orthogroup containing at least two genes from at least one other species, thus facilitating their potential recovery by the OrthoFiller method. Genes that are not contained in such an orthogroup are unlikely to be recoverable using an orthogroup-based homology approach, and so were not assessed here. These depleted sets simulate genomes with limited annotations. The model species were used here because their annotated gene models are the most likely to be correct, and therefore any recovery can be judged accurately. Furthermore, they provided the best training sets for Augustus.

For each depletion level (10% or 90%), 5 different data sets were produced for *A. thaliana* and 10 different data sets were produced for *S. cerevisiae*. Each of the 10% depleted datasets were disjoint in terms of the sets of genes removed. Similarly, each of the 90% depleted data sets were disjoint in terms of the sets of genes they

had remaining. OrthoFiller was run on each of these 30 data sets together with the full sets of genes from the other species in the relevant species sets.

4.3.6 Assessment of recovered genes

The quality of the predicted genes for the simulated depleted genomes was assessed by considering two scores: the *orthogroup F-score* and the *protein F-score*. Each score captures a different aspect of gene recovery accuracy, and measures the relatedness of a removed gene to genes that "recover" it, i.e. predicted genes which overlap the genomic region defined by the original removed gene. The *protein F-score* is a measure of how well the recovered gene model matches the gene model that was originally specified in the annotation. The *orthogroup F-score* measures the extent to which the orthogroup into which a particular recovered gene is clustered matches the orthogroup containing the original gene annotation. The protein F-score is defined as

$$pF\left(S, S'\right) = \frac{2 \cdot pP\left(S, S'\right) \cdot pR\left(S, S'\right)}{pP\left(S, S'\right) + pR\left(S, S'\right)}$$

where S is the original amino acid sequence, S' is the amino acid sequence of the recovered gene, and

$$pP\left(S, S'\right) = \frac{|S \cap S'|}{|S|}; \qquad pR\left(S, S'\right) = \frac{|S \cap S'|}{|S'|}$$

where the intersection is defined to be the sum of identical amino acids in an alignment of the two sequences (obtained using MAFFT L-INS-i as described in Chapter 2, Methods Section 2.7.4). The orthogroup F-score is defined as

$$oF\left(O, O'\right) = \frac{2 \cdot oP\left(O, O'\right) \cdot oR\left(O, O'\right)}{oP\left(O, O'\right) + oR\left(O, O'\right)}$$

where O is the orthogroup into which the gene is placed when no depletions have been made, O' is the orthogroup into which the gene is placed when orthogroups are calculated using the OrthoFiller results, and

$$oP\left(O, O'\right) = \frac{|O \cap O'|}{|O|}; \qquad oR\left(O, O'\right) = \frac{|O \cap O'|}{|O'|}$$

where the cardinality of the orthogroups takes into account only genes which were present in the input set of genome annotations, i.e. not counting any newly discovered genes.

4.4 Chapter discussion

In this chapter I presented OrthoFiller, an automated method for improving the completeness of genome annotations. It leverages information from multiple taxa, clustering genes into orthogroups and finding genes that are conserved between species but that have escaped detection. OrthoFiller is designed to be stringent, conservatively identifying genes that can be confidently considered to be missing members of existing orthogroups. Specifically, to pass the filtration criteria for detection by OrthoFiller, genes must be members of orthogroups conserved in multiple species. Thus OrthoFiller will not find genes that lack homologs in other species. These stringent criteria mean that not all genes that could be detected will be detected by the algorithm, but rather that the user should have confidence in the validity of genes identified by the method.

OrthoFiller is intended to be run after a genome annotation is considered by the user to be complete or near-complete. OrthoFiller is designed with small scale genome sequencing projects in mind and is provided to enable users without significant resources for comprehensive RNA-seq-based genome annotation to leverage information from related species to improve their genome annotations. However, OrthoFiller is equally suited for use in large-scale genome comparisons, reliably filling gaps in gene sets prior to large scale comparative genomics investigations. Application of OrthoFiller in these cases will enable genes to be analysed in downstream analyses that would otherwise have been classified as absent.

4.4.1 OrthoFiller detected hundreds of new genes

I demonstrated the utility of OrthoFiller on both plant and fungal genome datasets, both in its ability to successfully find missing genes, and in the effectiveness of its filters in eliminating low-quality gene predictions. Application of this method to small groups of plant and fungal genomes resulted in the identification of 861 and 32 previously unannotated genes respectively. These genes are conserved in one or more species but were absent from the genome annotation in which they were predicted. It is anticipated that application of OrthoFiller to larger datasets will likely result in further genome annotation improvement. For the fungi, I compared the new genes to alternative annotations housed at `gryc.inra.fr` (*Genome resources for yeast chromosomes*). Although the genome annotations differ to the ones used in this study, 24 of the new fungal genes predicted by OrthoFiller were also not present in these alternative genome annotations.

4.4.2 OrthoFiller filters out spurious gene predictions

The quality of genes found by OrthoFiller was assessed by artificial removal and recovery of subsets of genes from a single genome, treating the original gene models as true, and evaluating the quality of those genes that were recovered by comparison to the removed genes. In the absence of the OrthoFiller filtration steps, the proportion of poor-quality genes that are recovered is considerably higher, furthermore OrthoFiller avoids the over-prediction of genes endemic to many of its *de novo* counterparts.

OrthoFiller is designed for use on genomes that have already undergone some basic level of annotation. As can be seen by comparing the 10% and 90% removal cases in the two data sets, application to very poorly annotated genomes can result in more genes of dubious quality, both from a sequence and an orthogroup perspective. It is worth noting that many of the genes with lower quality scores, particularly those with only one of the scores being low, can be explained by

alternate gene models (in the protein F-score) and shifting of orthogroups due to expansion of annotated proteome sets (in the orthogroup F-score case). In all cases, in the absence of OrthoFiller filtration considerably higher numbers of genes were predicted that didn't resemble the genes that they were supposed to, indicating that they are erroneous. The results of OrthoFiller were consistent across all expanded and reduced reference species sets that we considered, with the exception of when only two species were used as input. It is therefore recommended that at least three reasonably closely related species are used as input for OrthoFiller.

4.4.3 More careful gene model construction could improve recall

Most of the genes predicted by OrthoFiller were relatively simple in terms of their intron-exon structure, most likely because complex gene models had difficulty passing the stringent hint and orthogroup membership filtering steps employed by OrthoFiller. It is possible that, with more concentrated efforts to ensure the gene models constructed after OrthoFiller's hit filtration step are optimal, greater retention could be achieved, and thus greater applicability to the finding of unannotated genes.

Finally, although OrthoFiller ensures that the gene models it outputs bear a significant similarity to other members of their orthogroups, it is still possible that some of the genes outputted by OrthoFiller are pseudogenes. These could be dead genes, or genes that are in the process of losing function, or have gained another function, and are still being transcribed by the cell. Again, the best way to double-check for the presence of pseudogenes in OrthoFiller results is to take additional steps to optimise the outputted gene models, and without such additional assessments it is difficult to assess the prevalence of pseudogenes in OrthoFiller results.

4.4.4 Summary

The OrthoFiller algorithm is designed to run on a Unix system with python and a minimal number of standard additional tools (Hmmer, BedTools, Augustus, R, OrthoFinder). The software, along with installation and implementation instructions, can be downloaded from `https://github.com/mpdunne/orthofiller`. The speed of the algorithm is principally dependent on the speed of Augustus and Hmmer, however processing time can be decreased by parallelising these steps of the method over multiple CPUs. OrthoFiller discovered hundreds of unannotated genes in published genomes, and demonstrated a considerable reduction in the number of spurious hits found when compared with the Augustus gene finder alone, in the absence of OrthoFiller filtration. Thus, OrthoFiller is able to improve the completeness of genome annotations without introducing large numbers of spurious genes.

5

OMGene: Mutual improvement of gene models through optimisation of evolutionary conservation

The content of this chapter was accepted for publication in BMC Genomics in April 2018 (Dunne and Kelly 2018). A copy of the published version is included for reference as Appendix B.

5.1 Chapter introduction

5.1.1 Background

In Chapter 3 I discussed methods for determining the accuracy of annotated gene models. These included experimental methods (i.e. comparison to RNA-seq data) and homology-based methods (i.e. identification of discrepancies in multiple sequence alignments of orthologs from multiple species). These analyses revealed that putative errors in gene models could be spotted via certain classes of discrepancies in multiple sequence alignments, and that combined with RNA-seq data these could provide estimates for the relative accuracy levels of genome annotations for two sets of species.

Although I observed a correlation between discrepant genes and imperfect extrinsic gene model scores, orthogroup alignment discrepancies are not necessarily indicative of the presence of a better gene model. The amino acid sequence that would optimise the intra-orthogroup agreement of a multiple sequence alignment may not be available as a valid gene model in a species' genome sequence. In addition, the transcript that optimises a multiple sequence alignment may not always be the one that is the most abundant in an organism's transcriptome: in particular, the presence of alternative splicing means multiple possible transcripts may be valid. However, in the absence of extrinsic information, it is parsimonious to choose gene models that maximise intra-orthogroup amino acid alignment agreement.

On an evolutionary scale, genomic regions of genuine biological function are much more likely to be conserved than the rest of a genome (Cooper et al. 2008). This observation has been leveraged in a handful of gene prediction tools (Wiehe and Gebauer-Jung 2001; Gross and Brent 2006; Gross et al. 2007; Alexandersson et al. 2003), all of which combine data from multiple species for parallel *de novo* prediction. By considering orthogroups of related genes, one can optimise the similarity of gene models across species by seeking conserved sequence structure across the various taxa. In this chapter I present OMGene (Optimise My Gene): an algorithm designed to improve genome annotations by optimising the agreement between gene models for orthologous genes in multiple species. OMGene is distinct from parallel *de novo* predictors in that it utilises high-quality existing annotations from multiple species to optimise gene models. It is intended to function without the need for additional empirical data, utilising only the local genome sequences for the genes in question, and works on existing predicted gene models.

OMGene aims to find the most consistent set of representative gene models for a set of inputted genes by seeking to maximise the agreement of their aligned amino acid sequences, returning the single best gene model for each gene. The inputted genes are assumed to belong to a single orthogroup. The algorithm constructs gene

models according to the canonical feature markers: ATG for start codons, GT or GC for splice donor sites, AG for splice acceptor sites, and TAA, TGA, or TAG for stop codons. Other features such as codon bias or poly-pyrimidine tracts are not considered. OMGene can also use non-canonical translation initiation and splice sites if inputted by the user as a command-line option. A standalone implementation of the algorithm is available under the GPLv3 licence at `https://github.com/mpdunne/omgene`. The algorithm is available as a python script, instructions for which, along with example data sets, are included in the GitHub repository.

5.1.2 Chapter aims and outline

In this chapter I describe the OMGene algorithm, and demonstrate its capacity to improve gene models in both publicly available annotations and simulated *de novo* annotated genomes. The content of this chapter was accepted for publication in BMC Genomics on 27th April 2018, and the published version is included for reference as Appendix B. The chapter is structurally the same as the published version (Results, Methods, Discussion), and the content is the same with a few minor differences. Firstly, two minor bugs were identified since the original submission of the manuscript: a bug in BedTools, which affected some of the gene change statistics, and a bug in the coverage score calculator. Each of these bugs were fixed or accounted for in the analysis presented in this chapter, had very minor effects, and did not change the overall conclusions. Secondly, methods and data that were included in the manuscript but were introduced in previous chapters of this thesis are not repeated here, and where applicable I explicitly reference their locations in previous chapters. Finally, some extra details are provided in Section 5.3.6 for the additional mammal species set introduced in this chapter. While this species set was included in the published manuscript, the additional details provided here were not, and are included to provide context for these species in line with Chapter 2, Section 2.2.

5.2 Results

5.2.1 Algorithm overview and evaluation criteria

An overview of the OMGene algorithm is provided in Figure 5.1. The input for OMGene is a set of gene model files in General Transfer Format (GTF – a variant on GFF2 with a more standardised structure than GFF3) and a set of corresponding FASTA genome files. The set of GTF files and their corresponding genome files are inputted to OMGene using a tab-delimited CSV file: each line in the input should contain first the file path to a gene model's GTF file, followed by the path to the genome for that species. Each GTF file should contain coordinates for one gene only. The FASTA file can contain one or multiple sequences but must contain the chromosome/region nucleotide sequence referenced by the GTF file. If the GTF contains multiple transcript variants then these are considered together as variants of a single gene. The suggested pipeline for using OMGene is to determine orthogroups using OrthoFinder (Emms and Kelly 2015), and to apply OMGene to a chosen subset of orthogroups. Further information regarding input formats can be found in the OMGene GitHub repository.

OMGene uses Exonerate (Slater and Birney 2005) as an initial step to cross-align amino acid sequences from all user-supplied genes to all of the genomic regions in which they are contained, in order to find nucleotide regions that correspond to conserved features (see Methods Section 5.3.1). It then combines this information with the original gene models to produce an initial set of prototype gene models, for optimisation. These prototype gene models do not necessarily follow the rules of correct gene model construction, but provide indications for the rough genomic coordinates of the locations of exon candidates. The amino acid sequences for these prototype gene models are then aligned, and the constituent gene parts are split into *adjacency groups* based on overlaps in the alignment (Methods Section 5.3.2). These adjacency groups provide instructions on which parts of the prototype gene

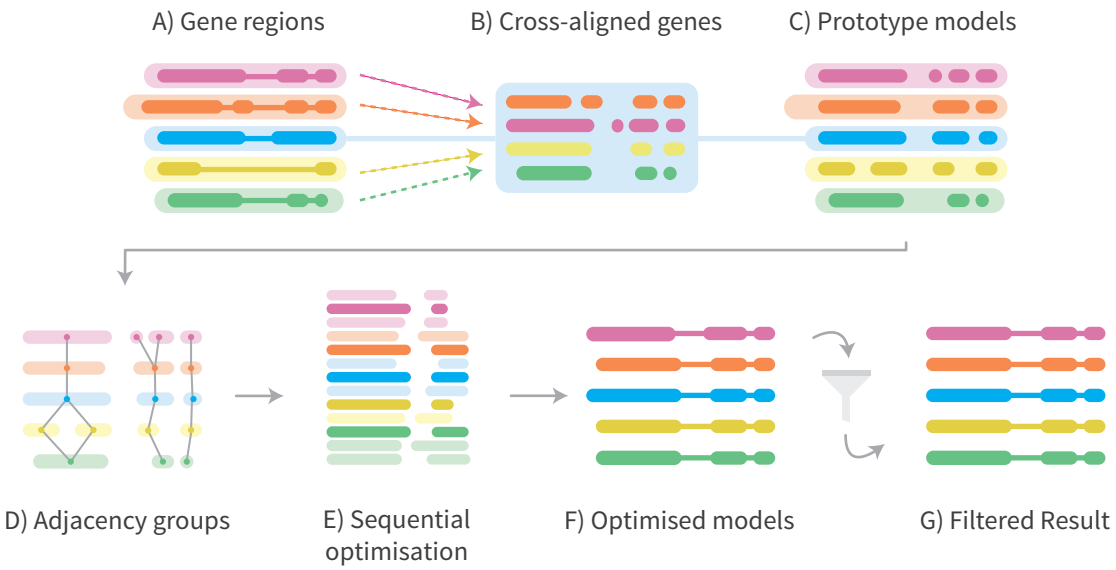


Figure 5.1: Simplified OMGene workflow. A) Gene regions corresponding to the inputted gene models are extracted from the genome sequences; B, C) Exonerate is used to cross-align all constituent coding exons to construct basic prototype "gene models"; D) The exonic regions from these prototype gene models are sorted into adjacency groups. E) The adjacency groups are then sequentially optimised using the multipartite choice function; F) Results are compared against the original gene models to incorporate potentially overlooked combinations, and filtered under various criteria to produce results. G) Optimised filtered gene models are returned to the user.

models for each species or region correspond to one another, and thus which parts should be optimised in tandem. Valid gene models are constructed sequentially by iteratively appending adjacency groups, and the genetic coordinates are recursively adjusted and assessed to optimise the agreement of the amino acid sequences using a multipartite choice function (Methods Section 5.3.3). The resultant gene models are then subjected to stringent filtering criteria (Methods Sections 5.3.4 and 5.3.5) before the finalised set of gene models are presented as sets of GTF coordinates, amino acid FASTA and CDS FASTA sequences.

To demonstrate the utility of OMGene, it was applied to orthogroups formed from three sets of test species: the five plant and five fungal species as introduced in Chapter 2, and an additional set of five mammal species. In addition, OMGene was run on the same sets of species but with all genes from three representative

model organisms – *A. thaliana*, *H. sapiens* and *S. cerevisiae* – replaced with *de novo* predicted genes obtained by running the Augustus (Stanke and Morgenstern, 2005) gene finder on those genomes (see Methods Section 5.3.8). These species were chosen as they have the best annotated genomes and thus the existing gene models will provide the best possible training set for Augustus *de novo* prediction. This *de novo* prediction analysis was done to simulate a typical small-scale genome sequencing project where a user has generated a well-trained set of gene models solely using computational prediction.

OMGene was assessed in three ways: RNA-seq data were used to compare the quality of genes before and after application of OMGene, from both RNA-seq coverage (i.e. the proportion of the predicted gene that is encompassed by reads mapped from RNA-seq data) and splice junction perspectives. To assess the accuracy of start codon predictions, OMGensparee-modified gene models were subjected to subcellular localisation prediction and the results were evaluated for consistency across each orthogroup. In all cases, OMGene showed considerable improvement for most of the genes that had sufficient data for assessment.

5.2.2 OMGene changed thousands of existing annotations

OMGene was applied to all SCU orthogroups from Species Sets P and F, as well as Species Set M, a set of mammal species. The full list of species contained in Set M can be found in Methods Section 5.3.6, along with an analysis of some of the genome and gene characteristics for this species set. Briefly, mammals have large genomes with very large introns and numerous small exons, and thus provide an additional challenge for OMGene. Orthogroups were formed from primary transcript sets using OrthoFinder, as described in Methods Section 5.3.7.

A full breakdown of the numbers of changes made by OMGene, and magnitudes of each change for each species can be found in Table 5.1. The table also shows the percentage of the new gene models that were already present in the annotation as

a non-primary, alternative gene model. Figures 5.2 and 5.3 respectively show the numbers of changed genes for each species and the magnitudes of changes made. Table 5.2 and Figure 5.4 show the distributions of the types of changes made for each species. GTF coordinates for all changed genes, along with various scores calculated for these as discussed below, are included in the Supplementary Material at `https://doi.org/10.5281/zenodo.1198798`, directory `ch5`.

The full set of primary plant gene annotations produced 3,885 SCU orthogroups, containing 19,425 genes. Application of OMGene to this test set resulted in gene model changes to one or more genes in 1169 (30.1%) of these orthogroups. In total, 1,421 of the inputted genes (7.3%) were altered. Of these altered versions, 154 genes (10.8% of 1421) were present in the original annotation as alternative (non-primary) transcripts for the inputted gene. Figure 5.5 shows examples of various types of gene model alteration for genes in *A. thaliana*.

The plant species that experienced the highest number of changes were *C. papaya* and *T. cacao*, which is consistent with them being less well-studied genomes: in particular it is in line with the results observed in Chapter 3. For all species, more nucleotides were removed than were added. In terms of the types of changes made, exon deletion and exon contraction were by far the most commonly seen changes, followed by start codon adjustment (Figure 5.4). This is in line with the observation that gene prediction programs tend to over-predict exons (Guigó et al. 2000), and the fact that commonly used RNA-seq gene model determination methods do not help the annotation of translation start sites. It should be noted that exon deletion events also encapsulate the separation of erroneously fused gene models, which can contribute many exon deletion events simultaneously.

For the mammal data set, 9,251 SCU orthogroups were considered, containing 46,255 genes. Of these, 2,122 orthogroups (22.9%) saw some change, with 2,718 genes (5.9%) undergoing alterations. Most gene changes occurred in *C. lupus*, *M. domestica*, and *O. cuniculus*, which, compared to *M. musculus* and *H. sapiens*, are

	Species	Changed genes	Nucs added/removed (means per change) + (mean)	- (mean)	Net (mean)	Annotated as alt. trans
Plants	*A. thaliana*	138	1648 (42)	-12896 (-63)	-11248 (-46)	53 (30.3%)
	B. rapa	67	1769 (59)	-9122 (-101)	-7353 (-61)	4 (4.1%)
	C. papaya	475	21246 (61)	-51641 (-73)	-30395 (-29)	0 (0.0%)
	C. rubella	255	6301 (71)	-28792 (-81)	-22491 (-51)	2 (0.7%)
	T. cacao	486	3547 (44)	-86073 (-79)	-82526 (-71)	95 (17.1%)
	TOTAL	1421	34511 (59)	-188524 (-77)	-154013 (-51)	154 (9.2%)
	A. thaliana de novo	549	12755 (43)	-139450 (-89)	-126695 (-68)	N/A
Mammals	*C. lupus*	698	8046 (57)	-97384 (-95)	-89338 (-76)	375 (53.7%)
	H. sapiens	397	4412 (59)	-41315 (-88)	-36903 (-68)	218 (54.9%)
	M. domestica	819	8907 (52)	-101626 (-83)	-92719 (-67)	349 (44.3%)
	M. musculus	270	9660 (119)	-19173 (-68)	-9513 (-26)	81 (30.0%)
	O. cuniculus	534	8236 (42)	-68510 (-86)	-60274 (-61)	243 (45.5%)
	TOTAL	2718	39261 (59)	-328008 (-87)	-288747 (-56)	1266 (47.1%)
	H. sapiens de novo	2907	250708 (79)	-952309 (-102)	-701601 (-56)	N/A
Fungi	*E. gossypii*	43	0 (0)	-3806 (-89)	-3806 (-89)	N/A
	D. hansenii	12	0 (0)	-1888 (-157)	-1888 (-157)	N/A
	K. lactis	11	0 (0)	-1308 (-119)	-1308 (-119)	N/A
	S. cerevisiae	9	92 (92)	-511 (-57)	-419 (-42)	N/A
	Y. lipolytica	17	115 (29)	-1997 (-143)	-1882 (-105)	N/A
	TOTAL	92	207 (41)	-9510 (-107)	-9303 (-99)	N/A
	S. cerevisiae de novo	17	597 (119)	-4814 (-321)	-4217 (-211)	N/A

Table 5.1: Quantities and magnitudes of changes made for each species after application of OMGene. In total, thousands of gene model changes were made, and significantly more changes were made for organisms with complex gene models (plants and mammals) than for species with simple gene models (fungi). In all cases, more nucleotides were removed than were added. For the mammal species set especially, many of the updated gene models were already present in the annotation as an alternative, non-primary transcript.

			Exon boundary		Exon		Intron		Moved
	Species	No. changes	Contraction	Extension	Add	Del	Add	Del	start
Plants	*A. thaliana*	245	93	22	4	88	5	12	21
	B. rapa	120	33	13	9	35	2	8	20
	C. papaya	1056	362	196	91	286	17	43	61
	C. rubella	445	109	30	37	87	1	20	161
	T. cacao	1164	436	57	9	544	10	12	96
	TOTAL	3030	1033	318	150	1040	35	95	359
	A. thaliana de novo	1874	756	235	46	747	2	10	78
Mammals	*C. lupus*	1164	300	72	38	418	9	4	323
	H. sapiens	544	90	43	29	223	0	0	159
	M. domestica	1388	342	94	58	478	20	3	393
	M. musculus	362	58	29	42	111	6	6	110
	O. cuniculus	980	284	123	43	245	18	9	258
	TOTAL	4438	1074	361	210	1475	53	22	1243
	H. sapiens de novo	12484	4206	1129	1866	4217	16	17	1033
Fungi	*E. gossypii*	43	0	0	0	0	0	0	43
	D. hansenii	12	0	0	0	1	0	0	11
	K. lactis	11	0	0	0	0	0	0	11
	S. cerevisiae	10	1	0	0	0	0	1	8
	Y. lipolytica	18	0	0	0	2	4	0	12
	TOTAL	94	1	0	0	3	4	1	85
	S. cerevisiae de novo	20	2	2	0	4	0	2	10

Table 5.2: Distributions of types of changes made to each species. For the fungi, the most common change by far was a moved start codon. For the plants, te most common changes were exon contraction and deletion, followed by exon extension and moved start codons. For the mammals, the most common changes were exon deletion, moved start codons, and exon contraction. In the *H. sapiens de novo* case, many genes were completely reworked, causing very large numbers of changes.

less well studied and more recently annotated. Again, by far the most common alterations here were removed exons, exon contractions and moved start codons: it is likely that these are all partially due to the choice of representative gene model (the longest).

For the full fungal data set, 2,771 SCU orthogroups were considered, containing 13,855 genes. Of these, 85 orthogroups (3.1%) exhibited some change, and 94 genes (0.7%) were altered. In this case, *E. gossypii* was the most commonly altered proteome, consistent again with it being one of the lesser-studied species on the list. By far the most common change type in the fungal data set was a moved start codon, consistent with the fact that splicing is a rare event in fungal genes (on average 5.09 exons for plants, 11.23 exons for mammals, 1.08 exons for fungi).

5.2.3 *De novo* annotations are strongly affected by OMGene

To simulate a *de novo* genome annotation project, OMGene was also applied to the same three data sets with *de novo* predicted gene models for representative species, *A. thaliana*, *H. sapiens* and *S. cerevisiae*. These species were chosen as they have the most complete annotations of their respective data sets, and therefore these genes are likely to be the most reliable for training a gene finding algorithm. The *de novo* genome annotation tool used was Augustus as before (see Methods Section 5.3.8).

For the plant data set with Augustus predictions for *A. thaliana*, 3,716 SCU orthogroups were considered. Of these, 549 (14.8%) saw some change in an *A. thaliana* gene. Similarly, for the mammal data set, 7,753 SCU orthogroups were considered, of which 2907 (37.5%) saw changes in a *H. sapiens* gene. For both *A. thaliana* and *H. sapiens* most changed genes underwent several individual changes: in particular, many of the *H. sapiens* genes were completely reworked, causing many individual feature changes. For the fungal data set, 2,715 SCU orthogroups were considered. Of these, 17 (0.6%) saw some change in a *S. cerevisiae* gene. As expected, in all cases, the total number of changes and the average size of change

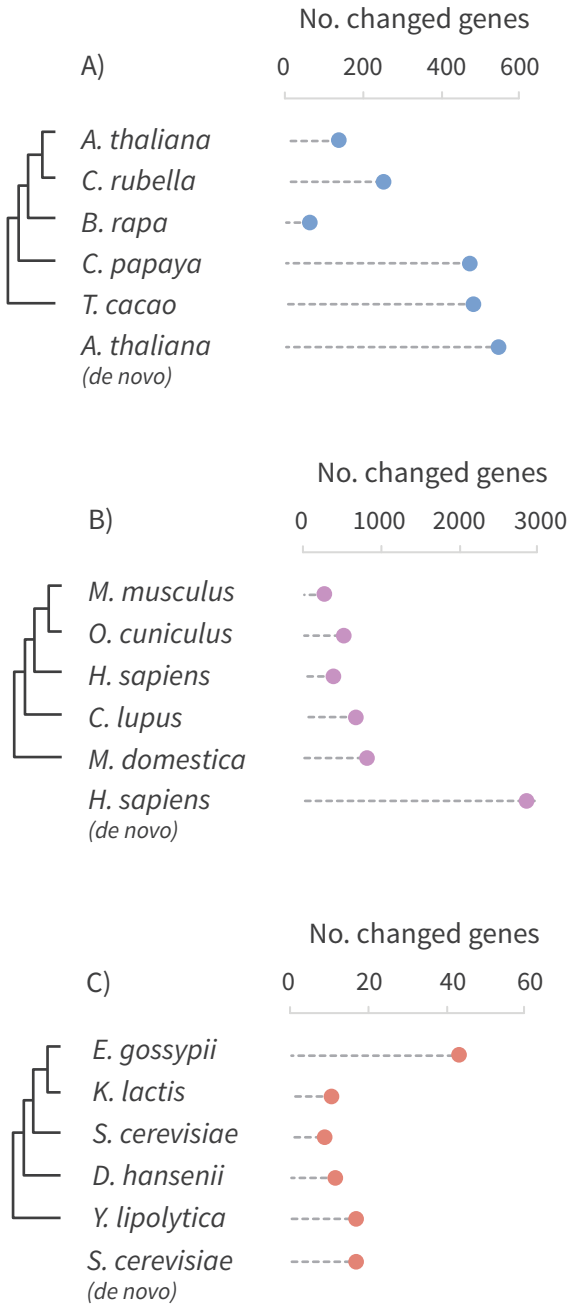


Figure 5.2: Number of genes changed by OMGene for each species. A) *C. papaya* and *T. cacao* experienced the most changed genes in the plant data set. Three times as many genes underwent changes in the *de novo* version of the *A. thaliana* gene set than the publicly available one; B) The number of changed genes in mammals was roughly the same as in plants. As expected, *M. musculus* and *H. sapiens* experienced the fewest changed genes. The *de novo* version of *H. sapiens* underwent considerably more changes than any of the publicly available data. C) The number of changed genes was significantly less for the fungal data set. As in the plants, genes in the *de novo version* of the representative species *S. cerevisiae* were changed more frequently than in the downloaded version.

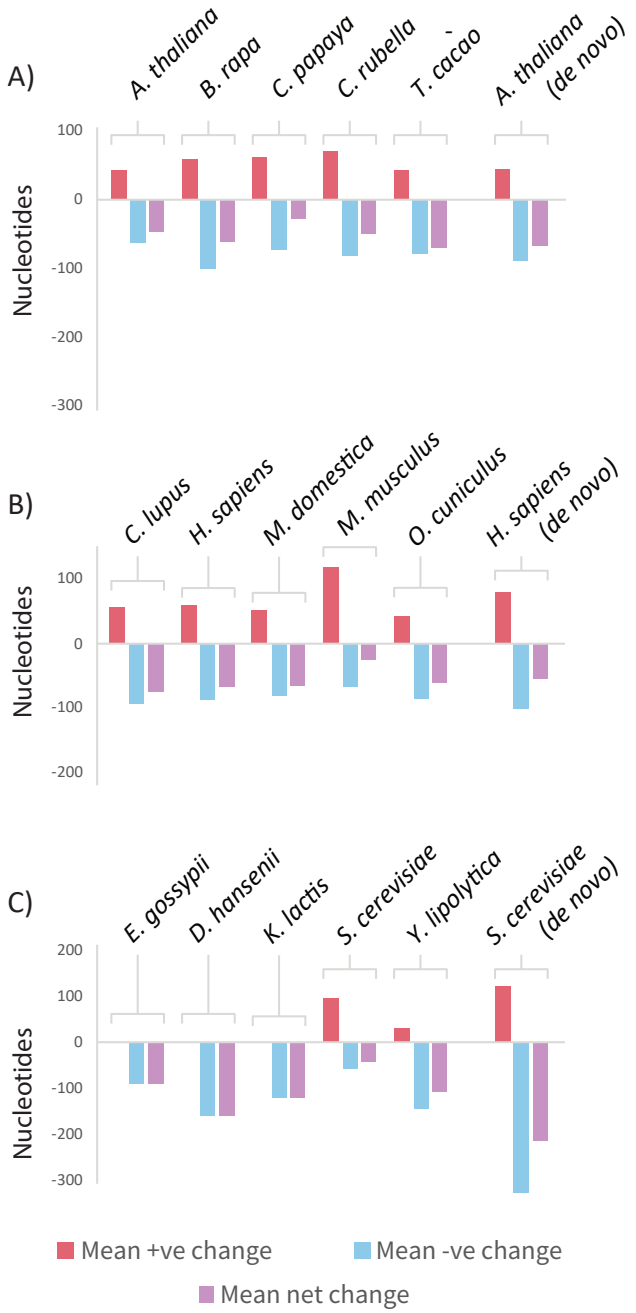


Figure 5.3: Average magnitudes of changes made by OMGene. A) Average magnitudes of each change for plants; B) Average magnitudes for changes made to mammal genes; C) Average magnitudes for changes made to fungal genes. In all cases, the average net change was a removal of genetic material.

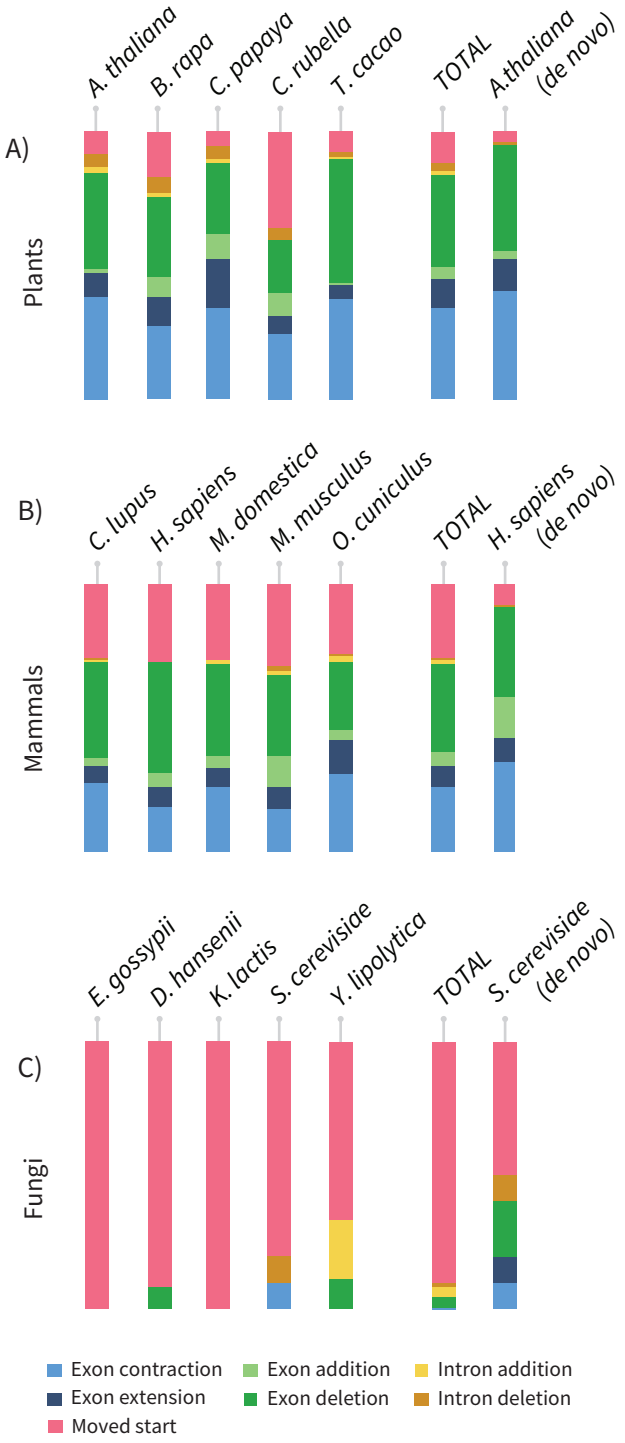


Figure 5.4: Distribution of types of changes made in the three data sets. A) The most common change in plants was exon deletion, followed by exon contraction. B) Moved start codons, contracted exons, and removed exons were most common in mammals. C) In fungi, the most common change was overwhelmingly a moved start codon.

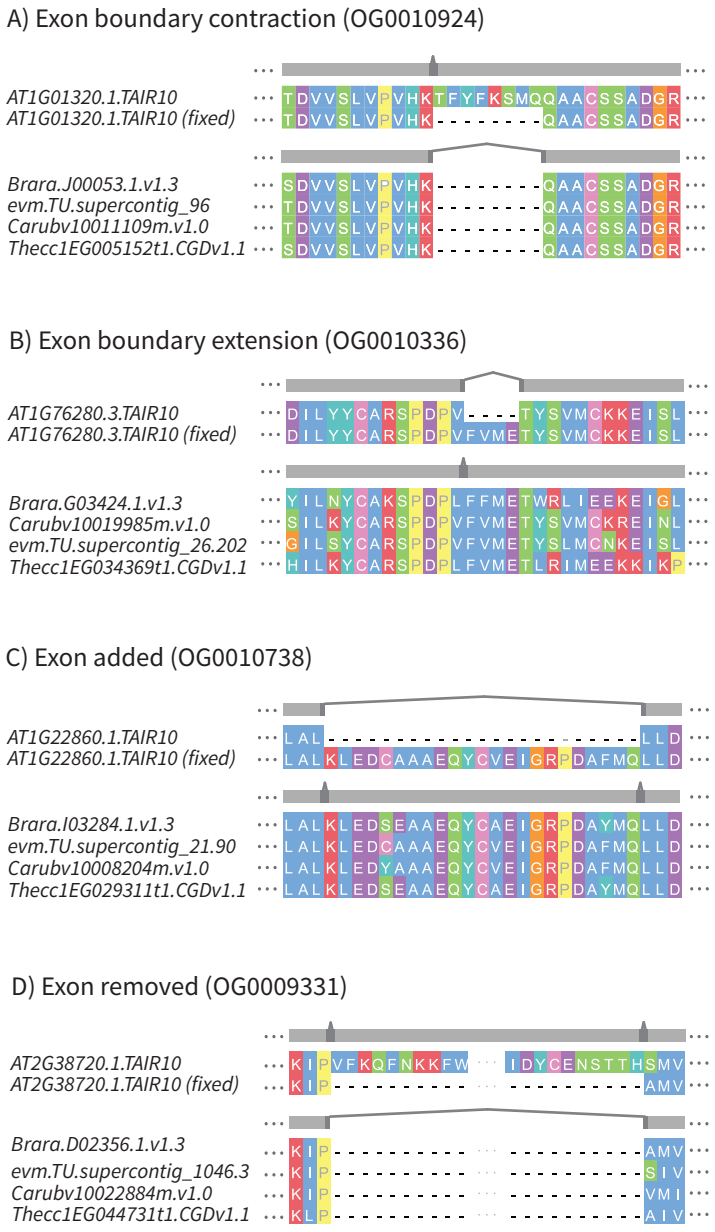


Figure 5.5: Examples of individual gene model changes for genes in *A. thaliana*. A) *AT1G01320.1.TAIR10*, orthogroup OG0010924, exon extension, splice acceptor side; B) *AT1G76280.3.TAIR10*, orthogroup OG10336, exon contraction, splice acceptor side; C) *AT1G22860.1.TAIR10*, orthogroup OG0010738, novel exon introduced; D) *AT2G38720.1.TAIR10*, orthogroup OG0009331, removed exon; *(continued on page 193)*

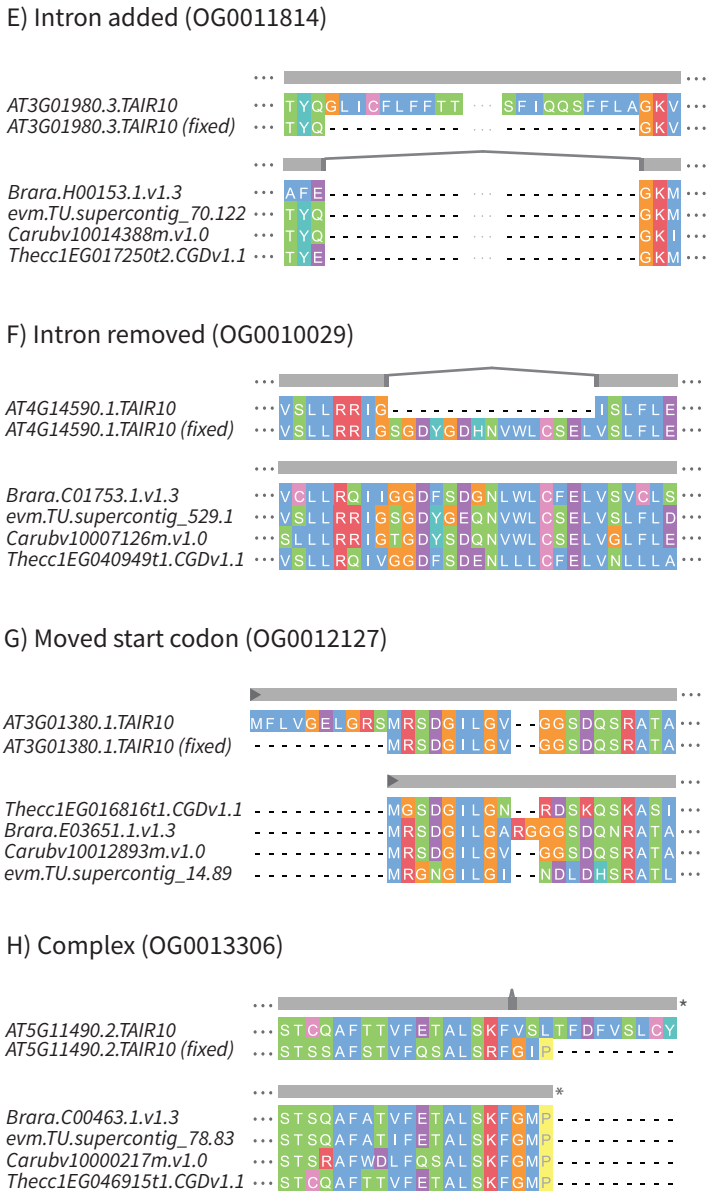


Figure 5.5: *(continued from page 192)* E) *AT3G01980.3.TAIR10*, orthogroup OG0011814, novel intron introduced; F) *AT4G14590.1.TAIR10*, orthogroup OG0010029, intron removed; G) *AT3G01380.1.TAIR10*, orthogroup OG0012127, moved start codon; G) *AT5G11490.2.TAIR10*, orthogroup OG0013306, complex event: exon has been removed and the previous exon boundary has been extended to include the stop codon.

made is greater for the *de novo* predicted gene models than the curated gene models. However, the distribution of types of changes made remained roughly the same.

5.2.4 Complex gene models are improved by OMGene

To assess the validity of changes made by OMGene, both the original and the updated gene model sets were compared using publicly available RNA-seq data from the NCBI Sequence Read Archive (Leinonen et al. 2010) (see Methods Section 5.3.9). Data were downloaded for all species except *E. gossypii* and *D. hansenii*, for which no adequate data were available. Each amended gene was assessed in two ways relative to this data: firstly by comparing the exact splice junction locations with RNA-seq mapping junctions; secondly by evaluating the coverage of exonic regions with RNA-seq.

To control for unreliable data, some "unassessable" genic regions were omitted from this analysis, illustrated in Figure 5.6. Gene regions where the mapped reads contained indels relative to the reference genome were omitted. Such regions prevent the formation of valid gene models, as they frequently disrupt the frame of genes. Genes that were within 1,000 bp of either end of a contig or scaffold, or that contained 10 or more contiguous "N" nucleotide bases were also omitted from the analysis. Such regions prevent the formation of valid genes due to insufficient data being available. Finally, gene regions with very low coverage were also omitted from the junction score analysis: gene regions required at least one base with a read depth of 10 or more. Genes that fail any of these tests are difficult to assess accurately, and so were not included in order to improve the reliability of the results.

Gene models outputted by OMGene were assessed on whether or not their junction F-scores and coverage scores (see Methods Section 5.3.11) had improved or been reduced, relative to mapped RNA-seq data. The full results can be seen in Table 5.3. For the plant data set, splice junctions were improved for 729 assessable genes, while 125 assessable gene models exhibited reduced agreement



Figure 5.6: Assembly issues can prevent accurate gene model formation. A) Genes annotated at the beginnings of scaffolds are often only partial transcripts; B) Genes whose gene regions contain stretches of "N" bases cannot be accurately assessed, as any valid gene model will skip these regions, causing inaccurate introns; C) Indels often cause frameshift errrors, which are corrected by inaccurate small introns in gene models.

(85.3% of changed genes with changed scores were an improvement). Similarly, when assessing RNA-seq coverage of gene models OMGene improved the agreement of the models with the data for 995 genes, while 152 genes exhibited reduced agreement (86.7% improved). For the *de novo* predicted *A. thaliana* genes, the success rates were essentially the same as for the public data (87.3% and 90.3% improved by junction F-score and coverage score respectively), but the absolute quantity of genes exhibiting a changed score increased roughly four-fold. This difference represents the considerable effort and evidence-based curation that has been invested in the *A. thaliana* genome annotation.

For the mammals, 948 assessable changed genes (79.2%) had an improved junction score compared to the original gene model. In addition, 82.4% of assessable genes had improved RNA-seq coverage scores. For the *de novo* data, 89.4% of assessable predicted *H. sapiens* genes had an improved splice junction score after application of OMGene, and 89.3% had an improved coverage score. The numbers of changed genes were again considerably more for the *de novo* predicted *H. sapiens* data, again indicating the high degree of attention that has been afforded to the annotation of this species.

The results for the fungal data set were not as good. Notably very few gene models showed any change in junction F-score, with only 7 genes exhibiting a changed score. This is due to the relatively simple exon structure of fungal genes,

		Splice junction F-score		Coverage score	
	Species	Better	Worse	Better	Worse
Plants	*A. thaliana*	94 (87.8%)	13 (12.1%)	107 (92.2%)	9 (7.8%)
	B. rapa	24 (63.1%)	14 (36.8%)	30 (61.2%)	19 (38.8%)
	C. papaya	246 (82.2%)	53 (17.7%)	338 (83.5%)	67 (16.5%)
	C. rubella	90 (89.1%)	11 (10.8%)	180 (94.7%)	10 (5.3%)
	T. cacao	275 (88.9%)	34 (11.0%)	340 (87.9%)	47 (12.1%)
	TOTAL	729 (85.3%)	125 (14.6%)	995 (86.7%)	152 (13.3%)
	A. thaliana de novo	422 (87.3%)	61 (12.6%)	465 (90.3%)	50 (9.7%)
Mammals	*C. lupus*	323 (85.9%)	53 (14.1%)	479 (86.5%)	75 (13.5%)
	H. sapiens	102 (82.9%)	21 (17.1%)	257 (87.4%)	37 (12.6%)
	M. domestica	239 (72.2%)	92 (27.8%)	434 (76.4%)	134 (23.6%)
	M. musculus	71 (62.3%)	43 (37.7%)	126 (70.4%)	53 (29.6%)
	O. cuniculus	213 (84.2%)	40 (15.8%)	354 (86.8%)	54 (13.2%)
	TOTAL	948 (79.2%)	249 (20.8%)	1650 (82.4%)	353 (17.6%)
	H. sapiens de novo	832 (89.4%)	99 (10.6%)	2031 (89.3%)	244 (10.7%)
Fungi	*K. lactis*	0 (N/A)	0 (N/A)	10 (100.0%)	0 (0%)
	S. cerevisiae	0 (N/A)	0 (N/A)	4 (80.0%)	1 (20.0%)
	Y. lipolytica	2 (28.5%)	5 (71.4%)	7 (63.6%)	4 (36.4%)
	TOTAL	2 (28.5%)	5 (71.4%)	21 (80.8%)	5 (25.0%)
	S. cerevisiae de novo	4 (100%)	0 (0%)	10 (90.9%)	1 (9.1%)

Table 5.3: Quantities of assessable genes whose coverage scores and junction F-scores were changed by OMGene. In most cases, changes made caused both improved junction and coverage scores. The results for the fungi were less clear, as data were sparse.

for which splicing is very rare, and splicing events predicted by OMGene are much less likely to be correct. In this case 2 genes had an improved score, and 5 had a reduced score (28.5% success), with all 5 of the losing genes coming from *Y. lipolytica*. The most common change made to fungal genes was a moved start codon, which, although not detectable in the junction F-score, can sometimes be detectable in the coverage score. This is reflected in the results, where 21 genes

showed an improved coverage score and 5 genes showed a worse coverage score (80.8% improved). In the *de novo* case, again the numbers increased, but here the success improved, with 4 (100%) genes improving by junction for *S. cerevisiae* and 10 (90.9%) improving by coverage score. The highly compact nature of fungal genomes, with few exons and limited space between genes means that the accuracy of *de novo* predicted genes is higher than in plants. Thus the utility of OMGene on these comparatively simpler genomes is limited.

Many of the cases for which OMGene results differ from RNA-seq evidence are attributable to real biological variability that confounds the evaluation criteria of the algorithm. For example, there are some instances where the most evolutionary conserved splice site was not the splice site observed in the RNA-seq data. Such events, by definition, cannot be detected by OMGene. Furthermore, RNA-seq mapping errors also contributed to reduced scores, as did artefacts resulting from spliced UTRs, and jagged read profiles, particularly in the fungal data, which made some coverage scores difficult to calculate reliably. Finally, the presence of multiple transcript isoforms within the RNA-seq data can reduce the score for a valid transcript even if it is the best choice for that particular gene. While users of OMGene should be aware of these confounding factors, the above data demonstrate that, in general, OMGene is much more likely to improve a given gene model than not.

5.2.5 OMGene improves subcellular localisation consistency

As explained in Chapter 3, RNA-seq analyses are typically unable to detect errors in the chosen start codon for a given gene. Start codon choice can be assessed using subcellular localisation prediction. Genes belonging to orthogroups changed by OMGene were assessed to determine whether the changes resulted in increased consistency of their predicted subcellular localisation characteristics of all genes in the orthogroup. Targeting predictions were made using TargetP (Emanuelsson

	Category	No. OGs w/ changed localisation predictions	Entropy score		
			Better	Same	Worse
Pla.	Public data	49	39 (79.6%)	3 (6.1%)	7 (14.3%)
	A. thaliana de novo	10	8 (80.0%)	0 (0%)	2 (20.0%)
Mam.	Public data	509	444 (87.2%)	19 (3.7%)	46 (9.0%)
	H. sapiens de novo	527	458 (86.9%)	23 (4.4%)	46 (8.7%)
Fun.	Public data	6	6 (100.0%)	0 (0%)	0 (0.0%)
	S. cerevisiae	1	0 (0%)	0 (0%)	1 (100%)

Table 5.4: Improvements in subcellular localisation prediction consistency for each species set. In all cases, subcellular localisation predictions were vastly improved after application of OMGene, except in the *S. cerevisiae de novo* case, for which data were scarce.

et al., 2000), and Shannon entropy was calculated to assess the consistency of the predictions within the orthogroups (see Chapter 3, Methods Sections 3.7.2 and 3.7.3). Entropy scores were compared only for orthogroups in which at least one gene model was altered by OMGene. Results of subcellular localisation prediction improvements after application of OMGene can be found in Table 5.4.

The 1,169 plant orthogroups in which one or more genes were altered were subjected to subcellular localisation prediction analysis (Table 5.4). Of these, gene model changes made by OMGene resulted in changes in predicted subcellular localisation for one or more constituent members of 49 orthogroups. In total, 39 improved agreement between gene models (79.6%), 3 remained the same (6.1%), and 14.3% increased entropy and thus increased disagreement between gene models. The effect was more profound for the mammal data: 509 orthogroups showed a change in localisation predictions for one or more of their genes, 444 of which (87.2%) resulted in an increased entropy score. The data for fungi were sparser: only 6 out of 85 changed orthogroups exhibited a change in subcellular localisation prediction, with all of these changes improving the consistency of localisation prediction.

Similar results were obtained for the simulated *de novo* annotation analysis in plants, although again the data were sparse here. Orthogroups containing the *de novo* predicted *A. thaliana* gene were considered together with the four original genes for the other species. Here, 10 of the *A. thaliana* genes experienced a change in subcellular localisation following application of OMGene. Of the 10 orthogroups containing these, 8 improved consistency (80.0%) and 2 reduced the consistency (20.0%). For *de novo H. sapiens* predictions, 527 of the start codon changes resulted in a change in subcellular localisation prediction change for *H. sapiens*. In 458 (86.9%) of these cases, this improved the consistency of the orthogroup. For the fungal data set, the data were extremely sparse, with only one gene experiencing a change in its targeting prediction, which reduced the consistency for its parent orthogroup. Thus, although data were sparse for the fungal dataset, in each of the plant, mammal, and fungal datasets the consistency of gene models was on the whole improved from a subcellular targeting perspective.

5.3 Methods

5.3.1 Cross-aligning genes to find missing exonic regions

For each inputted gene, the algorithm defines its *gene region* to be the region spanning the first and last base of any of its corresponding gene models (including UTRs), with a user-selected number of buffer bases either side. The default buffer size is 600bp: for genomes with large introns a larger buffer size should be used in order to find missing exons at the starts and ends of genes. The initial step of OMGene is to cross-align the amino acid sequences from each gene with the gene regions of the other genes, using Exonerate (Slater and Birney 2005). The rationale behind this step is to find exonic regions that are present in one or more of the gene models but absent from one or more of the other annotated gene models. Exonerate is run with options `--frameshift -10000` and `--model protein2genome`.

Exonerate is applied three times: first by cross-aligning the input protein sequences against all gene regions, second by cross-aligning the protein sequences that have been found in the first step against all gene regions, and finally by cross-aligning all individual exon sequences from the first step. This three-step process mitigates against lack of detection due to gene model errors in one or more of the input genes. The results of this stage, together with the exons from the original gene sequences, provides a set of potential gene parts, which may overlap and which may be incompatible in reading frame (i.e. overlapping regions that are in different frames). Compatible regions are merged; incompatible regions need to be resolved.

For each set of incompatible regions, multiple options are formed in two ways: firstly by keeping one of the other regions, and secondly by iteratively bisecting the regions to include half of one region and half of the other, until a compatible set is produced. Compatible combinations of gene parts produced in this way (of which there can be many) are strung together to form putative prototype gene models. The set of prototype gene models (one from each gene region) that has the highest alignment score α (see Chapter 3, Section 3.4.1) is carried forward to the next step. If multiple alternatives have the same alignment score, the choice is made arbitrarily.

5.3.2 Adjacency group calculation

OMGene builds genes sequentially by iteratively adding in putative exons to multiple genes simultaneously. Care must be taken to ensure the putative exons are added in a way conducive to vertical comparison of relevant regions. It cannot always be assumed that exon numbers for each gene match up vertically within the alignment, given the possible disruptive effects of either erroneously or biologically missing or present exons or introns. A decision strategy therefore needs to be implemented to decide which exons "match" with each other across the various genes in the alignment, and thus which should be considered together in the optimisation process. In OMGene, gene parts are considered in sequential *adjacency groups*

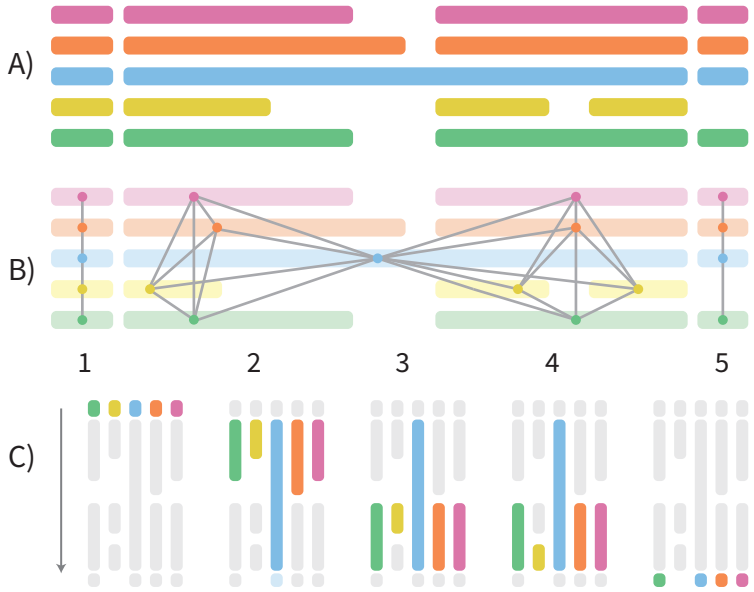


Figure 5.7: Calculation of adjacency groups. A) Amino acid sequences for individual putative exons are strung together and aligned. B) A graph is formed with vertices formed by gene parts (or exons), and edges drawn when the overlap between two parts is greater than or equal to two thirds the length of one of them. C) Cliques are extracted and then ordered lexicographically to form the adjacency groups.

based on their coordinates in a multiple sequence alignment. Exons are added in sequentially to these adjacency groups, and at each stage a valid gene model is sought on the left-hand side of the gene (i.e. starting at the start codon and seeking to adjoin exons in valid donor-acceptor pairs).

Adjacency groups are obtained as follows. Prototype gene models obtained as described in Methods Section 5.3.1 are formed by stringing together amino acid sequences for individual putative exons for each gene region, which are then aligned. The putative exons are grouped by considering them as a graph: each putative exon is a node on the graph, and two exons are connected by an edge if and only if one of the exons overlaps the other by a third or more of its length. The adjacency groups are then defined to be cliques in this graph. Cliques are determined with the python implementation of the NetworkX package (Hagberg et al. 2008), using the function `find_cliques`. Figure 5.7 illustrates the process.

5.3.3 A choice function for efficient best subset acquisition

Adjacency groups of putative exons are added sequentially and optimised at each junction (the term "junction" will be used here to also include initial exons, which are optimised for start codon choice). The optimisation process involves three stages: option generation, choice, and re-assessment. At each junction, options are generated by recursively seeking out, or "wiggling" viable splice junctions (or start codons) in each frame either side of the existing junction locations for both the left-hand and right-hand exon. For each junction, this generates a set of splice donor-acceptor pairs (or start codons), one of which must be selected for each gene as the jointly optimal choice.

The optimal set of junction choices for the set of genes is defined to be the set which maximises the alignment score α of the resulting set of gene models. A naïve way to determine this optimal choice would be to compute the alignment score for all possible options, and simply pick the best. However, even for small sets of sequences this becomes combinatorially explosive and prohibits computation. A more reasonable approach would be to compute only *pairwise* alignment scores, and use a method such as the Hungarian algorithm (Kuhn 1955) or similar for selecting the optimal subset. Still, for many gene combinations, the number of computations required becomes prohibitive.

Instead, one can exploit the nature of the sequences in question, namely the fact that there are at most 20 possible values present in any one column of an alignment (and typically far fewer given that these sequences are well aligned and that many of them are constructed from the same source), and that alignment scores can be calculated equally sequence-wise or column-wise. The *multipartite choice function* (Figure 5.8) aims, for a set of k gene regions and a set of $L = \sum_{i=1}^{i=k} l_i$ gene model variants for each gene region, to choose an optimal set containing one gene model variant from each gene region such that the alignment score is

maximised. This problem is equivalent to finding the heaviest maximal clique in an edge-weighted complete multipartite graph (where the sequences are the vertices and edge lengths are pairwise alignment scores). This function replaces the naïve approach of calculating pairwise alignment scores between all options and choosing the most coherent subset, which, while optimal, is computationally infeasible for large sets of variants.

To reduce further the complexity of the problem, options are chosen by comparison with a reference consensus alignment, produced by taking the most consistent set of amino acids for each column in a global alignment individually (Figure 5.8A-B). This column-wise optimisation is fast, and provides a basis for the sequence-wide optimisation. To produce the consensus, the set of L options is aligned to the reference (the original alignment) using MAFFT with the `--add` option (Katoh and Standley 2013). The inconsistent regions are then isolated and re-aligned using the more accurate but more computationally intensive MAFFT L-INS-i. For each column in the alignment, the set of amino acid choices (one for each gene region) that optimises the alignment score for that column is chosen as the consensus.

For each option i a binary string $H_i = \{h_1^i, \ldots, h_n^i\}$ (where n is the length of the alignment) is produced describing for each position in the alignment whether or not that option matches the consensus (Figure 5.8C). The chosen subset will be the set of options that globally maximises agreement with the consensus. If the strings $\{H_i\}_i$ are stacked vertically, such that they can be read as columns $\{V_j\}_{j=1}^n$ then the task is equivalent to finding a columnar binary string V with one nonzero entry for each gene region such that $|V_i : V \subseteq V_i|$ is maximised. For illustration, the string that achieves this is presented on the right-hand side of Figure 5.8C.

Here we introduce, abstractly, a definition: for sets S_1, S_2, and a set of "checkpoint" sets R, the set S_1 is *compatible* with S_2 with respect to $R = \{R_i\}_i$ if the binary intersection $S_1 \cap S_2 \cap R_i$ is nonzero for all i. Define G to be the set of binary strings which have values 1 for all sequences in one gene region, and

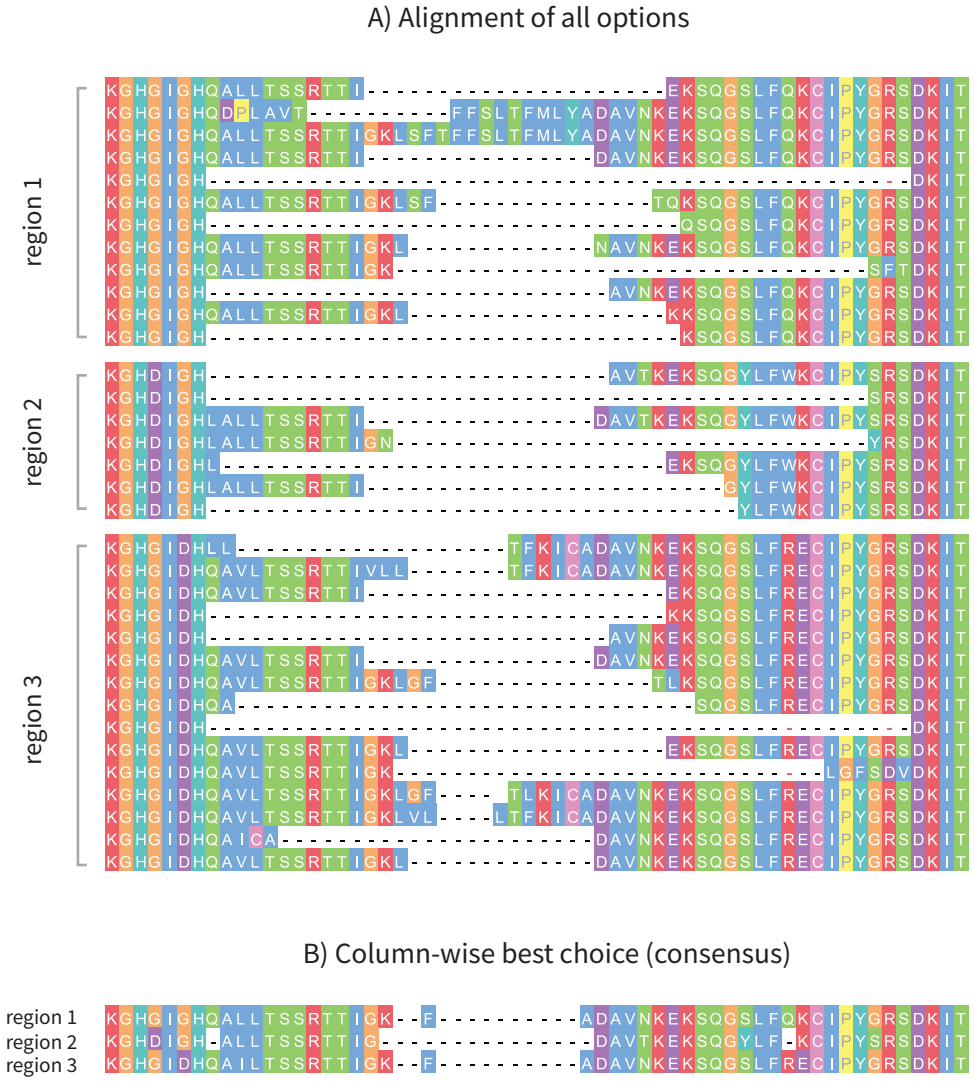


Figure 5.8: Illustration of multipartite choice function. The multipartite choice function aims to find optimal variants from a set of protein sequences. A) Sequence variants are aligned; B) A consensus alignment is produced: on a column-by-column basis the choice of amino acid for each sequence that optimises the alignment score for that column is chosen as a representative; *(continued on page 205)*

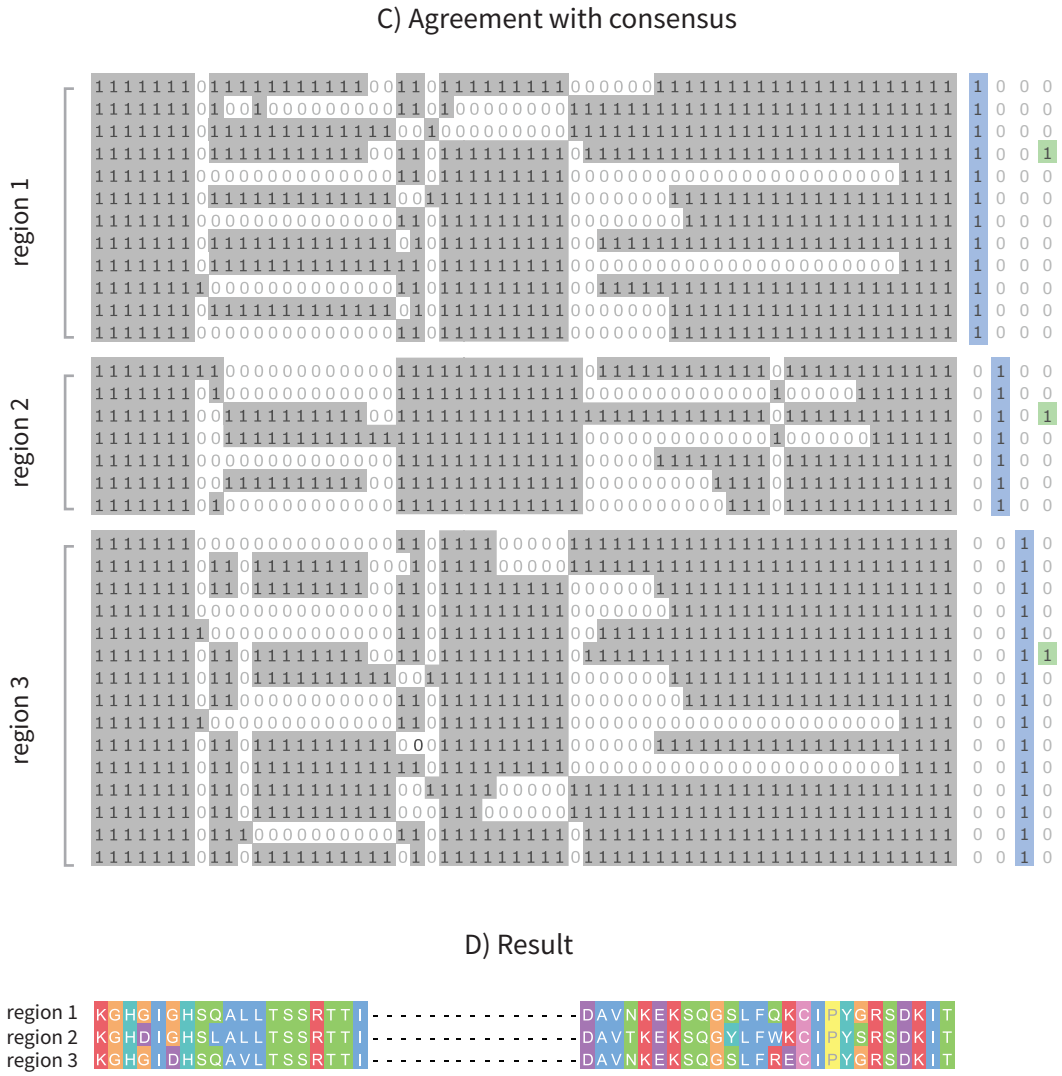


Figure 5.8: *(continued from page 204)* C) A binary representation is produced from the original alignment: for each base in alignment, a 1 is assigned if the base matches the consensus, and a 0 is assigned if it does not. This leaves a sequence of vertical binary strings. The aim is to find a single vertical binary string that agrees with (i.e. is a bitwise subset of) as many as possible of these, and that is also compatible with the category constraints. The best such string in this case is shown to the right in green. D) The result.

0 for all others (illustrated in blue in Figure 5.8C). G will be the checkpoint set for the purposes of comparing binary strings.

Given the set $A_0 = \{V_j\}_{j=1}^{n}$, an optimal subset is deduced by sequential random sampling. Ignoring all-1 strings, an initial $W_0 = V_k$ is chosen at random from A_0. Define $A_n = \{a \cap W_{n-1} : a \in A_{n-1}$ and a, W_{n-1} compatible w.r.t $G\}$, at each stage choosing W_n at random from A_n. The process $A_0, A_1, A_2, \ldots$ eventually converges on a single binary string. This reduction is performed a user-selected number of times, the default being 1,000. The result that is a subset of the largest number of V_is is declared the winner. In the event that the result still contains more than one option for each gene region, subsets of options are calculated and their multiple alignment score α is calculated, the winner being the subset with the highest α. In the event that multiple subsets exhibit the same maximal α, a subset is chosen arbitrarily from them.

Once a choice has been made, the results are assessed again to check that a better alignment cannot be produced by removing an exon. For example, the process above can only change the sizes of exons, it cannot remove them, and if removal is necessary then the above process will have reduced the size of a putative exon as far as possible. In the event that a particular exon is very small (< 40bp), or does not yield any valid junction sites, both that exon and the one before it are probed for removal, and the variant with the removed exon is compared against the other partial gene models in the evaluation step. In addition, any exons that were removed in the previous step are re-added, to see if their inclusion was only prohibited by the inclusion of an invalid subsequent exon. This, and the choice and optimisation processes, are iterated for each exon junction. Once this recursive step ceases to produce new gene models, the gene model set with the highest alignment score is declared the winner, and the next putative exons from the next adjacency group are added. If multiple alternatives have the same alignment score, the choice is made randomly. This is repeated until there are no further exons to add.

5.3.4 Resulting genes are double-checked against the input

To ensure that the optimisation process did not overlook potentially better variants in the user-supplied gene models, the process of sequentially adding and optimising putative gene parts is repeated. This time, instead of varying exon start and end sites, the only allowable options for each exon start and end are the original and new junctions, i.e. no "wiggling". The aim of this is to find the optimal combination of old and new junctions.

5.3.5 A final filtering step accounts for over-correction

The final step involves filtering the changes based on a selection of categories that have been observed to over-fix gene models: that is, changes that are made which are consistent with the objectives of OMGene, but which typically do not result in an improved gene model relative to extrinsic data. These filters are used to rule out certain changes to the original gene models, and revert those changes back to the original observed model.

The most basic requirement is that a given gene model change must not have caused a local reduction in amino acid alignment. The alignment score α of a 10 amino acid region either side of each change must have either remained the same or improved. The objective of the multipartite choice function is to locally optimise junctions in terms of their α score, so this requirement should be met in most cases, however it is still possible for exceptions to make it through. The choice of 10 amino acids here is arbitrary, but was considered a sensible value based on observations.

Similarly, the alignment score of a region surrounding a change must be of a reasonable quality to begin with. Regions with low alignment scores α are difficult to optimise, as they do not provide strong evidence for evolutionarily conserved regions. For 5 unchanged amino acid regions either side of the change, the adjusted alignment score $\bar{\alpha}$ must be 3 or higher (or all gaps) for some subset of three sequences containing the sequence of interest. Similarly the resulting score for

the changed region must also be higher than 3 or all gaps. Again, these values are based on observation; the score of 3 is however less than the expected perfect alignment score described in Chapter 3 Section 3.4.2.

Changes that have opened gaps in the alignment of three or more of the sequences are not allowed: this is a common occurrence due to sequences proximal to exon termini that by chance feature valid splice junction sequences that are in frame with the adjacent exons and are evolutionarily conserved. These tend not to be correct.

Finally, very small changes are forbidden: changes that have resulted in two or fewer amino acids being changed in a gapless region of the alignment, such that the new alignment is also gapless, are ignored. Similar changes to larger regions require an α increase of 4 or more. This is to avoid changes that reflect multiple choices of donor-acceptor pairs for essentially identical sequences.

Exon boundaries that do not pass the filters are discarded and the genes are reconstructed a final time using the adjacency group and multipartite choice function method, allowing only the surviving boundaries and those that were present in the original gene. These filtering criteria are very strict, and will cause many valid gene model changes to be overlooked. However, it was deemed more important here that spurious changes to existing models were minimised, and high-confidence changes are brought forward. The resultant genes are outputted in GTF, amino acid FASTA and CDS FASTA format.

5.3.6 An additional species set for algorithm verification

As an additional test for OMGene, an extra species set is introduced here: Species Set M, a set of mammalian genomes. To inform an appreciation of the difference between this set and those introduced previously, a brief analysis of the characteristics of these genomes is presented here. The full list of species contained in Species Set M can be found in Table 5.5. Briefly, the species are *Canis lupus familiaris* (domestic dog), ***Homo sapiens*** **(human)**, *Monodelphis domestica* (grey

	Tree	Species name	Source	Version	Tax. ID	References
Mammal species		*Mus musculus*	CCDS	GRCm38.p4 / CCDS21	10090	[1]
		Oryctolagus cuniculus	NCBI	OryCun2.0 / AR102	568996	[2,3]
		Homo sapiens	CCDS3	GRCh38.p7 / CCDS20	9606	[1]
		Canis lupus familiaris	NCBI	CanFam3.1 / AR105	6915	[2,4]
		Monodelphis domestica	NCBI	MonDom5 / AR103	13616	[2,5]

Table 5.5: List of mammal species used, a phylogenetic species tree for each group, sources, and references. All data sets were downloaded from NCBI Genbank (Benson et al. 2013), except for *H. sapiens* and *M. musculus*, whose data were downloaded from CCDS (Pruitt et al. 2009). The two entries for version are assembly / annotation. References: [1]Pruitt et al. 2009, [2]Benson et al. 2013, [3]Alföldi et al. 2009, [4]Lindblad-Toh et al. 2005, [5]Mikkelsen et al. 2007.

short-tailed possum), *Mus musculus* (house mouse), and *O. cuniculus* (European rabbit). As was the case for the plant and fungal data sets, a well-studied model organism from Species Set M was chosen as a reference point. In this case, either of *M. musculus* or *H. sapiens* would have sufficed for these purposes: *H. sapiens* was the species used here.

As before, it is useful to appreciate some of the genome characteristics of the species under consideration. Table 5.6 shows genome and assembly structure for the mammal genomes in Species Set M. These genomes are on average an order of 5-10 times larger than the plant genomes, and around 100 times larger than any of the fungal genomes. In addition, with the exceptions of *H. sapiens* and *M. musculus*, all the genome assemblies are highly fragmented and unfinished. A visual representation of the different assembly sizes alongside estimated genome sizes from the other species sets can be seen in Figure 5.9.

A breakdown of the proportions of each of the mammal genomes that are genic, coding, and intronic can be found in Table 5.7, and Figure 5.10 shows a graphical comparison between the coding percentages for each genome. The average proportion of coding bases in the mammal genomes here is 1.3%, which

	Species	Chromosomes	Ploidy	Regions	Assembly size
Mammals	*C. lupus*	$2 \times 38 + X/Y + M$	Diploid	3268	2.4Gbp
	H. sapiens	$2 \times 22 + X/Y + M$	Diploid	25	3.1Gbp
	M. domestica	$2 \times 8 + X/Y + M$	Diploid	5016	3.6Gbp
	M. musculus	$2 \times 19 + X/Y + M$	Diploid	22	2.7Gbp
	O. cuniculus	$2 \times 22 + X/Y + M$	Diploid	3241	2.7Gbp

Table 5.6: Genome sizes for various mammals. All are of the order of 3Gbp. X/Y are the sex chromosomes, M is the mitochondrial genome. These genomes are large, and mostly highly fragmented.

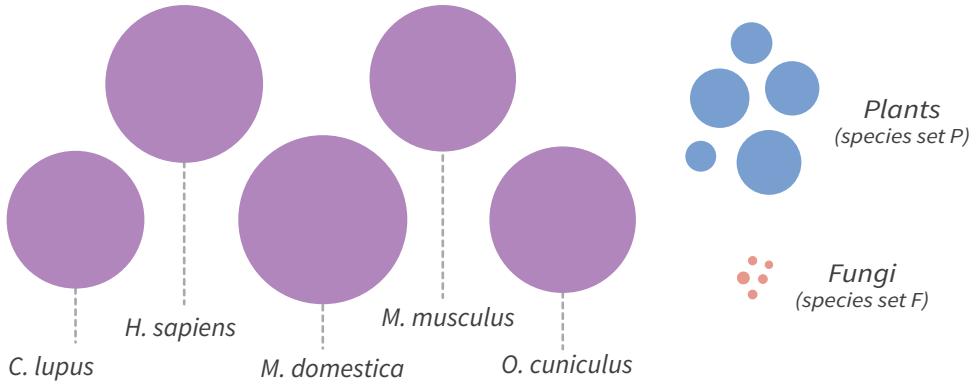


Figure 5.9: Mammals have large genomes. The mammal genomes are an order of magnitude larger than the plant genomes, which are in turn an order of magnitude larger than the fungal genomes.

is much lower than the corresponding statistics for the plant (11.3% coding) and fungal (61.2% coding) species sets.

In addition to having larger introns, mammal genes typically have smaller and more numerous exons, as shown in Table 5.8. On average, 91.6% of genes are multi-exon. In addition, an average of 39.2% of genes have alternate transcripts marked. Thus, mammal genomes provide a distinct additional challenge to the success of OMGene compared with application to plant and fungi genomes.

	Species	No. genes	% of genome which is		
			Genic	Coding	Intronic
Mammals	*C. lupus*	19321	41.9%	1.4%	39.3%
	H. sapiens	18859	34.1%	2.1%	33.1%
	M. domestica	20518	38.0%	1.0%	37.1%
	M. musculus	20322	29.1%	1.2%	28.0%
	O. cuniculus	16032	29.2%	0.9%	27.6%

Table 5.7: Coding percentages for mammal species. Large proportions of mammal genomes are intronic, and very small percentages of mammal genomes are coding. Note also that the mammal species here tend to have fewer genes annotated than the plants.

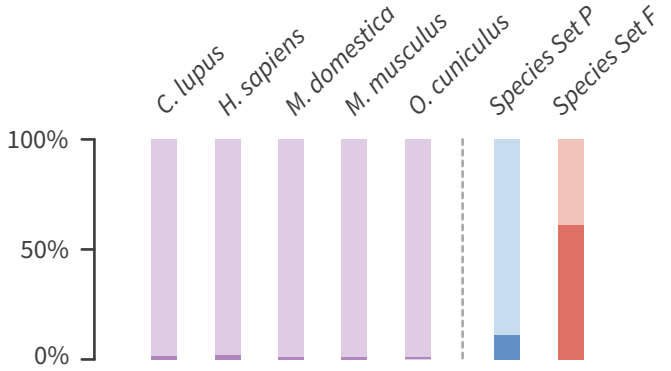


Figure 5.10: Most mammal DNA is non-coding. Coding regions make up a tiny proportion of mammal genomes, even less than plants, and vastly less than fungi.

	Species	Mean				% multi-exon genes	% of genes with alternate transcripts
		Exon size	Intron size	No. exons	Transcript size		
Mammals	*C. lupus*	162	5504	12.3	2005	93.5%	52.7%
	H. sapiens	162	7756	10.6	1710	93.5%	40.1%
	M. domestica	172	7344	11.3	1943	91.8%	45.6%
	M. musculus	169	4845	9.9	1672	88.0%	18.3%
	O. cuniculus	167	5822	11.0	1838	91.2%	39.1%

Table 5.8: Gene structure statistics for mammal genes. The typical mammal gene has many very small exons, and single-exon genes are rare. Alternative splicing is common in mammals, and this is reflected in the high level of marked alternate transcripts.

5.3.7 Acquisition and choice of orthogroups

OMGene was applied to genes from Species Sets P, M, and F in two ways. Firstly, OMGene was applied to single-copy ubiquitous orthogroups formed from genes present in the public annotations of the species. For the plant data set, where multiple transcript variants were available, the primary transcript was used as listed in Phytozome (Goodstein et al. 2012). For the mammal data set, no primary transcripts were marked, and so for each gene the longest available transcript was used. For the fungal data, only two genes (two genes in *D. hansenii* had alternate trancripts, and the transcript with the lower-numbered gene ID was used. Orthogroups were inferred using OrthoFinder (Emms and Kelly 2015) with the default parameters.

5.3.8 Simulated *de novo* data sets

As an additional assessment, genes were predicted *de novo* for each of the representative model species from Sets P, M, and F, and OMGene was applied to the set of orthogroups calculated using the predicted genes from these alongside the original annotations from the other species in the sets. *De novo* gene predictions were made using Augustus (Stanke and Morgenstern 2005) version 3.2.2. Augustus (Stanke and Morgenstern 2005) was trained separately for *A. thaliana*, *H. sapiens*, and *S. cerevisiae* on all protein-coding gene locations extracted from their respective GTF files, using the included `autoAugTrain.pl` package. Augustus was then run on each genome with the parameters obtained from the training process, with no additional non-default parameters. Orthogroups were again inferred using OrthoFinder (Emms and Kelly, 2015).

5.3.9 RNA-seq data sources

For the plants and fungi, the RNA-seq data used were as listed in Chapter 2, Methods Section 2.7.7, Table 2.13. For the mammal species, to improve the

Species	SRA ID	Instrument/details
C. lupus	ERR266386	Illumina Genome Analyzer II, paired, brain frontal cortex, male
	ERR266355	Illumina Genome Analyzer II, paired, frontal cortex, female
	ERR266382	Illumina Genome Analyzer II, paired, frontal cortex, male
H. sapiens	SRR5938455	Illumina HiSeq 2000, paired, dorsolateral prefrontal cortex, male
M. domestica	SRR500906	Illumina HiSeq 2000, paired, brain
	SRR500925	Illumina HiSeq 2000, paired, brain
M. musculus	SRR5441717	Illumina HiSeq 2000, paired, brain (striatum)
	SRR6269591	Illumina NovaSeq 6000, paired, cerebellum
O. cuniculus	ERR266399	Illumina Genome Analyzer II, paired, frontal cortex, female
	SRR400990	Illumina Genome Analyzer II, paired, frontal cortex
	SRR401040	Illumina Genome Analyzer II, paired, frontal cortex
	SRR401041	Illumina Genome Analyzer II, paired, frontal cortex
	SRR401042	Illumina Genome Analyzer II, paired, frontal cortex

Table 5.9: Sources of mammal RNA-seq data used to assess hits. All data were downloaded from the Sequence Read Archive (Leinonen et al. 2010). For consistency, brain samples were chosen for all species.

resolution of the mapped reads, multiple sequencing data samples were used. These are listed in Table 5.9. Sequencing sources were again chosen with a preference for the same source tissue (leaf for plants, brain for mammals) to maximise the comparability of the results. All RNA-seq data sets for mammals were downloaded from the Sequence Read Archive (Leinonen et al. 2010), as described in Chapter 2, Methods Section 2.7.7.

5.3.10 Alignment and quantification of RNA-seq data

RNA-seq reads were mapped as described in Chapter 2, Methods Section 2.7.8. For the mammals, the mapped reads from the separate samples were combined using `samtools merge`, and all further analyses were performed as described in Chapter 2, Methods Section 2.7.8.

5.3.11 Coverage and splice junction scoring metrics

Genes were assessed in relation to their resemblance to RNA-seq data, both in terms of coverage and splice junctions. The coverage score used was the same as the one introduced in Chapter 3. The splice junction assessment used was a variant on the splice junction recall score implemented in Chapter 3. The score described in Chapter 3 was a *recall* score, and was only one-directional, to prevent alternative transcripts and other present junctions from affecting the score. Here, since direct comparisons are being made between two possible transcript choices, it makes sense to use a two-way score, the *junction F-score*: any unrelated transcripts present in a given gene region will detriment both the original and fixed scores equally.

The junction F-score for a gene is a measure of how well the splice junctions in an annotated gene model match the mapping junctions present in aligned RNA-seq reads, and *vice versa*. In contrast to the splice junction recall score, the junction F-score can detect missing introns and exons. For a gene model G and a corresponding genomic region R, define J_G to be the set of individual intron beginning and end coordinates in the gene model, and define J_R to be the set of mapping junction beginning and end coordinates in the mapped RNA-seq data. A minimum of 10 reads is required for a given RNA-seq junction to be counted. We may then define the junction F-score as:

$$jF\left(J_G, J_R\right) = \frac{2 \cdot jP\left(J_G, J_R\right) \cdot jR\left(J_G, J_R\right)}{jP\left(J_G, J_R\right) + jR\left(J_G, J_R\right)}$$

where

$$jP\left(J_G, J_R\right) = \frac{|J_G \cap J_R|}{|J_G|}; \qquad jR\left(J_G, J_R\right) = \frac{|J_G \cap J_R|}{|J_R|}.$$

The direction of each junction site (start or end of a junction) is taken into account when considering the intersection of the two sets.

5.4 Chapter discussion

In this chapter I presented OMGene, an automated method for improving the consistency of gene model annotations across species. OMGene is intended for use in computational *de novo* genome annotation projects where no empirical data (such as RNA-seq data) are available to train or correct gene model predictions, or to assist the construction of gene models for genes that are not expressed in the data available. OMGene is also designed to help users who wish to leverage conservation information to correct gene models of a single gene of interest across a set of species. Thus OMGene is suitable for both large and small scale analyses. OMGene is run as a python script with a tab-delimited input file, each line of which contains the path to a GTF file, containing coordinates for a single gene, and the path to its corresponding genome FASTA file. Full instructions for running OMGene can be found at the OMGene GitHub repository (`https://github.com/mpdunne/omgene`).

5.4.1 OMGene results reflect differences in gene model complexity

To demonstrate the utility and performance characteristics of OMGene, it was applied to three separate datasets of well-annotated plant, mammal, and fungal genomes. When applied to the plant data set, OMGene altered the gene models of one or more genes in 30.1% of the orthogroups that were evaluated; when applied to the mammal data set, 22.9% of SCU orthogroups saw some change. In contrast, only 3.1% of orthogroups were subject to modification in the fungal data set. This result reflects the differences in gene model complexity between the three species groups. Specifically, gene models in mammals and plants tend to have more exons than fungi (mean = 5.09 exons for plants, 11.23 for mammals, 1.08 exons for fungi) and thus there is considerably more potential for gene model variation in plants and mammals than in fungi. In light of this it was unsurprising that the most frequently observed change made in fungi was a change in choice of start codon. This is also

reflected in the high number of removed exons from plant and mammal genes, which is contributed to partly by the separation of erroneously fused adjacent genes.

5.4.2 OMGene works well on complex gene models

The changes made by OMGene were assessed relative to splice-mapped RNA-seq data to assess the level to which the gene models had been improved. For the plant data set, the results from OMGene clearly resembled the empirical data much more closely on the whole, with 85.4% and 86.7% of genes with changed scores improving in terms of their splice junctions and their coverage respectively. The profiles were different for different species, with many more changes being made for *C. papaya* and *T.cacao*; in addition the number of successes for *B. rapa* was slightly lower than for the other species.

In mammals, 79.2% and 82.4% of changed genes with changed scores has improved junction and coverage scores respectively. The three non-model species, *C. lupus*, *M. domestica*, and *O. cuniculus* faired well in terms of their junction scores. *H. sapiens* genes also typically scored highly, but *M. musculus* genes did not do so well. The reason why mouse genes changes were less successful than human gene changes is unknown. These two species also saw considerably fewer gene model changes than the other species. In all cases, mammal genes were mostly contracted via the removal of exons or by moving the start codon – this is at least partly due to biases in the choice of longest gene model as representatives for these genes.

The number of junction changes made for the fungal data set was considerably lower: only 7 changed genes had an altered junction F-score, 71.4% of which become worse after OMGene. Though this is far less than the plant and mammal data sets, it should be noted that the resolution of this data set does not lend itself to accurate conclusions about the general validity of changes made to fungal genes. The resolution and success rate for fungal genes from a coverage perspective was slightly higher, with 80.8% of the genes with changed scores improving. The low

resolution of junction data for fungal genes reflects the rarity of complex gene models in these species, and thus the low likelihood that deviations from simple, single-exon gene models are correct. Thus, while OMGene does not always produce gene models that agree optimally with RNA-seq data, it does improve the overall quality of gene model annotations even for relatively simple fungal genomes.

The improvements in gene model accuracy made by OMGene for the *de novo* predicted proteomes were much the same as for the publicly available, curated gene models. However, the number of changes made to the *de novo* predicted set was much greater, indicating that the considerable labour that has been applied to these model organisms has successfully controlled for potential errors. It should be noted that, although OMGene managed to improve many of the gene models outputted by Augustus, the two agreed in most cases (95.2%, 60.3%, and 99.4% for plants, mammals, and fungi respectively), indicating that the basic implementation of a well-trained Augustus *de novo* prediction produces genes that are highly consistent with their orthogroups.

5.4.3 OMGene improves start site identification

In addition to assessment of splice junctions, gene models were assessed by the consistency of their predicted subcellular localisation. Given that the orthogroups used in this analysis comprise ubiquitously conserved single copy genes, it is logical to assume that these genes should generally have the same subcellular localisation. For the full plant data set, of all orthogroups whose genes had different subcellular targeting predictions after application of OMGene, 79.6% had improved intra-orthogroup consistency, with 85.7% either improving or remaining the same. In mammals, 87.2% of orthogroups whose subcellular localisation predictions were changed by OMGene showed improved consistency. For the full fungal data set, although the data were sparse, all of the orthogroups considered had improved consistency.

The results for the plant data set were similar for the *de novo* annotated set (80.0% improvement), as was also the case for mammals (86.9% improvement). For fungal orthogroups containing *de novo* predicted *S. cerevisiae* genes, the only gene whose localisation prediction changed caused the consistency of its orthogroup to decrease, however the resolution of the data in this case is not sufficient to draw any conclusions. Thus, application of OMGene improves the accuracy of start codon specification in *de novo* predicted gene models.

5.4.4 Summary

When applied to publicly available plant, mammal, and fungal data sets, OMGene demonstrates proficiency in improving gene models from multiple perspectives. Due to stringent filtering criteria, it does not fix all errant gene models, however the gene models that it does fix represent an improvement the majority of the time. The overall improvement is larger for genomes with complex gene models. Thus, OMGene is potentially of considerable use to the annotation community, and should be both applied to novel genome projects as well as being used to update existing annotations.

6 General discussion

6.1 Overview

The overall objective of this thesis was to develop approaches for assessing the completeness and accuracy of genome annotations at a phylogenetically local scale in the absence of extrinsic data, and subsequently to develop computational tools to improve those annotations under the same constraints. This objective was achieved by seeking to optimise the consistency of orthogroups of genes from groups of related species.

In the first part of the thesis, comprising Chapters 2 and 3, I defined several extrinsic and orthogroup-based experiment-free approaches to assessing genome annotation completeness and accuracy. The main results were a filtered orthogroup HMM approach for estimating the quantity of missing genes, and a method to assess global trends in gene model accuracy by searching for discrepancies in otherwise coherent orthogroup multiple sequence alignments. Given that there were no objective, perfectly annotated reference annotations against which to compare the results of these methods, in both chapters it was necessary to define and utilise extrinsic approaches for method assessment. Each of the approaches that I defined in Chapters 2 and 3, extrinsic and non-extrinsic alike, has advantages and

drawbacks depending on the use case, and I will discuss the relative benefits of each in Section 6.2. All the methods were broadly consistent, however, indicating that each is valid in revealing underlying annotation issues.

In the second part of the thesis, comprising Chapters 4 and 5, I introduced two software packages leveraging the ideas of Chapters 2 and 3, namely OrthoFiller and OMGene. Through simulated depleted data sets, comparison to *de novo* predictions, application to publicly available data, and assessment relative to RNA-seq data using the methods defined in Chapters 2 and 3, I showed that each of these software packages offers significant improvements on both *de novo* predicted genes and on publicly available annotations. Furthermore, application of the methods to these data sets yielded improved versions of each set, thus directly supporting the effort to improve public genome databases.

In this general discussion, I will briefly recapitulate some of the work described in the four chapters of this thesis. In doing so, I will consider the relative advantages, drawbacks, and potential improvements for the approaches described, and suggest ways in which they can be combined and leveraged to tackle additional problems otherwise not considered here. Finally, I will place the algorithms in the broader context of the current genome annotation landscape, considering how they can be integrated into small and large genome projects and genome databases at various scales of automation.

6.2 Some annotations are better than others

As noted in Chapter 1, generalised approaches for assessing the extent of genome annotation errors are few and far between (Eilbeck et al. 2005; Yandell and Ence 2012; Mudge and Harrow 2016). In particular, existing efforts either rely on appeal to universally conserved genes (Veeckman et al. 2016), common protein domains (Holt and Yandell 2011), and RNA-seq data. Approaches that depend on universal features are circular given that these are the genes that were likely used to inform

the initial predictions for a genome, and extrinsic methods suffer from limited availability of comprehensive data sets. This was the motivation for generating phylogenetically local, experiment-free approaches for assessing the completeness and accuracy of genome annotations, which were the basis for Chapters 2 and 3. These approaches demonstrated clear signals for genome annotation deficiencies in some species, notably *C. papaya* and *T. cacao*, and affirmed the relatively high quality of annotations for the two model species, *A. thaliana* and *S. cerevisiae*. This section will discuss the results of these methods, and discuss some of their relative benefits and drawbacks.

6.2.1 A novel orthogroup HMM search and filtration system identifies under-annotated genomes

In Chapter 2 I introduced the concept of an *orthogroup* and used these to construct an evidence-based genome annotation assessment metric relative to small groups of species. In the discourse, I discussed a potential naïve approach involving simply counting absent orthogroup members, and I demonstrated why this was in general ineffective. However, I noted that in some cases strong signals of genome under-annotation could shine through this analysis, as was the case for *C. papaya*, which was the only species absent from otherwise well conserved orthogroups in more cases than expected. Thus, this gave an initial clue to annotation deficiencies in *C. papaya*.

The main result of Chapter 2 was a completeness metric, defined relative to small sets of closely related species. The method used to obtain this metric involved directly seeking out absent orthogroup members by searching genomes for matching genes, using a filtered HMM approach based on orthogroup amino acid alignments. In order to refine HMM hits to minimise the effects of spurious results, I implemented two filtering methods on this analysis: first, I applied an evidence-based hit score filter using threshold values learned from hits to existing annotations; second, I applied a hit-cleaning approach that aimed to reduce the

effects of hits to and from repetitive sequence regions. By considering high-quality recovered hits to genomes from orthogroups not already containing a gene in that genome, I defined three completeness metrics for various classes of orthogroup. These involved expressing the number of potentially missed gene families as a proportion of the number of gene families that would be present in the species if each of the hits were included in the genome annotation.

All species received scores above 95% by these metrics, indicating either that the genomes already have high levels of completeness, or that the approach underestimates the level to which these genomes are missing annotations due to strict filtering criteria. It is likely that a combination of these is true. The worst-performing species, in line with prior observations, was *C. papaya*, which recieved a score of 96.1% completeness for single-copy orthogroups. If this is accurate and can be scaled up as a completeness metric for the whole *C. papaya* proteome, this would equate to 1,076 unannotated genes: a significant number. Both the model species received high scores. *S. cerevisiae* received perfect scores, as no genes were found from orthogroups that did not already contain a gene from *S. cerevisiae*. *A. thaliana* scored 99.3% in the worst case, but still showed evidence for several missing genes. *B. rapa* and *C. rubella* also showed signals for under-annotation, albeit less than *C. papaya*, at 98.2% and 97.7% complete respectively for single-copy orthogroups. *T. cacao* score highly in all analyses. I suggest that this particular genome is *over-annotated*, since observation reveals that the annotated genes frequently overlap and many genes are contained within introns from other transcripts. The fungal species generally scored highly, which is unsurprising given that fungal genomes are small and have easier-to-predict gene models. The worst fungal performers here were *K. lactis* and *Y. lipolytica*, which had completeness levels of 97.7% and 98.6% for the stricter category of SCU-1 orthogroups.

6.2.2 Putative missing genes are supported by RNA-seq evidence

The nature of the evidence-based HMM approach facilitates a direct comparison of potential missing genes to extrinsic data. While such data can say nothing about the recall statistics of the completeness estimate, they can indicate its precision, i.e. whether the genes indicated by strong HMM hits should be considered likely candidates for missing genes. Here we must proceed with some caution: the hits accepted through this approach could correspond to pseudogenes or other retrotransposed genic elements. *A priori*, the problem of distinguishing pseudogenes from true protein-coding genes is very difficult. Pseudogenes are sometimes transcribed and can act as regulators for protein-coding transcripts (Tarazona et al. 2011). Pseudogenes are however transcribed at a far lower rate than their protein-coding cousins, at least for the organisms so far assessed – about 2-3% for human and 0.5-1% for mouse according to one study (Yano et al. 2004). Even if a transcript represents a pseudogene, if it is transcribed it is much more likely to be of biological interest than an untranscribed "genomic fossil" pseudogene, and so identifying such features is still worthwhile. Thus, if the candidate genes identified using the filtered HMM method in Chapter 2 demonstrate transcriptional activity, then this is evidence for their validity as either genuine protein-coding genes or, at worst, biologically relevant pseudogenes.

In all cases where data were available, the candidate missing genes obtained in Chapter 2 showed much higher RNA-seq coverage than background intergenic regions, frequently on a par with the existing annotated genes. For the fungi, the average read coverage for accepted candidate hits was 102.6, versus 125.9 for annotated genic regions; for plants, the average read coverage for accepted candidate hits was 57.7, compared with 54.3 for annotated genic. In some cases, the mean coverage was lower for the candidate hits, notably for *A. thaliana* and *B. rapa*. Highly expressed genes are more likely to have been annotated

already, and so identification of lowly or rarely expressed genes could explain these figures. The RNA-seq analyses here provide strong evidence that these genes are genuine, and the uniformity of the results suggests it is likely that most of them correspond to true protein-coding genes rather than pseudogenes. This suggests that the approach does indeed provide a conservative measure of the prevalence of missing genes in these genomes.

To summarize this section, assessing genome annotation completeness is difficult. Genomes and transcriptomes are messy. Gene duplication events and gene loss events can obscure the true evolutionary trajectories of ancestral genes, and retrotransposition of genic RNA can spread red herring homology signatures. RNA-seq data are often unrepresentative and not available, and universally conserved genes do not provide an accurate picture of annotation completeness due to circularity issues. The approach detailed in Chapter 2 and discussed here aims to circumvent these issues by 1) not relying on extrinsic data, and 2) restricting analyses to small sets of evolutionarily close species. The approach is not perfect, and should be used to complement other analyses, not to replace them. In particular, it is likely that the HMM-based approach both underestimates missing genes and is polluted slightly by the identification of pseudogenes and other non-functional elements. On the whole, however, these measures are well supported, and provide a good initial measure of genome annotation completeness deficiencies.

6.2.3 Extrinsic gene model scores identify poor annotations

In Chapter 3, I focussed on providing a method for assessing the accuracy of gene models, again at a phylogenetically local scale and in the absence of extrinsic data, by searching for clear discrepant regions in otherwise well-aligned orthogroups. In order to assess this approach, I needed to formalise several approaches for assessing the quality of gene models in reference to RNA-seq data, as well as in terms of

their subcellular localisation predictions. I used these assessment metrics to assess the capacity of discrepant orthogroups to identify inaccurate gene models.

There are no standard approaches for assessing the correctness of gene models relative to the mapping profiles of their RNA-seq-obtained transcript models. In order to provide a reference against which to compare my orthogroup-based approach, I needed to formalise these. The two scores I used were the *coverage score* and the *splice junction recall score*. Each score is dependent on RNA-seq reads of a reasonable depth, and so analyses were restricted to transcripts with an average read depth of 10 or more reads across the length of their annotated model. Each score has benefits and drawbacks. Many valid gene models will not receive a perfect coverage score, due to inconsistent RNA-seq read profiles, presence of ncRNA transcription in the introns of genes, and pollution of results via alternative transcripts. Transcripts which receive very low coverage scores, however, are likely to be due to incorrect gene models. The splice junction recall score is designed to compare splice junctions present in a gene model to what is observed in the RNA-seq data. As a one-way score, this forgives alternate transcripts and other pollutants in the same locus as a gene, but cannot detect missing features such as absent introns or exons. Thus, although neither score is perfect, each should provide an indication of the prevalence of any significant errors in annotated transcripts.

The fungal species were not amenable to RNA-seq analysis. This was because fungi lack complex gene models, and thus splicing errors are likely to be exceedingly rare. Manual observation also revealed that RNA-seq mapping profiles to the fungal genomes tended to be somewhat jagged, which may have reduced the reliability of the coverage score in these cases. Even though the data were sparse, however, *Y. lipolytica* showed the highest incidence of inaccurate junction scores: 1.5% of highly expressed genes had splice junction errors. This was in line with previous observations regarding the lesser annotation completeness of this species. In terms of coverage score, all species again had a low incidence of poor scores,

though there was no clear result for these species over different stratifications of coverage score thresholds. That said, *S. cerevisiae* had the highest number of genes with perfect coverage scores, at 85.0%. Thus, although extrinsic analyses are not particularly useful for fungal genes, they support the notion that the model species has the best gene models.

For the plant species, the data were clearer. *C. papaya* and *T. cacao* both performed poorly with regards to both the coverage and splice junction recall scores. Again, this is in line with prior observations about the completeness of these species' annotations, with 36.4% and 21.7% respectively of highly expressed genes having imperfect *jR* scores. The same pattern was observed for the coverage scores, with 34.0% and 20.1% respectively of *C. papaya* and *T. cacao* transcripts receiving scores less than 0.95. It is worth noting that, while the completeness assessments could have been affected by the phylogenetic structures of the underlying species trees, these extrinsic gene model assessments were independent of the relationships between the species, and so the measures are independent. The results for the scores were highly correlated ($R^2 = 0.983$, $P = 0.990$) for the plant data. Given that each score reveals a different aspect of the underlying gene models, the correlation here supports the fact that both are revealing underlying gene model inadequacies. Thus, these extrinsic scores are likely to provide a good measure of gene model accuracy.

6.2.4 Alignment discrepancies betray erroneous gene models

As a non-extrinsic measure, I defined – somewhat strictly – the notion of a discrepant orthogroup (DOG). These are orthogroups whose alignments contain regions of gaps or non-gaps for a stretch of five amino acids or more, which in turn are flanked by ungapped regions of high-scoring alignment. These are very strict criteria, and the number of such orthogroups present in both the plant set and the fungi set was small. These are the kinds of sequences that one would identify by eye as being suspicious, and my goal here was to demonstrate that these were often lacking

in terms of their extrinsic scores as well, thus indicating that updating the gene models to generate more consistent orthogroup alignments could reliably yield improvements relative to the true transcriptome.

Extrinsic scores were compared to DOGs. In all cases for the plants, DOGs were shown to have much lower splice junction recall scores than the background. It should be noted that the "background" here also likely contained wrong gene models that were not DOGs, thus "background" does not mean "correct". In addition, DOGs that are caused by absent introns and exons will not necessarily have poor splice junction recall scores, due to the one way nature of the score. In light of these facts, the difference between DOGs and the background almost certainly underestimates the prevalence of inaccurate gene models in DOGs. Similarly, and again only for the plants, for which the extrinsic scores were relevant, discrepant orthogroups were shown to have lower coverage scores than the background. Thus, clear discrepancies in orthogroup alignments are a strong indicator of inaccurate gene models.

To summarise the results of the discrepancy analysis, a score was defined by expressing the number of discrepant SCU orthogroups as a proportion of all SCU orthogroups for that species set. Here, in line with the extrinsic analyses, *C. papaya* and *T. cacao* performed the worst (7.0% and 3.8% discrepant). The two model species scored among the best, although *A. thaliana* (1.2%) was beaten by *B. rapa* (1.0%). It may be that *B. rapa* has particularly well-constructed gene models, or that it is not amenable to discrepancy analysis for some reason.

6.2.5 Non-extrinsic methods reveal hidden errors

As noted, fungal orthogroups are not amenable to gene model analysis by appeal to RNA-seq, given that fungal genes rarely contain introns. The most likely gene model error in a fungal gene is an inaccurate start codon, an error that can occur in a gene from any species due to the presence of multiple viable

methionine residues in any given open reading frame (ORF). Given that ORFs and upstream 5′ UTRs are also transcribed into RNA, RNA-seq provides very little help in determining the most accurate start codon for a gene. To test such cases, I assessed the starts of genes in terms of their subcellular localisation consistency, an entropy score that assesses the agreement between subcellular localisation characteristics predicted by analysing amino acid motifs at the 5′ ends of transcripts. For both fungi and plants, subcellular localisation predictions were shown to be less consistent for initial discrepant orthogroups (iDOGs), indicating that orthogroup discrepancies can be used to detect gene model errors at the starts of genes. Thus, in addition to being able to identify splicing errors, iDOG analysis can be used to identify start codon errors.

6.2.6 Many primary transcripts are suboptimal

As an additional analysis, all plant primary transcripts were compared to their alternative gene models, under each score discussed. For example, alternate transcript choices improved the consistency of the amino acid alignment in 39.4% of cases, and 7.2% of orthogroups with imperfect subcellular localisation consistency could be made consistent with an alternative transcript. Furthermore, 16.9% of highly expressed alternate transcripts had better splice junction recall scores, and 32% of marked primary transcripts had suboptimal coverage scores compared with alternatives. Choosing a representative primary transcript for a gene is of critical importance for accurate biological analysis involving the assessment of protein-coding genes, and transcript choice can have a dramatic effect on downstream inferences. Thus the results of Chapter 3 have two purposes: not only do they facilitate the estimation of gene model error prevalence in a genome, they also provide the means to ensure the best representative transcript is chosen from the full set of gene model options.

6.2.7 Discrepancy score likely underestimates the problem

Given the strict nature of DOG assessment, it will considerably underestimate the prevalence of gene model errors. However, I have shown that DOGs are usually indicative of underlying gene model errors, and thus can be used to infer global properties about genome annotations. Furthermore, DOGs, as defined, identify cases in which only one gene is discrepant. Thus they rely on all of the genomes under consideration to be mostly correct. This is often not the case, and less strict characterisations of DOGs should be developed in order to widen the scope of applicability for this flavour of analysis. In general however, the discrepancy score can be used as a lower estimate for deficiencies in gene model annotations, and thus should be applied to public databases to flag up annotations that are in need of care and attention.

6.3 OrthoFiller and OMGene improve annotations

The concepts employed in Chapters 2 and 3 lent themselves directly to constructive approaches to genome annotation improvement, in addition to assessment. In the second part of the thesis, Chapters 4 and 5, I described two self-contained software programs designed strategically to fix the kinds of errors identified in Chapters 2 and 3. Each algorithm was tested on the plant and fungal species sets introduced in Chapter 2, as well as on additional supplemental species introduced during the review process. This section will discuss briefly the relative successes and limitations of each program.

6.3.1 OrthoFiller detects missing genes

When applied to two sets of publicly available genome annotations OrthoFiller identified 896 previously unannotated genes across the 10 species. In particular, 137 novel genes were identified in the two model species: 136 genes in *Arabidopsis thaliana* and one gene in *Saccharomyces cerevisiae*. The distributions of genes

recovered was roughly in line with the projected incompleteness levels obtained in Chapter 2. In total, 60.0% of these genes were supported by RNA-seq read evidence from the selected samples. Although this support level was lower than what was observed for the already annotated genes, it is important to note that highly and frequently expressed genes are much more likely to have already been annotated than lowly or rarely expressed genes. Thus OrthoFiller provides separate strong evidence for the existence of these genes in the absence of RNA-seq data.

6.3.2 OrthoFiller proceeds with caution

With some basic rules, for example splicing constraints, start and stop codons, open reading frames, GC content and other gene evidence, the full complement of *possible* protein-coding genes in any given genome is astronomical. Gene prediction algorithms frequently *over-predict* genes (Guigó et al. 2000; Denton et al. 2014). The naïve approach of predicting all possible protein-coding genes would certainly fill in many of the orthogroups with missing members, but would also leave vast quantities of junk data. The goal of OrthoFiller is, therefore, caution. OrthoFiller is, in essence, a filtration system, to distinguish gene finder results that are likely to be genuine from those for which there is insufficient homology evidence for them to be asserted as true genes.

OrthoFiller uses three filtering steps, applied to HMM analysis and Augustus gene prediction, to considerably reduce the incidence of spurious gene predictions and predict only genes that sufficiently match existing orthogroups. In doing so, many genuine genes will be discarded. This is evidenced by the relatively low recall of genes removed from *S. cerevisiae* and *A. thaliana* data sets: 45.4% and 31.3% of removed genes were recovered for plants and fungi respectively, compared with 94.1% and 88.7% respectively in the *de novo* case (for 10% and 90% cases combined). However, OrthoFiller had much greater precision (i.e. the percentage of results that were contained in the original removed gene set) than Augustus alone: 94.3% and

99.3% for plants and fungi respectively, compared with 60.5% and 88.4% in the *de novo* case. Given that these species should be assumed to be largely complete, it is unlikely that many of these extra genes found with the *de novo* approach are real.

Furthermore, genes recovered using OrthoFiller were of a higher quality than those recovered without the filtering steps. In the *de novo* case, 45.3% of plant results received dually low recovery scores ($oF, pF < 0.5$), compared with 2.4% of genes recovered using OrthoFiller. For the fungi, no genes found using OrthoFiller were of dually low quality, and 2.6% of genes recovered *de novo* were of dually low quality. Fungal genes, which are small and have few introns, are easier to predict than plant genes, and thus it is not surprising that the *de novo* predictor did a good job on these genomes, however it still returned considerable numbers of likely spurious genes. Thus, genes returned by OrthoFiller are more likely to be correct than genes returned *de novo*.

6.3.3 Strategies for improving OrthoFiller

There are a number of avenues for potential improvement of OrthoFiller, which I will consider here. Firstly, in assessing the results of OrthoFiller, I noticed that many of the outputted gene models tended to be simpler than expected in terms of their intron-exon structure. If this is an artefact of poor gene predictions, it is likely that improvements in the individual gene model prediction steps could improve the output of OrthoFiller. This could be done in a number of ways, for example by using multiple gene prediction programs to build genes in the regions indicated by HMM hints, and to pick the ones that most match the amino acid alignment used to find them. In addition, OMGene, the gene model optimisation program introduced in Chapter 5 could be used to optimise gene models at this stage. This is likely to increase the recall of OrthoFiller relative to the simulated depleted data sets used for assessment, as I suspect many more complex gene models are being filtered out due to gene model prediction errors. Using multiple prediction algorithms would

likely require each to be trained and run independently, increasing the runtime of OrthoFiller but improving the results. Similarly, I would expect greater recall could be achieved by supplementing the initial search with individual protein BLAST searches, as is done in some gene prediction pipelines already (Yandell and Ence 2012). This would again increase the run time of the algorithm. The filtration steps that are the essence of OrthoFiller would remain the same under these analyses.

Genes found using OrthoFiller are in general of a much higher quality than those found using *de novo* methods alone. Despite having stringent criteria for acceptance of putative gene models, OrthoFiller is not immune to the discovery of pseudogenes. Any pseudogenes found using OrthoFiller will have displayed high levels of sequence similarity to the other genes in the orthogroup, and so are likely to still be biologically useful, however post-OrthoFiller analyses could be implemented to double-check these. Improved gene model prediction as described in the paragraph above would also facilitate a basic pseudogene assessment step. If gene models can be optimised, an additional filtration step that ensures outputted gene models have a certain degree of similarity to the orthogroup alignment used to predict them could be implemented. Predicted genes that fail to reach this degree of similarity would be rejected.

OrthoFiller will not find all biologically present genes that are missing from an assembly's annotation, however the genes that it does find are more likely to be correct than without the OrthoFiller filtration steps. Especially for highly under-annotated genomes, OrthoFiller demonstrates a strong capacity to reliably recover missing annotations, without adding hundreds or thousands of spurious additional genes. Thus, OrthoFiller will be of great value to any annotation or re-annotation effort, especially for genomes that have been observed to be under-annotated.

6.3.4 OMGene improves gene models

In Chapter 5 I introduced OMGene, an automated strategy for systematically improving gene models via mutual optimisation of amino acid sequences in a multiple sequence alignment. Across the 15 species assessed, OMGene changed 4,231 gene models. OMGene made fewer changes to the model organisms and more changes to species previously identified as having lower-quality annotations (*C. papaya*, *T. cacao*, *E. gossypii*, *Y. lipolytica*). When applied to *de novo* annotated genomes, OMGene returned even larger numbers of changes, with 3,473 changes made across three species.

In total, 81.6% of assessable changed gene models were improved from an RNA-seq splice junction perspective, 83.9% were improved from an RNA-seq coverage perspective, and 86.2% of assessable changed orthogroups were improved from the perspective of subcellular localisation consistency. While OMGene worked well for plants and mammals, the success was lower for simple gene models such as those in fungi, for which splice junction errors are rare.

Generally, OMGene removed more genic material than it added. This was expected for the mammal species, given that the transcripts were chosen such that peptide length was maximised. In the other cases, this is in line with the observation that computational gene finders tend to over-predict, rather than under-predict exons (Guigó et al. 2000). On the whole, OMGene demonstrated high potential for improving the gene models of publicly available gene annotations, and demonstrated even higher potential for improving the gene models of *de novo* predicted annotations.

6.3.5 Strategies for improving OMGene

Although OMGene improved gene models in a large number of cases, both quantitative and observational analyses revealed situations in which the algorithm could be improved. Firstly, although around 85% of genes analysed were improved using

OMGene, roughly 15% of gene models were made worse. In some of these cases, the algorithm has done its job and significantly improved the multiple sequence alignment, however the underlying sequences have not followed the presumed tenet that the optimal (and thus transcribed) gene model is the one that optimises the multiple sequence alignment. Little can be done about these cases. In other cases, due either to evasion of the stringent OMGene filters or due to implementation errors, the gene model has been worsened by OMGene. These are the cases that need to be investigated and improved upon. These can happen, for example, due to the sequential optimisation nature of OMGene: genes are optimised from the start codon forward. This can yield sequentially optimal genes that fail to obtain a global optimum in terms of their gene model. These types of errors need to be carefully considered, as it would be preferred that OMGene does not reduce the quality of existing annotations.

The major potential for improvement for OMGene is to tackle "easy" alignment discrepancies first, before implementing the global optimisation scheme. These types of errors would be similar to the ones identified as DOGs in Section 3, but with less stringent requirements, given that additional evidence will be sought to confirm the existence of a possible better gene model. OMGene implements several strict filtering steps to mitigate the effects of observed over-predictions. While many gene models are improved, simple observations of the results identify cases where a gene model change should have been made, but was not. Indeed, many such positive changes were present before implementation of the filters, but did not make it into the final result. Such cases are an unfortunate consequence of caution, and it was considered better to be safe by omitting some of the potential positive changes than over-fixing the genes. This could be mitigated by going for the "easy-fixes" first.

Given that OMGene does not fix all gene models, some caution should be employed when applying it to established gene annotations, as some of the gene models may have their accuracy reduced. For genomes that have had less manual

attention however, OMGene demonstrated considerable success in improving gene models for large number of genes. As a layer on *de novo* predictions, and for laboratories preparing new genomes for annotation, however, the improvements yielded by OMGene will be of great value to the annotation community.

6.4 The future

The work described in this thesis provides a significant step toward readily applicable software for the improvement of genome annotations. In particular, OrthoFiller and OMGene can be used by individual laboratories to improve genome annotations without the need for large amounts of experimental data or manual curation, and gene sets can be optimised before the point of submission to pipelines such as the one provided by the NCBI (Thibaud-Nissen et al. 2013). This final section will consider more wide-ranging applications of the ideas of this thesis, both making use of and separate to OrthoFiller and OMGene, and will consider their place in light of potential future developments in sequencing and annotation technology.

6.4.1 An OrthoFiller/OMGene hybrid could increase recall

OrthoFiller suffers in part from relatively poor recall, due to stringent filtration steps to avoid over-prediction of genes. Two key steps in the OrthoFiller algorithm are the comparison of Augustus-predicted gene models to the hits that were used to obtain them, and the subsequent placement of those genes into orthogroups. Both of these steps can suffer if the initial gene model is not adequately determined. As such, a modification of OrthoFiller that optimises the gene model at this internal step may improve the recall and accuracy of the algorithm. Such an improvement could be obtained by using multiple gene model predictors alongside or instead of Augustus, or implementing an internal OMGene optimisation step in OrthoFiller, thus hybridising the two algorithms. The optimisation step would be implemented after the *de novo* prediction of genes (currently carried out using

Augustus). This would have the additional benefit of providing higher-confidence gene models for detecting pseudogenisation, and thus OMGene could be used to help classify OrthoFiller results as likely pseudogenes or true protein-coding genes by optimising the final gene model and implementing a cutoff for alignment similarity relative to the rest of its orthogroup.

6.4.2 OrthoFiller/OMGene will reach singularities

One of the advantages of the phylogenetically local assessments of Chapters 2 and 3 is that they avoid the circularity inherent in comparing annotated gene sets to the universal features that will almost certainly have formed the basis for gene finding and algorithm training. *The singularity* of an annotation completeness or accuracy metric is the point at which the specific type of evidence used to inform the assessment is already fully represented within the data. This would cause the metric to return 100% completeness or accuracy.

The methods described in Chapters 2 and 3 define measures of completeness and accuracy *relative* to small sets of related species, and if OMGene and OrthoFiller do their job properly, after applying them these measures too will yield 100%. This won't necessarily mean that the genomes are absolutely 100% complete and accurate. After filling in and adjusting all the orthogroups and genes in the input species set, the singularity will be approached for this input set. Thus, at this point, a new approach will need to be found for assessing completeness and accuracy, and the cycle will start over. This could be avoided in part by repeatedly seeking different species sets, however it is important that the relative nature of these measures be kept in mind.

6.4.3 A combined approach could identify non-canonical features

The standard model of gene composition – for eukaryotes at least – is to begin with a methionine (ATG) codon, splice between GT and AG dinucleotides spanning a

polypyrimidine tract, and end with a TAA, TGA, or TAG stop codon. OMGene, OrthoFiller, and practically all *ab initio* and automated genome annotation pipelines are designed to stick to these rules. In nature, however, exceptions to the rules are far from uncommon. Translation can be initiated in eukaryotes with CTG codons via a Leu-tRNA pathway (Starck et al. 2012). Furthermore, translation initiation has been observed in at least 47 of the 64 possible codons in prokaryotes (Hecht et al. 2017), and it is not beyond the realms of feasibility that some of these could exist in eukaryotes, albeit in very rare cases. A-to-I and C-to-U RNA editing can yield cryptic start codons such as ATA, ACA, and ACG (Pachter 2012), similarly cryptic stop codons can be encoded by CAA, CAG, and CGA, and apparent stop codons (TAA, TAG, or TGA) can be cryptically translated as tryptophan (TGG). GC-AG splice sites are not uncommon (roughly 0.6% of splice sites in eukaryotes), and at least eight other splice pairs have been observed (Burset et al. 2000). Some of these may again be cryptic canonical features revealed after RNA-editing events. Finally, TGA codons can sometimes be translated as selenocysteine instead of indicating translation termination (Böck et al. 1991).

Non-canonical features such as the ones described above are usually manually marked after consideration of experimental evidence, and they are typically grossly underrepresented in newly annotated genomes. If a non-canonical feature is present in a gene, either the gene model will be designed to circumvent it (for example, splicing "bridges" over selenocysteine codons), or the gene will be omitted entirely, or may be assumed to be a pseudogene. Through homology, such features can be identified. Genes that may have been missed for these reasons can be identified using the HMM step of OrthoFiller. The alignment step of OMGene can then be used to identify potential non-canonical features, such as selenocysteines, or cryptic splice sites and stop codons. The type of change could then be fed to OMGene as a parameter, and the more permissive version of OMGene could then be run on the same gene to optimise it under the new constraints.

6.4.4 What to do with all the data?

Large online genomic data repositories such as GenBank (Benson et al. 2013) and Ensembl (Curwen et al. 2004; Aken et al. 2016) are excellent resources for submitting and obtaining genome information. However, proposed assembly and annotation updates for any given species are difficult to upload for anyone except the team who submitted the original data for that species. This is one reason why it is so important for teams to ensure the original annotations are of as high a quality as possible, and moreover to accept the burden of upkeep of the resources that they provide. Mark Yandell compared the release of a genome assembly to parenthood: "annotation responsibilities do not end with birth" (Yandell and Ence 2012). For researchers who were not involved in the original project, the current best approach upon discovery of an annotation error is to contact the team responsible for uploading the data. On a larger scale, however, it is impractical for all updates to be channelled in this way, and person-hour bottlenecks will naturally inhibit the flow of information onto databases such as GenBank and Ensembl. Alternative tactics for genome resource database management have been suggested. In 2007, Steven Salzberg posited a "wiki" style solution of open authorship and large-scale peer review (Salzberg 2007), like the system used in Wikipedia. This is the approach that I would favour, yet ten years on the dream remains unrealised. The provisions for robustly crowdsourcing annotation updates such as the ones that will be suggested by OrthoFiller and OMGene need to improved and made more accessible, and approaches need to be standardised.

6.4.5 The future of sequencing

Updates to sequencing technologies like the ones already emerging will undoubtedly change the face of genome annotation. Ultra-long read sequencing approaches such as NanoPore (Jain et al. 2017) will allow resolution of repetitive genome regions, and will vastly improve the quality of genome assemblies, minimising their negative

effects on annotations. Nanopore RNA sequencing will improve the mapping of transcripts and eliminate amplification biases (Garalde et al. 2018), making gene models easier to read directly from sequenced transcripts. Undoubtedly, machine learning and artificial intelligence approaches will assume the role of manual curators in optimising gene models based on extrinsic data. However, as noted in this thesis, extrinsic data does not always tell the whole story, and homology-based approaches such as those employed in OrthoFiller and OMGene will always be needed.

6.5 Overall summary

Genome annotation errors are common and can cause significant problems for biological analyses. Genome annotations are frequently taken at face value, and errors are often propagated between species and analyses, with the source of error becoming progressively obscured. As such, every care needs to be taken by those carrying out genome annotation projects, and even anyone using genome annotation data, that the gene sets being used are as complete and accurate as possible. The genome assembly and annotation community need to pay serious attention to the upkeep of public genome annotations.

In this thesis I contributed to improving genome annotations in three significant ways. Firstly, I provided scalable approaches to assessing the relative levels of completeness and accuracy of publicly available genome annotations, in the absence of extrinsic data and without leveraging universally conserved genes. These approaches will allow researchers to flag which genome annotations need particular attention, especially for genes that are not highly expressed in RNA-seq samples. In analysing the completeness and accuracy of the genome annotations, I obtained conservative lower estimates for the prevalence of annotation errors, and in the worst case (*C. papaya*) concluded that at least 3.9% of genes are missing and at least 7.0% have inaccurate gene models, based on non-extrinsic methods alone. Secondly, I introduced OrthoFiller, a software tool for improving the completeness

of genome annotations. Given that prior analyses, as well as the ones carried out in this thesis, indicated that many genomes are under-annotated, such a tool was lacking and in high demand. Using OrthoFiller, I was able to find 896 high-confidence genes in plants and fungi that were previously unannotated. Thirdly, I introduced OMGene, a tool to automatically improve gene models relative to other members of their orthogroups. Again, given that gene models are highly error-prone, such a tool was in high demand, was lacking, and will be of great use to the annotation community. Across the plant and fungal species, OMGene updated 1,770 gene models, with 85.4% of changed junction scores and 86.0% of changed coverage scores being improvements.

In summary, OrthoFiller and OMGene should be incorporated into automated genome reannotation pipelines to help minimise the spread of false information, or at the very least be used by any team assembling and annotating a genome to improve annotations before submission. Overall, the algorithms in this thesis demonstrate the deficiencies in public annotations, and provide approaches to improve them, thus providing a novel, important, and significant contribution to tackling the problem of inaccurate genome annotations.

Appendices

A

OrthoFiller (published version)

This appendix contains the original manuscript (Dunne and Kelly 2017) of the work contained in Chapter 4, which was accepted for publication in BMC Genomics on 18th May 2017.

Dunne and Kelly *BMC Genomics* (2017) 18:390
DOI 10.1186/s12864-017-3771-x

BMC Genomics

SOFTWARE Open Access

OrthoFiller: utilising data from multiple species to improve the completeness of genome annotations

Michael P. Dunne and Steven Kelly*

Abstract

Background: Complete and accurate annotation of sequenced genomes is of paramount importance to their utility and analysis. Differences in gene prediction pipelines mean that genome annotations for a species can differ considerably in the quality and quantity of their predicted genes. Furthermore, genes that are present in genome sequences sometimes fail to be detected by computational gene prediction methods. Erroneously unannotated genes can lead to oversights and inaccurate assertions in biological investigations, especially for smaller-scale genome projects, which rely heavily on computational prediction.

Results: Here we present OrthoFiller, a tool designed to address the problem of finding and adding such missing genes to genome annotations. OrthoFiller leverages information from multiple related species to identify those genes whose existence can be verified through comparison with known gene families, but which have not been predicted. By simulating missing gene annotations in real sequence datasets from both plants and fungi we demonstrate the accuracy and utility of OrthoFiller for finding missing genes and improving genome annotations. Furthermore, we show that applying OrthoFiller to existing "complete" genome annotations can identify and correct substantial numbers of erroneously missing genes in these two sets of species.

Conclusions: We show that significant improvements in the completeness of genome annotations can be made by leveraging information from multiple species.

Keywords: Genome annotation, Gene prediction, Orthology, Orthogroup

Background

Genome sequences have become fundamental to many aspects of biological research. They provide the basis for our understanding of the biological properties of organisms, and enable extrapolation and comparison of information between species. Owing to the increasing availability and affordability of whole-genome sequencing technology [1, 2], genomic data sets are now produced at a rate at which it is infeasible to rely entirely on careful manual curation to annotate a new genome; rather it is taken as given that a considerable portion of the process must be automated.

There has been substantial methodology development in the area of automated gene prediction, with the production of several effective algorithms for identifying genes in *de novo* sequenced genomes [3]. In general, these methods predict genes by learning species-specific characteristics from training sets of manually curated genes. These characteristics include the distribution of intron and exon lengths, intron GC content, exon GC content, codon bias, and motifs associated with the starts and ends of exons (splice donor and acceptor sites, poly-pyrimidine tracts and other features). These characteristics are then used to identify novel genes in raw nucleotide sequences. These prediction methods vary in their performance, as demonstrated by considerable disagreement in the genes and gene models that they predict [3, 4]. For example, one study [4] comparing Augustus, Fgenesh, GENSCAN and MAKER, looked at the number of genes predicted on a sample set of *D. melanogaster* assemblies with varying numbers of scaffolds. At the extreme end, with 707 scaffolds, the

* Correspondence: steven.kelly@plants.ox.ac.uk
Department of Plant Sciences, University of Oxford, South Parks Road, Oxford OX1 3RB, UK

BioMed Central

© The Author(s). 2017 **Open Access** This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated.

most frugal prediction (MAKER, with 12687 predicted genes) was almost doubled by the most generous prediction (GENSCAN, with 22679 predicted genes). Thus it is to be expected that genome annotations generated by different research groups using different methodologies will differ considerably in the complement of genes that they contain. This disparity is exemplified by a recent study [5] that analysed 12 published plant genomes, assessing them for completeness relative to highly conserved gene sets such as BUSCO [6] and CEGMA [7]. The study found strong evidence for universal eukaryotic genes which appeared to be present in the genomes but had no corresponding gene annotations. This indicates that many genomes likely lack gene annotations even for highly conserved genes.

Absent or inaccurate gene models can not only contribute to oversights in biological investigations, they can also lead to false assertions in large-scale genome and cross-species analyses [8]. For example, incorrectly missing gene annotations can be mistakenly interpreted as gene loss, and such interpretations can lead to mistaken inferences about the biological or metabolic properties of an organism. Similarly, missing gene models can lead to errors in gene expression analyses that map and quantify RNA-seq reads using predicted gene models. Here, reads derived from erroneously missing genes, as they have no reference to map to, have the potential to map to the wrong gene leading to errors in transcript abundance estimation.

Much of the cost and effort involved in *de novo* genome annotation can be reduced by leveraging data from other taxa. Moreover, data from disparate taxa have the potential to be used simultaneously to improve a cohort of genome annotations in a mutualistic framework. A number of approaches have been developed to utilise data from other species to improve or assist the process of genome annotation. For example, an automated alignment-based fungal gene prediction (ABFGP) method [9] has been developed for fungal genomes. While this method works well on fungal genomes, it cannot be applied to other taxa and thus has limited general utility.

OrthoFiller aims to simultaneously leverage data from multiple species to mutually improve the genome annotations of all species under consideration, using the concept of orthogroups. It is designed specifically to find "missing" genes in sets of predicted genes from multiple species. That is, to identify those genes that should be present in a genome's annotation, whose existence can be verified through comparison with known gene families. A standalone python implementation of the algorithm is available under the GPLv3 licence at https://github.com/mpdunne/orthofiller. Example datasets and instructions for running the algorithm are included in the git repository.

Results

Problem definition, algorithm overview and evaluation criteria

OrthoFiller aims to find genes that are present in a species' genome, but which have no predicted gene model in the genome annotation for that species. It takes a probabilistic, orthology-based approach to gene identification, leveraging information from multiple species simultaneously to improve the completeness of the genome annotations for all species under consideration. OrthoFiller is not designed for *ab initio* gene prediction and requires that each genome under consideration possesses a basic level of annotation, taken to be at least 100 annotated genes. The genomes should ideally be from a set of related species from the same taxonomic group (genus, family, order or class).

OrthoFiller makes use of the concept of orthogroups, an extension of the pairwise concepts of orthologous and paralogous genes. For a given set of species, an orthogroup is the set of genes descended from a single ancestral gene in the last common ancestor of those species [10]. By definition, orthogroups may contain paralogous as well as orthologous genes. In the case of OrthoFiller, orthogroups provide the basis for the gene searching process, the aim being specifically to find unannotated members of existing orthogroups. As a result, OrthoFiller is able to find genes that are paralogous as well as orthologous to known genes.

A workflow for OrthoFiller is shown in Fig. 1. The basic input for the algorithm is a set of genome annotation files in general transfer format (GTF) and a set of corresponding genome sequence files in FASTA format. The output from the algorithm is a set of FASTA format amino acid sequences for each inputted species, and a set of GTF annotations containing the new genes. OrthoFiller uses OrthoFinder [10] to cluster the inputted genes into orthogroups. Advanced users may also specify their own orthogroups, as described in the software documentation.

Protein sequences are extracted from the genome FASTA files using the coordinates in the GTF files and a user-selected translation table. OrthoFiller then uses OrthoFinder to cluster the genes from the species into orthogroups. The protein sequences of each orthogroup are aligned and the source nucleotide sequences for these proteins are threaded back through the protein multiple sequence alignment to create multiple sequence alignments of the nucleotide sequences of each orthogroup. Each nucleotide alignment is used to build a hidden Markov model (HMM) that is used to search the complete genome sequence of each species under consideration. The scores of these HMMs are used to learn the score distributions of true positive and false positive HMM hits (see Implementation). Each hit to an HMM

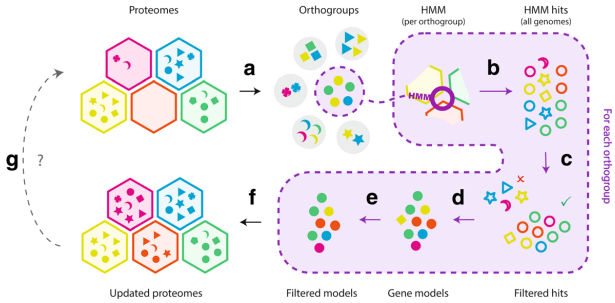


Fig. 1 Workflow diagram for the OrthoFiller algorithm. **a** Proteomes are subdivided into orthogroups using OrthoFinder. **b** Protein sequences in each orthogroup are subject to multiple sequence alignment, back-translated to DNA and used to create hidden Markov models (HMMs). These HMMs are used to search each genome in the set. **c** The set of hits are evaluated and filtered to remove low quality hits. **d** Gene models are constructed around each retained hit using Augustus. **e** The new gene models are compared to the hints that were used to generate them, and filtered to remove those which bear insufficient similarity to the hints. **f** The filtered genes are clustered into orthogroups and genes that are successfully placed into the orthogroup that was used to identify them are retained. **g** The process may be run once, or iteratively until no further genes are found

that does not overlap with an existing predicted gene is subject to filtration using species-specific parameters that have been learned for true and false positive hits. Each hit that survives this filtration is considered to be a potential genic region, or *hint*. The algorithm then attempts to build gene models around these hints, using the Augustus [11] gene finder. Gene models constructed by Augustus are subject to two successive rounds of assessment and filtration. Firstly, the predicted gene models are compared against the hints that were used to inform them: if the gene model and its source hint are not sufficiently similar (see Implementation), the gene model is considered to be unrelated to the hint, and thus to the orthogroup used to inform its prediction. Secondly, the newly predicted genes that satisfy the first criterion are subject to orthogroup inference using the full set of existing and newly predicted genes. Those newly predicted genes that are clustered in an orthogroup whose HMM was used to predict them are then accepted as *bona fide* genes and added to the genome annotation. Thus, genes predicted by OrthoFiller satisfy stringent orthology-based criteria for their existence.

To demonstrate the utility of OrthoFiller on real data it was applied independently to two sets of species. Set A comprised five fungal genomes (Table 1) and Set B comprised five plant genomes (Table 2), sourced from the Joint Genome Institute (JGI) and the Saccharomyces Genome Database (SGD) [12–15]. OrthoFiller was assessed using these datasets in two ways: first via simulating an incomplete genome annotation by randomly removing entries from the genome annotation of one species from each set, and assessing the accuracy of OrthoFiller in recovering the removed genes; second by application of OrthoFiller to the complete datasets and validating the novel detected genes through analysis of publicly available RNA-seq data.

Two measures were used to assess the quality of recovered genes: the protein F-score and the orthogroup F-score, both defined in the Implementation section. These scores were calculated for all genes identified by OrthoFiller, by comparing the recovered gene with the removed gene and assuming that the original removed gene model was correct. Genes that are unique to the test species that lack homologues in other species were not analysed in this test, as OrthoFiller was designed to find evolutionarily conserved genes. As there were no publicly-available comparable methods that perform the same task as OrthoFiller, the method was assessed in comparison to performing the analysis without conducting the OrthoFiller evaluation and filtration steps. i.e. using unfiltered HMM hits from the orthogroups as hints for the *de novo* gene finder Augustus, and accepting all identified gene models that did not overlap an existing gene.

Evaluation of OrthoFiller on *S. cerevisiae* after removal of 10% of gene annotations

Figure 2 and Table 3 show the results of running OrthoFiller on the set of fungal species shown in Table 1 after random removal of 10% of "discoverable" genes (genes that were contained in an orthogroup with at least one

Table 1 Species Set A, fungal species used for algorithm validation

Species name	Source	Strain	Taxonomy ID	References
Eremothecium gossypii	JGI[a]	*ATCC10895*	284811	[12]
Debaromyces hansenii	JGI	*CBS767*	284592	[13, 14]
Kluyveromyces lactis	JGI	*CLIB210*	284590	[13]
Saccharomyces cerevisiae	SGD[b]	*S288C*	559292	[26]
Yarrowia lipolytica	JGI	*CLIB122*	284591	[13]

[a]Joint Genome Institute; [b]Saccharomyces Genome Database

Table 2 Species Set B, plant species used for algorithm validation

Species name	Source	Version	Taxonomy ID	References
Arabidopsis thaliana	JGI	TAIR10	3702	[15]
Brassica rapa	JGI	v1.3	3711	[15]
Carica papaya	JGI	ASGPBv0.4	3649	[15]
Capsella rubella	JGI	v1.0	81985	[15]
Theobroma cacao	JGI	v1.1	3641	[15]

gene from another species) from the predicted complement of genes in *S. cerevisiae* (i.e. 513 nuclear encoded gene annotations were deleted from a total set of 5129 discoverable genes). This was performed 10 times, each time with a different disjoint subset removed. The full details of detection of the deleted genes at different stages in the OrthoFiller algorithm are shown in Additional file 1: Figure S1.

After running OrthoFiller, an average of 160 genes were predicted in the genome of *S. cerevisiae* that were not present in the submitted depleted genome annotation file. Of these, 98.9% overlapped with genes that were deleted from the original annotation. 96.1% of genes were recovered to high accuracy (protein F-score ≥ 0.95). The mean protein F-score of the remaining genes of lower accuracy (protein F-score < 0.95) was still high at 0.89 (Fig. 2b). All of the genes that had lower gene model accuracy were placed in exactly the same orthogroup as expected when the sequences were subjected to orthogroup inference. Thus, although typically around six of the gene models differed from the original reference gene model, this difference was not sufficient to disrupt downstream identification of orthogroups.

To provide a comparison, in the absence of the OrthoFiller evaluation steps an average of 582 genes were identified, of which only 79.2% overlapped with genes that were deleted from the original annotation and an average of 120 genes had not been predicted as genes in the original *S. cerevisiae* genome (Fig. 2a), indicating many of these were erroneous. In addition, on average 11 removed genes were split into multiple individual genes upon recovery. Of the predicted genes that overlapped with removed genes, 89.3% were recovered with high accuracy (protein F-score ≥ 0.95) and the mean protein F-score of those recovered to a lower accuracy was 0.57 (Fig. 2b), considerably lower than with OrthoFiller. Of the lower-quality genes, an average of 42.1% had an orthogroup F-score less than or equal to 0.95, compared with 4.0% for OrthoFiller. Moreover, 39.4% of the lower-quality genes were sufficiently mis-predicted that they failed to be placed in an orthogroup, or were placed in an orthogroup that shared no members with the orthogroup that contained the original gene. Thus in the absence of OrthoFiller filtration there was an increase in the percentage of gene prediction errors and a reduction in the accuracy of orthogroup inference.

Figure 2c-d show the distribution of orthogroup F-scores versus protein F-scores obtained following application of OrthoFiller to this test dataset. In these figures, results from all 10 runs have been pooled together. The majority of genes recovered with OrthoFiller had both high protein and orthogroup F-scores (Fig. 2c): 94.4% had both F-scores ≥ 0.95. This indicates that the majority of predicted genes were identical (or nearly identical) to the original removed gene and that when subjected to orthogroup inference they were clustered in the correct orthogroup. Imperfect protein F-scores can be explained by discrepancies in intron/exon and start/stop codon choices between the removed and recovered gene models. Imperfect orthogroup F-scores were due to fluctuations in orthogroup membership. Figure 2d shows the results in the absence of OrthoFiller processing. In this case, 85.3% were of dually high quality. However,

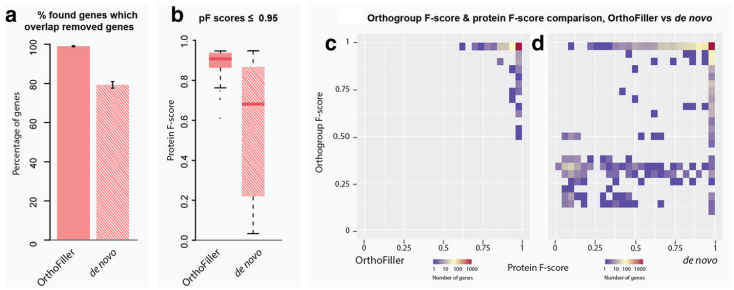


Fig. 2 Performance of OrthoFiller on *S. cerevisiae* genome with 10% of annotated genes removed. **a** Using OrthoFiller an average of 158 of 160 found genes (99.0%) had genomic locations matching any of the 513 deleted genes. In the absence of OrthoFiller filtration an average of 447 of 582 (79.2%) of the found genes matched any of the deleted genes. **b** Boxplot of protein F-scores for genes predicted using OrthoFiller, versus predictions made in the absence of OrthoFiller filtration, that had a protein F-score of ≤0.95. **c** Density plot showing the protein and orthogroup F-scores for all recovered genes using OrthoFiller. **d** Density plot showing the protein and orthogroup F-scores for all recovered genes in the absence of OrthoFiller filtration

Table 3 Recovery of removed genes in *S. cerevisiae*, averaged over 10 runs

	10% annotations removed				90% annotations removed			
	OrthoFiller		*de novo*		OrthoFiller		*de novo*	
No. genes removed	513		513		4615		4615	
Total genes found[a]	160	(11.4)	582	(21.2)	1473	(47.5)	4270	(28.2)
Found genes which overlap removed genes[a]	158	(11.2)	462	(11.9)	1472	(47.4)	4163	(18.3)
Total recovered genes[a]	158	(11.1)	447	(9.6)	1471	(47.3)	4074	(22.4)
Number of split genes[a]	0	(0.3)	11	(4.5)	1	(0.7)	89	(8.49)
Mean pF-score of found genes[a]	0.99	(<0.01)	0.95	(0.01)	0.99	(<0.01)	0.96	(<0.01)
Mean oF-score of found genes[a]	0.99	(<0.01)	0.96	(0.01)	0.99	(<0.01)	0.96	(<0.01)
High-quality found genes (pF-score ≥0.95)[a]	152	(9.1)	412.3	(9.4)	1401.7	(44.76)	3766	(33.6)
Lower-quality found genes (pF-score <0.95) [a]	6.2	(3.2)	49.3	(12.2)	70.6	(5.6)	396.7	(23.1)
Mean pF-score of lower-quality genes[a]	0.89	(0.04)	0.57	(0.07)	0.88	(<0.01)	0.63	(0.01)
% of lower-quality genes with oF-score <0.95[a]	4.0	(8.1)	42.1	(9.3)	4.8	(1.4)	38.7	(1.1)

[a]Numbers shown are rounded mean values from 10 disjoint removed subsets of genes, with standard deviations bracketed

3.2% of predicted genes had both a low (<0.5) protein and orthogroup F-score, indicating those predicted genes were sufficiently incorrect to cause errors in orthogroup inference. Thus, although OrthoFiller does not recover all deleted genes (30.8% of removed genes), application of OrthoFiller resulted in the recovery of high-quality gene annotations that contain few (in this example there are none) incorrectly predicted genes.

Evaluation of OrthoFiller on *S. cerevisiae* after removal of 90% of gene annotations

Figure 3 and Table 3 show the performance statistics for OrthoFiller using a version of *S. cerevisiae* genome where 90% of gene annotations were removed (4615 annotations). This represents an extreme case where a genome has minimal annotation. The full details of detection of the deleted genes at different stages in the OrthoFiller algorithm are shown in Additional file 2: Figure S2. Again, the experiment was run 10 times, this time with disjoint subsets of 10% of genes left remaining. Here, application of OrthoFiller resulted in the identification of on average 1473 genes, of which 99.9% overlapped with the removed genes. Of the found genes, 95.2% were recovered with a protein F-score of 0.95 or greater. Of the genes with lower protein F-scores (Fig. 3b), only 4.7% had an orthogroup F-score < 0.95. As before, although these gene models differed from the original reference gene model, this difference was not sufficient to disrupt downstream orthogroup inference.

In the absence of OrthoFiller filtration, an average of 4270 genes were found, of which 97.5% overlapped the removed genes. An average of 89 removed genes were split into multiple individual genes upon recovery. On average, 90.4% of the found genes had a protein F-score ≥ 0.95. Of the genes with lower protein F-scores, 39.4% had an orthogroup F-score lower than 0.95, and

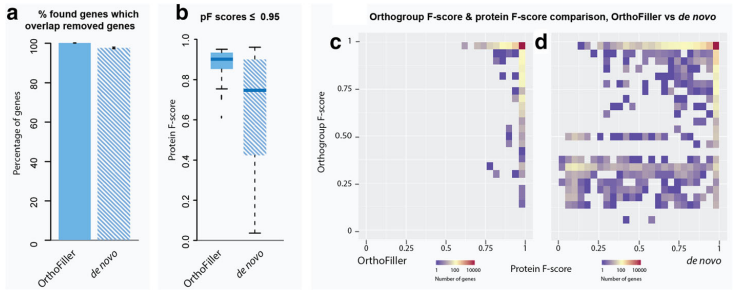


Fig. 3 Performance of OrthoFiller on S. cerevisiae genome with 90% of annotated genes removed. **a** Using OrthoFiller an average of 1472 of 1473 found genes (>99.9%) matched any of the 4615 deleted genes. In the absence of OrthoFiller filtration an average of 4163 of 4267 (97.5%) of the found genes matched any of the deleted genes. **b** Boxplot of protein F-scores for genes predicted using OrthoFiller, versus predictions made in the absence of OrthoFiller filtration, that had a protein F-score of ≤0.95. **c** Density plot showing the protein and orthogroup F-scores for all recovered genes using OrthoFiller. **d** Density plot showing the protein and orthogroup F-scores for all recovered genes in the absence of OrthoFiller filtration

32.3% were sufficiently mis-predicted that they failed to be placed in any orthogroup at all, or in an orthogroup completely different to the one that was used to find them.

Figures 3c-d show the distribution of orthogroup F-scores versus protein F-scores for recovery in the 90% removal case. Figure 3c shows that most genes were recovered well, with 89.8% genes predicted correctly and placed in the correct orthogroup when subject to orthogroup inference (protein F-score ≥ 0.95, orthogroup F-score ≥ 0.95). Interestingly, there are many genes that are predicted correctly but are placed into a slightly different orthogroup to what was expected. This is due to changes in orthogroup membership caused by the many still-missing genes.

Thus, although the input datasets are dramatically different the performance characteristics of OrthoFiller on the 10 and 90% datasets are broadly consistent (e.g. 30.8 and 31.9% recovery of removed genes respectively, of which 94.1 and 89.8% were high-accuracy predictions).

Evaluation of OrthoFiller on *A. thaliana* after removal of 10% of gene annotations

As it could be argued that fungal genomes present an easier challenge, an additional demonstration of the utility of OrthoFiller on an alternative group of organisms was also conducted. Here the analogous test of the method was applied to a set of five land plant genomes (Table 2). OrthoFiller was run five times with 10% of discoverable *A. thaliana* genes removed, with a different disjoint random subset removed each time. Table 4 and Fig. 4 show average performance statistics from these runs.

An average of 1233 genes were discovered by OrthoFiller, 89.7% of which overlapped removed genes. Of these, 39.0% were recovered to very high accuracy (protein F-score ≥ 0.95), and the mean protein score of the lower-quality genes was 0.61. 95% of these lower-quality genes were placed in the correct orthogroup after orthogroup inference, indicating that the lower protein scores were typically a product of inaccurate gene model rather than the gene itself being incorrect, and orthogroup inference remained reliable.

In the absence of OrthoFiller filtration, an average of 13184 genes were discovered, considerably more than the number of genes removed, indicating that a large number of these may be erroneous: only 37.3% of these genes overlapped removed genes. Of those that overlapped removed genes, 24.7% represented removed genes that were split into multiple parts, and only 13.0% were high quality. The average protein F-score of the lower-quality genes was 0.31, and 81.1% of these lower-quality genes also had a low (<0.95) orthogroup F-score. Thus in the absence of OrthoFiller filtration, large numbers of unreliable genes were found.

Figure 4c-d show the distribution of orthogroup F-scores versus protein F-scores for recovery in the 10% removal case for *A. thaliana*. Using OrthoFiller, 37.1% of genes had both a high (≥0.95) protein and orthogroup F-score. Furthermore, in the *de novo* case, 52.4% of recovered genes scored poorly on both metrics (<0.5), compared with 1.4% of genes found using OrthoFiller. Thus using OrthoFiller reduces the proportion of found genes which are erroneous.

Evaluation of OrthoFiller on *A. thaliana* after removal of 90% of gene annotations

Average performance statistics for the application of OrthoFiller to the 90% depleted *A. thaliana* genome (21683 genes removed) can be seen in Table 4 and Fig. 5. Statistics are averaged over five runs using different disjoint subsets of discoverable genes left in the annotation.

Table 4 Recovery of removed genes in A. *thaliana*

	10% annotations removed				90% annotations removed			
	OrthoFiller		*de novo*		OrthoFiller		*de novo*	
No. genes removed	2410		2410		21683		21683	
Total genes found[a]	1233	(37.5)	13184	(426.7)	11480	(96.2)	42504	(223.5)
Found genes which overlap removed genes[a]	1107	(37.5)	4918	(130.4)	11344	(89.0)	35609	(149.5)
Total recovered genes[a]	1036	(31.72)	2269	(16.4)	10380	(59.7)	20431	(33.0)
Number of split genes[a]	67	(5.4)	1214	(23.6)	945	(34.4)	7451	(37.7)
Mean pF-score of found genes[a]	0.75	(0.01)	0.45	(<0.01)	0.70	(<0.01)	0.55	(<0.01)
Mean oF-score of found genes[a]	0.94	(<0.01)	0.51	(<0.01)	0.89	(<0.01)	0.64	(<0.01)
High-quality found genes (pF-score ≥0.95)[a]	432.0	(29.3)	640.8	(27.4)	3419.8	(21.3)	9079.6	(99.5)
Lower-quality found genes (pF-score <0.95)[a]	674.6	(32.1)	4277.2	(147.8)	7923.8	(99.5)	26529.8	(201.3)
Mean pF-score of lower-quality genes[a]	0.61	(0.02)	0.31	(<0.01)	0.57	(<0.01)	0.33	(<0.01)
% of lower-quality genes with oF-score < 0.95[a]	34.6	(1.9)	81.1	(0.9)	43.4	(0.6)	73.9	(0.1)

[a]Numbers shown are rounded mean values from 5 disjoint removed subsets of genes, with standard deviations bracketed

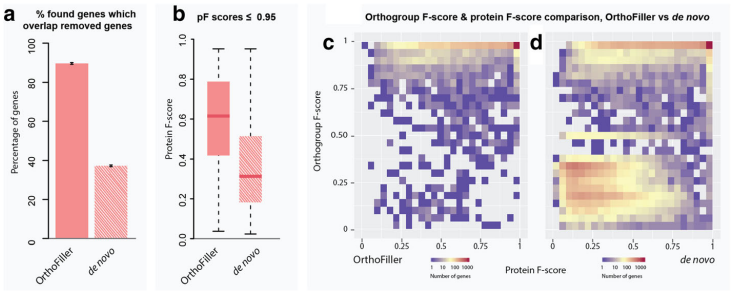


Fig. 4 Performance of OrthoFiller on *A. thaliana* genome with 10% of annotated genes removed. **a** Using OrthoFiller 89% of found genes overlapped removed gene annotations, compared with 37% without filtration; **b** Boxplot of protein F-scores for genes predicted using OrthoFiller, versus predictions made in the absence of OrthoFiller filtration, that had a protein F-score of ≤0.95. Scores were typically lower for the *de novo* runs. **c** Density plot showing the protein and orthogroup F-scores for all recovered genes using OrthoFiller. **d** Density plot showing the protein and orthogroup F-scores for all recovered genes in the absence of OrthoFiller filtration

Using OrthoFiller, 11480 genes were discovered, of which 98.8% overlapped genes that had been removed. Of these, 30.1% had a high protein F-score, and the average protein F-score for those of lower quality remained relatively high at 0.57. 56.5% of genes with lower protein F-scores ended up in the right orthogroup after orthogroup inference.

In the absence of OrthoFiller filtration, many more genes were found than removed, with an average of 42504 genes found. Of these, 83.8% overlapped removed genes, however many removed genes (20%) were split into multiple parts. Only 25% of the overlapping genes were recovered to high accuracy (protein F-score ≥ 0.95), and the average protein F-score for the lower-quality genes was low at 0.33, indicating that a large number of these genes bore little resemblance to the removed gene they overlapped. In addition, 73.9% of lower-quality genes had low-quality orthogroup inference, indicating that the proteins were sufficiently mis-predicted that they disrupted orthogroup inference. This shows that, although more genes were recovered in the absence of OrthoFiller filtration, considerably more noise and erroneous predictions are produced.

Figure 5c-d show the distribution of orthogroup F-scores versus protein F-scores for recovery in the 90% removal case for *A. thaliana*. On average, 38.1% of genes recovered without OrthoFiller were of dually low quality (protein F-score, orthogroup F-score <0.5), compared with only 3.4% of genes recovered with OrthoFiller. Thus, similar to the fungal data set analysis, the performance characteristics of OrthoFiller on the 10 and 90% plant datasets are broadly consistent (e.g. 42.9 and 47.9% recovery respectively, 39.0 and 30.1% high-accuracy recoveries

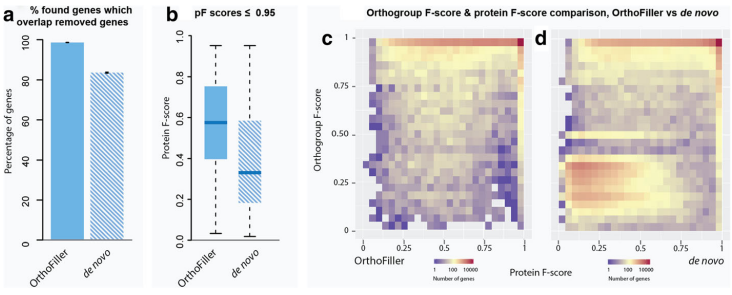


Fig. 5 Performance of OrthoFiller *on A.* thaliana genome with 90% of annotated genes removed. **a** Using OrthoFiller 98.8% of found genes overlapped removed gene annotations, compared with 83.7% without filtration; **b** Boxplot of protein F-scores for genes predicted using OrthoFiller, versus predictions made in the absence of OrthoFiller filtration, that had a protein F-score of ≤0.95. **c** Density plot showing the protein and orthogroup F-scores for all recovered genes using OrthoFiller. **d** Density plot showing the protein and orthogroup F-scores for all recovered genes in the absence of OrthoFiller filtration

respectively), and both contain a considerably smaller proportion of erroneous genes than would be found without filtering.

Evaluation of OrthoFiller on *S. cerevisiae* and *A. thaliana* with different species sets

In order to test the behaviour of OrthoFiller under different species subsets, the algorithm was run on expanded and reduced versions of species sets A and B. The species were selected to test the sensitivity of OrthoFiller to inclusion of additional close and distant relatives, and to test the behaviour of the method when different numbers of species are provided. In each case, three additional species were chosen. One species was an in-group to the original species set and the other two were progressively more distantly related outgroups to the original set. The additional species are listed in Tables 5 and 6 as species sets A+ and B+ respectively.

For both the plant and fungal datasets, the behaviour of OrthoFiller was robust to the inclusion of additional species (Figs. 6 and 7). However, in both cases the method performed poorly when only two species were included. It is therefore recommended that users include at least three species in their analysis.

Evaluation of the SGD reliability classification of recovered genes

The genes recovered in *S. cerevisiae* when applying OrthoFiller after 10 and 90% removal were compared with reliability classifications provided by the SGD. In this genome version, the SGD classified 5103 genes as verified, 784 as dubious, and 806 as uncharacterised. Of the 4766 verified genes that were clustered into orthogroups and hence discoverable by OrthoFiller, 1857 genes were recovered. Of the dubious genes, only three were clustered into orthogroups, and none of these were recovered by OrthoFiller. Of the 806 uncharacterised genes, 357 were clustered into orthogroups, and 53 were recovered by OrthoFiller. Thus dubious and uncharacterised genes are recovered at lower rates than verified genes.

OrthoFiller detects hundreds of conserved genes not present in the reference genome annotations

In addition to testing the ability of OrthoFiller to recover already predicted genes, the algorithm was applied to both of the sets of complete genomes listed in Tables 1 and 2, to assess the potential for novel genes to be discovered. In total, 34 novel genes were found across the fungal species evaluated, and 830 novel genes were found across the plant species. The number of genes found for each species in each set is listed in Tables 7 and 8. All newly predicted genes described in this study are publicly available for download from the Zenodo Research Data Repository (see Implementation).

To be detected as a novel gene OrthoFiller requires genes to pass rigorous sequence similarity tests to genes in other species (including empirical evaluation of sequence similarity scores to distinguish real from spurious hits), which in itself provides evidence for the existence of predicted genes through homology. To provide additional evidence for the existence of the novel predicted genes they were subjected to analysis using publicly available RNAseq data from the Sequence Read Archive (SRA) [16]. The datasets used for this analysis are listed in Tables 9 and 10, and examples of RNAseq coverage on a selection of novel genes can be found in Fig. 8. The tables also show the percentage of the novel genes found that had evidence for their existence in the RNAseq data. For most genomes, most genes predicted by OrthoFiller are highly supported by RNAseq evidence, with the all the novel fungal genes exhibiting RNAseq evidence, and on average 67.2% of novel plant genes being supported. Given that the plant RNAseq datasets come from single tissue samples under a single condition it is not expected that all genes will be detected in these samples. For example, similar detection statistics were obtained for the original predicted genes from the source datasets, shown in Tables 9 and 10. It should also be noted that genes that are present in RNAseq data are more likely to have been annotated already, given that many genome annotation pipelines rely on such data to perform their analyses [17].

Table 5 Species Set A+, additional fungal species used for algorithm validation

Species name	Source	Strain	Taxonomy ID	References
Aspergillus nidulans	JGI	FGSC A4	227321	[27]
Schizosaccharomyces pombe	JGI	972 h-	4896	[28, 29]
Zygosaccharomyces rouxii	JGI	CBS732	559307	[30]

Table 6 Species Set B+, additional plant species used for algorithm validation

Species name	Source	Version	Taxonomy ID	References
Brassica oleraceae capitata	JGI	v1.0	3716	[31]
Citrus clementina	JGI	V1.0	85861	[32]
Populus trichocarpa	JGI	v3.1	3694	[33]

Discussion

Here we present OrthoFiller, an automated method for improving the completeness of genome annotations. It leverages information from multiple taxa, clustering genes into orthogroups and finding genes that are conserved between species but that have escaped detection. OrthoFiller is designed to be stringent, conservatively

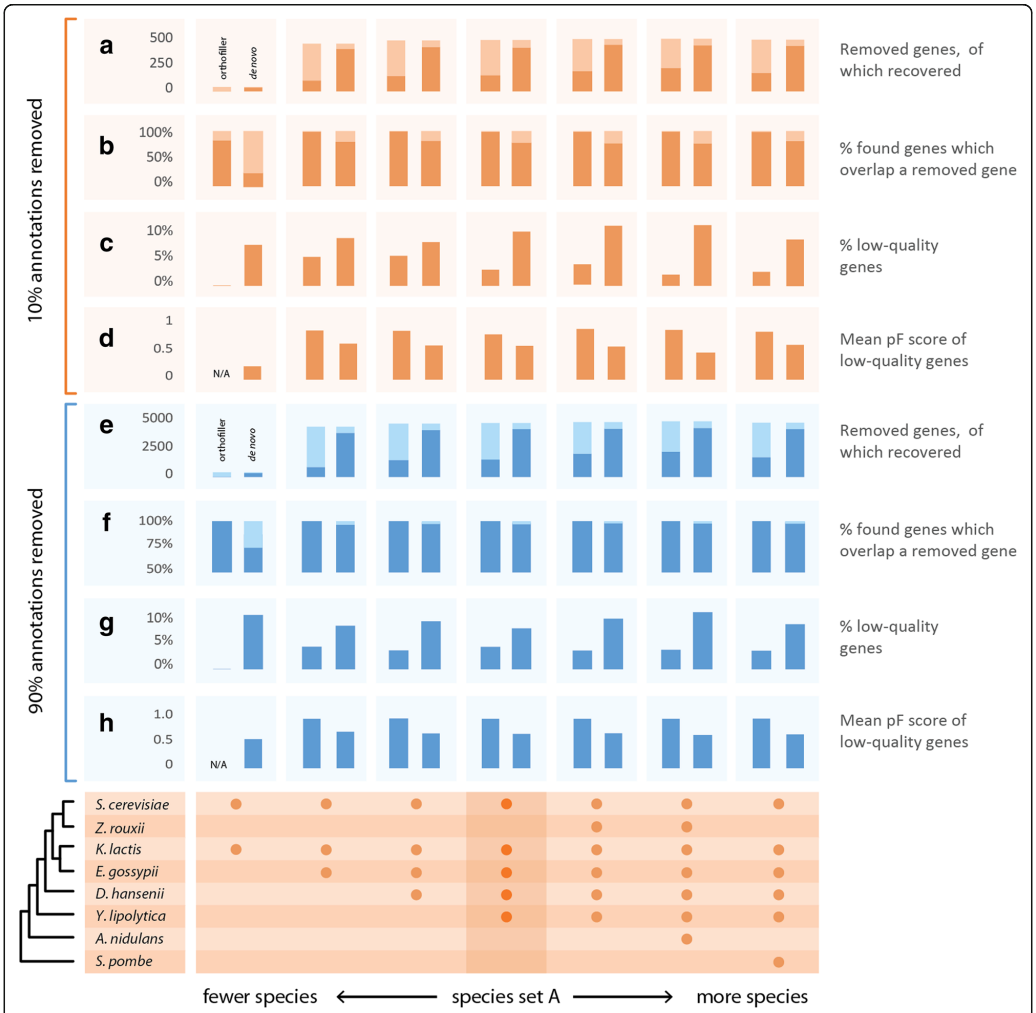


Fig. 6 Analysis of OrthoFiller on different subsets of fungal species. Species subsets are indicated by *bullets* on the bottom panel. Behaviour of OrthoFiller is consistent across different numbers and relatedness of species, with the exception of when only two species are used. **a**, **e** Number of genes removed from *S. cerevisiae* (*light* + *darker colour*), and the proportion recovered (*darker colour*). **b**, **f** Proportion of genes found which overlap removed genes (*darker colour*) is consistently higher for OrthoFiller than for its *de novo* counterpart. **c**, **g** Percentage of recovered genes of low quality, i.e. with protein F-score <0.95, is consistently lower in OrthoFiller than the *de novo* version. **d**, **h** Mean protein F-score of lower-quality genes is consistently higher for OrthoFiller than for the *de novo* version

identifying genes that can be confidently identified as missing members of existing orthogroups. Specifically, to pass the filtration criteria for detection by OrthoFiller, genes must be members of orthogroups conserved in multiple species. Thus OrthoFiller will not find genes that lack homologues in other species. These stringent criteria mean that not all genes that could be detected will be detected by the algorithm, but rather that the user should have confidence in the validity of genes identified by the method.

OrthoFiller is intended to be run after a genome annotation is considered by the user to be complete or near-complete. OrthoFiller is designed with small scale genome sequencing projects in mind and is provided to enable users without significant resources for comprehensive RNAseq-based genome annotation to leverage

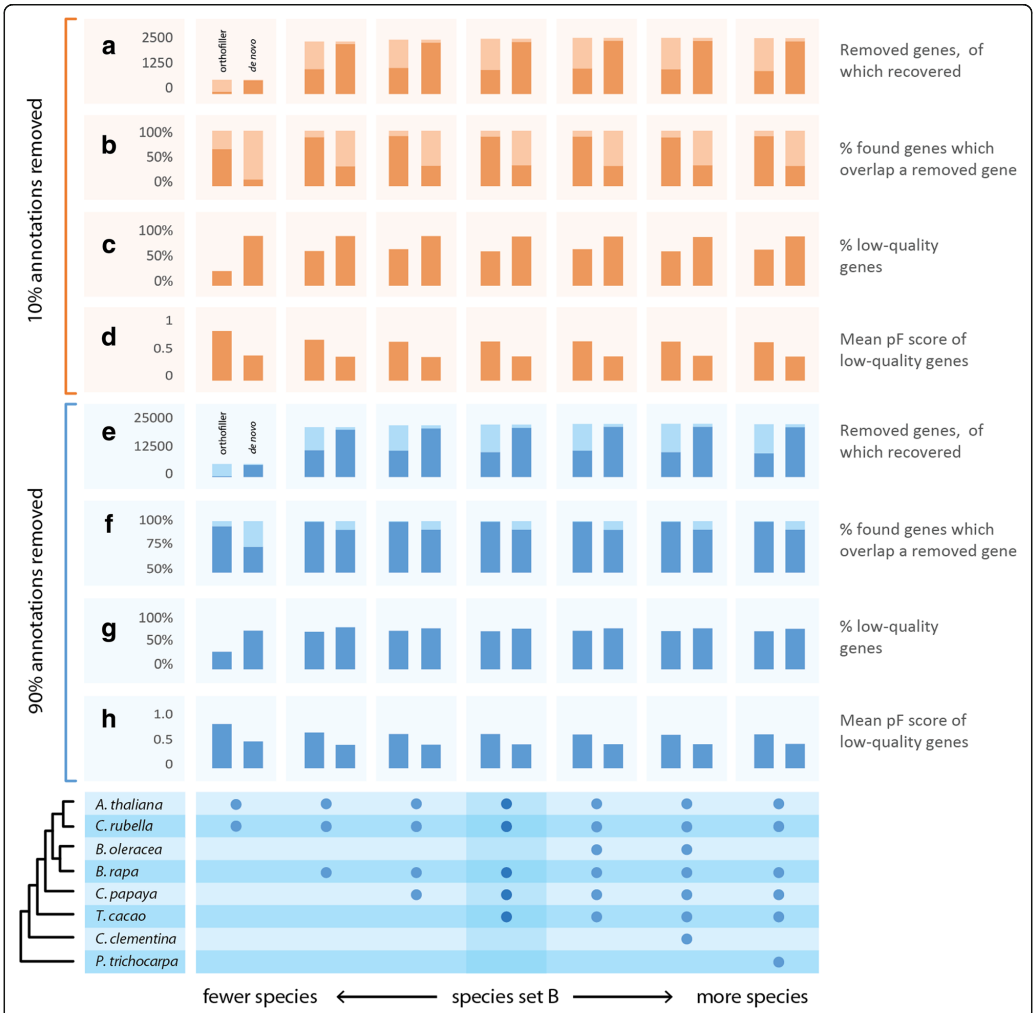


Fig. 7 Analysis of OrthoFiller on different subsets of plant species. Species subsets are indicated by *bullets* on the bottom panel. Behaviour of OrthoFiller is consistent across different numbers and relatedness of species, with the exception of when only two species are used. **a**, **e** Number of genes removed from *A. thaliana* (*light* + *darker colour*), and the proportion recovered (*darker colour*). **b**, **f** Proportion of genes found which overlap removed genes (*darker colour*) is consistently higher for OrthoFiller than for its *de novo* counterpart. **c**, **g** Percentage of recovered genes of low quality, i.e. with protein F-score <0.95, is consistently lower in OrthoFiller than the *de novo* version. **d**, **h** Mean protein F-score of lower-quality genes is consistently higher for OrthoFiller than for the *de novo* version

information from related species to improve their genome annotations. However, OrthoFiller is equally suited for use in large-scale genome comparisons, reliably filling gaps in gene sets prior to large scale comparative genomics investigations. Application of OrthoFiller in these cases will enable the downstream analysis of genes that would otherwise have been classified as absent.

The utility of OrthoFiller is demonstrated on both fungal and plant genome datasets, both in its ability to successfully find missing genes, and in the effectiveness of its filters in eliminating low-quality gene predictions. Application of this method to small groups of plant and fungal genomes resulted in the identification of 34 and 830 genes respectively. These genes are conserved in one or more species but were absent from the genome

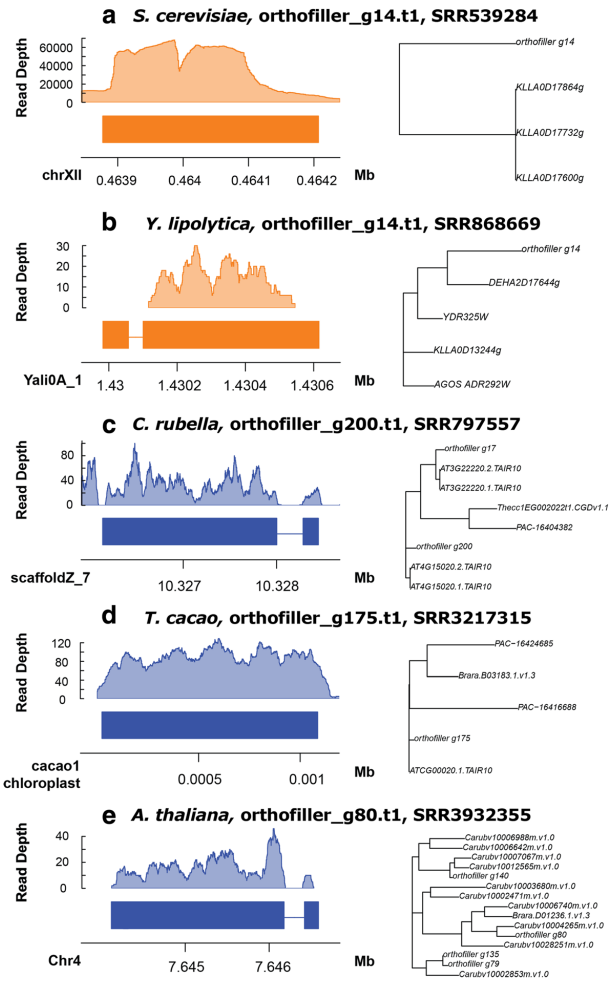


Fig. 8 Five representative examples of RNAseq coverage on genes predicted using OrthoFiller. Coverage plots demonstrate RNAseq reads mapping to locations of new genes, and phylogenetic trees demonstrate the relationship of the newly predicted gene to other genes in the orthogroup. **a** Species *S. cerevisiae*, gene orthofiller_g14.t1, RNAseq data from SRR539284. **b** Species *Y. lipolytica*, gene orthofiller_g14.t1, RNAseq data SRR868669. **c** Species *C. rubella*, gene orthofiller_g200.t1, RNAseq data SRR797557. **d** Species *T. cacao*, gene orthofiller_g175.t1, RNAseq data SRR321315. **e** Species *A. thaliana*, gene orthofiller_g80.t1, RNAseq data SRR3932355

Table 7 Novel genes in fungal species

Species name	Genome size (Mbp)	No. pre-existing genes	No. new genes.
E. gossypii	9.10	4768	2
D. hansenii	12.15	6272	13
K. lactis	10.69	5076	8
S. cerevisiae	12.16	6572	2
Y. lipolytica	20.50	6447	9

annotation in which they were predicted. We anticipate that application of OrthoFiller to larger datasets will likely result in further genome annotation improvement. Alternative publicly available genome versions for the fungal species exist such as those housed at gryc.inra.fr [18]. Although the genome annotations differ to the ones used in this study, 24 of the genes predicted by OrthoFiller were also not present in these alternative genome annotations (details available in Zenodo

Table 8 Novel genes in plant species

Species name	Genome size (Mbp)	No. pre-existing genes	No. new genes.
A. thaliana	119.67	27416	116
B. rapa	315.05	40492	10
C. papaya	342.68	27751	382
C. rubella	134.83	26521	228
T. cacao	346.16	29452	94

dataset that accompanies this manuscript, see Implementation). The quality of genes found by OrthoFiller was assessed by artificial removal and recovery of subsets of genes from a single genome, treating those original gene models as true, and evaluating the quality of those genes that were recovered by comparison to the removed genes. In the absence of the OrthoFiller filtration steps, the proportion of poor-quality genes that are recovered is considerably higher, furthermore OrthoFiller avoids the over-prediction of genes that can occur in many of its *de novo* counterparts.

OrthoFiller is mainly designed for use on genomes that have already undergone some basic level of annotation. As can be seen by comparing the 10% and 90% removal cases in the two data sets, application to very poorly annotated genomes can result in more genes of dubious quality, both from a sequence and an orthogroup perspective. It is worth noting that many of the genes with lower-quality scores, particularly those with only one of the scores being low, can be explained by alternate gene models (in the protein F-score) and shifting of orthogroups due to altered proteome sets (in the orthogroup F-score case). In all cases, in the absence of OrthoFiller filtration considerably higher numbers of genes were predicted that didn't resemble the genes that they were supposed to, indicating that they are erroneous. The results of OrthoFiller were robust to addition or removal of species. However, OrthoFiller performed poorly when only two species were used as input. It is therefore recommended that at least three species are used as input for OrthoFiller.

The OrthoFiller algorithm is designed to run on a Unix system with python and a minimal number of standard additional tools (HMMer, BedTools, Augustus, R, OrthoFinder). The software can be downloaded from https://github.com/mpdunne/orthofiller, where installation and implementation instructions can also be found. The speed of the algorithm is principally dependent on the speed of Augustus and HMMer, however processing time can be decreased by running OrthoFiller on multiple CPUs and thus parallelising these steps of the method.

Accurate and complete genome annotation is of paramount importance to the effective analysis of genomic and transcriptomic data, as well as for phylogenetic inference from genomic data. As the quantity of published genomes increases, care must be taken to ensure accuracy and quality of genome annotations are maintained. Automated methods that leverage publicly available information from multiple species to improve the annotation of newly sequenced genomes will help improve the accuracy and completeness of these resources and thus the quality of all analyses that utilise them.

Conclusions

OrthoFiller is an algorithm for improving the completeness of genome annotations. It leverages data from multiple species to identify conserved genes that have escaped detection, correct these detection errors and thus improve the genome completeness of all species under consideration.

Implementation

Data sources

For algorithm development and evaluation, a set of five well-annotated fungal genomes (Table 1) and a set of five well-annotated plant genomes (Table 2) were selected. Evaluation of the algorithm focussed on *S. cerevisiae* and *A. thaliana*, as the gene models in these genomes have historically been subject to extensive improvement and revision and are the most likely to be correct.

Table 9 SRA RNA-seq data coverage for novel genes in fungal genomes

Species	SRA ID	Instrument/details	Genes in original annotation			Novel genes		
			Total	W/ reads	%	Total	W/ reads	%
E. gossypii	*N/A*[a]	*N/A*	4768	*N/A*	*N/A*	*2*	*N/A*	*N/A*
D. hansenii	SRR1296968	Illumina HiSeq 2000, paired end	5781	6272	92.2%	13	13	100%
K. lactis	SRR1200528	Illumina Genome Analyzer II, single	5075	5076	100%	8	8	100%
S. cerevisiae	SRR539284	Illumina HiSeq 2000, paired end	6560	6572	99.8%	2	2	100%
Y. lipolytica	SRR868669	Illumina HiSeq 2000, single	6432	6447	99.8%	9	9	100%

[a]No publically available data found for this species/strain

Table 10 SRA RNA-seq data coverage for novel genes in plant genomes

Species	SRA ID	Instrument/details	Genes in original annotation			Novel genes		
			Total	W/ reads	%	Total	W/ reads	%
A. thaliana	SRR3932355	Illumina HiSeq 2500, paired end. Wild type Columbia rep1	27416	26110	95.2	116	89	76.7
B. rapa	SRR2984945	Illumina HiSeq 2000, paired end. ga-deficient dwarf (gad1-2) + GA rep2	40492	35793	88.4	10	4	40.0
C. papaya	SRR3509576	Illumina HiSeq 2500, paired end. SunUp/Sunset cultivar, young hermaphrodite leaf	27751	24589	88.6	382	327	85.6
C. rubella	SRR797557	Illumina Genome Analyzer IIx, paired end	26521	21239	80.1	228	137	60.1
T. cacao	SRR3217315	Illumina HiSeq 2000, paired end. Flower/leaf sample	29452	25758	87.5	94	69	73.4

Algorithm overview

OrthoFiller proceeds in five stages summarised in Fig. 1 and described in detail in the following sections. In brief, the algorithm begins by inferring a set of orthogroups from the protein coding genes of the set of species submitted to OrthoFiller (Fig. 1a). The protein sequences in these orthogroups are subject to multiple sequence alignment, converted to nucleotide sequences and used to build HMMs. These HMMs are used to search the genomes of each species under consideration (Fig. 1b) and the resultant HMM hits are subject to stringent filtering (Fig. 1c) before being used as hints for gene model construction (Fig. 1d). The gene models are subject to additional filtering (Fig. 1e) and only those gene models that pass all filters are added to the revised genome annotation. The revised genome annotations are then subject to orthogroup inference (Fig. 1f) and resultant orthogroups are analysed to confirm the identity of the newly predicted genes. The complete details for each step of this algorithm are described in the sections below.

Inference of Orthogroups and construction of HMMs

Orthogroups are inferred using OrthoFinder [10]. If a gene from the source annotation is not included in an orthogroup with at least one other sequence, it is classed as a *singleton*, and is not considered in downstream analyses. This is consistent with the problem definition of OrthoFiller, that is to identify unannotated genes that are conserved between species. Amino acid sequences from the orthogroups are aligned with MAFFT [19], using the L-INSI algorithm, and the resultant multiple sequence alignments are back-translated using the source nucleotide sequences. Aligning in this way, i.e. by amino acid sequence first, increases the robustness of the alignment under synonymous mutations and thus provides higher alignment quality [20]. The resulting nucleotide alignments are converted to Hidden Markov models (HMMs) using HMMer [21], each of which is then searched against each input genome in turn to generate a set of hits per HMM per species.

Evaluation of HMM search results

Due to the probabilistic nature of HMM searches, there is considerable variation in the quality of the relationship between a hit region and the set of sequences used to generate the source HMM. One expects a large amount of "background noise", that is sequence regions which pass the thresholds of the HMM but whose relevance is dubious. Each HMM hit has an associated bit score, an aggregated base-by-base similarity score between the hit and the aligned sequences used to generate it: we use this score to assess the quality of the hit. The bit score is strongly dependent on the hit length, thus to prevent gene length from biasing downstream analyses the bit score of a hit is divided by the hit length, to generate the *adjusted score* for a hit *h:*

$$score_{adj}(h) = \frac{score(h)}{length(h)}$$

The adjusted score is related to the e-value. However, the e-value calculation enforces a strict lower limit of $\sim 1 \times 10^{-300}$, all lower scores being rounded down to zero. Thus use of e-values would introduce irreversible length bias and would lead to downstream errors, as has been shown previously [10]. As bit scores do not have a threshold value, and they have been previously shown to be capable of facilitating accurate inference of phylogenetic trees [22], and length-corrected bit scores are used as the basis of the scoring scheme in OrthoFinder [10], they were used here.

For each species, a threshold value for hit acceptance or rejection based on a hit's adjusted score is created, by considering the distribution of hits which overlapped known genes. Anything above this threshold is considered to be genuine, and anything below this threshold is considered to be noise. An HMM hit is classed as *good* if it overlaps any gene from the orthogroup used to create the HMM, *bad* if it only overlaps genes from orthogroups other than the one used to create the HMM, and *candidate* if it overlaps no known gene at all. Here candidate hits are the potential new genes of interest, and the *good* and *bad*

genes are used to inform our judgement about the reliability of the candidate hits.

Distributions of adjusted scores for good and bad hits to the *S. cerevisiae* genome from all HMMs generated by the species in Table 1 are provided as an example in Additional file 3: Figure S3. Distributions for good and bad hits are clearly delineated into two distinct distributions. Note that in this case there are relatively few candidate hits, since the genome under inspection is already well annotated and is expected to have few missing gene predictions. Skew-t distributions are fit separately to the good and bad score distributions using *gamlss* [23]. Skew distributions were chosen because they allow flexibility in location, shape and scale of the underlying data and are commonly used for estimating parameters such as location and scale, while allowing the same distribution type to be used to fit both the good and bad hits. A separate skew-t distribution for the good and bad hits is fit for each species. In the event that there are insufficient good and bad hits to fit distributions, good and bad hits from the other species are aggregated and a threshold value is calculated from this.

For a given adjusted score x, the distributions of the *good* and *bad* hits are used to estimate both the absolute probabilities of a hit being genuine or being a mistake. We can estimate

$$P(genuine|x) = \frac{P(x|genuine)P(genuine)}{P(x|genuine)P(genuine) + P(x\ |mistake)P(mistake)}$$

$$P(mistake|x) = \frac{P(x|mistake)P(mistake)}{P(x|genuine)P(genuine) + P(x\ |mistake)P(mistake)}$$

and then retain the hit depending on whether it has a higher probability of being genuine than being a mistake, based on its adjusted score. The probabilities $P(genuine)$ and $P(mistake)$ are estimated by considering the proportion of good/bad hits which are good and bad respectively. The probability density functions $P(x|genuine)$ and $P(x|mistake)$ are determined using the fitted distributions as described above.

Acquisition and evaluation of putative predicted genes

Hits which survive the hit filtration step are passed to the gene-finding program Augustus [11] as *hints* specified as exon parts. Only predicted genes that have a non-zero overlap with these hints are retained. These predicted genes are then subjected to a *hint filter*, which aims to separate those genes which have genuinely arisen from the hint from those that overlap the hint by chance. The hint filter evaluates a *hint F-score* for each predicted gene, by comparing against the hints from a particular orthogroup which overlap it. The hint F-score is a measure of how well the found gene corresponds to the hints used to inform its discovery. Each predicted gene G will have at least one *hint region* corresponding to it, which is a set of non-overlapping coordinates obtained from merging all hints that overlap G, and which are all derived from the same orthogroup. For a hint region H and a predicted gene G, the hint F-score is defined as:

$$hf(H,\ G) = \frac{2 \cdot hP(H,G) \cdot hR(H,G)}{hR(H,G) + hP(H,G)}$$

where

$$hP(H,G) = \frac{|H \cap G|}{|H|};\ hR(H,G) = \frac{|H \cap G|}{|G|}$$

The filter uses a threshold hint F-score value of 0.8 (i.e. on average 80% of the length of the predicted gene is covered by the hit and vice versa), below which potential gene models are discarded. This value was chosen based on an analysis of hint F-scores of good and bad hits (as defined above) versus the Augustus output corresponding to them. Example distributions for hint F-scores for the good and bad hits can be seen in Additional file 4: Figure S4, in which it can be clearly seen that practically all genuine hints pass the threshold value of 0.8.

Once gene models have been filtered, they are fed once again into OrthoFinder, to cluster them into orthogroups. The orthogroup of each newly predicted gene is compared with the orthogroup(s) which were used to predict that gene. It is possible that multiple orthogroups informed the prediction of the same gene; similarly, there may be fluctuations in orthogroup membership between the original and new genomes. It is therefore only required that the new orthogroup into which the gene is clustered has non-zero overlap with at least one of the orthogroups used to predict it, and genes which do not fulfil this criterion are discarded.

Algorithm evaluation

Recovery of removed genes

The test set of species from Table 1 was used to analyse the effectiveness of OrthoFiller for genomes of various levels of completion. Altered versions of the *S. cerevisiae* genome annotation were constructed with 10 and 90% of genes randomly removed, and the level of recovery of the removed genes upon implementation of OrthoFiller was assessed, where a gene G was considered to be *recovered* if OrthoFiller predicted a gene such that G and G' have non-zero overlap.

The quality of the predicted genes was assessed by considering two scores: the orthogroup F-score and the protein F-score. The protein F-score is defined as

$$pF(S,S') = \frac{2 \cdot pP(S,S') \cdot pR(S,S')}{pR(S,S') + pP(S,S')}$$

where S is the original amino acid sequence and S' is the amino acid sequence of the recovered gene, and

$$pP(S,S') = \frac{|S \cap S'|}{|S|}; pR(S,S') = \frac{|S \cap S'|}{|S'|}$$

where the sequence length is the number of amino acids in the sequence, and the intersection length is defined to be the sum of identical amino acids in an alignment (MAFFT L-INSI) of the two sequences. The orthogroup F-score is defined as

$$oF(S,S') = \frac{2 \cdot oP(O,O') \cdot oR(O,O')}{oR(O,O') + oP(O,O')}$$

where O is the orthogroup that the gene is placed when no deductions have been made, O' is the orthogroup into which the gene is placed when OrthoFinder is run on the OrthoFiller results, and

$$oP(O,O') = \frac{|O \cap O'|}{|O|}; oR(O,O') = \frac{|O \cap O'|}{|O'|}$$

where cardinality of the orthogroups takes into account only genes which were present in the input set of genome annotations, i.e. not counting the newly discovered genes.

Evaluation of novel predicted genes

RNA-seq data was downloaded from the Sequence Read Archive, and aligned to the genome with HISAT2 [24] using default parameters. Coverage was calculated using BedTools coverage [25].

Availability and requirements

The software is available under the GPLv3 licence at https://github.com/mpdunne/orthofiller. All newly predicted genes and gene models are available for download from the Zenodo Research Data Repository doi.org/10.5281/zenodo.546090.

Additional files

Additional file 1: Figure S1. Recovery of removed genes from *S. cerevisiae* after 10% removal: Representation of removed genes at each stage, OrthoFiller vs. *de novo*. **A**) The number of deleted genes that obtained hits from one or more orthogroup HMMs. **B**) The number of deleted genes that had hits after OrthoFiller hint filtration. **C**) No hint filtration. **D**) The number of deleted genes for which a gene prediction was made using Augustus that satisfied OrthoFiller filtration tests. **E**) The number of deleted genes that for which a gene prediction was made using Augustus in the absence of OrthoFiller filtration. **F**) The number newly predicted genes that were retained or discarded based on the orthogroup assignment filter step in OrthoFiller. (TIF 562 kb)

Additional file 2: Figure S2. Recovery of removed genes from *S. cerevisiae* after 90% removal: Representation of removed genes at each stage, OrthoFiller vs. *de novo*. **A**) The number of deleted genes that obtained hits from one or more orthogroup HMMs. **B**) The number of deleted genes that had hits after OrthoFiller hint filtration. **C**) No hint filtration. **D**) The number of deleted genes for which a gene prediction was made using Augustus that satisfied OrthoFiller filtration tests. **E**) The number of deleted genes for which a gene prediction was made using Augustus in the absence of OrthoFiller filtration. **F**) The number newly predicted genes that were retained or discarded based on the orthogroup assignment filter step in OrthoFiller. (TIF 573 kb)

Additional file 3: Figure S3. hit score distributions for *good*, *bad* and *candidate* hits. Hits are to the *S. cerevisiae* genome, using HMMs from all orthogroups from the five fungal species described in Table 1. **A**) Length normalised bit scores of HMM hits to regions of the genome that contained genes that were not part of the orthogroup used to generate the HMM (bad hits). **B**) Length normalised bit scores of HMM hits to regions of the genome that do contain the gene used to generate the HMM (good hits). **C**) Length normalised bit scores of HMM hits to regions of the genome that do not contain any previously annotated genes (candidate novel gene hits). **D**) All distributions overlaid. (TIF 143 kb)

Additional file 4: Figure S4. Example distribution of hint F-scores for good vs. bad hints. Here, Augustus has been allowed to predict genes that are already present in the input genome, hence we can consider separately the good and bad hits as hints. Hits and hints are from running OrthoFiller on the five fungal species described in Table 1. Shown are the distributions of hint F-scores for good (green) and bad (red) hits respectively, demonstrating that practically all of the genuine hints have a hint F-score of 0.8 or higher. (TIF 1424 kb)

Abbreviations

GTF: General transfer format; HMM: Hidden Markov model; oF: orthogroup F-Score; pF: protein F-score

Acknowledgements

Thanks to Bastian Greshake for assisting with early testing of the software.

Funding

SK is a Royal Society University Research Fellow. This work was supported by the European Union's Horizon 2020 research and innovation programme under grant agreement number 637765. MPD is supported by an EPSRC studentship through EP/G03706X/1. The funders had no role in the design of the study or the collection, analysis and interpretation of the data, or in writing the manuscript.

Author's contributions

SK conceived the project. MPD developed the algorithm. SK and MPD analysed the data and wrote the manuscript. Both authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Consent for publication

Not applicable.

Ethics approval and consent to participate

Not applicable.

Article Note

The original version of this article was revised. In Table 1, Saccharomyces and Kluyveromyces were spelt incorrectly. In Table 4 of the original article, the standard deviation was not indicated in brackets, while the Table 4 legend was incorrect. These have all been corrected in the updated version.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Received: 5 December 2016 Accepted: 8 May 2017
Published online: 18 May 2017

References

1. Hayden EC. The $1,000 genome. Nature. 2014;507:295.
2. Wetterstrand, KA. DNA sequencing costs: data from the NHGRI Genome sequencing program (GSP). 2016. URL: http://www.genome.gov/sequencingcosts.

3. Yandell, Mark, and Daniel Ence. A beginner's guide to eukaryotic genome annotation. Nature Reviews Genetics. 2012;13.5:329–342.
4. Denton JF, Lugo-Martinez J, Tucker AE, Schrider DR, Warren WC, Hahn MW. Extensive Error in the Number of Genes Inferred from Draft Genome Assemblies. PLoS Comput Biol. 2014;10:12.
5. Veeckman E, Ruttink T, Vandepoele K, Are we there yet? Reliably estimating the completeness of plant genome sequences. Plant Cell. 2016;28:1759–68.
6. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. Genome analysis BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015;31(19):3210–2.
7. Parra G, Bradnam K, Korf I. Genome analysis CEGMA : a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics. 2007;23(9):1061–7.
8. Koonin EV, Galperin MY. Genome Annotation and Analysis. In: Sequence - Evolution - Function: Computational Approaches in Comparative Genomics. Boston: Kluwer Academic; 2003. Chapter 5.
9. van der Burgt A, Severing E, Collemare J, de Wit P. Automated alignment-based curation of gene models in filamentous fungi. BMC Bioinformatics. 2014;15(1):19.
10. Emms DM, Kelly S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biol. 2015;16(1):157.
11. Stanke M, Morgenstern B. AUGUSTUS: A web server for gene prediction in eukaryotes that allows user-defined constraints. Nucleic Acids Res. 2005;33 Suppl 2:465–7.
12. Dietrich FS, Voegeli S, Brachat S, Lerch A, Gates K, Steiner S, Mohr C, Pöhlmann R, Luedi P, Choi S, Wing RA, Flavier A, Gaffney TD, Philippsen P. The Ashbya gossypii genome as a tool for mapping the ancient Saccharomyces cerevisiae genome. Science (80-). 2004;304:304–7
13. Dujon B, Sherman D, Fischer G, Durrens P, Casaregola S, Lafontaine I, De Montigny J, Blanchin S, Beckerich J-M, Beyne E, Bleykasten C, Babour A, Boyer J, Cattolico L, Confanioleri F, De Daruvar A, Despons L, Fabre E, De Montigny J, Marck C, Neuvéglise C, Talla E, Goffard N, Frangeul L, Aigle M, Anthouard V, Babour A, Barbe V, Barnay S, Blanchin S, Beckerich J-M, Beyne E, Bleykasten C, Boisramé A, Boyer J, Cattolico L, Confanioleri F, De Daruvar A, Despons L, Fabre E, Fairhead C, Ferry-Dumazet H, Groppi A, Hantraye F, Hennequin C, Jauniaux N, Joyet P, Kachouri R, Kerrest A, Koszul R, Lemaire M, Lesur I, Ma L, Muller H, Nicaud J-M, Nikolski M, Oztas S, Ozier-Kalogeropoulos O, Pellenz S, Potier S, Richard G-F, Straub M-L, Suleau A, Swennen D, Tekaia F, Wésolowski-Louvel M, Westhof E, Wirth B, Zeniou-Meyer M, Zivanovic I, Bolotin-Fukuhara M, Thierry A, Bouchier C, Caudron B, Scarpelli C, Gaillardin C, Weissenbach J, Wincker P, Souciet J-L. Genome evolution in yeasts. Nature. 2004;430(6995):35–44.
14. Sacerdot C, Casaregola S, Lafontaine I, Tekaia F, Dujon B, Ozier-kalogeropoulos O. Promiscuous DNA in the nuclear genomes of hemiascomycetous yeasts. FEMS Yeast Res. 2008;8:846–57.
15. Goodstein DM, Shu S, Howson R, Neupane R, Hayes RD, Fazo J, Mitros T, Dirks W, Hellsten U, Putnam N, Rokhsar DS. Phytozome: A comparative platform for green plant genomics. Nucleic Acids Res. 2012;40(D1):1178–86.
16. Leinonen R, Sugawara H, Shumway M. The sequence read archive. Nucleic Acids Res. 2011;39.
17. Yandell M, Ence D. A beginner's guide to eukaryotic genome annotation. Nat Rev Genet. 2012;13(May):329–42.
18. Neuvéglise C, Devillers H. Genome Resources for Yeast Chromosomes. [Online]. Available: http://gryc.inra.fr/. Accessed 06 Apr 2017.
19. Katoh K, Standley DM. MAFFT Multiple Sequence Alignment Software Version 7 : Improvements in Performance and Usability Article Fast Track. Mol Biol Evol. 2013;30(4):772–80.
20. Wernersson R, Pedersen AG. RevTrans: multiple alignment of coding DNA from aligned amino acid sequences. Nucleic Acids Res. 2003; 31(13):3537–9.
21. Finn RD, Clements J, Eddy SR. HMMER web server: interactive sequence similarity searching. Nucleic Acids Res. 2011;39:W29–W37.
22. Huson DH, Scornavacca C. Dendroscope 3: an interactive tool for rooted phylogenetic trees and networks. BMC Bioinformatics. 2012;61(6):1061–7.
23. Stasinopoulos DM, Rigby RA. Generalized additive models for location scale and shape (GAMLSS) in R. Appl. Statist. 2005;54(3):507–554.
24. Daehwan K, Langmead B, Salzberg S. HISAT: a fast spliced aligner with low memory requirements. Nature. 2015;12(4):357–360.
25. Quinlan AR. BEDTools: a flexible suite of utilities for comparing genomic features. 201;26(6):841–842.
26. Gnerre S, Maccallum I, Przybylski D, Ribeiro FJ, Burton JN, Walker BJ, Williams L, Nicol R, Gnirke A, Nusbaum C, Lander ES, Jaffe DB, Sharpe T, Hall G, Shea TP, Sykes S, Berlin AM, Aird D, Costello M, Daza R, Williams L, Nicol R, Gnirke A, Nusbaum C, Lander ES, Jaffe DB. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. Proc Natl Acad Sci USA. 2011;108(4):1513–8.
27. Galagan JE, Calvo SE, Cuomo C, Ma L, Wortman JR, Batzoglou S, Spevak CC, Clutterbuck J, Kapitonov V, Jurka J, Scazzocchio C, Farman M, Butler J, Purcell S, Harris S, Braus GH, Draht O, Busch S, Enfert CD, Bouchier C, Goldman GH, Denning DW, Caddick M, Hynes M, Paoletti M, Fischer R, Miller B. Sequencing of Aspergillus nidulans and comparative analysis with A. fumigatus and A. oryzae. Nature. 2005; 438(December):1105–15.
28. Wood V, Gwilliam R, Rajandream M, Lyne M, Lyne R, Stewart A, Sgouros J, Peat N, Hayles J, Baker S, Basham D, Bowman S, Brooks K, Brown D, Brown S, Chillingworth T, Churcher C, Collins M, Connor R, Cronin A, Davis P, Feltwell T, Fraser A, Gentles S, Goble A, Hamlin N, Harris D, Hidalgo J, Hodgson G, Holroyd S, Hornsby T, Howarth S, Huckle EJ, Hunt S, Jagels K, James K, Jones L, Jones M, Leather S, Mcdonald S, Mclean J, Mooney P, Moule S, Mungall K, Murphy L, Niblett D, Odell C, Oliver K, Neil SO, Pearson D, Quail MA, Rabbinowitsch E, Rutherford K, Rutter S, Saunders D, Seeger K, Sharp S, Skelton J, Simmonds M, Squares R, Squares S, Stevens K, Taylor K, Taylor RG, Tivey A, Walsh S, Warren T, Whitehead S, Woodward J, Volckaert G, Aert R, Robben J, Grymonprez B, Weltjens I. The genome sequence of Schizosaccharomyces pombe. Nature. 2002;415(6874):871–80.
29. Wood V, Harris MA, Mcdowall MD, Rutherford K, Vaughan BW, Staines DM, Aslett M, Lock A, Kersey PJ, Oliver SG. PomBase: a comprehensive online resource for fission yeast. Nucleic Acids Res. 2012;40(October 2011):695–9.
30. Génolevures Consortium. Comparative genomics of protoploid Saccharomycetaceae. 2009. p. 1696–709.
31. Liu S, Liu Y, Yang X, Tong C, Edwards D, Parkin IAP, Zhao M, Ma J, Yu J, Huang S, Wang X, Wang J, Lu K, Fang Z, Bancroft I, Yang T, Hu Q, Wang X, Yue Z, Li H, Yang L, Wu J, Zhou Q, Wang W, King GJ, Pires JC, Lu C, Wu Z, Sampath P, Wang Z, Guo H, Pan S, Yang L, Min J, Zhang D, Jin D, Li W, Belcram H, Tu J, Guan M, Qi C, Du D, Li J, Jiang L, Batley J, Sharpe AG, Park B, Ruperao P, Cheng F, Waminal NE, Huang Y, Dong C, Wang L, Li J, Hu Z, Zhuang M, Huang Y, Huang J, Shi J, Mei D, Liu J, Lee T, Wang J, Tang X, Liu W, Wang Y, Zhang Y, Lee J, Kim HH, Denoeud F, Xu X, Liang X, Hua W, Wang X, Wang J, Chalhoub B, Paterson AH. The Brassica oleracea genome reveals the asymmetrical evolution of polyploid genomes. Nat Commun. 2014;5(May):1–11.
32. Wu GA, Prochnik S, Jenkins J, Salse J, Hellsten U, Murat F, Perrier X, Ruiz M, Scalabrin S, Terol J. Sequencing of diverse mandarin, pummelo and orange genomes reveals complex history of admixture during citrus domestication. Biotechnol, Nat. 2015;32(7):656–62.
33. Tuskan G, Difazio S, Jansson S, Bohlmann J, Grigoriev I, Hellsten U, Putnam N, Ralph S, Rombauts S, Salamov A, Schein J, Sterck L, Aerts A, Bhalerao R, Bhalerao R, Blaudez D, Boerjan W, Brun A, Brunner A, Busov V, Campbell M, Carlson J, Chalot M, Chapman J, Chen G, Cooper D, Coutinho P, Couturier J, Covert S, Cronk Q, Cunningham R, Davis J, Degroeve S, Déjardin A, Depamphilis C, Detter J, Dirks B, Dubchak I, Duplessis S, Ehlting J, Ellis B, Gendler K, Goodstein D, Gribskov M, Grimwood J, Groover A, Gunter L, Hamberger B, Heinze B, Helariutta Y, Henrissat B, Holligan D, Holt R, Huang W, Islam-Faridi N, Jones S, Jones-Rhoades M, Jorgensen R, Joshi C, Kangasjärvi J, Karlsson J, Kelleher C, Kirkpatrick R, Kirst M, Kohler A, Kalluri U, Larimer F, Leebens-Mack J, Leplé J, Locascio P, Lou Y, Lucas S, Martin F, Montanini B, Napoli C, Nelson D, Nelson C, Nieminen K, Nilsson O, Pereda V, Peter G, Philippe R, Pilate G, Poliakov A, Razumovskaya J, Richardson P, Rinaldi C, Ritland K, Rouzé P, Ryaboy D, Schmutz J, Schrader J, Segerman B, Shin H, Siddiqui A, Sterky F, Terry A, Tsai C, Uberbacher E, Unneberg P, Vahala J, Wall K, Wessler S, Yang G, Yin T, Douglas C, Marra M, Sandberg G, Van Y, Peer de, Rokhsar D. The Genome of Black Cottonwood. Science. 2006;313(September):1596–1605.

B

OMGene (published version)

This appendix contains the original published manuscript (Dunne and Kelly 2018) of the work contained in Chapter 5, which was accepted for publication in BMC Genomics on 27th April 2018.

Dunne and Kelly *BMC Genomics* (2018) 19:307
https://doi.org/10.1186/s12864-018-4704-z

BMC Genomics

SOFTWARE Open Access

OMGene: mutual improvement of gene models through optimisation of evolutionary conservation

Michael P. Dunne and Steven Kelly*

Abstract

Background: The accurate determination of the genomic coordinates for a given gene – its *gene model* – is of vital importance to the utility of its annotation, and the accuracy of bioinformatic analyses derived from it. Currently-available methods of computational gene prediction, while on the whole successful, frequently disagree on the model for a given predicted gene, with some or all of the variant gene models often failing to match the biologically observed structure. Many prediction methods can be bolstered by using experimental data such as RNA-seq. However, these resources are not always available, and rarely give a comprehensive portrait of an organism's transcriptome due to temporal and tissue-specific expression profiles.

Results: Orthology between genes provides evolutionary evidence to guide the construction of gene models. OMGene (Optimise My Gene) aims to improve gene model accuracy in the absence of experimental data by optimising the consistency of multiple sequence alignments of orthologous genes from multiple species. Using RNA-seq data sets from plants, mammals, and fungi, considering intron/exon junction representation and exon coverage, and assessing the intra-orthogroup consistency of subcellular localisation predictions, we demonstrate the utility of OMGene for improving gene models in annotated genomes.

Conclusions: We show that significant improvements in the accuracy of gene model annotations can be made, both in established and in *de novo* annotated genomes, by leveraging information from multiple species.

Keywords: Genome annotation, Annotation errors, Orthogroups, Orthology, Gene model

Background

The utility of any given genome is dependent on the comprehensiveness and accuracy of its gene model - annotations. Inaccuracies in the annotated locations and structures of protein coding genes can lead to myriad downstream errors. These include misinformed conclusions about the biological properties of an organism, as well as errors in transcript quantification, phylogenetic tree inference, protein localisation, and protein structure predictions. It is therefore vital to downstream analysis, both computational and experimental, to ensure that gene annotations are as accurate as possible.

The absolute quantity of publicly available genomic data has grown exponentially over the past two decades, as has the number of taxa represented [1–3], owing to the consistently decreasing costs of acquiring whole genome sequences [4, 5]. Accordingly, the feasibility of manual gene model annotation has diminished progressively, with a corresponding increase in reliance on computational gene prediction software. As such there are numerous tools available for the *de novo* and data-assisted prediction of genes [6]. These tools typically rely on genetic signatures such as GC content, codon bias, feature length distributions, and various conserved DNA sequence motifs. Though many of these tools are highly proficient at gene prediction, mistakes are common. Gene prediction tools often disagree on the quantity of genes that they predict [7–9]. Furthermore, even when gene predictors agree on the location of a gene, the predicted intron-exon structure for that gene can vary considerably between the different methods [10]. Common errors include erroneous exon/intron retention/omission, inaccurate exon/intron

* Correspondence: steven.kelly@plants.ox.ac.uk
Department of Plant Sciences, University of Oxford, South Parks Road, Oxford OX1 3RB, UK

© The Author(s). 2018 **Open Access** This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated.

boundaries, frame errors, misplaced start codons, and fragmentation/fusion of gene models.

When available, the use of extrinsic empirical data, most notably RNA-seq, is the most reliable currently available method for procuring gene models. For example, single contiguous RNA-seq reads obtained from mRNA sequencing can be split across multiple loci when mapped to the genome, providing evidence for the locations of splice junctions. Unfortunately, empirical data is generally not available for all genes in a given species: many genes are expressed in a cell-type or cell-cycle specific manner and for organisms with many disparate tissue types it can be difficult to obtain RNA-seq data that cover the full breadth of the transcriptome [11, 12]. In addition, not all gene sequences are amenable to reliable and accurate alignment, in particular identical duplicate genes and genes that contain repetitive regions found in multiple other genes [13]. Furthermore library preparation protocols and other statistical factors can make reliable gene model inferences difficult [14–16]. Finally, there are some aspects of gene models that are simply not revealed by RNA-seq analysis: for example the presence of 5'UTR sequences or internal methionine residues mean that there can often be multiple plausible start codons locations for a given open reading frame (ORF), and so start codon location cannot be inferred from RNA-seq data.

Feature locations (splice sites, exons, transcription start sites) have been shown to be highly conserved across evolutionary timescales, often more so than the constituent amino acid sequences they encapsulate [17, 18], despite alternative splicing being a driver of divergence [19]. Given various gene model predictions, if multiple highly similar (in sequence and structure) gene models exist for a gene across multiple taxa, they are more likely to be biologically correct than disparate alternatives. Several de novo gene prediction algorithms have utilised this concept to constrain gene searches by predicting genes in multiple genomes simultaneously: notably SLAM [20], SGP-1 [21], TWINSCAN/N-SCAN/CONTRAST [22, 23], and the most recent version of Augustus [24]. However, no tool currently exists that can systematically and automatically improve the annotations of already predicted genes by leveraging annotated gene models from other species.

By considering *orthogroups* of related genes, one can optimise the similarity of gene models across species by seeking conserved structure across the various taxa. In the absence of extrinsic data, it is parsimonious to choose gene models that maximise intra-orthogroup amino acid alignment agreement. OMGene (Optimise My Gene) aims to improve genome annotations by optimising the agreement between gene models for orthologous genes in multiple species. It is designed to function without the need for additional empirical data, utilising only the local genome sequences for the genes in question, and works on existing predicted gene models. A standalone implementation of the algorithm is available under the GPLv3 licence at https://github.com/mpdunne/omgene. The algorithm is available as a python script, instructions for which, along with example data sets, are included in the git repository.

Implementation

Algorithm description

The input for OMGene is a set of gene model files in general transfer format (GTF) and a set of corresponding FASTA genome files. GTF is a variant of GFF3 with a more standardised structure. As many public gene datasets use GFF3 format, a tool for extracting well-formed gene entries from GFF3 files is included in the GitHub repository. Each GTF file should contain coordinates for one gene only. The FASTA file can contain one or multiple sequences but must contain the nucleotide sequence referenced by the GTF file. The set of GTF files and their corresponding genome files are inputted to OMGene using a tab-delimited CSV file: each line in the input should contain first the file path to a gene model's GTF file, followed by the path to the genome for that species. If the GTF contains multiple transcript variants then these are considered together as variants of a single gene. Further information regarding input formats can be found in the OMGene GitHub repository.

For each inputted gene, the algorithm defines its *gene region* to be the region spanning the first and last base of any of its corresponding gene models, with a user-selected number of buffer bases either side (default value is 600 bp). The initial step of OMGene is to cross-align the amino acid sequences from each gene with the gene regions of the other genes using Exonerate [25]. The rationale behind this step is to find exonic regions that are present in one or more gene models but absent from one or more annotated gene model. This is performed three times: first by cross-aligning the input protein sequences against all gene regions, second by cross-aligning the protein sequences that have been found in the first step against all gene regions, and finally by cross-aligning all individual exon sequences from the first step. This three-step process mitigates against lack of detection due to gene model errors in one or more of the input genes. This, together with the exons from the original gene sequences, comprises a set of potential gene parts, which may overlap and which may be incompatible in reading frame. Compatible combinations of gene parts (i.e. without frame-shift errors) are strung together to form a putative gene model. Many such putative gene models may exist: the set of putative gene models with the highest alignment score (see Alignment

score calculation below) is carried forward to the next step. If multiple alternatives have the same alignment score, the choice is made randomly.

The set of putative gene models from the previous step are aligned, and the set of putative exons from all genes is divided into *adjacency groups*: sets of exons that overlap each other in the alignment (see below). Exons are added in sequentially in these adjacency groups, and at each stage a valid gene model is sought on the left hand side of the gene (i.e. starting at the start codon and seeking to adjoin exons in valid donor-acceptor pairs). Multiple options for each gene are produced at each new junction, by recursively seeking out, or "wiggling" splice junctions (or start codons) in each frame either side of the existing exons start and end points. This produces a set of junction options for each pair of exon ends. A multipartite choice function is then used to choose the optimal set of exons across all secies, as described below. In the event that a particular exon is very small (< 40 bp), or does not yield any valid junction sites, both that exon and the one before it are probed for removal, and the variant with the removed exon is compared against the other partial gene models in the evaluation step. Once this recursive step ceases to produce new gene modes, the gene model set with the highest alignment score is declared the winner, and the next putative exon from the next adjacency group is added. If multiple alternatives have the same alignment score, the choice is made randomly. This is repeated until there are no further exons to add.

To ensure that the optimisation process did not overlook potentially better variants in the user-supplied gene models, the process above is repeated. This time, instead of varying exons start and end sites, the set of newly created junctions are compared against the original junctions, aiming to find the optimal combination of new and old junctions.

The final step involves filtering the changes based on a selection of categories that have been observed to over-fix gene models. Firstly, we require the alignment score α of a 10 amino acid region each side of the change to have either remained the same or improved. This is a basic requirement which should be met in most cases due to the way in which sequence variants are chosen. Secondly, changes that have opened gaps in the alignment in three or more of the sequences are not allowed: this is a common occurrence due to sequences proximal to exon termini that by chance feature valid splice junction sequences that are in frame with the adjacent exons and are evolutionarily conserved. These tend not to be correct. Thirdly, very small changes are forbidden: changes that have resulted in two or fewer amino acids being changed in a gapless region of the alignment, such that the new alignment is also gapless, are ignored. Similar changes to larger regions require an α (see Alignment score below) increase of 4 or more. This is to avoid changes that reflect multiple choices of donor-acceptor pairs for essentially identical sequences. Finally, the alignment in the region of the change must be of reasonable quality: for unchanged 5 amino acid regions either side of the region under consideration, the adjusted alignment score $\overline{\alpha}$ must be 3 or higher (or all gaps) for some subset of three sequences containing the sequence of interest. Similarly the resulting score for the changed region must also be higher than 3 or all gaps. Exon boundaries that do not pass the filters are discarded and the genes are reconstructed a final time, allowing only the surviving boundaries and those that were present in the original gene. The resultant genes are outputted in GTF, amino acid FASTA and CDS FASTA format.

Data sources

For algorithm development and evaluation, three species sets were selected (Table 1): a set of five fungal genomes, a set of five plant genomes, and a set of five mammal genomes. Each species set contained at least one well-annotated model organism. Orthogroups were inferred using OrthoFinder [26]. For the plant data set, where multiple transcript variants were available, the primary transcript was used as listed in Phytozome [27]. For the mammal data set, no primary transcripts were marked, and so for each gene the longest available transcript was used. RNA-seq data sources are listed in Table 2. and were downloaded from the Sequence Read Archive [28].

De novo gene prediction

De novo gene predictions were made using Augustus [24] version 3.2.2. Training was performed using all well-formed gene models from each species, and using the autoAugTrain.pl script included with the software. Augustus was run individually on each genome with the default settings.

Alignment score

An amino acid alignment can be considered as an ordered sequence $A = (C_n)_{n=1}^{n=l}$ of columns $C_n = (c_1^n, \ldots, c_l^n)$. The *column score* γ for a column C_n is defined as the average pairwise Blosum62 score for amino acids in that column:

$$\gamma(C_n) = \frac{\sum_{1 \le i < j \le l} Blos\left(c_i^n, c_j^n\right)}{l}$$

The Blosum62 matrix was used as it is the most widely used amino acid substitution matrix. The *alignment score* α for an alignment A is constructed column-wise as:

$$\alpha(A) = \sum_{n=1}^{l} \gamma(C_n)$$

The *adjusted alignment score* $\overline{\alpha}$ is defined as $\overline{\alpha} = \frac{\alpha}{l}$, where l is the alignment length.

Table 1 Species sets used for algorithm validation

	Species Name	Source	Version/Strain	Taxonomy ID	References
Plant species	*Arabidopsis thaliana*	JGI[1]	TAIR10	3702	[39, 40]
	Brassica rapa	JGI	v1.3	3711	[27]
	Carica papaya	JGI	ASGPBv0.4	3649	[39, 41]
	Capsella rubella	JGI	v1.0	81,985	[39, 42]
	Theobroma cacao	JGI	v1.1	3641	[39, 43]
Mammalian species	*Canis lupus familiaris*	NCBI	CanFam3.1/AR105	6915	[44, 45]
	Homo sapiens	CCDS[3]	GRCh38.p7/CCDS20	9606	[46]
	Monodelphis domestica	NCBI	MonDom5 / AR103	13,616	[45, 47]
	Mus musculus	CCDS	GRCm38.p4/CCDS21	10,090	[46]
	Oryctolagus cuniculus	NCBI	OryCun2.0/AR102	568,996	[45, 48]
Fungal species	*Eremothecium gossypii*	JG	*ATCC10895*	284,811	[49]
	Debaryomyces hansenii	JGI	*CBS767*	284,592	[50, 51]
	Kluyveromyces lactis	JGI	*CLIB210*	284,590	[50]
	Saccharomyces cerevisiae	SGD[4]	*S288C*	559,292	[52]
	Yarrowia lipolytica	JGI	*CLIB122*	284,591	[50]

[1]*Joint Genome Institute;* [2]*National Centre for Biotechnology Information;* [3]*Consensus Coding Sequence Project;* [4]*Saccharomyces Genome Database*

Table 2 SRA RNA-seq data sources

	Species	SRA ID	Instrument/details
Plants	*A. thaliana*	SRR3932355	Illumina HiSeq 2500, paired end. Wild type Columbia
	B. rapa	SRR2984945	Illumina HiSeq 2000, paired end. ga-deficient dwarf (gad1–2) + GA rep2
	C. papaya	SRR3509576	Illumina HiSeq 2500, paired end. SunUp/Sunset cultivar, young hermaphrodite leaf
	C. rubella	SRR3993756	Illumina HiSeq 2000, paired end. Leaf sample
	T. cacao	SRR3217315	Illumina HiSeq 2000, paired end. Flower/leaf sample
Mammals	*C. lupus*	ERR266386	Illumina Genome Analyzer II, paired end, brain frontal cortex, male
		ERR266355	Illumina Genome Analyzer II, paired end, brain frontal cortex, female
		ERR266382	Illumina Genome Analyzer II, paired end, brain frontal cortex, male
	H. sapiens	SRR5938455	Illumina HiSeq 2000, paired end, dorsolateral prefrontal cortex, male
	M. domestica	SRR500906	Illumina HiSeq 2000, paired end, brain
		SRR500925	Illumina HiSeq 2000, paired end, brain
	M. musculus	SRR5441717	Illumina HiSeq 2000, paired end, brain (striatum)
		SRR6269591	Illumina NovaSeq 6000, paired end, cerebellum
	O. cuniculus	ERR266399	Illumina Genome Analyzer II, paired end, brain frontal cortex, female
		SRR400990	Illumina Genome Analyzer II, paired end, brain frontal cortex
		SRR401040	Illumina Genome Analyzer II, paired end, brain frontal cortex
		SRR401041	Illumina Genome Analyzer II, paired end, brain frontal cortex
		SRR401042	Illumina Genome Analyzer II, paired end, brain frontal cortex
Fungi	*K. lactis*	SRR1200528	Illumina Genome Analyzer II, single
	S. cerevisiae	SRR539284	Illumina HiSeq 2000, paired end
	Y. lipolytica	SRR868669	Illumina HiSeq 2000, single

Multipartite choice function

The multipartite choice function (Fig. 1) aims, for a set of k gene regions and a set of l_k gene model variants for each gene region, to choose an optimal set containing one gene model variant from each gene region such that the alignment score is maximised. This problem is equivalent to finding the heaviest maximal clique in an edge-weighted complete multipartite graph. This function replaces the naïve approach of calculating pairwise alignment scores between all options and choosing an optimal subset, which, while optimal, is computationally infeasible.

To reduce the complexity of the problem, options are chosen by comparison with a reference consensus alignment, produced by taking the most consistent set of amino acids for each column in a global alignment individually (Fig. 1a-b). This column-wise optimisation is fast, and provides a basis for the sequence-wide optimisation. To produce the consensus, The set of Σl_k options is aligned to the reference (the original alignment) using MAFFT –add [29]. The inconsistent regions are then isolated and re-aligned using the more accurate but more computationally intensive MAFFT L-INS-i. For each column in the alignment, the set of amino acid choices (one for each gene region) that optimises the alignment score for that column is chosen.

For each option i a binary string $H_i = \{h_1^i, \ldots, h_n^i\}$ is produced describing for each position in the alignment whether or not that option matches the consensus (Fig. 1c). The chosen subset will be the set of options that globally maximises agreement with the consensus. If the strings $\{H_i\}_i$ are stacked vertically, such that they can be read as columns $\{V_j\}_{j=1}^n$ then the task is equivalent to finding a columnar binary string V with one nonzero entry for each gene region such that $|V_i : V \quad V_i|$ is maximised.

Given the set $A_0 = \{V_j\}_{j=1}^n$, an optimal subset is deduced by sequential random sampling. Ignoring all-1 strings, an initial $W_0 = V_k$ is chosen at random from A_0. For sets S_1, S_2 and a set of "checkpoints" R, the set S_1 is *compatible with* S_2*with respect to* $R = \{R_i\}_i$ if the binary intersection $S_1 \cap S_2 \cap R_i$ is nonzero for all i. Define $A_n = \{a \cap W_{n-1} : a, W_{n-1}$ compatible w. r. t $G\}$, where G is the set of binary strings which are zero for all but one gene region, at each stage choosing W_n at random from A_n. The process A_0, A_1, A_2, … eventually converges on a single binary string. This reduction is performed a user-selected number of times, the default being 1000. The result that is a subset of the largest number of V_i is declared the winner. In the event that the result still contains more than one option for each gene region, subsets of options are calculated and their multiple alignment score α is calculated, the winner being the subset with the

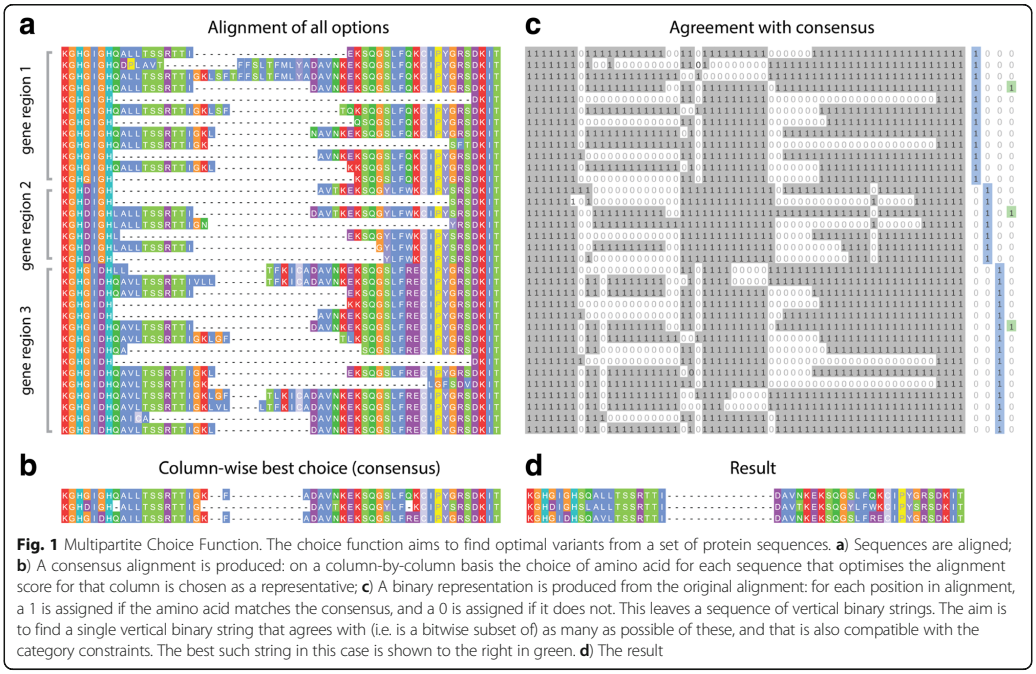


Fig. 1 Multipartite Choice Function. The choice function aims to find optimal variants from a set of protein sequences. **a**) Sequences are aligned; **b**) A consensus alignment is produced: on a column-by-column basis the choice of amino acid for each sequence that optimises the alignment score for that column is chosen as a representative; **c**) A binary representation is produced from the original alignment: for each position in alignment, a 1 is assigned if the amino acid matches the consensus, and a 0 is assigned if it does not. This leaves a sequence of vertical binary strings. The aim is to find a single vertical binary string that agrees with (i.e. is a bitwise subset of) as many as possible of these, and that is also compatible with the category constraints. The best such string in this case is shown to the right in green. **d**) The result

highest α. In the event that multiple subsets exhibit the same maximal α, a subset is chosen arbitrarily from them.

Adjacency group calculation

OMGene builds genes sequentially by iteratively adding in putative exons to multiple genes simultaneously. Care must be taken to ensure the gene parts (which in turn become exons once gene models are constructed) are added in a way conducive to vertical comparison of relevant regions (see Fig. 2). In OMGene, gene parts are considered in sequential *adjacency groups* based on their coordinates in a multiple sequence alignment. Prototype gene models are formed by stringing together amino acid sequences for individual putative exons for each gene region: these are then aligned, and a graph is formed from this alignment. Each putative exon is a node on the graph, and two exons are connected by an edge if and only if one of the exons overlaps the other by a third or more of its length. The adjacency groups are then defined to be cliques in this graph. Cliques are determined using the python implementation of the NetworkX package [30].

Junction F-score

The *junction F-score* for a gene is a measure of how well the splice junctions observed in mapped RNA-seq data are represented in the gene model. For a gene model G and corresponding gene region R, define J_G to be the set of individual intron beginning and end coordinates in the gene model, and define J_R to be the set of mapping junction beginning and end coordinates in the mapped RNA-seq data. A minimum of 10 reads is required for a given RNA-seq junction to be counted. We may then define the junction F-score as:

$$jF(J_G, J_R) = \frac{2 \cdot jP(J_G, J_R) \cdot jR(J_G, J_R)}{jR(J_G, J_R) + jP(J_G, J_R)}$$

where

$$jP(J_G, J_R) = \frac{|J_G \cap J_R|}{|J_R|}; \quad jR(J_G, J_R) = \frac{|J_G \cap J_R|}{|J_G|}$$

The direction of each junction site (start or end of a junction) is taken into account when considering the intersection of the two sets.

Coverage score

The *coverage score* is a measure of how well RNA-seq data represents a given gene. Given that gene expression levels can vary considerably and irregularly across the length of a transcript [13–16], care must be taken to ensure the expression profile for a gene region is properly interpreted. For example, sample preparation methods can bias coverage towards the centre and 3′ ends of the transcript; furthermore, jagged read profiles and transcription of antisense regions [31] and other intronic ncRNAs can cause expression profiles to be highly non-binary. To mitigate this, a rolling threshold

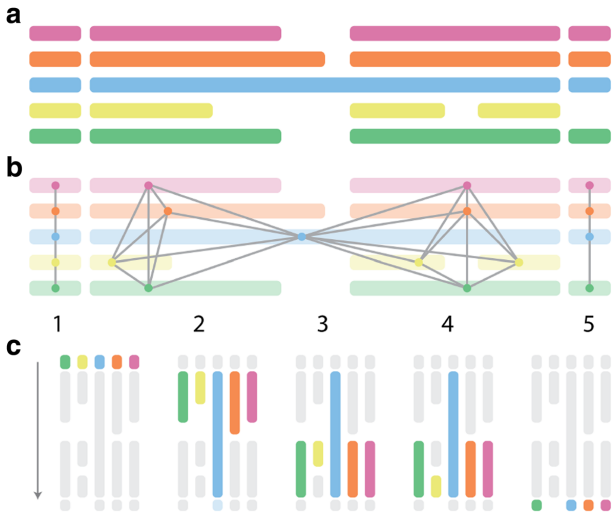


Fig. 2 Calculation of adjacency groups. **a**) Amino acid sequences for individual putative exons are strung together and aligned. **b**) A graph is formed with vertices formed by gene parts (or exons), and edges drawn when the overlap between two parts is greater than or equal to one third the length of one of them. **c**) Cliques are extracted and then ordered lexicographically to form the adjacency groups

approach is used. For a gene region R, and a genomic coordinate $x \in R$, the expression characteristic χ is defined as:

$$\chi(x) = \min(\max(\{\rho(y) : y \in R, y < x\}), \max(\{\rho(y) : y \in R, y > x\}))$$

Where $\rho(y)$ is the read count at genomic coordinate y. Bases in the gene region to which the RNA-seq data has been mapped are categorised based on whether they are likely to correspond to exonic or non-exonic regions: a base x is considered to be exonic or *on* (i.e. likely included in the mature mRNA) if $\rho(x) > \frac{\chi(x)}{5}$, and intronic or *off* (i.e. likely not included in the mature mRNA) if $\rho(x) < \frac{\chi(x)}{5}$. The coverage score for a gene model $G = \{G_1, \ldots, G_n\}$, where the G_i are alternately exons and introns, is defined as:

$$C(G) = \frac{1}{n}\left(\sum_{G_i \text{ exonic}} \frac{|\{x \in G_i : x \text{ on}\}|}{|G_i|} + \sum_{G_j \text{ intronic}} \frac{|\{x \in G_i : x \text{ off}\}|}{|G_j|}\right)$$

that is, the average length-adjusted coverage score for each individual feature (exon or intron) in the gene model.

RNA-seq data

RNA-seq data were downloaded from the Sequence Read Archive, assessed using FastQC [32], trimmed using SeqTK [33] and aligned to the genome with HiSAT2 [34, 35] using default parameters and single or paired-end methods as appropriate. Per-base coverage was calculated using SAMtools mpileup [36].

Subcellular localisation analysis

Subcellular localisation predictions for all datasets were obtained using TargetP [37]. For the plant dataset only, TargetP was run with the –P option to predict chloroplast targeting sequences. The localisation consistency for an orthogroup O was calculated as an entropy score across the categories for each gene:

$$H(O) = -\frac{1}{|O|}\sum_{C \in \mathcal{C}(O)} \frac{|C|}{|O|} \cdot \log\left(\frac{|C|}{|O|}\right)$$

where $\mathcal{C}(O) = \{C_1, \ldots, C_n\}$ is the partition of genes in O into their localisation categories.

Results

Problem definition, algorithm overview and evaluation criteria

An overview of the OMGene algorithm is provided in Fig. 3. OMGene aims to find the most consistent set of representative gene models for a set of inputted genes by seeking to maximise the agreement of their aligned amino acid sequences, returning the single best gene model for each gene. The algorithm constructs gene models based on relatively simple constraints: AUG for start codons; GU or GC for splice donor sites, AG for splice acceptor sites, and UAA, UGA, or UAG for stop codons. Other features such as codon bias or polypyrimidine tracts are not considered. OMGene can also use non-canonical translation initiation and splice sites if inputted by the user as a command-line option.

The input for OMGene is a user-selected set of gene models, in general transfer format (GTF), which are assumed to belong to a single orthogroup. For a given set of species, an orthogroup is the set of genes descended from a single ancestral gene in the last common ancestor of those species [26]: this set may contain paralogous as well as orthologous genes. The suggested pipeline for using OMGene is to determine orthogroups using OrthoFinder [26], and to apply OMGene to each or a chosen subset of these orthogroups.

OMGene uses Exonerate [25] as an initial step to cross-align amino acid sequences from all user-supplied genes to the genomic regions of the genes from all species in question, in order to find conserved translatable features. It then combines this information with the original gene models to produce an initial set of prototype exonic regions, or *gene parts*, for optimisation. The amino acid sequences for these prototype gene models are then aligned, and the constituent gene parts are split into adjacency groups based on overlaps in the alignment (see Implementation). Adjacency groups are sequentially appended to the gene models, and the genetic coordinates are recursively adjusted and assessed to optimise the agreement of the amino acid sequences between species. The resultant gene models are then subject to stringent filtering criteria before the finalised set of gene models are presented as sets of GTF coordinates, amino acid FASTA and coding sequence (CDS) FASTA sequences.

To demonstrate the utility of OMGene, it was applied to orthogroups formed from three sets of test species: a set of five plant species, a set of five mammal species and a set of five fungal species (Table 1). Prokaryotes were not considered, as their protein coding genes lack introns and intergenic regions are either very small or absent and thus gene prediction is comparatively trivial. For plants and fungi, 600 buffer bases either side of each gene model were included, for mammals, which have larger introns, 2000 buffer bases were used. OMGene was applied to orthogroups that contained exactly one gene from each species, referred to as single-copy ubiquitous (SCU) orthogroups. In addition, OMGene was run on the same set but with all genes from three representative species – *A. thaliana*, *H. sapiens* and *S. cerevisiae* – replaced with *de novo* predicted genes, obtained

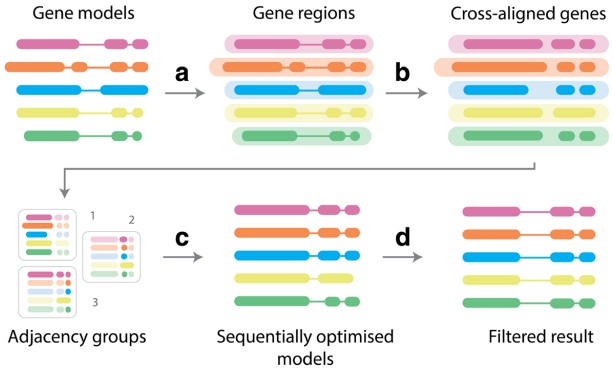


Fig. 3 Simplified overview of OMGene workflow. **a**) Gene regions are extracted from around the gene model; **b**) Exonerate is used to cross-align all constituent exons and full open reading frames to construct basic prototype gene models; **c**) The exonic regions from these prototype gene models are sorted into adjacency groups, which are then sequentially optimised using the multipartite choice function; **d**) Results are compared against the original gene models to incorporate potentially overlooked combinations, and filtered under various criteria to produce results

by running the Augustus [24] gene finder on those genomes. These species were chosen as they have the best annotated genomes and thus the existing gene models will provide the best possible training set for Augustus *de novo* prediction. This *de novo* prediction analysis was done to simulate a typical genome-sequencing project where a user has generated a well-trained set of gene models solely using computational prediction.

OMGene was assessed in three ways: RNA-seq data were used to compare the accuracy of gene models before and after application of OMGene, from both coverage (i.e. the proportion of the predicted gene that is encompassed by reads mapped from RNA-seq data) and splice junction perspectives. To assess the accuracy of start codon prediction, OMGene-modified gene models were subject subcellular localisation prediction and the results were evaluated for consistency across the orthogroup. The RNA-seq data used to assess the success of OMGene were downloaded from the NCBI Sequence Read Archive [28] and are listed in Table 2.

Application of OMGene to publicly available datasets

Quantities and nature of changes made

The full plant data set contained 3694 SCU orthogroups, containing 18,470 genes. Application of OMGene to this test set resulted in gene model changes to one or more genes in 1543 (41.8%) of these orthogroups. In total, 2017 of the inputted genes (10.9%) were altered. Of these altered versions, 154 genes (7.6% of 2017) were present in the original genome annotations as alternative (non-primary) transcripts for the inputted gene. Figure 4 shows examples of various types of gene model alteration for genes in *A. thaliana*. A full breakdown of per-species change quantities for all species sets can be found in Table 3, Figs. 5 and 6; Table 4 and Fig. 7 show the distribution of the types of changes made. All gene models that were changed by OMGene are included in the supplementary material as a set of GTF files.

The plant species that experienced the highest number of changes were *C. papaya* and *T. cacao*, which is consistent with them being more recently published and less well-studied genomes. For all species, more nucleotides were removed than were added, indicating either that gene models predictions tend to be over-cautious or that OMGene is more proficient at removing material than at adding it in. In terms of the types of changes made, exon deletion was by far the most commonly seen change, followed by moved start codon and exon boundary adjustment (Fig. 7). It should be noted that exon deletion events also encapsulate the separation of erroneously fused gene models, which can contribute many exon deletion events simultaneously.

For the mammal data set, 8771 SCU orthogroups were considered, containing 43,855 genes. Of these, 2100 orthogroups (23.9%) saw some change, with 2686 genes (6.1%) undergoing alterations. Most gene changes occurred in *C. lupus*, *M. domestica*, and *O. cuniculus*, which, compared to *M. musculus* and *H. sapiens*, are less well studied and more recently annotated. By far the most common alterations here were removed exons and moved start codons, which is likely partially due to the choice of representative gene model for input to OMGene – the longest.

For the full fungal data set, 2710 SCU orthogroups were considered, containing 13,550 genes. Of these, 100 orthogroups (3.7%) exhibited some change, and 109

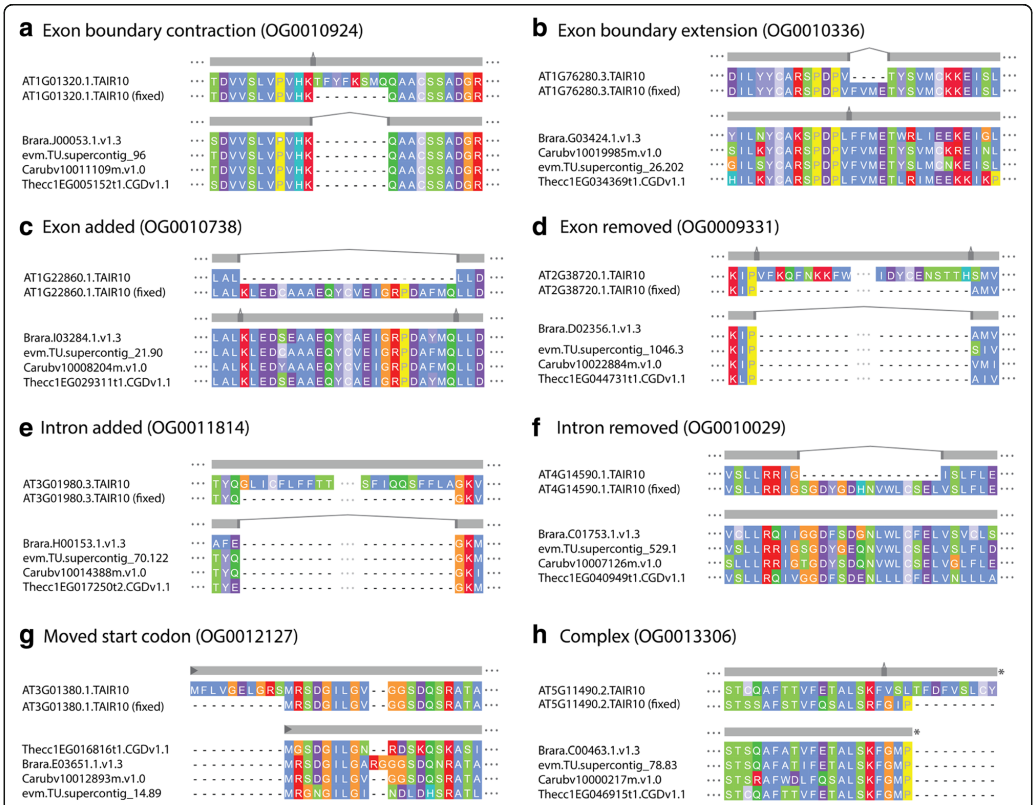


Fig. 4 Examples of individual gene model changes for genes in *A. thaliana*. **a**) AT1G01320.1, orthogroup OG0010924, exon extension, splice acceptor side; **b**) AT1G76280.3.TAIR10, orthogroup OG10336, exon contraction, splice acceptor side; **c**) AT1G22860.1, orthogroup OG0010738, novel exon introduced; **d**) AT2G38720.1, orthogroup OG0009331, removed exon; **e**) AT3G01980.3, orthogroup OG0011814, novel intron introduced; **f**) AT4G14590.1, orthogroup OG0010029, intron removed; **g**) AT3G01380.1, orthogroup OG0012127, moved start codon; **h**) AT5G11490.2, orthogroup OG0013306, complex event: exon has been removed and the previous exon boundary has been extended to include the stop codon

genes (0.8%) were altered. In this case, the genes in *E. gossypii* were the most commonly altered, consistent again with it being one of the lesser-studied species on the list. By far the most common change type in the fungal data set was a moved start codon, consistent with the fact that splicing is a rare event in fungal genes (on average 5.09 exons for plants, 11.23 exons for mammals, 1.08 exons for fungi).

To simulate a *de novo* genome annotation project, OMGene was also applied to the same three data sets with *de novo* predicted gene models for representative species, *A. thaliana, H. sapiens* and *S. cerevisiae*. These species were chosen as they have the most complete annotations of their respective data sets, and therefore these genes are likely to be the most reliable for training a gene finding algorithm. The genome annotation tool used was Augustus (see Implementation) as it is one of the best and most frequently used gene prediction algorithms.

For the plants data set with Augustus predictions for *A. thaliana*, 3694 SCU orthogroups were considered. Of these, 598 (16.2%) saw some change in an *A. thaliana* gene. Similarly, for the mammal data set, 7311 SCU orthogroups were considered, of which 2907 saw changes in a *H. sapiens* gene. For both *A. thaliana* and *H. sapiens* most changed genes underwent several individual changes. For the fungi data set, 2710 SCU orthogroups were considered. Of these, 19 (0.7%) saw some change in a *S. cerevisiae* gene. Tables 3 and 4 show a full breakdown of the types and amounts of changes made. As expected, in both cases, the total number changes and the average size of change made is greater

Table 3 Per-species gene change breakdown

	Species	No. changed genes	Nucleotides added/removed (means per change)			In original annotation as alternative "non-primary" gene model
			+ (mean)	+ (mean)	+ (mean)	
Plants	*A. thaliana*	175	1749 (43)	−23,747 (−118)	−22,139 (−92)	53 (30.3%)
	B. rapa	97	1787 (58)	−25,740 (−250)	− 23,953 (− 179)	4 (4.1%)
	C. papaya	540	23,820 (65)	−72,053 (− 128)	−48,233 (−52)	0 (0.0%)
	C. rubella	298	6568 (71)	−55,005 (− 170)	− 48,437 (−117)	2 (0.7%)
	T. cacao	556	3700 (45)	−120,984 (−118)	−117,284 (− 124)	95 (17.1%)
	TOTAL	1666	37,624 (61)	− 297,529 (−145)	− 259,905 (−97)	154 (9.2%)
	A. thaliana de novo	598	13,623 (42)	−167,038 (−35)	−51,177 (−57)	N/A
Mammals	*C. lupus*	698	8993 (64)	−98,467 (−117)	−89,474 (−91)	375 (53.7%)
	H. sapiens	397	4429 (59)	−41,401 (−101)	−36,972 (−76)	218 (54.9%)
	M. domestica	787	8637 (53)	−100,256 (−101)	−91,619 (−80)	349 (44.3%)
	M. musculus	270	9685 (120)	−19,236 (−79)	− 9551 (−29)	81 (30.0%)
	O. cuniculus	534	12,038 (61)	−72,398 (− 112)	−60,360 (−71)	243 (45.5%)
	TOTAL	2686	43,782 (67)	−331,758 (− 106)	−287,976 (−76)	1266 (47.1%)
	H. sapiens de novo	2907	251,344 (79)	− 952,864 (−167)	− 701,520 (−79)	N/A
Fungi	*E. gossypii*	46	0 (0)	− 4338 (−93)	−4338 (−93)	N/A
	D. hansenii	13	0 (0)	− 2080 (−149)	−2080 (−149)	N/A
	K. lactis	11	0 (0)	− 1314 (− 110)	−1314 (−110)	N/A
	S. cerevisiae	11	93 (93)	− 2483 (− 191)	− 2390 (−170)	N/A
	Y. lipolytica	23	117 (29)	− 4186 (− 199)	− 4069 (− 163)	N/A
	TOTAL	104	210 (42)	−14,401 (− 135)	− 14,191 (−127)	N/A
	S. cerevisiae de novo	19	601 (120)	− 5561 (− 347)	− 4960 (−236)	N/A

for the *de novo* predicted gene models than the curated gene models. However, the distribution of types of changes made remained roughly the same.

Splice junction and feature coverage analysis

To assess the validity of changes made by OMGene, both the original and the updated gene model sets were compared using publicly available RNA-seq data from the NCBI Sequence Read Archive [28] (see Implementation and Table 2). Data were downloaded for all species except *E. gossypii* and *D. hansenii*, for which no adequate data were available. Each amended gene was assessed in two ways relative to this data: firstly by comparing the exact splice junction locations with RNA-seq derived splice junctions; secondly by evaluating the coverage of exonic regions with RNA-seq. To control for unreliable data, some genic regions were omitted from this analysis. Gene regions in which the RNA-seq data suggested there were indels in the reference genome, or that were within 1000 bp of the end of a contig or scaffold, or that contained 10 or more contiguous "N" nucleotide bases were omitted from the analysis (see Implementation). Regions with these characteristics prevent the creation of reliable gene models, and so are deemed here "unassessable" and not useful for determining gene model accuracy. Gene regions with very low coverage were also omitted from the junction score analysis: gene regions required at least one base with a read depth of 10 or more.

Gene models outputted by OMGene were assessed on whether or not their junction and coverage F-scores (see Implementation) had improved or been reduced. The full results can be seen in Table 5. For the plant data set, OMGene improved the agreement of the gene model with the splice junctions inferred from RNA-seq data for 729 assessable genes, while 125 assessable gene models exhibited reduced agreement (85.3% improved). Similarly, when assessing RNA-seq coverage of gene models OMGene improved the agreement of the models with the data for 1026 genes, while 167 genes exhibited reduced agreement (86.0% improved). For the *de novo* predicted *A. thaliana* genes, the success rates were essentially the same as for the public data (87.3% and 91.1% improved by junction and coverage F-scores respectively), but the absolute quantity of genes exhibiting a changed score increased roughly four-fold. This difference represents the considerable effort and evidence-based curation that has been invested in the *A. thaliana* genome annotation.

For the mammals, 948 assessable changed genes (79.2%) had an improved junction score compared to the original

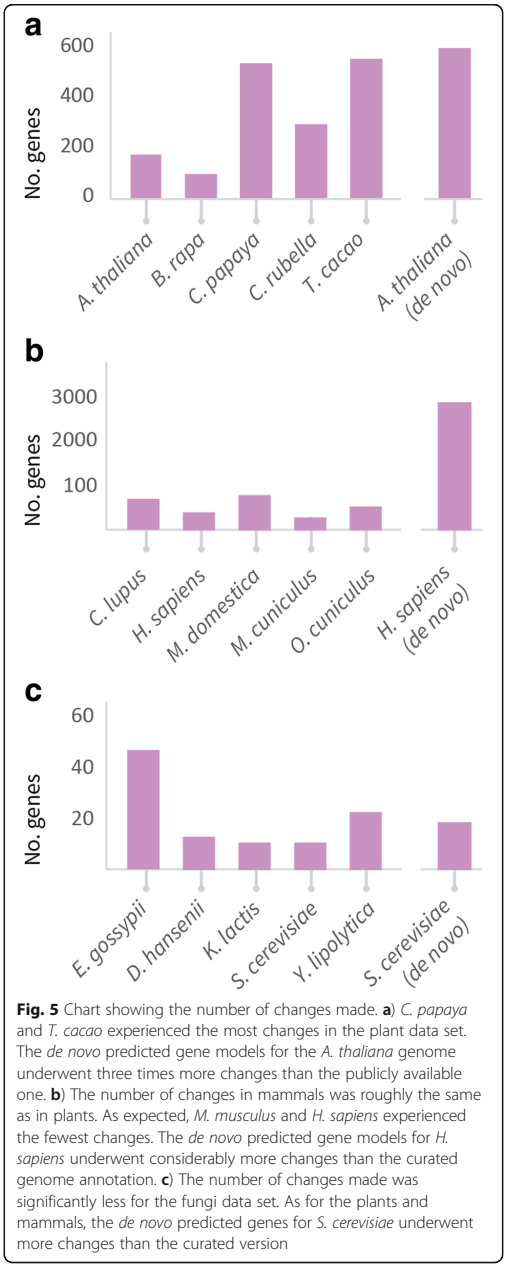


Fig. 5 Chart showing the number of changes made. **a**) *C. papaya* and *T. cacao* experienced the most changes in the plant data set. The *de novo* predicted gene models for the *A. thaliana* genome underwent three times more changes than the publicly available one. **b**) The number of changes in mammals was roughly the same as in plants. As expected, *M. musculus* and *H. sapiens* experienced the fewest changes. The *de novo* predicted gene models for *H. sapiens* underwent considerably more changes than the curated genome annotation. **c**) The number of changes made was significantly less for the fungi data set. As for the plants and mammals, the *de novo* predicted genes for *S. cerevisiae* underwent more changes than the curated version

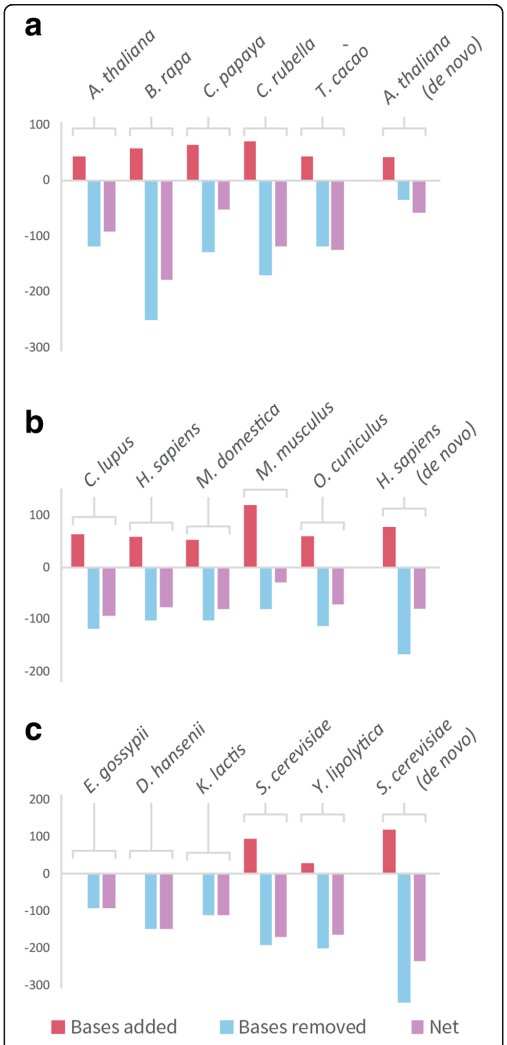


Fig. 6 The average number of nucleotides added or removed from gene models as a result of changes made by OMGene. The units of the y-axis are the number of nucleotides. **a**) Average magnitudes of each change for plants; **b**) Average magnitudes for changes made to mammal genes; **c**) Average magnitudes for changes made to fungal genes. In all cases, predicted gene sequences were shortened to a greater extent than they were lengthened.

gene model. In addition, 82.83% of assessable genes had improved RNA-seq coverage scores. For the *de novo* data, 89.4% of assessable predicted *H. sapiens* had an improved splice junction score after application of OMGene, and 88.74% had an improved coverage score. The numbers of changed genes were again considerably more for the de novo predicted *H. sapiens* data, again indicating the high degree of attention that has been afforded to the annotation of this species.

Table 4 Summary of gene model change categories

	Species	No. changes	Exon boundary		Exon		Intron		Moved start
			contraction	extension	add	del	add	del	
Plants	*A. thaliana*	242	47	23	4	117	5	13	33
	B. rapa	134	11	14	9	56	3	8	33
	C. papaya	928	148	205	95	345	18	42	74
	C. rubella	415	32	32	39	101	1	19	191
	T. cacao	949	117	59	9	624	10	13	117
	TOTAL	2668	355	333	156	1243	37	95	448
	A. thaliana de novo	1344	151	255	49	780	2	10	97
Mammals	*C. lupus*	980	116	72	38	418	9	4	323
	H. sapiens	485	31	43	29	223	0	0	159
	M. domestica	1148	122	89	56	472	19	3	387
	M. musculus	324	20	29	42	111	6	6	110
	O. cuniculus	834	139	122	43	245	18	9	258
	TOTAL	3771	428	355	208	1469	52	22	1237
	H. sapiens de novo	8883	606	1128	1866	4217	16	17	1033
Fungi	*E. gossypii*	46	0	0	0	1	0	0	45
	D. hansenii	13	0	0	0	1	0	0	12
	K. lactis	11	0	0	0	0	0	0	11
	S. cerevisiae	13	1	0	0	0	1	1	10
	Y. lipolytica	24	0	0	0	4	5	0	15
	TOTAL	107	1	0	0	6	6	1	93
	S. cerevisiae de novo	20	0	2	0	4	0	2	12

The results for the fungal data set (see Table 6) were not as good. Notably very few gene models showed any change in junction F-score, with only 8 genes exhibiting a changed score. This is due to the relatively simple exon structure of fungal genes, for which splicing is very rare, and splicing events predicted by OMGene are much less likely to be correct. In this case 3 genes had an improved score, and 5 had a reduced score (37.5% success), with all 5 of the losing genes coming from *Y. lipolytica*. The most common change made to fungal genes was a moved start codon, which, although not detectable in the junction F-score, can be detectable in the coverage F-score. This is reflected in the results, where 30 genes showed an improved coverage F-score and 10 genes showed a worse coverage F-score (75% improved). In the *de novo* case, again the numbers increased while the percentage success remained roughly the same, with 4 (100%) genes improving by junction for *S. cerevisiae* and 11 (64.7%) improving by coverage score. The highly compact nature of fungal genomes, with few exons and limited space between genes means that the accuracy of *de novo* predicted genes is higher than in plants and mammals. Thus the utility of OMGene on these comparatively simpler genomes is limited.

Many of the cases for which OMGene results differ from RNA-seq evidence are attributable to real biological variability that confounds the evaluation criteria of the algorithm. For example, there are some instances where the most evolutionary conserved splice site was not the splice site observed in the RNA-seq data. Such events, by definition, cannot be detected by OMGene. Furthermore, RNA-seq mapping errors also contributed to reduced scores, as did artefacts resulting from spliced UTRs, and jagged read profiles, particularly in the fungal data, that made some coverage scores difficult to calculate reliably. Finally, the presence of multiple transcript isoforms within the RNA-seq data can reduce the score for a valid transcript even if it is the best choice for that particular gene. While users of OMGene should be aware of these confounding factors, the above data demonstrates that, in general, OMGene is much more likely to improve a given gene model than not even for extensilvey curated genomes.

Assessment of subcellular localisation predictions for 5′ end analysis

Given that genes from the same orthogroup are, by definition, descended from a single ancestral gene, it is

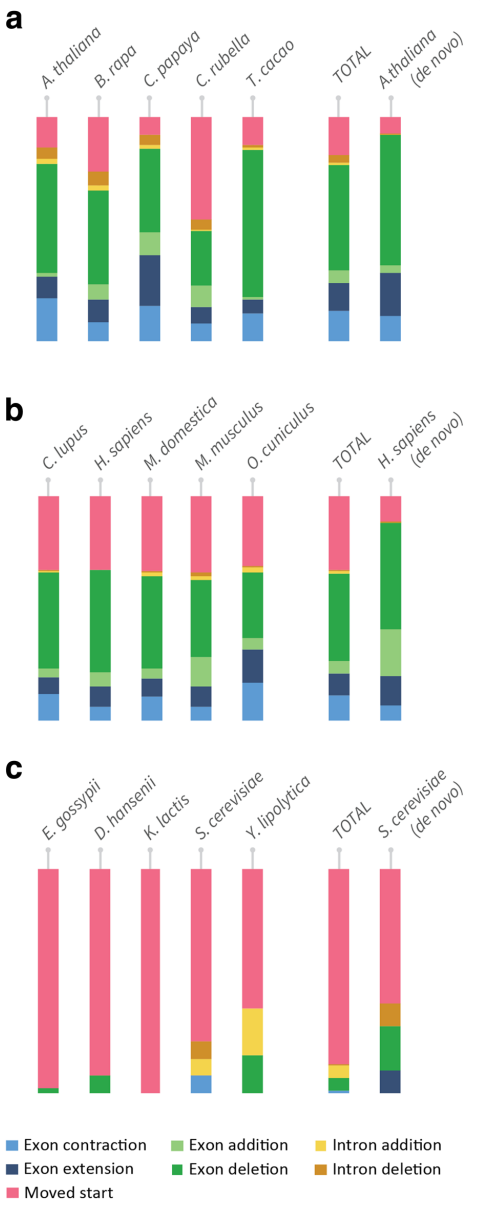


Fig. 7 Distribution of types of changes made in the three data sets. **a**) The most common change in plants was exon deletion. **b**) Moved start codons and removed exons were most common in mammals. **c**) In fungi, the most common change was overwhelmingly a moved start codon

reasonable to assume that they should be consistent in their predicted subcellular localisation. Several subcellular targeting sequences are located at the N-termini of genes [38], thus one expects genes with inaccurately predicted start codons to yield inaccurate results when assessing their targeting signals. Genes belonging to orthogroups changed by OMGene were assessed to determine whether the changes resulted in increased consistency of the predicted subcellular localisation of all genes in the orthogroup. Targeting predictions were made using TargetP [37], and Shannon entropy was calculated to assess the consistency of the predictions within the orthogroups (see Implementation). Entropy scores were compared only for orthogroups in which at least one gene model was altered by OMGene. An entropy score of 0 indicates that all members of the orthogroup are predicted to localise to the same sub-cellular compartment; the worst possible entropy score given five genes and four possible localisations identified by TargetP (chloroplast, mitochondrion, secreted, cytoplasmic) is $-\frac{2}{5}\log_2(\frac{1}{5})-\frac{3}{5}\log_2(\frac{1}{5}) \approx 1.92$, indicating that only two of the genes agree. An example orthogroup whose prediction entropy score has been improved by start codon adjustment can be seen in Fig. 8.

The 1543 plant orthogroups in which one or more genes were altered were subjected to subcellular prediction analysis (Table 6). Of these, gene model changes made by OMGene resulted in changes in predicted subcellular localisation for one or more constituent members of 55 orthogroups. In total, 74 improved agreement between gene models (74%), 13 remained the same (13%), and 13% increased disagreement between predicted subcellular localisation of gene models. The effect was more profound for the mammal data: 509 orthogroups showed a change in localisation predictions for one or more of their genes, 444 of which (87.2%) resulted in improved predicted localisation consistency. The data for fungi were sparser: only 7 out of 95 changed orthogroups exhibited a change in subcellular localisation prediction, with 6 of these changes improving the consistency of localisation prediction (85.7%) and 1 increasing disagreement (14.3%).

Similar results were obtained for the *de novo* annotation analysis in plants, although again the data were sparse here. Orthogroups containing the *de novo* predicted *A. thaliana* genes were considered together with the four original genes for the other species. Here, 11 of the *A. thaliana* genes experienced a change in subcellular localisation following application of OMGene. Of the 11 orthogroups containing these, 9 improved consistency (81.9%) and 2 reduced the consistency (18.2%). For de novo *H. sapiens* predictions, 527 of the start codon changes resulted in a change in subcellular localisation prediction change for *H. sapiens*. In 458 (86.9%) of these cases, this improved the consistency of

Table 5 RNA-seq coverage and junction F-scores

	Species	Junction F-score		Coverage F-score	
		Better	Worse	Better	Worse
Plants	*A. thaliana*	94 (87.8%)	13 (12.1%)	109 (91.5%)	10 (8.4%)
	B. rapa	24 (63.1%)	14 (36.8%)	29 (56.8%)	22 (43.1%)
	C. papaya	246 (82.2%)	53 (17.7%)	344 (83.9%)	66 (16.0%)
	C. rubella	90 (89.1%)	11 (10.8%)	186 (91.6%)	17 (8.3%)
	T. cacao	275 (88.9%)	34 (11.0%)	358 (87.3%)	52 (12.6%)
	TOTAL	729 (85.3%)	125 (14.6%)	1026 (86.0%)	167 (13.9%)
	A. thaliana de novo	422 (87.3%)	61 (12.6%)	475 (91.1%)	46 (8.8%)
Mammals	*C. lupus*	323 (85.9%)	53 (14.1%)	478 (86.28%)	76 (13.72%)
	H. sapiens	102 (82.9%)	21 (17.1%)	258 (87.46%)	37 (12.54%)
	M. domestica	239 (72.2%)	92 (27.8%)	439 (76.48%)	135 (23.52%)
	M. musculus	71 (62.3%)	43 (37.7%)	133 (71.12%)	54 (28.88%)
	O. cuniculus	213 (84.2%)	40 (15.8%)	353 (86.10%)	57 (13.90%)
	TOTAL	948 (79.2%)	249 (112.5%)	1661 (82.83%)	359 (17.77%)
	H. sapiens de novo	832 (89.4%)	99 (10.6%)	2017 (88.74%)	256 (11.26%)
Fungi	*K. lactis*	0 (N/A)	0 (N/A)	9 (100.0%)	0 (0%)
	S. cerevisiae	0 (N/A)	0 (N/A)	6 (75.0%)	2 (25.0%)
	Y. lipolytica	2 (28.5%)	5 (71.4%)	11 (64.7%)	6 (35.2%)
	TOTAL	3 (37.5%)	5 (62.5%)	30 (75.0%)	10 (25.0%)
	S. cerevisiae de novo	4 (100%)	0 (0%)	11 (64.7%)	6 (35.2%)

the orthogroup. For the fungal data set, the data were extremely sparse, with only one gene experiencing a change in its targeting prediction, which reduced the consistency for its orthogroup. Thus, although data were sparse for the fungal dataset, in each of the plant, mammal, and fungal datasets the consistency of gene models was on the whole improved from a subcellular targeting perspective.

Discussion

Here we present OMGene, an automated method for improving the consistency of gene model annotations across species. OMGene is intended for use in computational *de novo* genome annotation projects where no empirical data (such as RNA-seq data) is available to train or correct gene model predictions, or to assist the construction of gene models for genes that are not expressed in the data available. OMGene is also designed to help users who wish to leverage conservation information to correct gene models of a single gene of interest across a set of species. Thus OMGene is suitable for both large and small scale analyses.

OMGene is run as a python script with a tab-delimited input file, each line of which contains the path to a GTF file, containing coordinates for a single gene, and the path to its corresponding genome FASTA file. Full instructions for running OMGene can be found at the OMGene GitHub repository (https://github.com/mpdunne/omgene).

Table 6 Subcellular localisation predictions

	Category	No. orthogroups with changed localisation predictions	Entropy score		
			Better	Same	Worse
Plants	Public data	55	42 (76.4%)	5 (7.7%)	8 (14.5%)
	A. thaliana de novo	11	9 (81.9%)	0 (0%)	2 (18.2%)
Mammals	Public data	509	444 (87.2%)	19 (3.7%)	46 (9.0%)
	H. sapiens de novo	527	458 (86.9%)	23 (4.4%)	46 (8.7%)
Fungi	Public data	7	6 (85.7%)	0 (0%)	1 (14.3%)
	S. cerevisiae de novo	1	0 (0%)	0 (0%)	1 (100%)

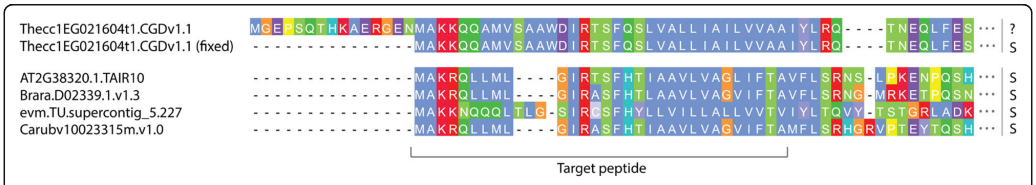


Fig. 8 Example change in subcellular localisation prediction for a gene. Thecc1EG021604t1.CGDv1.1 from *T. cacao* has undergone a change in start codon, revealing a signal peptide at its 5′ end. In this case, what was previously assumed to be cytosolic has been found to be targeted to the secretory pathway, the same as the other members of the orthogroup (OG0009265). In this case, the Shannon entropy score for the orthogroup has fallen from 0.72 to 0

OMGene results reflect differences in gene model complexity between species sets

To demonstrate the utility and performance characteristics of OMGene, it was applied to three separate datasets of well-annotated plant, mammal and fungal genomes. When applied to the plant data set, OMGene altered the gene models of one or more genes in 41.8% of the orthogroups that were evaluated; when applied to the mammal data set, 23.9% of SCU orthogroups saw some change. In contrast, only 3.7% of orthogroups were subject to modification in the fungal data set. This result reflects the differences in gene model complexity between the three species groups. Specifically, gene models in plants tend to have more exons than fungi (mean = 5.09 exons for plants, 11.23 for mammals, 1.08 exons for fungi) and thus there is considerably more potential for gene model variation in plants and mammals than in fungi. In light of this it was unsurprising that the most frequently observed change made in fungi was a change in choice of start codon. This is also reflected in the high number of removed exons from plant and mammal genes, which is contributed to partly by the separation of erroneously fused adjacent genes.

OMGene works well on complex gene models

The changes made by OMGene were assessed relative to splice-mapped RNA-seq data to assess the level to which it had improved the gene models. For the plant data set, the results from OMGene clearly resembled the empirical data more closely, with 85.4% and 86.0% of genes improving in terms of their splice junctions and their coverage respectively. The profiles were different for different species, with many more changes being made for *C. papaya* and *T.cacao;* in addition the number of successes for *B. rapa* was slightly lower than for the other species.

In mammals, the three non-model species, *C. lupus*, *M. domestica*, and *O. cuniculus* faired well in terms of their junction scores. *H. sapiens* genes also typically scored highly, but *M. musculus* genes did not do so well. The reason why mouse genes changes were less successful than human gene changes is unknown. These two species also saw considerably fewer gene model changes than the other species. In all cases, mammal genes were mostly contracted via the removal of exons or by moving the start codon – this is at least partly due to biases in the choice of longest gene model as representatives for these genes.

The number of junction changes made for the fungal data set was considerably lower: only 8 changed genes had an altered junction F-score, 62.5% of which become worse after OMGene. Though this is less than the plant data set, it should be noted that the resolution of this data set does not lend itself to accurate conclusions about the general validity of changes made to fungal genes. The resolution and success rate for fungal genes from a coverage perspective was slightly higher, with 75% of the genes with changed scores improving. The low resolution of junction data for fungal genes reflects the rarity of complex gene models in these species, and thus the low likelihood that deviations from simple, single-exon gene models are correct. Thus, while OMGene does not always produce gene models that agree with transcriptome data, it does improve the overall quality of gene model annotations even for relatively simple fungal genomes.

The improvements in gene model accuracy made by OMGene for the *de novo* predicted gene models were much the same as for the publicly available, curated gene models. However, the number of changes made to the *de novo* predicted set was much greater, indicating that the considerable labour that has been applied to these model organisms has successfully controlled for potential errors. It should be noted that, although OMGene managed to improve many of the gene models outputted by Augustus, the two agreed in most cases (86.1%, 60.3% and 98.6% for plants, mammals and fungi respectively), indicating that the basic implementation of a well-trained Augustus *de novo* prediction produces genes that are highly consistent between species.

OMGene improves the consistency of subcellular localisation predictions

In addition to assessment of coverage and splice junctions, gene models were assessed by the consistency of their predicted subcellular localisation. Given that the orthogroups used in this analysis comprise ubiquitously conserved single copy genes, it is logical to assume that these genes should generally have the same subcellular localisation. For the full plant data set, of all orthogroups whose genes had different subcellular targeting predictions after application of OMGene, 76.4% had improved intra-orthogroup consistency, with 85.5% either improving or remaining the same. In mammals, 87.2% of orthogroups whose subcellular localisation predictions were changed by OMGene showed improved consistency. For the full fungal data set, although the data were sparse, 85.7% of the orthogroups considered had improved consistency.

The results for the plant data set were similar for the *de novo* annotated set (85.7% improvement), as was also the case for mammals (86.9% improvement). For fungal orthogroups containing *de novo* predicted *S. cerevisiae* genes, the only gene whose localisation prediction changed caused the consistency of its orthogroup to decrease, however the resolution of the data in this case is not sufficient to draw any conclusions. Thus, application of OMGene improves the accuracy of start codon specification in *de novo* predicted gene models.

Conclusions

When applied to publicly available plant, mammal and fungal data sets, OMGene demonstrates proficiency in improving gene models from multiple perspectives. Due to stringent filtering criteria, it does not fix all errant gene models, however the gene models that it does fix represent an improvement the majority of the time. The overall improvement is larger for genomes with complex gene models.

Availability and requirements

Project name: OMGene.
Project home page: https://github.com/mpdunne/omgene
Operating system: Unix.
Programming languages: Python v2.7.x.
Other requirements: Bedtools, Exonerate, Python packages (scipy, numpy, pyfaidx, networkx, BioPython).
License: GPLv3.

Abbreviations

CDS: Coding sequence; GTF: General transfer format; SCU: Single copy, ubiquitous

Funding

SK is a Royal Society University Research Fellow. This work was supported by the European Union's Horizon 2020 research and innovation programme under grant agreement number 637765. MPD is supported by an EPSRC studentship through EP/G03706X/1.

Availability of data and materials

The software is available under the GPLv3 licence at https://github.com/mpdunne/omgene.

Authors' contributions

SK conceived the project. MPD developed the algorithm. SK and MPD analysed the data and wrote the manuscript. Both authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Received: 17 November 2017 Accepted: 19 April 2018
Published online: 27 April 2018

References

1. Cochrane G, Karsch-mizrachi I, Nakamura Y. The international nucleotide sequence database collaboration the international nucleotide sequence database collaboration. Nucleic Acids Res. 2011;39:D15–8.
2. Land M, Hauser L, Jun S, Nookaew I, Leuze MR, Ahn T, Karpinets T, Lund O, Kora G, Wassenaar T, Poudel S, Ussery DW. Insights from 20 years of bacterial genome sequencing. Funct Integr Genomics. 2015;15:141–61.
3. NCBI, "GenBank and WGS Statistics," 2017. [Online]. Available: https://www.ncbi.nlm.nih.gov/genbank/statistics/.
4. Hayden EC. The $1,000 genome. Nature. 2014;507:295.
5. K. Wetterstrand, "DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP)," Mar-2016. [Online]. Available: www.genome.gov/sequencingcosts.
6. M. Yandell and D. Ence, "A beginner's guide to eukaryotic genome annotation," *Nat Rev Genet*, vol 13, no May, pp. 329–342, 2012.
7. Denton JF, Lugo-Martinez J, Tucker AE, Schrider DR, Warren WC, Hahn MW. Extensive error in the number of genes inferred from draft genome assemblies. PLoS Comput Biol. 2014;10(12)
8. E. Veeckman, T. Ruttink, and K. Vandepoele, "Are We There Yet ? Reliably Estimating the Completeness of Plant Genome Sequences," *Plant Cell*, vol. 28, no. August, pp. 1759–1768, 2016.
9. Dunne MP, Kelly S. OrthoFiller: utilising data from multiple species to improve the completeness of genome annotations. BMC Genomics. 2017;18(390)
10. Nasiri J, Naghavi M, Rad SN, Yolmeh T, Shirazi M, Naderi R, Nasiri M, Ahmadi S. Gene identification programs in bread wheat: a comparison study. Nucleosides Nucleotides Nucleic Acids. 2013;32(10):529–54.
11. Sudmant PH, Alexis MS, Burge CB. Meta-analysis of RNA-seq expression data across species, tissues and studies. Genome Biol. 2015;16(287)
12. Danielsson F, James T, Gomez-Cabrero D, Huss M. Assessing the consistency of public human tissue RNA-seq data sets. Brief Bioinform. 2015;16(6):941–9.
13. Conesa A, Madrigal P, Tarazona S, Gomez-cabrero D, Cervera A, Mcpherson A, Szcze W, Gaffney DJ, Elo LL, Zhang X. A survey of best practices for RNA-seq data analysis. Genome Biol. 2016;17(13)
14. Wang L, Nie J, Sicotte H, Li Y, Eckel-passow JE, Dasari S, Vedell PT, Barman P, Wang L, Weinshiboum R, Jen J, Huang H, Kohli M, Kocher JA. Measure transcript integrity using RNA-seq data. BMC Bioinformatics. 2016;17(58)
15. Hansen KD, Brenner SE, Dudoit S. Biases in Illumina transcriptome sequencing caused by random hexamer priming. Nucleic Acids Res. 2010;38(12):e131.
16. Jiang H, Salzman J. A penalized likelihood approach for robust estimation of isoform expression. Stat Interface. 2015;8(4):437–45.
17. Abril JF, Castelo R, Guigó R. Comparison of splice sites in mammals and chicken. Genome Res. 2005;15(1):111–9.
18. Betts MJ, Guigó R, Agarwal P, Russell RB. Exon structure conservation despite low sequence similarity: a relic of dramatic events in evolution? EMBO J. 2001;20(19):5354–60.
19. Nurtdinov RN, Neverov AD, Favorov AV, Mironov AA, Gelfand MS. Conserved and species-specific alternative splicing in mammalian genomes. BMC Evol Biol. 2007;7:249.

20. Alexandersson M, Cawley S, Pachter L. SLAM: cross-species gene finding and alignment with a generalized pair hidden Markov model. Genome Res. 2003;13:496–502.
21. Wiehe T, Gebauer-Jung S. SGP-1: prediction and validation of homologous genes based on sequence alignments. Genome Res. 2001;11:1574–83.
22. Gross SS, Brent MR. Using multiple alignments to improve gene prediction. J Comput Biol. 2006;13(2):379–93.
23. Gross SS, Do CB, Sirota M, Batzoglou S. CONTRAST: a discriminative, phylogeny-free approach to multiple informant de novo gene prediction. Genome Biol. 2007;8:R269.
24. Stanke M, Morgenstern B. "AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints," *Nucleic Acids Res.*, vol. 33, no suppl_2. 2005:465–7.
25. Slater G, Birney E. Automated generation of heuristics for biological sequence comparison. BMC Bioinformatics. 2005;6(31)
26. Emms DM, Kelly S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biol. 2015;16(157)
27. Goodstein DM, Shu S, Howson R, Neupane R, Hayes RD, Fazo J, Mitros T, Dirks W, Hellsten U, Putnam N, Rokhsar DS. Phytozome: a comparative platform for green plant genomics. Nucleic Acids Res. 2012;40:D1178–86.
28. R. Leinonen, H. Sugawara, and M. Shumway, "The sequence read archive," *Nucleic Acids Res.*, vol. 39, no suppl_1, pp. D19–D21, 2010.
29. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability Article Fast Track. Mol Biol Evol. 2013;30(4):772–80.
30. Hagberg AA, Schult DA, Swart PJ. Exploring network structure, dynamics, and function using NetworkX. Proc 7th Python Sci Conf. 2008:11–6.
31. Pelechano V, Steinmetz LM. Gene regulation by antisense transcription. Nat Rev Genet. 2013;14:880–93.
32. S. Andrews, "FastQC: a quality control tool for high throughput sequence data," 2010. [Online]. Available: http://www.bioinformatics.babraham.ac.uk/projects/fastqc.
33. H. Li, "Seqtk: Toolkit for processing sequences in FASTA/Q formats," 2012. [Online]. Available: https://github.com/lh3/seqtk. [Accessed: 17-Jan-2018].
34. D. Kim, B. Langmead, and S. Salzberg, "HISAT2: graph-based alignment of next-generation sequencing reads to a population of genomes." 2017.
35. Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. Nat Methods. 2015;12(4):357–62.
36. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. The sequence alignment/map format and SAMtools. Bioinformatics. 2009;25(16):2078–9.
37. Emanuelsson O, Nielsen H, Brunak S, von Heijne G. Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. J Mol Biol. Jul. 2000;300(4):1005–16.
38. Schneider G, Fechner U. Advances in the prediction of protein targeting signals. Proteomics. Jun. 2004;4(6):1571–80.
39. Goodstein DM, Shu S, Howson R, Neupane R, Hayes RD, Fazo J, Mitros T, Dirks W, Hellsten U, Putnam N, Rokhsar DS. Phytozome: a comparative platform for green plant genomics. Nucleic Acids Res. 2012;40(D1):1178–86.
40. Lamesch P, Berardini TZ, Li D, Swarbreck D, Wilks C, Sasidharan R, Muller R, Dreher K, Alexander DL, Garcia-Hernandez M, Karthikeyan AS, Lee CH, Nelson WD, Ploetz L, Singh S, Wensel A, Huala E. The Arabidopsis information resource (TAIR): improved gene annotation and new tools. Nucleic Acids Res. 2012;40:D1202–10.
41. R. Ming, S. Hou, Y. Feng, Q. Yu, A. Dionne-laporte, J. H. Saw, P. Senin, W. Wang, B. V Ly, K. L. T. Lewis, L. Salzberg, L. Feng, M. R. Jones, R. L. Skelton, J. E. Murray, W. Qian, J. Shen, P. Du, M. Eustice, E. Tong, E. Lyons, R. E. Paull, T. P. Michael, K. Wall, D. W. Rice, H. Albert, M. Wang, Y. J. Zhu, M. Schatz, N. Nagarajan, A. Acob, P. Guan, A. Blas, C. M. Wai, C. M. Ackerman, Y. Ren, C. Liu, J. Wang, J. Wang, J. Na, V. Eugene, B. Haas, J. Thimmapuram, D. Nelson, X. Wang, E. John, W. Zhang, G. Presting, A. Windsor, R. Navajas-pérez, J. Torres, F. A. Feltus, B. Porter, Y. Li, a M. Burroughs, L. Liu, D. a Christopher, S. M. Mount, and P. H. Moore, "The draft genome of the transgenic tropical fruit tree papaya (Carica papaya Linnaeus)," Nature, vol. 452, no. 7190, pp. 991–996, 2008.
42. Slotte T, Hazzouri KM, Ågren JA, Koenig D, Maumus F, Guo Y, Steige K, Platts AE, Escobar JS, Newman LK, Wang W, Mandáková T, Vello E, Smith LM, Henz SR, Steffen J, Takuno S, Brandvain Y, Coop G, Andolfatto P, Hu TT, Blanchette M, Clark RM, Quesneville H, Nordborg M. The Capsella rubella genome and the genomic consequences of rapid mating system evolution. Nat Genet. 2013;45(7):831–5.
43. Motamayor JC, Mockaitis K, Schmutz J, Haiminen N, III DL, Cornejo O, Findley SD, Zheng P, Utro F, Royaert S, Saski C, Jenkins J, Podicheti R, Zhao M, Scheffler BE, Stack JC, Feltus FA, Mustiga GM, Amores F, Phillips W, Marelli JP, May GD, Shapiro H, Ma J, Bustamante CD, Schnell RJ, Main D, Gilbert D, Parida L, Kuhn DN. The genome sequence of the most widely cultivated cacao type and its use to identify candidate genes regulating pod color. Genome Biol. 2013;14:r53.
44. Lindblad-Toh K, Wade CM, Mikkelsen TS, Karlsson EK, Jaffe DB, Kamal M, Clamp M, Chang JL, Kulbokas EJ, Zody MC, Mauceli E, Xie X, Breen M, Wayne RK, Ostrander EA, Ponting CP, Galibert F, Smith DR, DeJong PJ, Kirkness E, Alvarez P, Biagi T, Brockman W, Butler J, Chin CW, Cook A, Cuff J, Daly MJ, DeCaprio D, Gnerre S, Grabherr M, Kellis M, Kleber M, Bardeleben C, Goodstadt L, Heger A, Hitte C, Kim L, Koepfli KP, Parker HG, Pollinger JP, Searle SMJ, Sutter NB, Thomas R, Webber C, Lander ES. Genome sequence, comparative analysis and haplotype structure of the domestic dog. Nature. 2005;438(7069):803–19.
45. Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. GenBank. Nucleic Acids Res. 2013;41:D36–42.
46. Pruitt KD, Harrow J, Harte RA, Wallin C, Diekhans M, Maglott DR, Searle S, Farrell CM, Loveland JE, Ruef BJ, Hart E, Suner MM, Landrum MJ, Aken B, Ayling S, Baertsch R, Fernandez-Banet J, Cherry JL, Curwen V, DiCuccio M, Kellis M, Lee J, Lin MF, Schuster M, Shkeda A, Amid C, Brown G, Dukhanina O, Frankish A, Hart J, Maidak BL, Mudge J, Murphy MR, Murphy T, Rajan J, Rajput B, Riddick LD, Snow C, Steward C, Webb D, Weber JA, Wilming L, Wu W, Birney E, Haussler D, Hubbard T, Ostell J, Durbin R, Lipman D. The consensus coding sequence (CCDS) project: identifying a common protein-coding gene set for the human and mouse genomes. Genome Res. 2009;19:1316–23.
47. Mikkelsen TS, Wakefield MJ, Aken B, Amemiya CT, Chang JL, Duke S, Garber M, Gentles AJ, Goodstadt L, Heger A, Jurka J, Kamal M, Mauceli E, Searle SMJ, Sharpe T, Baker ML, Batzer MA, Benos PV, Belov K, Clamp M, Cook A, Cuff J, Das R, Davidow L, Deakin JE, Fazzari MJ, Glass JL, Grabherr M, Greally JM, Gu W, Hore TA, Huttley GA, Kleber M, Jirtle RL, Koina E, Lee JT, Mahony S, Marra MA, Miller RD, Nicholls RD, Oda M, Papenfuss AT, Parra ZE, Pollock DD, Ray DA, Schein JE, Speed TP, Thompson K, VandeBerg JL, Wade CM, Walker JA, Waters PD, Webber C, Weidman JR, Xie X, Zody MC, Baldwin J, Abdouelleil A, Abdulkadir J, Abebe A, Abera B, Abreu J, Acer SC, Aftuck L, Alexander A, An P, Anderson E, Anderson S, Arachi H, Azer M, Bachantsang P, Barry A, Bayul T, Berlin A, Bessette D, Bloom T, Blye J, Boguslavskiy L, Bonnet C, Boukhgalter B, Bourzgui I, Brown A, Cahill P, Channer S, Cheshatsang Y, Chuda L, Citroen M, Collymore A, Cooke P, Costello M, D'Aco K, Daza R, De Haan G, DeGray S, DeMaso C, Dhargay N, Dooley K, Dooley E, Doricent M, Dorje P, Dorjee K, Dupes A, Elong R, Falk J, Farina A, Faro S, Ferguson D, Fisher S, Foley CD, Franke A, Friedrich D, Gadbois L, Gearin G, Gearin CR, Giannoukos G, Goode T, Graham J, Grandbois E, Grewal S, Gyaltsen K, Hafez N, Hagos B, Hall J, Henson C, Hollinger A, Honan T, Huard MD, Hughes L, Hurhula B, Husby ME, Kamat A, Kanga B, Kashin S, Khazanovich D, Kisner P, Lance K, Lara M, Lee W, Lennon N, Letendre F, LeVine R, Lipovsky A, Liu X, Liu J, Liu S, Lokyitsang T, Lokyitsang Y, Lubonja R, Lui A, MacDonald P, Magnisalis V, Maru K, Matthews C, McCusker W, McDonough S, Mehta T, Meldrim J, Meneus L, Mihai O, Mihalev A, Mihova T, Mittelman R, Mlenga V, Montmayeur A, Mulrain L, Navidi A, Naylor J, Negash T, Nguyen T, Nguyen N, Nicol R, Norbu C, Norbu N, Novod N, O'Neill B, Osman S, Markiewicz E, Oyono OL, Patti C, Phunkhang P, Pierre F, Priest M, Raghuraman S, Rege F, Reyes R, Rise C, Rogov P, Ross K, Ryan E, Settipalli S, Shea T, Sherpa N, Shi L, Shih D, Sparrow T, Spaulding J, Stalker J, Stange-Thomann N, Stavropoulos S, Stone C, Strader C, Tesfaye S, Thomson T, Thoulutsang Y, Thoulutsang D, Topham K, Topping I, Tsamla T, Vassiliev H, Vo A, Wangchuk T, Wangdi T, Weiand M, Wilkinson J, Wilson A, Yadav S, Young G, Yu Q, Zembek L, Zhong D, Zimmer A, Zwirko Z, Jaffe DB, Alvarez P, Brockman W, Butler J, Chin CW, Gnerre S, MacCallum I, Graves JAM, Ponting CP, Breen M, Samollow PB, Lander ES, Lindblad-Toh K. Genome of the marsupial Monodelphis domestica reveals innovation in non-coding sequences. Nature. 2007;447(7141):167–77.
48. Alföldi J, Di Palma F, Lindblad-toh K. The European rabbit genome. In: Rabbit biotechnology: rabbit genomics, Transgenesis, cloning and models; 2009. p. 129.
49. F. S. Fred S. Dietrich, S. Voegeli, S. Brachat, A. Lerch, K. Gates, S. Steiner, C. Mohr, P. Luedi, S. Choi, R. A. Wing, A. Flavier, T. D. Gaffney, P. Philippsen, P. Fred S.

Dietrich, Fred S Voegeli S Brachat S Lerch A Gates K Steiner S. and Mohr C Luedi P Choi S wing r a Flavier A and Gaffney T D and Philippsen, "the Ashbya gossypii genome as a tool for mapping the ancient Saccharomyces cerevisiae genome," *Science (80-.).*, vol. 304, no. April, 2004.

50. Dujon B, Sherman D, Fischer G, Durrens P, Casaregola S, Lafontaine I, De Montigny J, Blanchin S, Beckerich J-M, Beyne E, Bleykasten C, Babour A, Boyer J, Cattolico L, Confanioleri F, De Daruvar A, Despons L, Fabre E, De Montigny J, Marck C, Neuvéglise C, Talla E, Goffard N, Frangeul L, Aigle M, Anthouard V, Babour A, Barbe V, Barnay S, Blanchin S, Beckerich J-M, Beyne E, Bleykasten C, Boisramé A, Boyer J, Cattolico L, Confanioleri F, De Daruvar A, Despons L, Fabre E, Fairhead C, Ferry-Dumazet H, Groppi A, Hantraye F, Hennequin C, Jauniaux N, Joyet P, Kachouri R, Kerrest A, Koszul R, Lemaire M, Lesur I, Ma L, Muller H, Nicaud J-M, Nikolski M, Oztas S, Ozier-Kalogeropoulos O, Pellenz S, Potier S, Richard G-F, Straub M-L, Suleau A, Swennen D, Tekaia F, Wésolowski-Louvel M, Westhof E, Wirth B, Zeniou-Meyer M, Zivanovic I, Bolotin-Fukuhara M, Thierry A, Bouchier C, Caudron B, Scarpelli C, Gaillardin C, Weissenbach J, Wincker P, Souciet J-L. Genome evolution in yeasts. Nature. 2004;430(6995):35–44.
51. Sacerdot C, Casaregola S, Lafontaine I, Tekaia F, Dujon B, Ozier-kalogeropoulos O. Promiscuous DNA in the nuclear genomes of hemiascomycetous yeasts. FEMS Yeast Res. 2008;8:846–57.
52. Gnerre S, Maccallum I, Przybylski D, Ribeiro FJ, Burton JN, Walker BJ, Williams L, Nicol R, Gnirke A, Nusbaum C, Lander ES, Jaffe DB, Sharpe T, Hall G, Shea TP, Sykes S, Berlin AM, Aird D, Costello M, Daza R, Williams L, Nicol R, Gnirke A, Nusbaum C, Lander ES, Jaffe DB. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. PNAS. 2011;108(4):1513–8.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

BMC

References

Aken, Bronwen L. et al. (2016). "The Ensembl gene annotation system". *Database* **2016** (baw093).

Alexandersson, Marina, Simon Cawley, and Lior Pachter (2003). "SLAM: Cross-species gene finding and alignment with a generalized pair hidden Markov model". *Genome Research* **13**, pp. 496–502.

Alföldi, Jessica, Federica Di Palma, and Kerstin Lindblad-toh (2009). "The European rabbit genome". In: *Rabbit Biotechnology: Rabbit Genomics, Transgenesis, Cloning and Models*. Chap. 11, p. 129.

Alkan, Can, Saba Sajjadian, and Evan E. E. Eichler (2011). "Limitations of next-generation genome sequence assembly". *Nature Methods* **8** (1), pp. 61–65.

Allen, Jonathan E. and Steven L. Salzberg (2005). "JIGSAW: Integration of multiple sources of evidence for gene prediction". *Bioinformatics* **21** (18), pp. 3596–3603.

Altschul, S F et al. (1990). "Basic local alignment search tool". *Journal of Molecular Biology* **215** (3), pp. 403–410.

Andrews, Simon (2010). *FastQC: A quality control tool for high throughput sequence data*. http://www.bioinformatics.babraham.ac.uk/projects/fastqc.

Austin, Christopher M. et al. (2017). "*De novo* genome assembly and annotation of Australia's largest freshwater fish, the Murray cod (*Maccullochella peelii*) from Illumina and Nanopore sequencing read". *GigaScience* **6** (8), pp. 1–6.

Barrera, Nelson P. and Carol V. Robinson (2011). "Advances in the mass spectrometry of membrane proteins: From individual proteins to intact complexes". *Annual Review of Biochemistry* **80** (1), pp. 247–271.

Bennett, Michael D. et al. (2003). "Comparisons with *Caenorhabditis* (∼100 Mb) and *Drosophila* (∼175 Mb) using flow cytometry show genome size in *Arabidopsis* to be ∼157 Mb and thus ∼25% larger than the Arabidopsis Genome Initiative estimate of ∼125 Mb". *Annals of Botany* **91** (5), pp. 547–557.

Benson, Dennis A. et al. (2013). "GenBank". *Nucleic Acids Research* **41**, pp. D36–D42.

Bentley, David R. et al. (2008). "Accurate whole human genome sequencing using reversible terminator chemistry". *Nature* **456** (7218), pp. 53–59.

Bernal, Axel et al. (2007). "Global discriminative learning for higher-accuracy computational gene prediction". *PLoS Computational Biology* **3** (3), pp. 0488–0497.

Böck, A. et al. (1991). "Selenocysteine: The 21st amino acid". *Molecular Microbiology* **5** (3), pp. 515–520.

Boley, Nathan et al. (2014). "Genome-guided transcript assembly by integrative analysis of RNA sequence data". *Nature Biotechnology* **32** (4), pp. 341–346.

Bork, P and E V Koonin (1998). "Predicting functions from protein sequences – where are the bottlenecks?" *Nature Genetics* **18** (4), pp. 313–318.

Borodovsky, Mark and James Mcininch (1993). "GeneMark: Parallel gene recognition for both DNA strands". *Computers & Chemistry* **17** (2), pp. 123–133.

Boulesteix, Matthieu, Michèle Weiss, and Christian Biémont (2006). "Differences in genome size between closely related species: The *Drosophila melanogaster* species subgroup". *Molecular Biology and Evolution* **23** (1), pp. 162–167.

Bray, Nicolas L. et al. (2016). "Near-optimal probabilistic RNA-seq quantification". *Nature Biotechnology* **34** (5), pp. 525–527.

Brenner, Steven E. (1999). "Errors in genome annotation". *Trends in Genetics* **15** (4), pp. 132–133.

Burge, C (1997). "Identification of complete gene structure in human genomic DNA". *Journal of Molecular Biology* **268**, pp. 78–94.

Burgess, Bindu Nanduri et al. (2010). "Gene model detection using mass spectrometry". In: *Proteome Bioinformatics. Methods in Molecular Biology (vol. 604)*. Ed. by Simon Hubbard and Andrew Jones. Humana Press, pp. 137–144.

Burset, M, I A Seledtsov, and V V Solovyev (2000). "Analysis of canonical and non-canonical splice sites in mammalian genomes". *Nucleic Acids Research* **28** (21), pp. 4364–4375.

Cantarel, Brandi L. et al. (2008). "MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes". *Genome Research* **18** (1), pp. 188–196.

Cao, H et al. (2014). "Rapid detection of structural variation in a human genome using nanochannel-based genome mapping technology". *Gigascience* **3** (34).

Casaregola, Serge et al. (2000). "Genomic exploration of the hemiascomycetous yeasts: 17. *Yarrowia lipolytica*". *FEBS Letters* **487** (1), pp. 95–100.

Casola, Claudio and Tomasz E Koralewski (2017). "Pinaceae show elevated rates of gene duplication and gene loss that are robust to incomplete gene annotation". *bioRxiv*, https://doi.org/10.1101/192203.

CCDS (2018). *CCDS releases and statistics*. https://tinyurl.com/ccds-statistics (CCDS website).

Claros, Manuel Gonzalo et al. (2012). "Why assembling plant genome sequences is so challenging". *Biology* **1** (2), pp. 439–459.

Conesa, Ana et al. (2016). "A survey of best practices for RNA-seq data analysis". *Genome Biology* **17** (13).

Cooper, Gregory M et al. (2008). "Qualifying the relationship between sequence conservation and molecular function". *Genome Research* **18**, pp. 201–205.

Cox, Jürgen and Matthias Mann (2011). "Quantitative, high-resolution proteomics for data-driven systems biology". *Annual Review of Biochemistry* **80**, pp. 273–299.

Curwen, Val et al. (2004). "The Ensembl automatic gene annotation system". *Genome Research* **14** (5), pp. 942–950.

Dayhoff, M.O., R.M. Schwartz, and B.C. Orcutt (1978). "A model of evolutionary change in proteins". In: *Atlas of Protein Sequence and Structure*, pp. 345–352.

De Smet, R. et al. (2013). "Convergent gene loss following gene and genome duplications creates single-copy families in flowering plants". *Proceedings of the National Academy of Sciences* **110** (8), pp. 2898–2903.

Decaprio, David et al. (2007). "Conrad : Gene prediction using conditional random fields". *Genome Research* **17** (9), pp. 1389–1398.

Denton, James F. et al. (2014). "Extensive error in the number of genes inferred from draft genome assemblies". *PLoS Computational Biology* **10** (12), e1003998.

Deutsch, Michael and Manyuan Long (1999). "Intron-exon structures of eukaryotic model organisms". *Nucleic Acids Research* **27** (15), pp. 3219–3228.

Devos, Damien and Alfonso Valencia (2001). "Intrinsic errors in genome annotation". *Trends in Genetics* **17** (8), pp. 429–431.

Dietrich, Fred S et al. (2004). "The *Ashbya gossypii* genome as a tool for mapping the ancient *Saccharomyces cerevisiae* genome". *Science* **304** (5668), pp. 304–307.

Doležel, Jaroslav and Jan Bartoš (2005). "Plant DNA flow cytometry and estimation of nuclear genome size". *Annals of Botany* **95** (1), pp. 99–110.

Dönnes, Pierre and Annette Höglund (2004). "Predicting protein subcellular localization: past, present, and future." *Genomics Proteomics Bioinformatics* **2** (4), pp. 209–215.

Dujon, Bernard et al. (2004). "Genome evolution in yeasts". *Nature* **430** (6995), pp. 35–44.

Dunne, Michael and Steven Kelly (2018). "OMGene: Mutual improvement of gene models through optimisation of evolutionary conservation". *BMC Genomics* **19** (307).

Dunne, Michael P. and Steven Kelly (2017). "OrthoFiller: utilising data from multiple species to improve the completeness of genome annotations". *BMC Genomics* **18** (390).

Dwight, Selina S et al. (2004). "*Saccharomyces* Genome Database: Underlying principles and organisation". *Briefings in Bioinformatics* **5** (1), pp. 9–22.

Eddy, Sean R (2011). "Accelerated profile HMM searches". *PLoS Computational Biology* **7** (10), e1002195.

Edgar, Robert C. (2004). "MUSCLE: Multiple sequence alignment with high accuracy and high throughput". *Nucleic Acids Research* **32** (5), pp. 1792–1797.

Edman, P (1950). "Method for determination of the amino acid sequence in peptides". *Acta Chemica Scandinavica* **4**, pp. 283–293.

Eilbeck, Karen et al. (2005). "The Sequence Ontology: A tool for the unification of genome annotations". *Genome Biology* **6** (5), R44.

Eilbeck, Karen et al. (2009). "Quantitative measures for the management and comparison of annotated genomes". *BMC Bioinformatics* **10** (67).

Elsik, Christine G. et al. (2007). "Creating a honey bee consensus gene set". *Genome Biology* **8** (1), R13.

Elsik, Christine G. G. et al. (2006). "Community annotation: Procedures, protocols, and supporting tools". *Genome Research* **16** (11), pp. 1329–1333.

Emanuelsson, O et al. (2000). "Predicting subcellular localization of proteins based on their N-terminal amino acid sequence". *Journal of Molecular Biology* **300** (4), pp. 1005–1016.

Emms, David M. and Steven Kelly (2015). "OrthoFinder: Solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy". *Genome Biology* **16** (157).

English, Adam C. et al. (2012). "Mind the gap: Upgrading genomes with Pacific Biosciences RS long-read sequencing technology". *PLoS ONE* **7** (11), e47768.

Fiers, W. et al. (1976). "Complete nucleotide sequence of bacteriophage MS2 RNA: Primary and secondary structure of the replicase gene". *Nature* **260**, pp. 500–507.

Finn, Robert D. et al. (2014). "Pfam: The protein families database". *Nucleic Acids Research* **42** (Database issue), pp. D222–D230. arXiv: `arXiv:1011.1669v3`.

Florea, Liliana et al. (1998). "A computer program for aligning a cDNA sequence with a genomic DNA sequence". *Genome Research* **8** (814), pp. 967–974.

Foissac, Sylvain et al. (2008). "Genome annotation in plants and fungi: EuGène as a model platform". *Current Bioinformatics* **3** (2), pp. 87–97.

Galagan, James E et al. (2005). "Sequencing of *Aspergillus nidulans* and comparative analysis with *A. fumigatus* and *A. oryzae*". *Nature* **438** (7071), pp. 1105–1115.

Garalde, Daniel R et al. (2018). "Highly parallel direct RNA sequencing on an array of nanopores". *Nature Methods* **15** (July 2017), pp. 201–206.

García-Ortega, Luis Fernando and Octavio Martínez (2015). "How many genes are expressed in a transcriptome? Estimation and results for RNA-Seq". *PLoS ONE* **10** (6), e0130262.

GenBank (2017). *GenBank and WGS Statistics*. https://www.ncbi.nlm.nih.gov/genbank/statistics/.

Génolevures Consortium (2009). "Comparative genomics of protoploid Saccharomycetaceae". *Genome Research*, pp. 1696–1709.

Gnerre, Sante et al. (2011). "High-quality draft assemblies of mammalian genomes from massively parallel sequence data". *Proceedings of the National Academy of Sciences* **108** (4), pp. 1513–1518.

Goodstein, David M. et al. (2012). "Phytozome: A comparative platform for green plant genomics". *Nucleic Acids Research* **40** (Database issue), pp. D1178–D1186.

Green, M. L. and P. D. Karp (2005). "Genome annotation errors in pathway databases due to semantic ambiguity in partial EC numbers". *Nucleic Acids Research* **33** (13), pp. 4035–4039.

Greenleaf, William J. and Arend Sidow (2014). "The future of sequencing: Convergence of intelligent design and market Darwinism". *Genome Biology* **15** (303).

Gregory, T. Ryan et al. (2007). "Eukaryotic genome size databases". *Nucleic Acids Research* **35** (Database issue), pp. D332–D338.

Griffith, Malachi and Obi Griffith (2018). *RegTools*. https://regtools.readthedocs.io.

Grigoriev, Igor V. et al. (2014). "MycoCosm portal: Gearing up for 1000 fungal genomes". *Nucleic Acids Research* **42** (Database issue), pp. D699–D704.

Gross, Samuel S. and Michael R. Brent (2006). "Using multiple alignments to improve gene prediction". *Journal of Computational Biology* **13** (2), pp. 379–393.

Gross, Samuel S. et al. (2007). "CONTRAST: A discriminative, phylogeny-free approach to multiple informant *de novo* gene prediction". *Genome Biology* **8**, R269.

Gu, Zhenglong et al. (2004). "Duplicate genes increase gene expression diversity within and between species". *Nature Genetics* **36**, pp. 577–579.

Guigó, R et al. (2000). "An assessment of gene prediction accuracy in large DNA sequences." *Genome Research* **10** (10), pp. 1631–1642.

Gurevich, Alexey et al. (2013). "QUAST: Quality assessment tool for genome assemblies". *Bioinformatics* **29** (8), pp. 1072–1075.

Haas, Brian J. et al. (2008). "Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments". *Genome Biology* **9** (1), R7.

Haas, Brian J. et al. (2011). "Approaches to fungal genome annotation". *Mycology* **2** (3), pp. 118–141.

Hagberg, Aric A, Daniel A Schult, and Pieter J Swart (2008). "Exploring network structure, dynamics, and function using NetworkX". *Proceedings of the 7th Python in Science Conference (SciPy2008)*, pp. 11–16.

Han, Mira V. et al. (2013). "Estimating gene gain and loss rates in the presence of error in genome assembly and annotation using CAFE 3". *Molecular Biology and Evolution* **30** (8), pp. 1987–1997.

Hansen, Kasper D, Steven E Brenner, and Sandrine Dudoit (2010). "Biases in Illumina transcriptome sequencing caused by random hexamer priming". *Nucleic Acids Research* **38** (12), e131.

Hecht, Ariel et al. (2017). "Measurements of translation initiation from all 64 codons in *E. coli*". *Nucleic Acids Research* **45** (7), pp. 3615–3626.

Henderson, J and S Salzberg (1997). "Finding genes in DNA with a hidden Markov model". *Journal of Computational Biology* **4** (2), pp. 127–141.

Henikoff, S. and J. G. Henikoff (1992). "Amino acid substitution matrices from protein blocks." *Proceedings of the National Academy of Sciences* **89** (22), pp. 10915–10919.

Henikoff, S and J G Henikoff (1993). "Performance evaluation of amino acid substitution matrices." *Proteins: Structure, Function, and Genetics* **17** (1), pp. 49–61.

Holt, Carson and Mark Yandell (2011). "MAKER2 : An annotation pipeline and genome-database management tool for second- generation genome projects". *BMC Bioinformatics* **12** (491).

Hong, Xin, Douglas G. Scofield, and Michael Lynch (2006). "Intron size, abundance, and distribution within untranslated regions of genes". *Molecular Biology and Evolution* **23** (12), pp. 2392–2404.

Howe, K L, T Chothia, and R Durbin (2002). "GAZE: A generic framework for the integration of gene-prediction data by dynamic programming". *Genome Research* **12** (9), pp. 1418–1427.

Jacques, Noémie et al. (2010). "Population polymorphism of nuclear mitochondrial DNA insertions reveals widespread diploidy associated with loss of heterozygosity in *Debaryomyces hansenii*". *Eukaryotic Cell* **9** (3), pp. 449–459.

Jain, Miten et al. (2017). "Nanopore sequencing and assembly of a human genome with ultra-long reads". *Nature Biotechnology* **36** (4), pp. 338–345.

Jiang, Hui and Julia Salzman (2015). "A penalized likelihood approach for robust estimation of isoform expression". *Statistics and Its Interface* **8** (4), pp. 437–445.

Jiang, Huifeng, Lin Xu, and Zhenglong Gu (2011). "Growth of novel epistatic interactions by gene duplication". *Genome Biology and Evolution* **3**, pp. 295–301.

Kapusta, Aurélie, Alexander Suh, and Cédric Feschotte (2017). "Dynamics of genome size evolution in birds and mammals". *Proceedings of the National Academy of Sciences* **114** (8), E1460–E1469.

Katoh, Kazutaka and Daron M. Standley (2013). "MAFFT multiple sequence alignment software version 7: Improvements in performance and usability". *Molecular Biology and Evolution* **30** (4), pp. 772–780.

Kim, D, B Langmead, and S Salzberg (2017). *HISAT2: Graph-based alignment of next-generation sequencing reads to a population of genomes.* http://ccb.jhu.edu/software/hisat2/index.shtml.

Kim, Daehwan, Ben Langmead, and Steven L Salzberg (2015). "HISAT: A fast spliced aligner with low memory requirements". *Nature Methods* **12** (4), pp. 357–360.

Kitzman, Jacob O. (2016). "Haplotypes drop by drop". *Nature Biotechnology* **34** (3), pp. 296–298.

Korf, Ian (2004). "Gene finding in novel genomes". *BMC Bioinformatics* **5** (59).

Kuhn, H. W. (1955). "The Hungarian method for the assignment problem". *Naval Research Logistics* **2** (1-2), pp. 83–97.

Lamesch, Philippe et al. (2012). "The Arabidopsis Information Resource (TAIR): Improved gene annotation and new tools". *Nucleic Acids Research* **40** (Database issue), pp. D1202–1210.

Lander, E S et al. (2001). "Initial sequencing and analysis of the human genome". *Nature* **409** (6822), pp. 860–921.

Lee, Tae Ho et al. (2013). "PGDD: A database of gene and genome duplication in plants". *Nucleic Acids Research* **41** (Database issue), pp. D1152–D1158.

Leinonen, Rasko, Hideaki Sugawara, and Martin Shumway (2010). "The sequence read archive". *Nucleic Acids Research* **39** (Database issue), pp. D19–D21.

Lewis, S E et al. (2002). "Apollo: A sequence annotation editor." *Genome Biology* **3** (12), research0082.1–0082.14.

Li, Heng (2012). *Seqtk: Toolkit for processing sequences in FASTA/Q formats.* https://github.com/lh3/seqtk.

Li, Heng et al. (2009). "The Sequence Alignment/Map format and SAMtools". *Bioinformatics* **25** (16), pp. 2078–2079.

Lindblad-Toh, Kerstin et al. (2005). "Genome sequence, comparative analysis and haplotype structure of the domestic dog". *Nature* **438** (7069), pp. 803–819.

Liu, Qian et al. (2008). "Evigan: A hidden variable model for integrating gene evidence for eukaryotic gene prediction". *Bioinformatics* **24** (5), pp. 597–605.

Liu, Shengyi et al. (2014). "The *Brassica oleracea* genome reveals the asymmetrical evolution of polyploid genomes". *Nature Communications* **5** (3930).

Lynch, Michael and John S. Conery (2003). "The origins of genome complexity". *Science* **302** (5649), pp. 1401–1404.

Majoros, W. H., M. Pertea, and S. L. Salzberg (2004). "TigrScan and GlimmerHMM: Two open source *ab initio* eukaryotic gene-finders". *Bioinformatics* **20** (16), pp. 2878–2879.

Marchler-Bauer, Aron et al. (2011). "CDD: A conserved domain database for the functional annotation of proteins". *Nucleic Acids Research* **39** (Database issue), pp. D225–D229.

McTaggart, Alistair R. et al. (2016). "Fungal genomics challenges the dogma of name-based biosecurity". *PLoS Pathogens* **12** (5), pp. 10–12.

Mikkelsen, Tarjei S. et al. (2007). "Genome of the marsupial *Monodelphis domestica* reveals innovation in non-coding sequences". *Nature* **447** (7141), pp. 167–177.

Mills, Caitlyn L., Penny J. Beuning, and Mary Jo Ondrechen (2015). "Biochemical functional predictions for protein structures of unknown or uncertain function". *Computational and Structural Biotechnology Journal* **13**, pp. 182–191.

Milo, Ron and Rob Phillips (2018). *Cell biology by the numbers: How big are genomes?* http://book.bionumbers.org/how-big-are-genomes/.

Ming, Ray et al. (2008). "The draft genome of the transgenic tropical fruit tree papaya (*Carica papaya* Linnaeus)". *Nature* **452** (7190), pp. 991–996.

Moretti, Sébastien et al. (2014). "Selectome update: Quality control and computational improvements to a database of positive selection". *Nucleic Acids Research* **42** (Database issue), pp. D917–D921.

Motamayor, Juan C et al. (2013). "The genome sequence of the most widely cultivated cacao type and its use to identify candidate genes regulating pod color". *Genome Biology* **14**, r53.

Mudge, Jonathan M. and Jennifer Harrow (2016). "The state of play in higher eukaryote gene annotation". *Nature Reviews Genetics* **17** (12), pp. 758–772.

Muñoz, José Fernando et al. (2013). "The eukaryotic genome, its reads, and the unfinished assembly". *FEBS Letters* **587** (14), pp. 2090–2093.

Nasiri, Jaber et al. (2013). "Gene identification programs in bread wheat: A comparison study". *Nucleosides, Nucleotides and Nucleic Acids* **32** (10), pp. 529–554.

NCBI (2018a). *Eukaryotic genome annotation at NCBI*. https://www.ncbi.nlm.nih.gov/genome/annotation_euk/#graphs.

— (2018b). *How to submit data to GenBank*. https://www.ncbi.nlm.nih.gov/genbank/submit/.

Neuvéglise, Cécile and Hugo Devillers. *Genome resources for yeast chromosomes*. http://gryc.inra.fr/.

NHGRI (2004). *Saccharomyces Genome Database*. https://www.yeastgenome.org.

Otto, Thomas D. et al. (2011). "RATT: Rapid annotation transfer tool". *Nucleic Acids Research* **39** (9), e57.

Pachter, Lior (2012). "A closer look at RNA editing". *Nature Biotechnology* **30** (3), pp. 246–247.

Parra, G, E Blanco, and R Guigó (2000). "GeneID in *Drosophila*". *Genome Research* **10** (4), pp. 511–515.

Parra, Genis, Keith Bradnam, and Ian Korf (2007). "Genome analysis CEGMA : a pipeline to accurately annotate core genes in eukaryotic genomes". *Bioinformatics* **23** (9), pp. 1061–1067.

Parton, Richard M. et al. (2014). "Subcellular mRNA localisation at a glance". *Journal of Cell Science* **127** (10), pp. 2127–2133.

Patro, Rob, Geet Duggal, and Carl Kingsford (2015). "Salmon: Accurate, versatile and ultrafast quantification from RNA-seq data using lightweight-alignment". *bioRxiv*, http://biorxiv.org/content/early/2015/06/27/021592.

Pearson, William R (2013). "Selecting the right similarity-scoring matrix." *Current Protocols in Bioinformatics* **43**, pp. 3.5.1–3.5.9.

Pelechano, Vicent and Lars M Steinmetz (2013). "Gene regulation by antisense transcription". *Nature Reviews Genetics* **14**, pp. 880–893.

Price, Morgan N., Paramvir S. Dehal, and Adam P. Arkin (2009). "FastTree: Computing large minimum evolution trees with profiles instead of a distance matrix". *Molecular Biology and Evolution* **26** (7), pp. 1641–1650.

— (2010). "FastTree 2 - Approximately maximum-likelihood trees for large alignments". *PLoS ONE* **5** (3), e9490.

Promponas, Vasilis J., Ioannis Iliopoulos, and Christos A. Ouzounis (2015). "Annotation inconsistencies beyond sequence similarity-based function prediction – phylogeny and genome structure". *Standards in Genomic Sciences* **10** (108).

Pruitt, Kim D., Tatiana Tatusova, and Donna R. Maglott (2007). "NCBI reference sequences (RefSeq): A curated non-redundant sequence database of genomes, transcripts and proteins". *Nucleic Acids Research* **35** (Database issue), pp. D61–D65.

Pruitt, Kim D. et al. (2009). "The consensus coding sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes". *Genome Research* **19**, pp. 1316–1323.

Quinlan, Aaron R and Ira M Hall (2010). "BEDTools: A flexible suite of utilities for comparing genomic features". *Bioinformatics* **26** (6), pp. 841–842.

Rahman, Atif and Lior Pachter (2013). "CGAL: Computing genome assembly likelihoods". *Genome Biology* **14** (1), R8.

Rhoads, Anthony and Kin Fai Au (2015). "PacBio sequencing and its applications". *Genomics, Proteomics and Bioinformatics* **13** (5), pp. 278–289.

Rigby, R and D Stasinopoulos (2005). "Generalized additive models for location scale and shape". *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **54** (3), pp. 507–554.

Robert, Christelle and Mick Watson (2015). "Errors in RNA-seq quantification affect genes of relevance to human disease". *Genome Biology* **16** (177).

Rutherford, Kim et al. (2000). "Artemis: Sequence visualization and annotation". *Bioinformatics* **16** (10), pp. 944–945.

Sacerdot, Christine et al. (2008). "Promiscuous DNA in the nuclear genomes of hemiascomycetous yeasts". *FEMS Yeast Research* **8**, pp. 846–857.

Salamov, Asaf A et al. (2000). "*Ab initio* gene finding in *Drosophila* genomic DNA". *Genome Research* **10** (4), pp. 516–522.

Salzberg, Steven L. (2007). "Genome re-annotation: A wiki solution?" *Genome Biology* **8** (102).

Schatz, Michael C., Arthur L. Delcher, and Steven L. Salzberg (2010). "Assembly of large genomes using second-generation sequencing". *Genome Research* **20** (9), pp. 1165–1173.

Schneider, Gisbert and Uli Fechner (2004). "Advances in the prediction of protein targeting signals." *Proteomics* **4** (6), pp. 1571–1580.

Schnoes, Alexandra M. et al. (2009). "Annotation error in public databases: Misannotation of molecular function in enzyme superfamilies". *PLoS Computational Biology* **5** (12), e1000605.

Schweikert, Gabriele et al. (2009). "mGene: Accurate SVM-based gene finding with an application to nematode genomes". *Genome Research* **19** (11), pp. 2133–2143.

Simão, Felipe A et al. (2015). "BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs". *Bioinformatics* **31** (19), pp. 3210–3212.

Sjölander, Kimmen et al. (2011). "Ortholog identification in the presence of domain architecture rearrangement". *Briefings in Bioinformatics* **12** (5), pp. 413–422.

Slater, Guy and Ewan Birney (2005). "Automated generation of heuristics for biological sequence comparison". *BMC Bioinformatics* **6** (31).

Slotte, Tanja et al. (2013). "The *Capsella rubella* genome and the genomic consequences of rapid mating system evolution". *Nature Genetics* **45** (7), pp. 831–835.

Souvorov, A (2018). *Gnomon – The NCBI eukaryotic gene prediction tool*. https://www.ncbi.nlm.nih.gov/genome/annotation_euk/gnomon/.

Stanke, Mario and Burkhard Morgenstern (2005). "AUGUSTUS: A web server for gene prediction in eukaryotes that allows user-defined constraints". *Nucleic Acids Research* **33** (Web Server issue), W465–W467.

Starck, Shelley R et al. (2012). "Leucine-tRNA initiates at CUG start codons for protein synthesis and presentation by MHC class I". *Science* **336** (6089), pp. 1719–1723.

Sulston, J. E. and S. Brenner (1974). "The DNA of *Caenorhabditis elegans*." *Genetics* **77** (1), pp. 95–104.

Syed, Naeem H. et al. (2012). "Alternative splicing in plants – coming of age". *Trends in Plant Science* **17** (10), pp. 616–623.

Tang, Haibao (2017). "Disentangling a polyploid genome". *Nature Plants* **3** (9), pp. 688–689.

Tang, Haibao and Eric Lyons (2012). "Unleashing the genome of *Brassica* rapa". *Frontiers in Plant Science* **3** (172).

Tarazona, Sonia et al. (2011). "Differential expression in RNA-seq: A matter of depth". *Genome Research* **21** (12), pp. 2213–2223.

The Arabidopsis Genome Initiative (2000). "Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*". *Nature* **408** (6814), pp. 796–815.

Thibaud-Nissen, Françoise et al. (2013). "Eukaryotic Genome Annotation Pipeline". In: *The NCBI Handbook [Internet]. 2nd edition*, http://www.ncbi.nlm.nih.gov/books/NBK169439/.

Thomma, Bart P.H.J. P.H.J. et al. (2016). "Mind the gap; seven reasons to close fragmented genome assemblies". *Fungal Genetics and Biology* **90**, pp. 24–30.

Triant, Deborah A. and William R. Pearson (2015). "Most partial domains in proteins are alignment and annotation artifacts". *Genome Biology* **16** (99).

Tuskan, GA et al. (2006). "The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray)". *Science* **313** (5793), pp. 1596–1605.

Usuka, Jonathan, Wei Zhu, and Volker Brendel (2000). "Optimal spliced alignment of homologous cDNA to a genomic DNA template". *Bioinformatics* **16** (3), pp. 203–211.

Veeckman, Elisabeth, Tom Ruttink, and Klaas Vandepoele (2016). "Are we there yet? Reliably estimating the completeness of plant genome sequences". *Plant Cell* **28** (8), pp. 1759–1768.

Vogt, Gerhard, Thure Etzold, and Patrick Argos (1995). "An assessment of amino acid exchange matrices in aligning protein sequences: The twilight zone revisited". *Journal of Molecular Biology* **249** (4), pp. 816–831.

Wang, Liguo et al. (2016). "Measure transcript integrity using RNA-seq data". *BMC Bioinformatics* **17** (58).

Wang, Zhuo, Yazhu Chen, and Yixue Li (2004). "A brief review of computational gene prediction methods." *Genomics Proteomics Bioinformatics* **2** (4), pp. 216–221.

Wernersson, Rasmus and Anders Gorm Pedersen (2003). "RevTrans: Multiple alignment of coding DNA from aligned amino acid sequences". *Nucleic Acids Research* **31** (13), pp. 3537–3539.

Wetterstrand, KA (2016). *DNA sequencing costs: Data from the NHGRI genome sequencing program (GSP).* www.genome.gov/sequencingcosts.

Wiehe, Thomas and S Gebauer-Jung (2001). "SGP-1: Prediction and validation of homologous genes based on sequence alignments". *Genome Research* **11**, pp. 1574–1583.

Wilhelm, J. (2003). "Real-time PCR-based method for the estimation of genome sizes". *Nucleic Acids Research* **31** (10), e56.

Wingfield, Brenda D. et al. (2016). "Draft genome sequences of *Armillaria fuscipes*, *Ceratocystiopsis minuta*, Ceratocystis adiposa, *Endoconidiophora laricicola*, *E. polonica* and *Penicillium freii* DAOMC 242723". *IMA Fungus* **7** (1), pp. 217–227.

Wood, V et al. (2002). "The genome sequence of Schizo*Saccharomyces* pombe". *Nature* **415** (6874), pp. 871–880.

Wood, Valerie et al. (2012). "PomBase: A comprehensive online resource for fission yeast". *Nucleic Acids Research* **40** (Database issue), pp. D695–D699.

Wu, G Albert et al. (2014). "Sequencing of diverse mandarin, pummelo and orange genomes reveals complex history of admixture during citrus domestication". *Nature Biotechnology* **32** (7), pp. 656–662.

Wu, Yi Chieh, Matthew D. Rasmussen, and Manolis Kellis (2012). "Evolution at the subgene level: Domain rearrangements in the *Drosophila* phylogeny". *Molecular Biology and Evolution* **29** (2), pp. 689–705.

Xiong, Zijun et al. (2016). "Draft genome of the leopard gecko, *Eublepharis macularius*". *GigaScience* **5** (47).

Yandell, Mark and Daniel Ence (2012). "A beginner's guide to eukaryotic genome annotation". *Nature Reviews Genetics* **13** (5), pp. 329–342.

Yano, Yoshihisa et al. (2004). "A new role for expressed pseudogenes as ncRNA: Regulation of mRNA stability of its homologous coding gene". *Journal of Molecular Medicine* **82** (7), pp. 414–422.

Yeh, Ru-fang, Lee P Lim, and Christopher B Burge (2001). "Computational inference of homologous gene structures in the human genome". *Genome Research* **11**, pp. 803–816.

Zallot, Rémi et al. (2016). "Functional annotations of paralogs: A blessing and a curse". *Life* **6** (39).

Zhang, Simo V., Luting Zhuo, and Matthew W. Hahn (2016). "AGOUTI: Improving genome assembly and annotation using transcriptome data". *GigaScience* **5** (31).

Zhao, Chunzhao et al. (2013). "RNA-Seq analysis reveals new gene models and alternative splicing in the fungal pathogen *Fusarium graminearum*". *BMC Genomics* **14** (21).

Zhao, Shanrong and Baohong Zhang (2015). "A comprehensive evaluation of Ensembl, RefSeq, and UCSC annotations in the context of RNA-seq read mapping and gene quantification". *BMC Genomics* **16** (97).

Zmasek, Christian M. and Adam Godzik (2011). "Strong functional patterns in the evolution of eukaryotic genomes revealed by the reconstruction of ancestral protein domain repertoires". *Genome Biology* **12** (1), R4.

