Thesis submitted for the degree of Doctor of Philosophy

A method for the identification of biological pathways

Frantisek Honti

St Cross College, University of Oxford

Trinity term 2014

Abstract

Plenty of gene variants have been associated with disease, indicating widespread genetic heterogeneity, which leaves the molecular basis of complex diseases unclear. However, it is widely postulated that the products of genes whose mutations are implicated in the same disease function together in the same biological pathways and it is the disruption of these pathways that underlies the disease. Such pathways are not well defined and their identification could help elucidate disease mechanisms. To discern molecular pathways of relevance to complex disease, I have inferred functional associations between human genes from diverse data types and assessed these associations with a novel phenotype-based method. I could confirm the hypothesis that dysfunctions of genes associated with each other in terms of functional genomic and proteomic data tend to give rise to the same disease. Examining the functional association between disease-associated gene variants, I have found that genes implicated through *de novo* sequence variants are biased in their coding sequence length and that longer genes tend to cluster together in gene networks, leading to exaggerated p-values in functional studies. I have controlled for the confounding bias and, testing different data sources, found that an integrated phenotypic-linkage network offers superior power to detect functional associations among genes mutated in the same disease. Applying these methods to clinical phenotypes related to intellectual disability, I have observed an increased predictive potential in identifying

genes associated with these phenotypes. I have also performed case–control association analyses of variants from an exome-sequencing study of Parkinson's disease and tested the functional associations of the mutated genes. I have advanced a framework for the identification of biological pathways disrupted in complex disorders, also demonstrating the suitability of this method to functionally sub-cluster the gene variants underlying a complex disorder, with implications for the understanding of disease mechanisms.

Acknowledgments

I am grateful to my interviewers, Chris Ponting, Caleb Webber, Kay Davies and Ji-Long Liu, for the opportunities they have provided me, to my supervisors, Caleb Webber and Chris Ponting, for their support and to Viv Collins and Emily Frape for their help.

I thank Julia Steinberg, Tallulah Andrews, Steve Meader and Cynthia Sandor for valued debates, David Sims and Vlada Chalei for playing tennis with me, Giuseppe Gallone and Claudia Wehrspaun for regularly attending my film club, Man-Suen Chan for the IT support and Ana Marques for letting me revise this thesis on her payroll.

I am also grateful to the UK Medical Research Council and the EU FP7 GENCODYS programme for financial support.

Table of contents

Chapter 1: Introduction

1.1 Historical considerations

The causes of birth defects may have always puzzled people. Most children resembled their parents, but why were sometimes disabled children born to healthy parents? Hippocrates found the fault in a constriction of the womb or a contusion suffered by the mother (Lonie and Hippocrates 1981). He also stated that all body parts contributed to the seminal fluid which thus transmitted the characteristics of all parts, sometimes including their incidental abnormalities. This notion was still shared by Charles Darwin, who explained diversity as a consequence of changes in life conditions affecting the reproductive system and thus inducing variability (Darwin 1859).[1] When John Langdon Down described the congenital disorder now known as Down syndrome, he believed it to be caused by tuberculosis in the parents (Down 1866).

The term mutation, introduced by Hugo de Vries, initially referred to sudden leaps giving rise to new species and varieties from existing forms. Concurrently with the re-discovery of Gregor Mendel's laws of heredity, chromosomes were found to be the bearers of hereditary factors and the subjects of mutation (Morgan 1915; Morgan 1925). Thomas Hunt Morgan's studies of fruit flies clarified how the maneuvers of the

[1] The transmission of acquired characters was refuted by August Weismann. He found the source of individual variability in sexual reproduction, deeming accidental alterations in the molecular structure of the "germ-plasm" to have no share in the production of hereditary characteristics. See Weismann A. (1889) *Essays Upon Heredity*. Clarendon Press, Oxford. Vol. I, p. 271.

chromosomes during cell division led to the inheritance of wing and eye phenotypes according to Mendel's laws and linked inheritance. They also observed hundreds of new mutations – some extensive, some barely noticeable – that, unless lethal, were transmitted to the descendants.

Human malformations were shown to be heritable similarly to the mutations observed in animals and plants. In his book, Inborn Errors of Metabolism, Archibald Garrod described six disorders and discussed their incidence and heredity. He noticed that although they were very rare in the general population, they tended to occur in several members of a family. He noted that "there appears to be a close connexion between the occurrence of an anomaly in several children of normal parents and consanguinity of the parents," and that "the mode of incidence of alcaptonuria finds a ready explanation if the anomaly in question be regarded as a rare recessive character in the Mendelian sense" (Garrod 1923, p. 21).

The Mendelian inheritance of many inborn disorders was confirmed in the next decades and the accumulation of knowledge on these prompted their classification. Victor McKusick catalogued human disorders and phenotypes in his encyclopedia, Mendelian Inheritance in Man, distinguishing autosomal dominant, autosomal recessive and X-linked traits (McKusick 1966). Although DNA had been identified as the genetic information-carrying component of chromosomes and its composition as well as structure had been solved, no direct knowledge of the individual genes was available yet. The identification and mapping of genes became possible with insights from linkage analyses and the discovery of chromosomal banding patterns and polymorphic DNA regions (Chial 2008). Subsequently, the development of molecular

cloning facilitated the invention of techniques for determining the nucleotide sequence of DNA fragments which enabled comparative analyses of genes.

The advance of DNA sequencing, fuelled by the Human Genome Project, has enabled the identification of numerous mutations in genes associated with human disorders. It has also facilitated the identification of orthologs of human genes in other organisms invaluable for functional characterisation studies. At the same time, the development of DNA microarrays has allowed the detection of copy-number variations often affecting multiple adjacent genes. The technological advances of recent times have greatly contributed to the elucidation of the genetic causes of human disorders, revealing a bewildering amount of deleterious sequence variants and structural changes that may leave one wondering how come birth defects are the exception, not the rule.

1.2 Prevalence and heritability of genetic disorders

It is estimated that about 70% of human conceptions do not develop successfully to term, mainly due to chromosomal instabilities (Macklon *et al*. 2002; Vanneste *et al*. 2009). Of those that survive until birth, approximately 2.5% are born with a recognisable malformation (Winter 1996). Further, about 18.6% of pediatric visits are by children with genetic disorders (Kumar *et al*. 2001). Many diseases have a considerable genetic component, the most common neurodevelopmental disorders being attention deficit hyperactivity disorder (ADHD), epilepsy, intellectual disability, autism and schizophrenia (**Table 1.1**).

Table 1.1: Estimated prevalence and heritability of common neurological disorders.

Disorder	Prevalence	Heritability
ADHD	5–7%[1]	76%[9]
Epilepsy	2–5%[2]	70–88% [10]
Intellectual disability	1–3%[3]	80%[11]
Autism	0.6–1.1%[4]	64–92%[12]
Schizophrenia	0.7%[5]	81%[13]
Bipolar disorder	1–2%[6]	73–77%[14]
Alzheimer's disease	1.6%[7]	58–79%[15]
Parkinson's disease	0.3%[8]	34–40%[16]

[1](Willcutt 2012), [2](Sander and Shorvon 1988), [3](Harris 2006), [4](Newschaffer *et al*. 2007), [5](McGrath *et al*. 2008), [6](Anderson *et al*. 2012), [7](Hebert *et al*. 2003), [8](de Lau and Breteler 2006), [9](Faraone *et al*. 2005), [10](Kjeldsen *et al*. 2001), [11](Plomin *et al*. 1994), [12](Ronald *et al*. 2006), [13](Sullivan *et al*. 2003), [14](Edvardsen *et al*. 2008), [15](Gatz *et al*. 2006), [16](Wirdefeldt *et al*. 2011)

Probably all diseases are influenced by both genetic and environmental factors, their relative importance representing how heritable a given disease is. Heritability, in statistical sense, is defined as the proportion of total variance in a population for a trait that is attributable to the variation in genetic factors (Visscher *et al*. 2008). It is estimated by comparing individuals with different degree of genetic relationship – often monozygotic and dizygotic twins – and hundreds or thousands of observations are necessary to obtain precise estimates. The high heritability of neurodevelopmental disorders establishes that they are brought about by genetic variation, the identification of which would be invaluable for diagnostics and therapeutics.

1.3 The genome and molecular causes of mutations

Heritable characteristics are shaped by the genome: the genetic composition of an organism. The human genome is made up of 23 pairs of chromosomes and we also carry copies of a small, circular mitochondrial chromosome in our cells. The main,

genetic information-carrying component of chromosomes is a double-stranded macromolecule of deoxyribonucleic acid, or DNA, in which the genetic information is encoded by the sequence of four nucleobases: adenine (A), guanine (G), thymine (T) and cytosine (C). Some segments of this sequence constitute genes which encode functional gene products, in the form of RNA molecules and proteins. About 1.5% of the human genome codes for proteins, the rest of the DNA sequence is termed non-coding, although it contains thousands of different genes for RNA molecules, as well as regulatory regions.

The genome is variable, both between and within individuals. The difference between the genomes of two persons can reach 1%, owing to structural variation affecting long pieces of DNA sequence in the form of insertions, deletions, block substitutions, inversions and copy number variants (CNVs) (Frazer *et al*. 2009). The scale of genetic mutations ranges from whole-chromosome duplications and deletions to single nucleotide variants, also called point mutations. The genomes of distinct cells within an individual can also differ between different cell lineages as a result of mutations accumulating during our lifetime, leading to somatic mosaicism (De 2011). However, only the mutations occurring in the germline – such as the sperm or egg cells – get passed on to the offspring.

Mutations can occur randomly and spontaneously in the genome, but their incidence is increased by known mutagenic factors. External factors, such as exposure to ultraviolet (UV) radiation, nicotine and reactive oxygen species, can damage DNA and generate mutations. Certain viruses invade the genome of the infected cells and can disrupt genes and regulatory regions. In addition, some endogenous factors also

induce mutations, these include mobile or transposable elements, DNA polymerase slippage during DNA replication, imperfect DNA repair, recombination and unbalanced chromosomal segregation during cell division (De 2011).

Active mobile genetic elements such as transposons can produce copies of themselves and thus move within and expand the genome. It is estimated that about half of the human genome originates from transposable elements, most of this buildup having been inactivated over time (Cordaux and Batzer 2009). The most obvious mutagenic effect of transposons is their disruption of genes and regulatory regions during their propagation, when they insert themselves in a new genomic location. Besides disrupting the sequence, transposable elements can also introduce cryptic splice sites and transcription factor-binding sites and thus alter exon composition and gene expression, respectively (Cordaux and Batzer 2009).

The high sequence similarity of the many copies of transposons can also promote non-allelic homologous recombination between pairs of chromosomes during cell division. Genetic recombination is involved in DNA repair by the transfer of genetic information between homologous sequences at the same position on related chromosomes. Identical copies of mobile elements or regions containing segmental duplications can be mistaken for the same chromosomal position and thus misaligned, resulting in the formation of structural genomic variation, or genomic rearrangements (Hastings *et al*. 2009).

During DNA replication, the presence of nucleotide repeats in the copied sequence can lead to DNA polymerase slippage, causing a series of nucleotides to be duplicated or deleted. Likewise, if the DNA polymerase is held up for any reason and the replication

fork is halted, the newly synthesised DNA strand can align to another exposed template sequence in an incident called fork stalling and template switching (Hastings *et al*. 2009). This could lead to the formation of chromosomal rearrangements and CNVs. Even when their activity is not hindered, DNA polymerases are not infallible, with polymerase α being especially error-prone (Reijns *et al*. 2015).

Mutability is also not uniform throughout the genome (Michaelson *et al*. 2012). Local mutation rates are influenced by intrinsic factors such as the hypermutability of CG dinucleotides (Ellegren *et al*. 2003), replication timing (Francioli *et al*. 2015), chromatin organisation (Schuster-Bockler and Lehner 2012), nucleosome occupancy (Prendergast and Semple 2011), regulatory protein binding (Reijns *et al*. 2015) and DNA mismatch repair (Supek and Lehner 2015). All the above-mentioned factors influencing the distribution of mutations shape the genome during evolution through genetic variation, resulting in adaptation, but also in heritable disorders.

1.4 Association of gene variants with genetic disorders

As genetic variants are major contributors to disease heritability, the genotype of affected persons is expected to differ from that of unaffected ones. Some specific mutations must be more common in people with the same disease than in the unaffected population. Thus, to identify the genetic causes of a disease, the genotypes of cases and unaffected controls need to be compared. Then, if a specific mutation is significantly more frequent in the cases, it can be associated with the disease. Such case–control studies, often described as association studies, depend on accurate identification of mutations.

Large chromosomal aberrations, such as trisomy of chromosome 21, can be detected by karyotyping through staining and light microscopy. Deletions and translocations of pieces of chromosomes can be identified by fluorescent in situ hybridisation (FISH), marking specific parts of chromosomes with fluorescent probes. Besides, submicroscopic CNVs can be detected by means of DNA microarrays (Carter 2007). The use of microarrays for this purpose is based on hybridisation, the tendency of complementary sequences to anneal and bind to each other. DNA microarrays consist of numerous minute spots, each containing copies of specific, single-stranded deoxy-ribonucleotide sequence, which can bind complementary pieces of DNA tagged by fluorescent dyes. Then, the fluorescence level of sequences along a chromosome allows the detection of CNVs (Carter 2007; Graves 1999).

DNA microarrays are also used for the identification of common genetic variants called single-nucleotide polymorphisms (SNPs) (Hacia and Collins 1999). Such polymorphisms are the most abundant form of genetic variation and the development of microarrays for assaying hundreds of thousands of SNPs triggered an explosion of genome-wide association studies (GWAS). The SNPs found to be associated with a disease are not necessarily the causative mutations, rather, they mark a region of the genome which affects the risk of disease. SNPs can be considered as markers of alleles because of genetic linkage or linkage disequilibrium, meaning that nearby loci tend to be inherited together over generations.[1]

[1] Such co-inheritance patterns occur mainly as a result of bottlenecks in the population history of humans. Population bottlenecks resulted in a reduction of genetic variation and in the appearance of long, identical chromosomal sequence intervals in the descendants.

A notable example of a common gene variant with large effect is the *APOE* ε4 allele, which increases the risk of developing Alzheimer's disease severalfold (Liu *et al*. 2013). However, most of the identified risk factors are of small effect and they only explain a small fraction of the heritability of a common disease, thus their diagnostic and prognostic value is limited (Ropers 2007). The rarity of major risk factors is likely to be a consequence of selection, as even mildly deleterious mutations are unlikely to occur at high frequency in populations (Kryukov *et al*. 2007). Consequently, most of the genetic susceptibility to disease is expected to be imparted by rare variants, which are not detected by traditional GWAS.

Rare variants can be identified by DNA sequencing, which enables the direct detection of causal mutations. Sanger sequencing, developed by Sanger and colleagues in 1977, pioneered genomics and is still widely used. It is based on the use of DNA polymerase for *in vitro* replication of the studied DNA, applying chain-terminating dideoxynucleotides labeled radioactively or with a fluorescent dye (Sanger *et al*. 1977). The incorporation of such a dideoxynucleotide containing a known base in the synthesised sequence ends the extension of DNA, so that when the newly synthesised DNA fragments are separated by size during gel electrophoresis, the pattern of bands shows the order of nucleobases in the sequence (Sanger *et al*. 1977).

Many other sequencing techniques have become available by now, revolutionasing the field by processing millions of sequence reads in parallel (Mardis 2008). As a result of the development of these next-generation sequencing techniques, the cost and duration of sequencing a genome has decreased by orders of magnitude since the Human Genome Project, allowing the sequencing of large cohorts (Foo *et al*. 2012).

Whole-exome and whole-genome sequencing is superseding microarray-based genotyping methods, but some of the limitations of GWAS still apply to sequencing studies. Thousands of cases and controls are required to robustly associate rare variants with a common disease and population stratification has to be controlled for (Foo *et al*. 2012; Kiezun *et al*. 2012).

A popular approach to implicate rare variants in disease is to search for *de novo* (newly arising) mutations in parent–offspring trios (Veltman and Brunner 2012). *De novo* mutations are more likely to be pathogenic, as their effect on reproductive success has not been assessed yet, and thus they have not been thoroughly exposed to selection. Probands with autism were shown to have more than twice as many *de novo* loss-of-function mutations as their unaffected siblings (Krumm *et al*. 2014) and a significant excess of genes with two or more *de novo* missense mutations has also been found in cases with autism and intellectual disability (Samocha *et al*. 2014). The genetic heterogeneity observed in these studies suggests that *de novo* mutations of different genes in different individuals together could account for a part of common neurodevelopmental disorders.

It appears that the prevalence of different genetic disorders correlates with the number of genes that can cause the disorder when mutated. Rare, Mendelian disorders can be explained by mutations of a single gene, often identified as segregating with the disease in families. On the other hand, disorders that are caused by mutations in any of many different genes in different individuals can have high incidence, in spite of the severely reduced fitness and fecundity (Veltman and Brunner 2012). As numerous variants can contribute to the disease susceptibility in this case,

their identification is correspondingly more complicated by the likely presence of functionally neutral or unrelated mutations.

1.5 Functional interpretation of mutations

Missense variants, causing the substitution of an amino acid, are expected to be less deleterious in general than nonsense or other loss-of-function mutations. Correspondingly, the significant enrichment of *de novo* missense mutations in probands with autism and intellectual disability has been less pronounced than the enrichment of loss-of-function mutations, suggesting that a smaller proportion of missense mutations is pathogenic (Krumm *et al*. 2014; Samocha *et al*. 2014). However, missense mutations occur several times more often and knowledge about their functional consequences would facilitate the identification of disease-causing variants. Multiple factors can aid the distinction of pathogenic variants from neutral ones, including recurrence in cases, the predicted impact on the protein, effect in model organisms and membership of the mutated gene in an implicated molecular pathway.

The recurrence of mutations in the same gene in multiple affected individuals, accompanied by the absence of such mutations in unaffected controls, is the most decisive factor in the association of a gene variant to a disease (Hoischen *et al*. 2014). Because of genetic heterogeneity, often large samples are necessary for the robust association of a gene's variants to a disorder, although targeted resequencing techniques, such as molecular inversion probes, can reinforce the search for mutations (O'Roak *et al*. 2012a). Overlap of the mutated gene with disease-associated copy-number variations further increases the likelihood of a mutation being deleterious.

The impact of a missense mutation on the functionality of the encoded protein can be mediated through various physicochemical features, including macromolecular stability,[1] conformational dynamics,[2] binding affinity and other physiologically important properties (Stefl *et al*. 2013). The Grantham difference score represents the chemical dissimilarity between amino acids based on properties such as polarity and molecular volume (Grantham 1974). Other approaches, including SIFT and GERP, estimate the impact of a missense mutation by assessing the degree of sequence conservation from multiple alignments of homologous sequences (Cooper *et al*. 2005; Ng and Henikoff 2003). PolyPhen-2 predicts the severity of an amino acid change based on a number of features involving sequence conservation, structural properties and the position of the substitution within the protein (Adzhubei *et al*. 2010).

Knowledge of the gene's function can also support the association of a mutation with a disease, when the function is related to the disease physiology. A large amount of functional information is available from phenotypes of model organisms, applicable because of homology and the profound similarities of developmental processes and pathways among animals. Spontaneous mutations and targeted knockouts of mouse genes have been characterised and compiled in online databases such as the Mouse Genome Informatics (Eppig *et al*. 2012). The knockout phenotypes of orthologs of human genes do not always resemble the human symptoms, occasionally as a consequence of differences in genetic background, gene regulation, or gene function.

[1] Hydrogen bonds and hydrogen bond networks are among the most important determinants of protein stability and the root of the pH dependence of many reactions. Even a conservative mutation that results in the removal or addition of a hydrogen donor or acceptor can have a significant impact on the hydrogen bond network, potentially affecting an active-site region (Stefl *et al*. 2013).

[2] Biochemical reactions are often accompanied by conformational changes.

Membership of the affected gene in disease-associated biological pathways can provide further links to a disease, as well as offer clues to the disease mechanism. It may be considered a special case of recurrence when mutations concentrate in a functional network of genes in the affected probands. Pathways composed of molecular interactions and reactions are mapped and assembled in databases such as KEGG (Kanehisa *et al*. 2010) and Reactome (Croft *et al*. 2011), or can be inferred from functional genomics data (Honti *et al*. 2014). Sequencing studies of genetic disorders can contribute to the identification of biological pathways and vice versa, in mutual feedback (Krumm *et al*. 2014).

1.6 Molecular disease mechanisms

Understanding the processes by which genetic mutations cause a disorder would be invaluable in guiding efforts to treat the condition. Gene products are known to function in concert, interacting in supramolecular complexes or through series of reactions, and the dysfunction of a gene should be interpreted in this context. As the convergence of mutations in genes suggests their association with disease, the molecular convergence of the affected gene products in a pathway implicates the dysfunction of the pathway in a genetically heterogeneous disorder.

The understanding of disease mechanisms therefore requires the study of pathways, their identification and functional characterisation. Achievement of the mentioned goals is hindered by the circumstance that biological pathways are abstract constructs, with no clear bounds. Many genes have an indirect impact on each other and most of them can be said to act together in shaping the organism's fitness. Nevertheless, some

gene's products are more closely connected in performing a function than others and mutations of these genes tend to result in the same or similar phenotypes – justifying the practicality of thinking of biochemical and physiological processes in terms of pathways (Oti and Brunner 2007).

Many examples exist of interacting proteins associated with the same genetic disorder, such as in Fanconi anemia, characterised by inborn malformations, progressive bone-marrow failure, anemia and cancer predisposition (Brunner and van Driel 2004). The disorder is caused by mutations of any of more than a dozen genes encoding proteins involved in the functioning of a DNA-repair pathway, eight[1] of which form a core complex, enabling the monoubiquitination of the FANCD2 protein. Each of the eight proteins is essential for the monoubiquitination of FANCD2, which then interacts with downstream components of the pathway, including BRCA2, in preventing replication-associated DNA damage (Alpi and Patel 2009).

Visceral heterotaxy results from abnormal left–right patterning during embryogenesis, leading to unusual spatial arrangement of internal organs and vasculature. Rare mutations in the genes *ACVR2B*, *CRYPTIC*, *LEFTY2* and *ZIC3* have been associated with the disorder, the former three of which participate in the transforming growth factor-β (TGF-β) signaling pathway (Epstein *et al*. 2008). The proteins encoded by these genes play roles in NODAL signaling, leading to an asymmetrical distribution of the secreted protein NODAL, critical in establishing of correct left–right asymmetry.

Hypohidrotic ectodermal dysplasia presents with missing or malformed teeth, sparse hair and a reduced ability to sweat, as a result of abnormal development of the oral

[1] FANCA, FANCB, FANCC, FANCE, FANCF, FANCG, FANCL and FANCM.

and epidermal ectoderm. The mutations of three genes, *EDA*, *EDAR*, and *EDARADD* – which participate in the ectodysplasin pathway within the tumor necrosis factor (TNF) superfamily – are responsible for the disorder. Ectodysplasin (EDA) is a ligand expressed by epithelial cells which binds to the receptor EDAR, transducing a signal through an interaction with the protein EDARADD, which links the ectodysplasin receptor complex with downstream signaling proteins (Epstein *et al*. 2008).

The lacrimo-auriculo-dento-digital (LADD) syndrome involves malformations of the tear-producing lacrimal system, cup-shaped ears and dental and digital anomalies. It is caused by mutations in *FGF10* and in the tyrosine kinase domains of *FGFR2* and *FGFR3*, forming part of the fibroblast growth factor signaling pathway. FGFR2b and FGFR3 are the binding receptors of the ligand FGF10 (Epstein *et al*. 2008). Some mutations of *FGF10* cause the related but less severe aplasia of lacrimal and salivary glands (ALSG) disorder, without limb anomalies.

Alagille syndrome is the most common form of bile duct deficiency, accompanied by anomalies of the heart, eye and skeleton and characteristic facial features. Most of the patients are found to have a mutation of the *JAG1* gene and a small fraction have a mutation of *NOTCH2*. JAG1 is a cell-surface protein that binds to the receptor NOTCH2, located on an adjacent cell. Both proteins function in the Notch signaling pathway; the binding of the JAG1 ligand induces cleavage of the receptor's intracellular domain, which then translocates into the nucleus and mediates downstream effects (Epstein *et al*. 2008).

Tuberous sclerosis is characterised by morphological defects and benign growths in the brain and on the skin, kidney, heart and other organs. Mutations in *TSC1* and *TSC2*

were linked to the disorder, both of which function in the PI3K-Akt-mTOR pathway controlling cell growth and migration. The proteins TSC1 and TSC2 form a highly stable complex which acts as a GTPase activator for the GTPase RHEB, regulating the mTOR kinase (Epstein *et al*. 2008). mTOR is part of a complex that regulates the assembly of the eIF3 translation initiation complex, leading to protein translation and cell growth.

The cardio-facio-cutaneous (CFC) syndrome presents with cardiac defects, characteristic craniofacial features, ectodermal anomalies, hypotonia and developmental delay. CFC is caused by mutations in *BRAF*, *KRAS*, *MEK1* and *MEK2*, all of which are components of the RAS/MAPK pathway (Epstein *et al*. 2008). Activation of KRAS leads to the activation of BRAF which is one of the first kinases in the MAPK cascade of kinases including MEK1 and MEK2. The latter two phosphorylate and activate ERK1 and ERK2, which phosphorylate numerous downstream targets, including transcription factors.

Cornelia de Lange syndrome is characterised by distinctive facial features, excessive hairiness, abnormalities of the upper limbs, gastroesophageal reflux and developmental delay. Mutations in the genes *NIPBL*, *SMC1A* and *SMC3*, which encode components and regulators of the cohesin complex, have been associated with the disorder. The proteins SMC1A and SMC3 form key parts of the cohesin ring complex ensuring sister chromatid cohesion after DNA replication, directed by regulatory proteins, including NIPBL (Nasmyth and Haering 2009). Other components of the cohesin complex have been implicated in developmental disorders with similar clinical features to de Lange syndrome (Epstein *et al*. 2008).

1.7 Pathway analysis

Pathway analysis, or functional-enrichment analysis, is a test for shared function in a set of genes. It is generally used to verify the functional coherence of a group of genes and to identify the shared biological functions. Thus, it is of ready application in the interpretation of gene variants implicated in a disorder by GWAS or sequencing studies, in the search for molecular processes or pathways the malfunction of which can be underlying the disorder. In other words, it is functional-enrichment analysis that can piece together the gene variants associated with a genetically heterogeneous disorder.

Functional enrichment can sometimes be identified from overrepresentation of one or more individual functional annotations corresponding to Gene Ontology (GO) terms (Ashburner *et al*. 2000) or other functional groupings of genes, compiled in databases such as the Kyoto Encyclopedia of Genes and Genomes (KEGG) PATHWAY database (Kanehisa *et al*. 2010), Reactome (Croft *et al*. 2011), Ingenuity (www.ingenuity.com), MetaCyc (Caspi *et al*. 2013) or Biocarta (www.biocarta.com). Enrichment of a functional annotation is tested by comparing the observed number of genes sharing this annotation with expected values based on a probability distribution (Khatri *et al*. 2012).

Other, rank-based approaches, order all the studied candidate genes by a gene-level statistic, such as the p-value from a GWAS or differential expression analysis (Ramanan *et al*. 2012). Then, genes are associated with functional annotations or pathways and the ranks or p-values of genes that are associated with a given pathway are compared to the overall distribution. Gene set enrichment analysis (GSEA) involves the

calculation of an enrichment score representing a pathway, this score being analogous to a weighted Kolmogorov–Smirnov-like statistic (Subramanian *et al*. 2005). The statistical significance of this score is assessed by a permutation test that generates a null distribution (Subramanian *et al*. 2005).

However, functional annotations are limited both in quantity and quality, as the function of many genes is unknown and their classification to pathways is scant or sometimes inferred automatically, without review. The tests assessing over-representation moreover assume that the pathways are independent of each other, which is clearly not the case (Khatri *et al*. 2012). Pathways can intersect and overlap, leading to pleiotropy, and they can even form a subset of another pathway. Their artificial demarcation and the disregard of their interrelations likely decrease the power of statistical tests to detect functional enrichment.

The overlapping nature of pathways is captured by gene networks, representing functional links among a large number of genes. The functional links are often based on protein–protein interactions or derived from large-scale transcriptomic data, such as gene co-expression measures. Any information suggesting a functional connection between genes can form or be incorporated in the framework of a network. A link in a gene network is meant to represent an inferred functional association between two genes, suggesting that the linked genes act together in some way (Fraser and Marcotte 2004). To test a set of implicated genes for functional enrichment, the interconnected-ness between these genes is compared to that of genes mutated in controls or random gene sets.

A case in point is DAPPLE (Disease Association Protein–Protein Link Evaluator), a method which identifies molecular pathways implicated through the physical connectivity among proteins encoded by the candidate genes (Rossin *et al*. 2011). DAPPLE takes advantage of a high-confidence protein interaction network (InWeb) which only includes interactions observed in multiple independent and preferentially low-throughput experiments (Lage *et al*. 2007). Within this network, the connectivity of the proteins encoded by the genes implicated in a disorder or trait is assessed through randomisations (Rossin *et al*. 2011). Another approach, termed NETBAG, is based on a combined gene network built from several data types (Gilman *et al*. 2011).

Some data types connect proteins (e.g. physical interactions), while others indicate functional relation between genes (co-annotations, co-citation) or transcripts (co-expression). Although the inferred links are commonly transferred between these – so that a gene network can be based on physical interactions – they are obviously different and there is not necessarily a one-to-one correspondence between them. For example, the choice as to which transcript represents a gene in a gene co-expression network affects the network topology and can influence the results of an analysis. Nevertheless, ultimately one wants to learn about how genes or proteins act together and diverse data types can be used to infer functional associations and pathways.

Multiple exome-sequencing studies have recently tested gene variants for functional enrichment by network analysis, implicating pathways associated with chromatin remodelling, β-catenin and synaptic function in autism (Neale *et al*. 2012; O'Roak *et al*. 2012b) and cortical neurogenesis and synaptic transmission in schizophrenia (Gulsuner *et al*. 2013). These studies used protein–protein interaction data from GeneMania

(Warde-Farley *et al*. 2010) or DAPPLE (Rossin *et al*. 2011) or co-expression data from BrainSpan (www.brainspan.org) as functional links forming gene networks. They assessed the interconnectedness of *de novo* mutated genes by comparing the number of links between these genes with expected values generated by randomisations.

An obvious confounder in these analyses is the number of selected genes, which influences the number of links, therefore the number of random genes in each simulation is kept equal to the number of implicated genes. O'Roak et al. also considered gene-specific mutation rate estimates, selecting random genes according to how many mutations they have accumulated since the human–chimpanzee divergence. However, these estimates represent the degree of sequence conservation rather than frequency of mutations, resulting in a tendency to select less conserved genes, which are less likely to be functionally influential. Neale *et al*. controlled for node degree, keeping the number of connections of individual proteins in a network constant during the randomisations. But then again, it is not clear whether the frequency of *de novo* mutations is a function of node degree.

Proper testing for shared function requires the consideration of confounding bias in order to ensure that a functional enrichment is associated with the shared phenotype and not some other factor. However, there is no consensus about what confounding factors affect these analyses. Likewise, the reliability of the data sources contributing links to a network, such as of the different protein–protein interaction assays, is uncertain (Deane *et al*. 2002). The results of the functional analyses are thus questionable for multiple reasons. Considering the surge of sequencing projects, the amount of effort spent on them, the prevalence of neurological disorders and the

significance of understanding their disease mechanisms, it would be very important to analyse the generated data in a sound and efficient manner.

1.8 Thesis scope and structure

In this thesis, I establish that sequencing studies are affected by coding-sequence length bias and I develop and apply a toolkit of superior power to identify functional relationships among genes mutated in the same disease.

In **chapter 2**, I describe an interdisciplinary method to extract more information about functional associations between human genes from diverse data types and assess the reliability of the inferred associations. I test different data types and networks and find that genes mutated in the same disease cluster more strongly in the here-developed integrated phenotypic-linkage network than in other available gene networks. Examining the functional association between *de novo* gene variants, I identify a confounding bias in coding-sequence length that I control for. This work has been published in PLoS Computational Biology.

Chapters 3 and **4** present two current applications of the described method.
I perform functional-enrichment analyses of gene sets associated with clinical features related to intellectual disability (ID) in **chapter 3**. I observe robust functional coherence in most classes and find an increased predictive potential in identifying genes associated with these phenotypes. These analyses form my contribution to a study of genes implicated in ID which has not been written up yet.

In **chapter 4**, I perform case–control association analyses of variants from an exome-sequencing study of Parkinson's disease (PD) and test the functional associations of

the mutated genes. I carry out association tests both of single variants and at the gene level but find none of the variants to be significantly associated with PD after multiple-testing correction. A manuscript based on this work is under preparation.

Finally, in **chapter 5**, I review and summarise the findings from the preceding chapters and discuss their implications and limitations.

Chapter 2: Unbiased functional clustering of gene variants with a phenotypic-linkage network

2.1 Abstract

Groupwise functional analysis of gene variants is becoming standard in next-generation sequencing studies. As the function of many genes is unknown and their classification to pathways is scant, functional associations between genes are often inferred from large-scale omics data. Such data types – including protein–protein interactions and gene co-expression networks – are used to examine the interrelations of the implicated genes. Statistical significance is assessed by comparing the interconnectedness of the mutated genes with that of random gene sets. However, interconnectedness can be affected by confounding bias, potentially resulting in false positive findings. I show that genes implicated through *de novo* sequence variants are biased in their coding-sequence length and longer genes tend to cluster together, which leads to exaggerated p-values in functional studies; I present here an integrative method that addresses these bias. To discern molecular pathways relevant to complex disease, I have inferred functional associations between human genes from diverse data types and assessed them with a novel phenotype-based method. Examining the functional association between *de novo* gene variants, I control for the heretofore unexplored confounding bias in coding-sequence length. I test different data types and networks and find that the disease-associated genes cluster more significantly in an integrated phenotypic-linkage network than in other gene networks. I present a tool of

superior power to identify functional associations among genes mutated in the same disease even after accounting for significant sequencing study bias and demonstrate the suitability of this method to functionally cluster variant genes underlying polygenic disorders.

2.2 Introduction

It is widely postulated that the products of genes whose variants are implicated in the same disease participate in the same biological function or process whose disruption leads to the disease (Brunner and van Driel 2004; Vockley *et al*. 2000). This concept is supported by examples of complex disease in which the proteins encoded by the implicated genes interact, form a molecular complex or function at different steps of the same biochemical pathway (Oti and Brunner 2007; Rossin *et al*. 2011). As there is limited power to associate rare variants with disease by case–control studies, the use of functional-enrichment approaches that identify a shared function in a set of mutated genes is becoming standard in the interpretation of variants (Neale *et al*. 2012; O'Roak *et al*. 2012b; Sanders *et al*. 2012; Webber 2011).

Since the function of many genes is not known and their classification to pathways is scant, functional associations between genes are often inferred from large-scale omics data (Gulsuner *et al*. 2013; Neale *et al*. 2012; O'Roak *et al*. 2012b; Rossin *et al*. 2011; Sanders *et al*. 2012). However, the suitability of such data types, including protein–protein interactions and gene co-expression networks, for functional-enrichment analysis remains unclear. Moreover, the inferred functional associations can be affected by confounding factors, potentially resulting in false positive findings. Thus, it

is important to identify any bias affecting the implicated genes and control for them. Multiple exome-sequencing studies currently test variants for functional enrichment and yet there is no consensus concerning what to control for (Gulsuner *et al*. 2013; Neale *et al*. 2012; O'Roak *et al*. 2012b; Sanders *et al*. 2012).

In this study, I have inferred functional associations between human genes from diverse data types and assessed the phenotypic agreement of the inferred gene–gene associations. I have examined different data types and networks and found that genes mutated in the same disease cluster more significantly in an integrated phenotypic-linkage network than in other gene networks. Examining the functional association between *de novo* gene variants, I have identified a confounding bias in coding-sequence length that I control for. I present a tool that identifies functional associations among genes mutated in the same disease even after accounting for significant sequencing study bias and demonstrate the power of this tool to functionally subcluster the gene variants underlying a polygenic disorder.

This chapter is based on my paper published in PLoS Computational Biology (Honti *et al*. 2014), which I wrote under the supervision of Caleb Webber. I carried out all the work presented in this chapter and in the publication, in accordance with the Author Contributions section presented therein. I am thankful to Stephen Meader for contributing the GNF2 gene expression data (Su *et al*. 2004) and to my supervisor for his support and patience.

2.3 Materials and methods

2.3.1 Inference of functional associations from diverse data types

To gain the most information about genes whose variants may be relevant to disease and to explore the functional relations between them, I collected large amounts of functional genomics data on human genes and their orthologs (**Table 2.1**). In order to derive information about the functional similarity of genes, I processed the data sets so that they indicated gene–gene links. For every data type except physical interactions, I derived a score characterising gene pairs, such as the correlation coefficient from expression profiles or semantic similarity from gene annotations. The physical interactions were scored by measuring the median semantic similarity of mouse phenotype annotations for all the gene pairs derived from the same assay and then applying this median value to score all the links in the given data set.

Table 2.1: Data sources included in the integrated phenotypic-linkage network.

Data type	Source	Form	#links
Physical interactions	BioGRID, IntAct, Corum, DICS, Reactome	Binary	493,951
Sequence patterns	InterPro	Semantic similarity	18,326
Gene expression	GNF2, GSE3594, MTAB-62, Alizadeh *et al*., Kampmann *et al*., Nielsen *et al*., Schaner *et al*., Shyamsundar *et al*.	Correlation coefficient	1,345,035
Co-citation	STRING[1]	Scores	577,887
Pathways, reactions	Reactome, KEGG	Semantic similarity	71,342
Biological process	Gene Ontology	Semantic similarity	1,100,459
Molecular function	Gene Ontology	Semantic similarity	16,379
Cellular location	Gene Ontology	Semantic similarity	10,757

[1] I used the co-citation of mouse orthologs of human genes in the human gene network.

2.3.2 Semantic similarity

All gene annotation data (such as GO, KEGG, Reactome, InterPro and mouse and human phenotype annotations) were processed in the form of semantic similarity, which is a measure of relatedness between two genes assessed by the similarity of their annotations (Pesquita *et al*. 2009). The terms used to annotate genes have an information content (IC) defined as:

$$\mathrm{IC}(a) = -\log_2 p(a)$$

where $p(a)$ is the proportion of genes annotated with term a or its descendent terms among all genes with an annotation.

I used Resnik's (Resnik 1995) measure together with the GraSM approach (Couto *et al*. 2007) to calculate the similarity of terms organized in a hierarchical ontology, defining the semantic similarity between any two terms t_1 and t_2 as the average IC of their disjunct common ancestor terms (see **Figure 2.1A**):

$$sim_{GraSM}(t_1, t_2) = \overline{\underset{a \in A}{\mathrm{IC}(a)}}.$$

To measure the functional relatedness of two genes, I compared their annotations with the maximum (max) and best match average (bma) methods (Pesquita *et al*. 2008). Let T_1 denote the set of terms annotated to gene g_1 and T_2 denote the set of terms annotated to gene g_2, the semantic similarity of their annotations according to the max approach is then given by:

$$sim_{max}(g_1, g_2) = \max_{t_1 \in T_1, t_2 \in T_2} sim_{GraSM}(t_1, t_2)$$

while the semantic similarity of their annotations according to the bma approach is defined as:

$$sim_{bma}(g_1, g_2) = \frac{\frac{\sum_{t_1 \in T_1} \max_{t_2 \in T_2} sim_{GraSM}(t_1, t_2)}{|T_1|} + \frac{\sum_{t_2 \in T_2} \max_{t_1 \in T_1} sim_{GraSM}(t_1, t_2)}{|T_2|}}{2}.$$

2.3.3 Data sources

Gene expression. I inferred gene–gene linkages from co-expression of genes. To measure the co-expression of genes, I calculated the Pearson's correlation coefficient of their expression profiles, requiring at least ten tissues in which both genes were expressed for the calculation of a correlation coefficient. I used expression data from GNF2 (Su *et al*. 2004), GSE3594 (Zapala *et al*. 2005), MTAB-62 (Lukk *et al*. 2010) and further five sets (Alizadeh *et al*. 2000; Kampmann *et al*. 2005; Nielsen *et al*. 2002; Schaner *et al*. 2005; Shyamsundar *et al*. 2005), calculating the Pearson's correlation coefficients within each, evaluating the inferred gene–gene links as described below (see **Figure 2.1B**), selecting and re-scoring the links that correlated with the semantic similarity of mouse phenotypes and integrating these links as described below to create a combined co-expression network. The resulting integrated network outperformed networks derived from the individual co-expression datasets both in terms of coverage and accuracy.

Physical interactions. Protein–protein interactions (PPI) provided a binary measure of functional linkage between genes, with all derived gene pairs receiving the same score. I measured the median semantic similarity of mouse phenotype annotations for all the gene pairs derived from the same assay and used this median value to score all the

functional linkages in the given data set. I used physical interaction data divided by assay types from BioGRID (Stark *et al*. 2011) v3.1.72, IntAct (Kerrien *et al*. 2012) (downloaded on July 29, 2011), CORUM (Ruepp *et al*. 2010) (downloaded on May 5, 2011), DICS (Dietmann *et al*. 2009) (downloaded on June 6, 2011) and Reactome (Croft *et al*. 2011) (downloaded on May 5, 2011). I also derived indirect links based on shared interaction partners, re-scored them as described below and integrated these with the direct links to create a combined PPI network.

Co-citation. Co-citation scores were extracted from STRING (Szklarczyk *et al*. 2011) v8.3. I used the co-citation scores of mouse orthologs of human genes.

Gene annotations. Gene annotations were obtained from the Gene Ontology (Ashburner *et al*. 2000) (GO, downloaded on July 29, 2011), using the annotations to human and mouse genes in the biological process (BP), molecular function and cellular location categories, with evidence codes IDA, IMP, TAS and IC. Pathway annotations of mouse genes were obtained from KEGG (Kanehisa *et al*. 2010) (downloaded on March 30, 2011), pathway annotations of human genes were downloaded from Reactome (Croft *et al*. 2011) (on March 23, 2011). Protein domain annotations were obtained from InterPro (Hunter *et al*. 2012) (downloaded on May 23, 2011). Mouse phenotype annotations were obtained from the Mouse Genome Database (Eppig *et al*. 2012) (downloaded on August 24, 2011), human phenotype annotations were downloaded from the Human Phenotype Ontology (Robinson *et al*. 2008) (on August 8, 2012). The annotation terms in KEGG, Reactome and InterPro are not organized in a deep ontology with many levels, therefore I only used direct matches between these gene

annotations with the maximum (max) method in calculating semantic similarity scores for gene pairs.

2.3.4 Evaluation of data sets

To estimate the reliability of the individual data sets, I assessed them by examining the semantic similarity between the phenotypes associated with the unique mouse orthologs of the genes they linked to each other. For each data set, I derived gene–gene linkages (gene pairs) with data-specific scores characterizing the strength of a linkage and ordered the gene pairs by their score from largest to smallest. Next, I calculated and plotted the median semantic similarity of mouse phenotype annotations for bins of 1,000 gene pairs (see **Figure 2.1B**).

I tested if the data types linked together genes whose knockouts affected the same phenotypes. When the strongest linkages derived from a data set did not correspond to higher semantic similarities of phenotypes than expected by chance, I did not include the links from the given set in the integrated gene network (**Appendix, Table A.1**).

2.3.5 Re-scoring and integration of data

I selected data sets that produced a positive correlation with the semantic similarity of mouse phenotypes (**Table 2.1**) and fitted regression curves in order to re-score the links so that any data-specific scores characterising the gene pairs were replaced with the semantic similarity of phenotypes that they corresponded to according to a linear regression function (**Figure 2.1B**). Thus all gene pairs that had an original data-specific score were re-scored, including those that did not have phenotypic annotations.

By re-scoring the data types with a universal benchmark, I weighted them in proportion to their relative accuracy. When multiple data sources suggested functional linkage between the same two genes, I summed their link weights (**Appendix, Figure A.1**) increasingly down-weighting less reliable data according to a formula used by Marcotte and colleagues (Lee *et al*. 2011):

$$WS = L_0 + \sum_{i=1}^{n} \frac{L_i}{D \times i}$$

where L represents a re-scored link weight from a single data set, L_0 being the largest link weight among all the links between the given two genes, i is the index of the remaining links ordered by their weights for the gene pair and D is a free parameter. I optimized the value of this parameter and used D =5 in integrating data types to create the final phenotypic-linkage network.

2.3.6 Controlling for coding-sequence length

In testing for functional enrichment in a set of genes, the degree of functional association between the genes can be compared to that calculated for randomised gene sets. As the degree of functional association can be affected by confounding bias, it is important to identify such bias affecting the studied gene set and control for them. To control for coding-sequence (CDS) length during the randomisations, I selected random genes the CDS length of which matched the CDS length of the studied (mutated) genes. For each of the studied genes in turn I assigned a list of 100 genes with the same or most similar CDS length, using the longest CDS of each gene. Random gene sets were then assembled by selecting one random gene from each of these lists.

2.4 Results

To test for functional associations among gene variants, I derived functional links between genes from diverse data types. For example, I calculated correlation coefficients from expression profiles, whereas gene annotation data were processed in the form of semantic similarity, which is a measure of relatedness between two genes assessed by the similarity of their annotations (Pesquita *et al*. 2009) (**Figure 2.1A**). The data were likely to include noise leading to false links and their reliability was unknown. To estimate and take into account the accuracy of the links, I evaluated the individual data types with a novel, phenotype-based method, by examining the semantic similarity between the mouse phenotypes of the genes they related to each other (**Figure 2.1B**). That is, each data type in turn indicated gene–gene linkages (gene pairs) and the accuracy of these links was assessed by considering the similarity of the phenotypes arising from the disruption of the unique mouse orthologs of these genes. The data types were expected to link together genes whose knockouts give rise to the same phenotypes, even if these mouse phenotypes were not necessarily expected to resemble human symptoms. The similarity of mouse phenotype annotations correlated with the similarity of human disease phenotypes ($\rho=0.223$, $P<2\times10^{-16}$; **Appendix, Figure A.2**) and mouse phenotypes have been assigned to 6169 unique orthologs of human genes, 3.4-fold more than the 1801 genes annotated by the Human Phenotype Ontology (HPO; downloaded in 2012) (Robinson *et al*. 2008). Consequently, I used the phenotype annotations from the Mouse Genome Database (Eppig *et al*. 2012) as the benchmark against which to evaluate other data types and set aside the HPO annotations for use as a test set for validation.

Integration of different data types into a combined network is expected to improve the accuracy of links and thus, in addition to considering individual data types, I also built an integrated gene network (Franke *et al*. 2006; Lee *et al*. 2004). For this, I selected data types that consistently linked together genes associated with similar mouse knockout phenotypes and that produced a positive correlation with the semantic similarity of mouse phenotypes (**Table 2.1**). For each data source suggesting functional links, I fitted regression curves in order to re-score the links so that any data-specific scores characterising the gene pairs were replaced with the semantic similarity that they corresponded to according to a regression function (see **Figure 2.1B**). When multiple data sources suggested functional linkage between the same two genes, I summed their link weights according to the approach of Marcotte and colleagues (Lee *et al*. 2011), thereby down-weighting less reliable data (see Methods). The resulting integrated gene network outperforms networks derived from the individual data types both in terms of coverage and accuracy (**Figure 2.1C**).

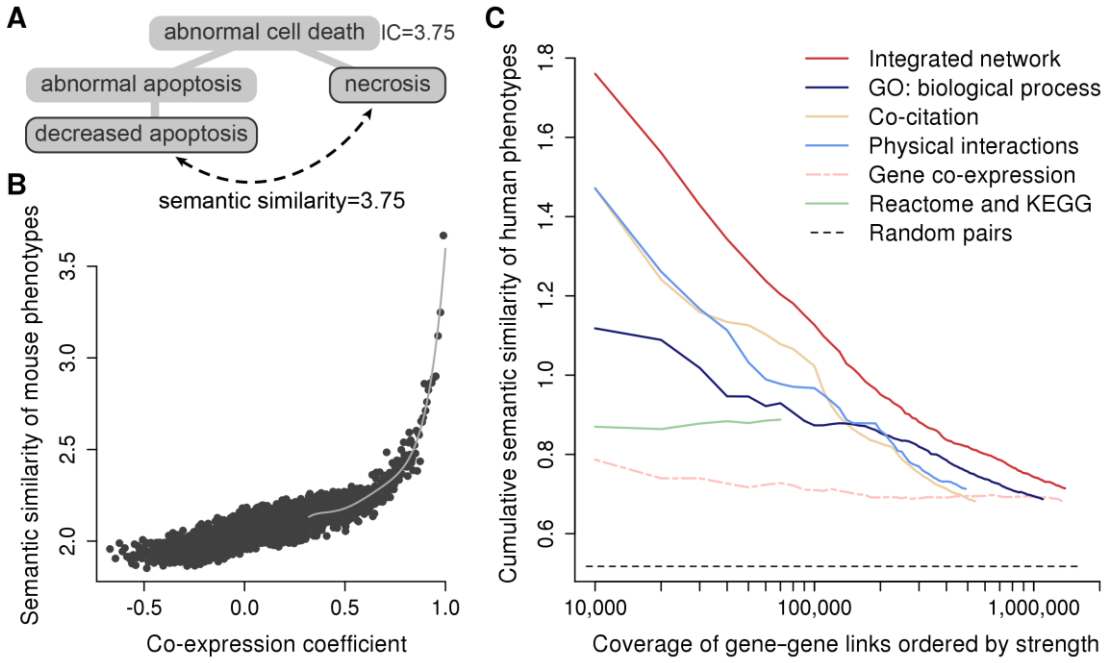


Figure 2.1: Processing and comparison of functional genomics data. (A) Terms in a phenotype ontology have an information content (IC) which is inversely proportional to the number of genes annotated with them. The semantic similarity between any two terms equals to the IC of their closest common ancestor term(s). (B) Gene–gene linkages derived from a data type are assessed and rescored according to the semantic similarity of the linked genes' mouse phenotype annotations. (C) The similarity in human phenotype annotations from the HPO is a benchmark on which all the data types can be compared, revealing their relative accuracy and coverage.

I corroborated the integrated phenotypic-linkage network by showing that genes whose perturbations are implicated in the same disease tend to be closely interlinked (**Appendix, Figures A.3–8**). It is possible that their tendency to be closely interconnected is due to shared functional annotations assigned to them because they were implicated in the same disease in the literature. Also, I cannot assume that the associations of genes to phenotypes – forming the test sets – were made independently of any data type. Consequently, I turned to recently reported *de novo* mutations associated with developmental disorders that were identified independently of the data types included in the network.

Genes with *de novo* substitutions in patients with the same disorder (Allen *et al*. 2013; de Ligt *et al*. 2012; Gulsuner *et al*. 2013; Neale *et al*. 2012; O'Roak *et al*. 2012b; Rauch *et al*. 2012; Sanders *et al*. 2012) showed a tendency to be more interconnected in the gene networks than random gene sets of the same size. However, as the interconnectedness of genes can be affected by confounding factors, it was important to identify any bias affecting the studied genes and control for them during the randomisations. I have found that the genes implicated through *de novo* sequence variants are biased in their coding-sequence (CDS) length, as longer genes are more likely to be mutated by chance (**Figure 2.2**). I also observe that genes with longer CDS tend to be interconnected (**Figure 2.2C**) and thus controlling for CDS length during the randomisations can significantly affect their relative degree of clustering (**Figure 2.3**).

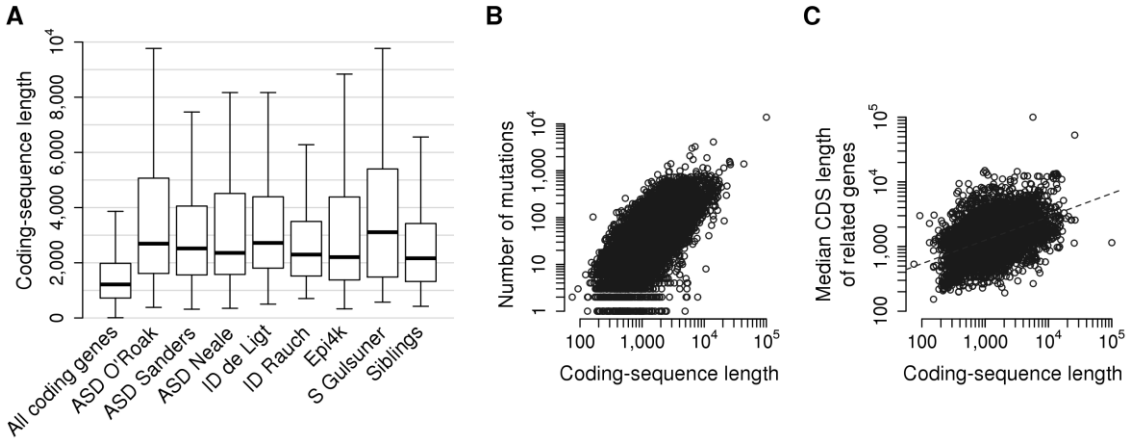


Figure 2.2: Coding-sequence (CDS) lengths of genes with *de novo* variants. (A) 'All genes' denotes all translated human genes, 'Siblings' denotes genes with *de novo* mutations in non-autistic siblings of ASD cases published by O'Roak *et al*. and Sanders *et al*. Even the genes mutated in the healthy siblings are significantly longer than all coding genes (Mann–Whitney U test, $P<2\times10^{-16}$). The box plots depict the values between the 1st and 3rd quartile of a distribution, the 2nd quartile (thick band) represents the median. (B) Mutational burden strongly correlates with coding sequence length in the Exome Variant Server (Spearman's $\rho=0.710$, $P<2\times10^{-16}$; http://evs.gs.washington.edu/EVS). All nonsynonymous mutations were considered across all human chromosomes. (C) The median CDS length of a gene's connections correlates with its CDS length (Spearman's $\rho=0.508$, $P<2\times10^{-16}$). I considered here the strongest 100,000 links from the integrated phenotypic-linkage network.

To control for CDS length during the randomisations, I have selected random genes the CDS length of which matched the CDS length of the studied candidate genes (see Methods). Node degree has been previously identified as a confounding factor in functional analyses, particularly where an increase in degree results from study bias (Gillis and Pavlidis 2011). However, controlling for node degree in a gene network does not correct the CDS length bias (**Figure 2.3B**). CDS length correlates very weakly with node degree (Spearman's ρ=0.050). The length bias are highly significant in all the studied gene sets (**Figure 2.2**), while the node degrees are significantly different only in some of the candidate gene sets and there is no correlation between the node degree and mutational burden of genes (**Appendix, Figure A.9**). Having examined different data types and networks (Lee *et al*. 2011; Szklarczyk *et al*. 2011; Warde-Farley *et al*. 2010), I find that the disease-associated genes cluster more significantly in the integrated phenotypic-linkage network than in other gene networks (**Figure 2.3 and Appendix, Figure A.10**).

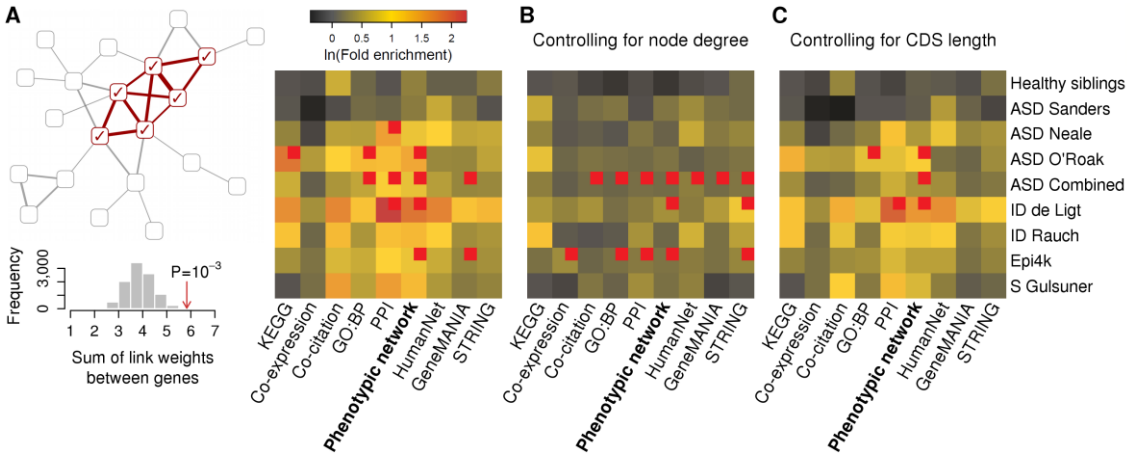


Figure 2.3: Clustering of genes hit by *de novo* nonsynonymous substitutions. (A) I have examined the network properties of whole sets of genes with nonsynonymous mutations implicated by recent exome-sequencing studies in autism (ASD), severe intellectual disability (ID), epilepsy or schizophrenia (S). I calculated the sum of link weights among genes from a set and compared this sum to that calculated for randomized gene sets in order to assess the degree of functional clustering. (B and C) The implicated genes tend to be significantly more strongly interconnected with each other by means of functional genomics data than random gene sets of the same size, but controlling for coding-sequence length considerably affects the p-values. The genes mutated in the same disease cluster most significantly in the integrated phenotypic-linkage network, while genes mutated in healthy controls do not cluster. Fold enrichments are presented on a logarithmic scale; a red square denotes statistical significance after Bonferroni correction.

2.5 Discussion

I have inferred functional-association networks of human genes from diverse data types and assessed the phenotypic agreement of the inferred links. Perhaps surprisingly, the links based on the semantic similarity of functional annotations from Reactome and KEGG performed worse than three other data types when assessed with the semantic similarity of human phenotypes (**Figure 2.1C**). This may have been a consequence of the less diversified ontology that is characteristic of these databases, but may also have been related to the lower coverage of these data.

Having examined different data types and networks, I have found that genes mutated in the same disease cluster more significantly in an integrated phenotypic-linkage network than in other gene networks (**Figure 2.3C**). Apparently, although a protein–protein interaction network (PPI) can produce higher fold enrichments, the integrated network (Phenotypic network) includes further informative data which probably increase its power in detecting a significant clustering.

Another gene network, called NETBAG, has been developed by Gilman and colleagues (Gilman *et al*. 2011). I could not access NETBAG for the performance comparison. Nevertheless, Gilman and colleagues have reported the use of shared disease associations among 478 human genes as the gold standard in their network construction (Gilman *et al*. 2011) and the used disease associations originated from a study published in 2001 (Jimenez-Sanchez *et al*. 2001). By comparison, my method takes advantage of over 100,000 mammalian genotype–phenotype relations and fully exploits bio-ontologies by means of semantic similarity, with both advances expected to enhance greatly the phenotypic-linkage network that I explicitly present here.

Examining the functional association between *de novo* gene variants, I have identified a confounding bias in coding-sequence length that I control for to avoid false positive findings. Numerous implicated variants are in fact expected to be neutral mutations but they are more likely to appear in genes with longer CDS, leading to a tendency of the implicated genes to be interconnected in gene networks (see **Figure 2.2**). These bias have confounded functional analyses and likely led to an overestimation of functional clustering in former studies.

Although I have found that genes with longer CDS tend to be interconnected (**Figure 2.2C**), there is only a very weak correlation between CDS length and node degree, longer proteins do not necessarily take part in larger numbers of interactions (**Appendix, Table A.2**). This may appear unexpected, as longer proteins can feature more binding sites and they may be multifunctional. Likewise, while I have observed that the CDS-length bias were highly significant in all the studied gene sets, including the unaffected siblings, the node degrees were not. The higher node degrees in some of the candidate gene sets may indicate a functional signal, as the same genes are significantly more conserved (**Appendix, Figure A.9**). I conclude that controlling for CDS length in functional analyses of gene variants is appropriate.

One way of controlling for CDS length is to compare the interconnectedness of the implicated genes with that of genes mutated in unaffected controls (Gulsuner *et al*. 2013). However, I observe that the control genes tend to be less interconnected than random genes (**Appendix, Figure A.11**), which suggests that my way of controlling for CDS length is more conservative.

The nature of the phenotypic-linkage network suggests that the clustered genes function together in the same disease-relevant cellular pathways (**Appendix, Figure A.12**). The functional convergence that I identify among the three sets of genes from independent exome studies of autism spectrum disorder demonstrates that the method is able to detect biological coherence among variant genes (**Figures 2.3 and A.12**). Throughout, I have considered the larger class of non-synonymous variants which is likely to possess a more diluted signal than nonsense variants. As with all clustering methods, my method is sensitive to the number of variants identified and to the likelihood of their causal relation. Half of my study sets included only 5–10 genes with nonsense variants, between which I either did not find any functional links or the sum of link weights was not significantly higher than expected after controlling for CDS length. For studies of rare or *de novo* variants derived from a single or small number of genomes, gene prioritizing methods based upon phenotypic similarity may be more appropriate (Robinson *et al*. 2014). Continuing efforts to systematically phenotype model organisms and to enrich the phenotype ontologies could further improve the resulting phenotypic-linkage networks that are constructed (Brown and Moore 2012). The integrated network toolkit is made available at http://wwwfgu.anat.ox.ac.uk/downloads/webber-resources/Network.

Chapter 3: Network analyses of genes associated with clinical phenotypes related to intellectual disability

3.1 Introduction

Intellectual disability (ID) is one of the most prevalent developmental disorders and a common comorbid condition of congenital malformations (Harris 2006). Many genes have already been associated with ID (Robinson *et al*. 2008) and more are being implicated by sequencing studies (de Ligt *et al*. 2012; Rauch *et al*. 2012). The extreme genetic heterogeneity of ID suggests that the disruption of any of multiple biological pathways can cause the condition, implying etiological diversity. The heterogeneity of underlying disease mechanisms is also consistent with the diversity of clinical presentations.

As ID is likely to result from diverse molecular and developmental defects, the identification and understanding of these processes is consequently hindered by their multitude. Thus, in order to study the separate disease mechanisms, it may be useful to focus on more specific clinical phenotypes. Dissection of ID into different subphenotypes could facilitate the uncovering of the underlying biological pathways (Ropers 2007) and the same applies to other complex disorders, as suggested elsewhere (Krumm *et al*. 2014). For this approach to be effective, the genes associated with the ID-related phenotypes have to be functionally coherent.

To analyse the genetic causes of ID, my collaborators introduced a clinical classification system based on clinical information. They manually classified genes according to the severity of their associated phenotypes and according to clinical features. My task was to perform network-based functional analyses of these gene sets in order to test three hypotheses: i) the ID-associated genes show biological coherence, ii) dividing these genes according to clinical ID classes increases biological coherence and iii) the genes belonging to the different classes are predictable. I have observed increased functional coherence in most classes and found an enhanced predictive potential in identifying genes associated with these phenotypes.

The work presented in this chapter derives from a collaboration between me, Korinna Kochinke, Christiane Zweier, Bonnie Nijhof, Michaela Fenckova, Pavel Cizek, Shivakumar Keerthikumar, Merel A.W. Oortweld, Tjitske Kleefstra, Jamie M. Kramer, Caleb Webber, Martijn A. Huynen and Annette Schenck; however, I carried out all the work in this chapter apart from the design of the classification systems described in section 3.2 and **Figure 3.1** and apart from the assembly of ID-associated genes. I carried out all the work presented in first-person singular.

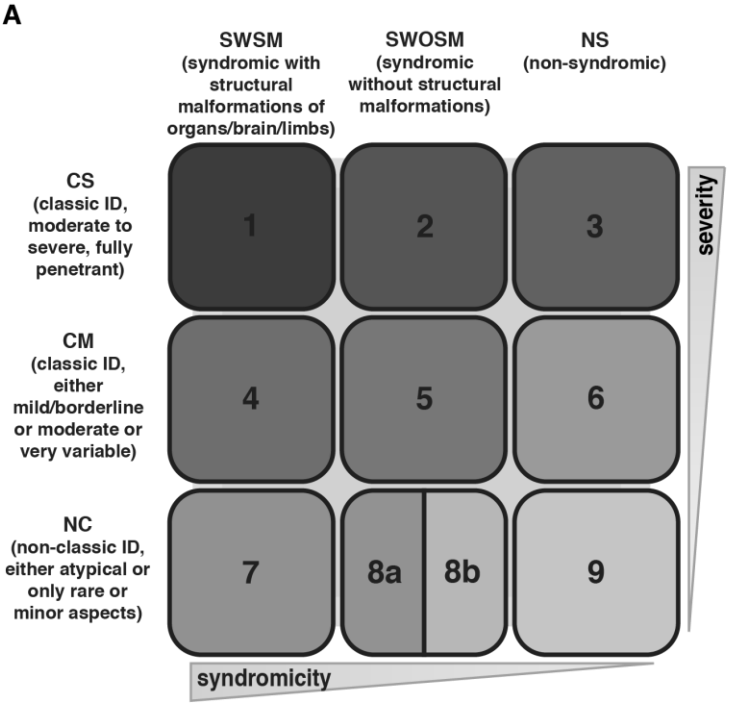
3.2 Materials and Methods

A manually-curated catalogue of 650 human ID-associated genes was assembled from the literature and OMIM (www.ncbi.nlm.nih.gov/omim). A phenotypic classification system has been designed in which the 650 genes were manually categorised according to “syndromicity” and the severity of their associated phenotypes (**Figure 3.1A**). Syndromic grading ranged from syndromic with structural malformations

through syndromic without structural malformations to non-syndromic, whereas severity ranged from severe to atypical. This classification formed 16 classes, which included between 1–407 genes from a total of 650 ID-associated genes. 77 of these 650 genes were assigned to more than one class. A separate classification was established of 27 ID-accompanying clinical phenotypes, associated with 22–298 genes from the total of 650 (**Figure 3.1B**).

Large-scale phenotype datasets of the fruit fly *Drosphila melanogaster* were generated by RNAi-mediated knockdown of 294 fly orthologs of the human ID-associated genes. Fly phenotypes were presented in two classifications, neuronal and wing phenotypes, with both classifications including 24 subgroups. The neuronal phenotypes resulting from the knockdown of the ID-associated orthologs were assessed by behavioural screening; lethality was evaluated throughout development, distinguishing early from late, as well as partial from total. Whenever I have used the classifications of fly phenotypes and the corresponding fly orthologs of the human ID-associated genes in the Results section, I have referred to “fly phenotypes”, in order to distinguish these from the human clinical classes presented in **Figure 3.1**. Otherwise all the analyses concerned the human clinical classes and human ID-associated genes.

A



B

Letter	Feature(s) or organ system
A	short stature
B	microcephaly
C	lethality
E	epilepsy
F	overgrowth/macrocephaly
G	progression/regression
H	neurological symptoms
I	malignancies
J	immunological anomalies
K	endocrine anomalies
L	L1: brain malformations L2: non-structural MRI anomalies
M	metabolic/mitochondrial anomalies
N	obesity
O	vegetative anomalies
P	behavioral anomalies
Q	myopathy (or: muscular anomalies)
R	blood cell anomalies
S	ectodermal anomalies
T	eye anomalies
U	U: skeletal anomalies Ua: limb anomalies Ub: vertebral/skull anomalies Uc: clefts
V	cardiac malformations
W	urogenital and renal anomalies
X	other malformations

Figure 3.1: A clinical classification system of ID. (A) Main clinical classes. Syndromic grading of ID: classes 1, 4 and 7 correspond to disorders that are syndromic with structural malformations (SWSM); classes 2, 5, and 8a/b include syndromic disorders without structural malformations (SWOSM); classes 3, 6, and 9 represent non-syndromic (NS) disorders related to ID. Severity axis of ID: classes 1, 2, and 3 include disorders with severe and fully penetrant manifestation (CS) of ID, classes 4, 5, and 6 include disorders with mild to moderate or very variable ID (CM), and classes 7, 8 and 9 comprise disorders with ID in a rare (8a) or atypical (8b) manifestation (NC). (B) ID-accompanying clinical phenotypes that occur with an estimated frequency of >20% within ID. Each of the 650 ID-associated genes has been assigned to at least one main clinical class. Accompanying phenotypes were associated with genes wherever appropriate. Figure courtesy of Prof. Annette Schenck.

In the network analyses, I used the integrated phenotypic-linkage network described in Chapter 2 and two co-expression networks. The first of the co-expression networks was the same that had been included in the integrated gene network in Chapter 2. This gene network was formed by the integration of eight co-expression data sets based on microarrays, as described in section 2.3.3. In addition to this, I also used a second co-expression network, based on the GTEx pilot data set (www.gtexportal.org). This second network was built from all gene pairs with a Pearson's correlation coefficient larger than 0.3. The value of the correlation coefficient served as weight of the corresponding gene–gene link in each case.

To test for functional-enrichment, I calculated the sum of link weights between genes that were associated with the same subgroup. I used this sum as a test statistic, comparing it to a null distribution formed of sums of link weights between the same number of random genes in the given network. I controlled for both node degree and coding-sequence (CDS) length during the randomizations, as described below. For the number-matched sub-sampling, I have calculated the sum of link weights in the GTEx network between the ID-associated genes belonging to the same subgroup and compared it to the sums of link weights between the same number of random genes belonging to any subgroup in the same classification (e.g. main classes or phenotypes, 650 genes in total).

To assess the co-clustering of two different gene sets, I calculated the sum of all the weighted links connecting these two gene sets. Then, I replaced one of the two gene sets with random genes controlled for both node degree and CDS length and measured the sum of link weights between these sets. After a sufficient number of

randomisations, I replaced the other gene set of the two with random genes controlled for both node degree and CDS length and repeated the measurements.

As the degree of functional association can be affected by confounding bias, it was important to identify such bias and control for them. To control for both node degree and CDS length during the randomizations, I selected random genes the node degree and CDS length of which matched the node degree and CDS length of the studied genes, respectively. To each of the studied genes in turn I assigned a list of 100 genes with the same or most similar node degree and CDS length, using the longest CDS of each gene.

I obtained these lists by calculating the difference in both node degree and CDS length between all genes in a given network and then selecting for each gene in turn 100 other genes that were the most similar to it. I normalised the node degrees and CDS lengths and calculated the Euclidean distance between genes based on these two measures. I used this Euclidean distance to form a list of the 100 most similar (that is, closest) genes to each gene. Random gene sets were then assembled by selecting one random gene from each of these lists. I used lists of 100 genes because it proved to be a sufficiently large set allowing efficient randomisations and at the same time the genes were still reasonably similar to each other in node degree and CDS length. I used the Euclidean distance between genes based on these two measures because it allowed a simple integration of differences in both node degree and CDS length in a single distance metric.

To calculate the expected precision values (and P-values) in the precision–recall analyses, I sub-sampled genes belonging to any subgroup from the same classification

(e.g. main classes or phenotypes, 650 genes in total). For the “All 650 genes” figure, I used random genes controlled for both node degree and CDS length from the phenotypic-linkage network to calculate the expected precision values.

3.3 Results

3.3.1 Functional coherence of ID-associated genes

To estimate the biological coherence of the ID-associated genes, their interconnectedness in a functional-linkage network can be compared to that of random genes. However, as the interconnectedness of genes can be affected by confounding factors, it was important to identify any bias affecting the studied genes and to control for them during the randomisations. I have found that the ID-associated genes are biased in both their node degree and coding-sequence (CDS) length (**Figure 3.2**).

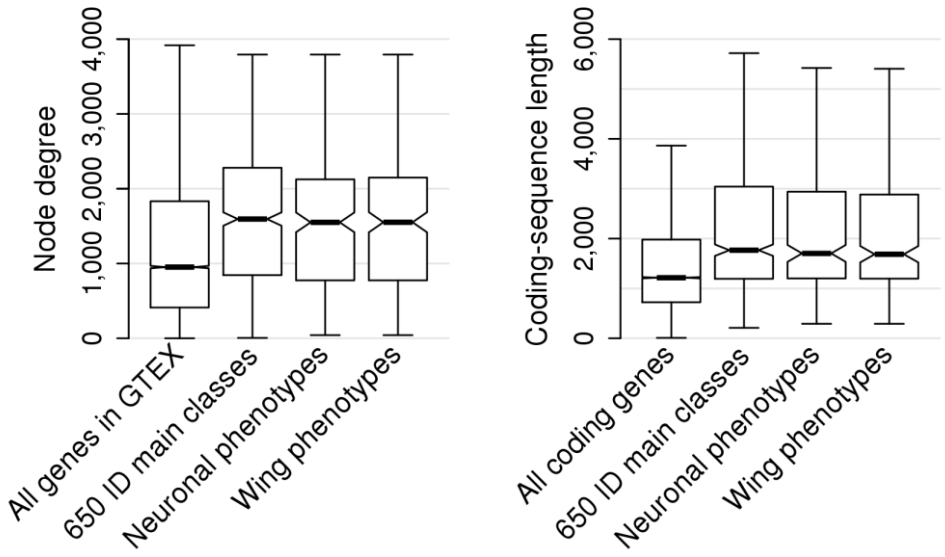


Figure 3.2: Node-degree and CDS-length bias of genes implicated in ID. The node degree of a gene represents the number of genes linked to this gene in the GTEx co-

expression network. The CDS length represents the longest CDS of a gene. Both the Neuronal phenotypes and Wing phenotypes represent fly phenotypes and the corresponding genes; all the examined distributions are significantly different from the expected distributions.

I have assessed the interconnectedness of the 650 human ID-associated genes controlling for node degree and CDS length and found that they significantly clustered in both the phenotypic-linkage and co-expression networks ($P<1.0\times10^{-4}$). I have repeated the same analysis with the different subgroups and found that they too tend to cluster (**Appendix, Tables A.4–7**). As the integrated phenotypic-linkage network includes functional information from the literature, its use in these analyses might constitute a case of circular reasoning. Therefore, I rely on results obtained with co-expression networks in the following.

To test whether the classification of the ID-associated genes into clinical ID classes increased their functional coherence, I performed number-matched sub-sampling, selecting random genes from the set of all 650 ID-associated genes (Methods). Assessing the interconnectedness of genes associated with the same subgroup in the GTEx co-expression network, I have found that these gene sets indeed tended to be functionally more coherent than the set of 650 ID-associated genes as a whole (**Tables 3.1** and **3.2, Figure 3.3**).

Table 3.1: Clustering of genes associated with the main clinical classes of ID.

Class	#genes	Σ_{GTEx}[1]	P-value
1	64	312.64412	0.05859
2	118	988.25832	0.07909
3	19	35.688667	0.02390
4	85	577.43336	0.01380
5	181	2183.7611	0.16928
6	71	348.47978	0.17658
7	30	51.542385	0.56144
8a	56	147.46212	0.94201
8b	87	547.52412	0.08709
9	1	0	1
SWSM	168	2056.9437	0.01920
SWOSM	407	10169.613	0.65873
NS	90	615.57770	0.03420
CS	196	2905.3897	0.00270
CM	329	7230.0565	0.06499
NC	167	1496.3272	0.95880

[1] Sum of link weights between the genes.

Table 3.2: Clustering of genes associated with ID-accompanying clinical phenotypes.

Letter	Feature	#genes	Σ_{GTEx}[1]	P-value
A	Short stature	140	1556.8737	0.00220
B	Microcephaly	143	1967.9804	0.00010
C	Lethality	133	1064.8048	0.60184
E	Epilepsy	216	2889.9795	0.51505
F	Overgrowth/macrocephaly	25	32.462219	0.68383
G	Progression/regression	125	999.75859	0.36016
H	Neurological symptoms	298	5593.8010	0.40186
I	Malignancies	22	57.596335	0.00230
J	Immunological anomalies	30	83.269694	0.01660
K	Endocrine anomalies	48	251.84223	0.00010
L1	Brain malformations	92	608.04978	0.08249
L2	Non-structural MRI anomalies	136	1196.3916	0.30977
M	Metabolic/mitochondrial anomalies	175	1785.9675	0.78482
N	Obesity	26	136.27003	0.00010
O	Vegetative anomalies	27	70.659166	0.01260
P	Behavioural anomalies	123	1220.7322	0.00220
Q	Myopathy or muscular anomalies	67	268.73373	0.55144
R	Blood cell anomalies	28	54.864273	0.22648
S	Ectodermal anomalies	86	528.48938	0.10349
T	Eye anomalies	139	1615.2409	0.00050
U	Skeletal anomalies	52	222.58986	0.02670
Ua	Limb anomalies	58	317.78992	0.00170
Ub	Vertebral/skull anomalies	28	41.341300	0.67543
Uc	Clefts	23	23.191930	0.84712
V	Cardiac malformations	59	248.41580	0.14249
W	Urogenital or renal anomalies	74	641.85449	0.00010
X	Other malformations	36	72.089949	0.64204

[1] Sum of link weights between the genes.

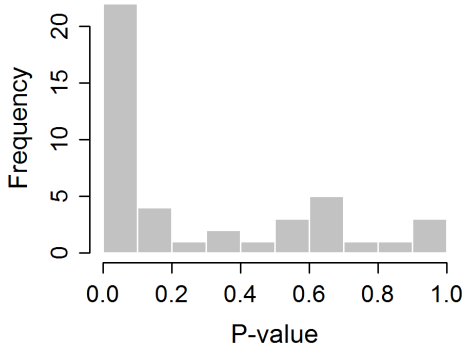


Figure 3.3: Distribution of P-values from the sub-sampling analyses with GTEx. The P-values obtained from the sub-sampling analyses are expected to be uniformly distributed under the null hypothesis. Their significant departure from a uniform distribution disproves the null hypothesis ($P=2.24\times10^{-7}$, Kolmogorov–Smirnov test).

Next, I wished to examine the coherence of these subgroups with corresponding gold-standard gene sets reported in the literature (Robinson *et al*. 2008). In order to avoid circularity, I used sets of genes recently implicated by *de novo* mutations in ID, epilepsy, autism and schizophrenia (Allen *et al*. 2013; de Ligt *et al*. 2012; Gulsuner *et al*. 2013; Neale *et al*. 2012; O'Roak *et al*. 2012b; Rauch *et al*. 2012; Sanders *et al*. 2012). Examining the clinical phenotypes, I have found that epilepsy and behavioural anomalies co-clustered with epilepsy and ID, but many of the other results were neither robust nor unambiguous (**Appendix, Figures A.13–15**).

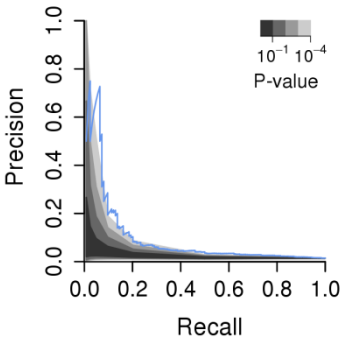
The relation of the main clinical classes to these reported gene sets was not clear and neither was the relevance of fly phenotypes to these sets (**Appendix, Figures A.16–19**). I also considered the relevance of the fly phenotypes to the clinical subgroups but have not noticed meaningful patterns (**Appendix, Figures A.20** and **A.21**).

3.3.2 Predictability of ID-associated genes

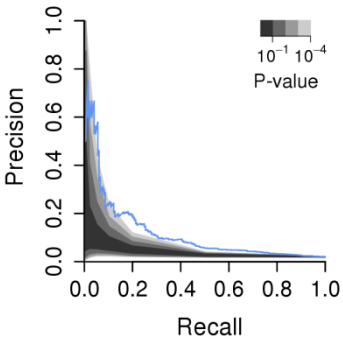
To test if the increased functional coherence of genes belonging to the clinical subgroups manifests in their increased predictability, I performed leave-one-out cross-validations. These analyses estimate how accurately one can predict genes relevant to ID based on known ID-associated genes. To assess the predictability of a clinical subgroup, all the 17,011 genes in the integrated network were rank-ordered by the sum of their link weights to the ten subgroup-associated genes closest to them and the highest ranking genes were predicted to be associated with the subgroup. The results of these analyses are represented by precision–recall curves, showing the proportion of true positives at different levels of coverage of known subgroup-associated genes (recall) in the predictions (**Figure 3.4**).

To show the significance of the results, I also estimated precision–recall curves with randomly selected genes sub-sampled from among the 650 ID-associated genes, the area they covered is colour-coded according to the corresponding P-value. The curves tended to be highly significant, suggesting that the sub-classified genes are better predictable than ID genes in general (**Figure 3.4**). However, these results were obtained with the integrated phenotypic-linkage network. The same analyses with the co-expression networks did not produce significant results. As the integrated network includes functional information from the literature, the findings regarding the predictability of ID-associated genes may be affected by circularity.

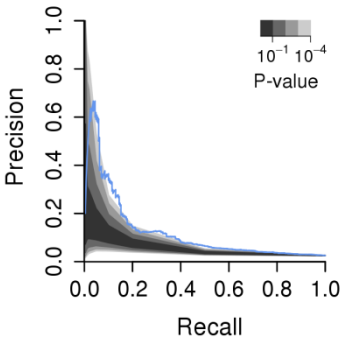
Microcephaly



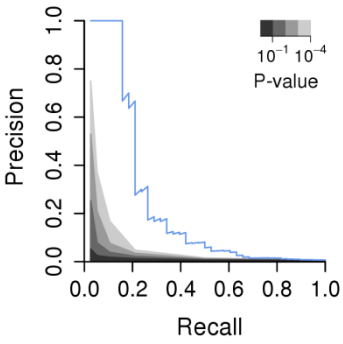
Epilepsy



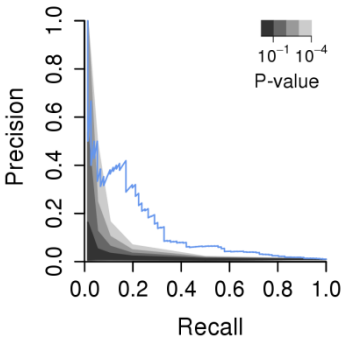
Neurological



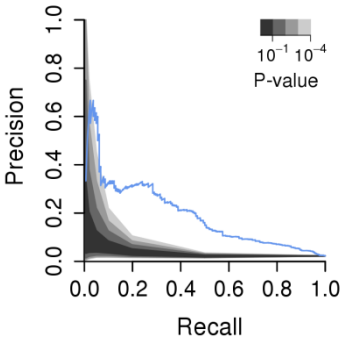
Endocrine anomalies



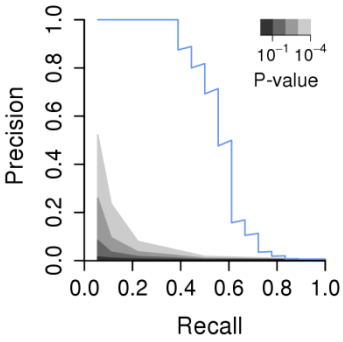
Brain malformations



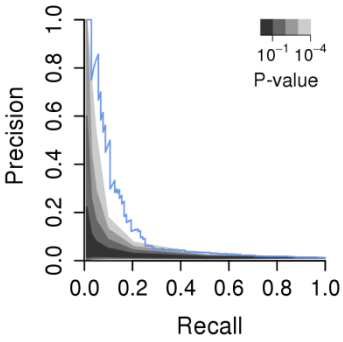
Metabolic



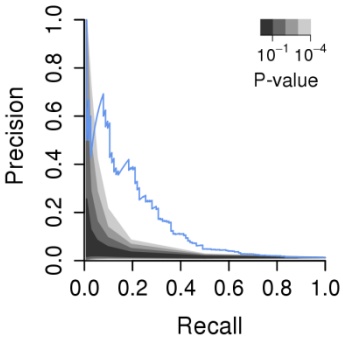
Obesity



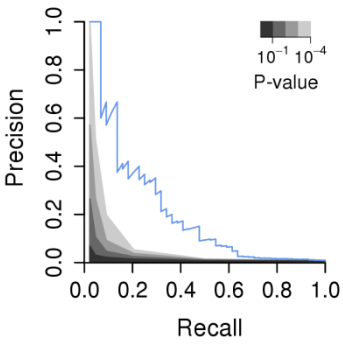
Behavioural anomalies



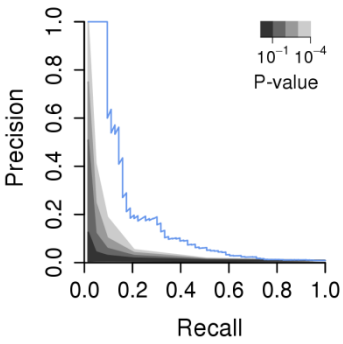
Eye anomalies



Limb anomalies



Urogenital/renal anomalies



All 650 genes

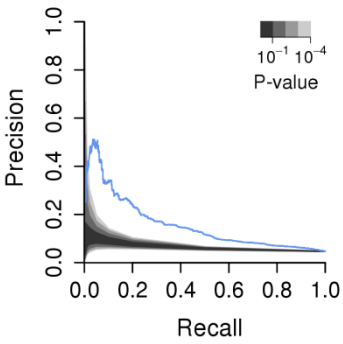


Figure 3.4: Prediction of genes associated with ID. The clustering of ID genes means that the genes most strongly linked to ID genes in the network are themselves likely to be associated with ID. To estimate how accurately one can predict ID genes based on known ID genes, I performed leave-one-out cross-validations. All the 17,011 genes in the integrated network were rank-ordered by the sum of their link weights to the ten subgroup-specific ID genes closest to them and the highest ranking genes were predicted to be associated with the subgroup. The precision–recall curves show the proportion of true positives at different levels of coverage of known subgroup-associated genes (recall) in the predictions. I also plotted precision–recall curves with randomly selected genes sub-sampled from among the 650 ID-associated genes, the area they covered is colour-coded according to the corresponding P-value.

3.4 Discussion

I have assessed the functional coherence of 650 ID-associated genes and found that they significantly cluster in multiple gene networks ($P<1.0\times10^{-4}$). Testing the inter-connectedness of genes within clinical subgroups, I have found an increased functional coherence in these subgroups, suggesting that the significant clustering of the 650 ID-associated genes is brought about by the presence of multiple functional clusters of genes among them (**Figure 3.3**). The increased functional coherence suggests that the introduced classification of ID-associated genes captures functional clusters, and thus this classification may be useful in uncovering distinct disease mechanisms.

I have also examined the coherence of the clinical subgroups with recently reported sets of genes implicated by sequencing studies and identified some consistency – such

as the co-clustering of epilepsy and behavioural anomalies with epilepsy and ID – but for the most part these results were inconclusive. The obscurity of these figures may have been caused by inadequacy of the used gene network, fault in the test, flaws in the classification, irrelevance of phenotypes and/or unreliability of the gene sets implicated by sequencing (Gratten *et al*. 2013).

I tested if the increased functional coherence of genes belonging to the clinical subgroups results in their increased predictability and obtained significant results with the phenotypic-linkage network (**Figure 3.4**). I made these predictions on the basis of guilt-by-association, which is subject to node-degree bias (Gillis and Pavlidis 2011). The classified genes had significantly high node degrees (**Figure 3.2**), but these bias were controlled for by the use of number-matched sub-sampling. However, the fact that the results were still significant may have partly been a consequence of circularity, namely, that some of the information contributing links to the network and some information used in the assembly of the clinical subgroups could be traced back to the same source.

Although the predictability of the classified genes is questionable, their increased functional coherence, observed with sub-sampling in a co-expression network, is clear. Thus, the gene sets forming the subgroups could be used in mapping biological pathways, which could eventually be targeted by molecular therapies during development (Hoischen *et al*. 2014). The sub-phenotyping of heterogeneous genetic disorders is likely to advance in cooperation with sequencing studies, resolving whether common genetic disorders are caused by many distinct mechanisms.

Chapter 4: Exome-sequencing of sporadic cases of Parkinson's disease

4.1 Introduction

The work presented herein derives from a collaboration between me, Sreeram Ramagopalan, Cynthia Sandor, Caleb Webber, Christopher Ponting, Richard Wade-Martins and the Oxford Parkinson's Disease Centre. I carried out all the work reported in the Results and Discussion sections, including everything presented in first-person singular. Sreeram Ramagopalan undertook an early analysis of the data and I inherited this project from him.

With a prevalence of approximately 0.3%, Parkinson's disease (PD) is the second most common neurodegenerative disorder after Alzheimer's disease (de Lau and Breteler 2006). Its most characteristic symptoms include shaking, rigidity, slowness of movement, depression and dementia. The motor symptoms result from the death of dopaminergic neurons, which can be temporarily counteracted by administration of Levodopa. However, primary PD is also termed as idiopathic – that is, having no known cause – which shows how little is known about its aetiology. Although not long ago a meta-analysis of data from genome-wide association studies (GWAS) identified 11 genome-wide significant loci (Nalls *et al*. 2011), in most PD cases the cause of the disease remains unknown. A role for rare variants, which are not captured by GWAS, can be hypothesised.

Rare variants are identified by DNA sequencing, including whole-exome sequencing, which involves the targeted sequencing of the protein-coding regions of the genome. The identification of causative variants would be invaluable for diagnostics and therapeutics, but thousands of cases and controls are required to robustly associate rare variants with a common disease (Foo *et al*. 2012; Kiezun *et al*. 2012). In addition, somatically acquired mutations have been implicated in PD and this sort of mutations is more difficult to identify (De 2011).

The aim of this study was to test rare variants identified through exome sequencing for association with PD. The case–control association analyses involved 230 case and 884 control samples after filtering. I carried out association tests of single variants and at the gene level but found none of the variants to be significantly associated with PD after multiple-testing correction. I also tested the functional associations of the genes with the strongest p-values but found no functional signal.

4.2 Methods

4.2.1 Clinical characteristics of PD subjects

Established in September 2010, the Oxford Discovery cohort (http://opdc.medsci.ox.ac.uk) includes patients with idiopathic PD diagnosed in the previous 3.5 years according to UK PD Society Brain Bank diagnostic criteria (Hughes *et al*. 1992), recruited from a 2.9 million-large Thames-valley population with the aim of following up the cohort over the history of their disease. PD patients were prospectively recruited over two years from secondary and primary care following ethics committee approval and verbal and written consent. Clinical information was

pre-screened to exclude subjects with atypical parkinsonian disorders. Patients underwent a two-hour assessment by a neurologist and a research nurse. Following standardised assessment, patients could be diagnosed as having PD, with an estimated clinical probability for this diagnosis being made by the research clinicians. Clinical assessments were performed while the subjects were taking their normal medications.

250 PD subjects were recruited from the Discovery cohort from September 2010, of whom 224 (90%) were selected on the following basis: i) *LRRK2* gene positive (n=4), ii) GBA gene positive (n=4), iii) subject had undergone MRI brain scan (n=20), iv) subject had undergone Cerebrospinal Fluid (CSF) examination (n=39), v) subject had undergone skin biopsy (n=16), vi) subject had at least 1 first-degree PD-affected relative (n=52), vii) subject had at least 1 second-degree PD-affected relative (n=21) and viii) subject had a Montreal Cognitive Assessment (MoCA) score ≤22, indicating significant cognitive impairment (n=68). Of the 250 PD subjects, the male-to-female ratio was 0.65:0.35, mean (and SD) age was 68.0 (9.6) years (age range 33.0–90.0 years) and mean disease duration was 3.7 years. Ethnicity of the PD subjects was: 92% white British, 2.6% white Irish, 2.9% white other, 0.4% Indian and 2.1% other.

4.2.2 Exome sequencing

120µl of 25ng/µl of DNA was fragmented to an average size of 150bp (75-300 bp), and subjected to Illumina DNA sequencing library creation with Bravo automated liquid handling. Adapter ligated libraries were amplified through 6 cycles of PCR. 500ng/µl of each amplified library was hybridized to RNA baits (SureSelect Human All Exon 50Mb) and post sequence capture processed following the Agilent SureSelect protocol. Enriched libraries were post capture indexed with a unique DNA barcode through 12

cycles of PCR. Equimolar pools of 8 libraries were sequenced by the Illumina HiSeq machine with the 75 base paired-end protocol.

4.2.3 Data processing and variant calling

Sequencing reads from each lane were aligned to the human reference genome used in the 1000 Genomes phase 2 (Abecasis *et al*. 2012). The GATK and SAMtools toolkits were used to process reads and merge the data into a single BAM file per individual (Li *et al*. 2009; McKenna *et al*. 2010). SNP and indel calls were made with SAMtools (Li 2011) by pooling the alignments from all individual exome BAM files. These were standard tools and procedures necessary for the identification of mutations.

4.2.4 Control Exome samples

The following exome sample datasets were obtained from the UK10K dataset to be used as controls in a case–control association analysis: NEURO_IOP_COLLIER (170 samples), NEURO_ABERDEEN (347 samples), NEURO_EDINBURGH (219 samples), RARE_NEUROMUSCULAR (114 samples), RARE_THYROID (65 samples) and NEURO_IMGSAC (114 samples), making a total of 1029 control exomes.

4.2.5 Annotation

The calls were annotated with dbSNP build 137 IDs which are standard SNP IDs used in reporting the results of association studies; population allele frequencies were integrated from the 1000 Genomes Phase 1 (Abecasis *et al*. 2012). Variant consequence annotations were added with the Ensembl Variant Effect Predictor (McLaren *et al*. 2010) release 74. These provided coding consequence predictions useful for the identification of non-synonymous variants as well as SIFT and PolyPhen annotations.

4.2.6 Association analysis

Population stratification was examined through principal component analysis and single-variant association testing was also carried out with PLINK (Purcell *et al*. 2007). VCF files were converted to PED format accepted by PLINK with VCFtools (Danecek *et al*. 2011). Rare variant testing was performed with SKAT, a collapsing method for gene-level association analysis (Wu *et al*. 2011).

4.2.7 Functional-enrichment analysis

To test for functional-enrichment, I calculated the sum of link weights between genes and used this sum as a test statistic, comparing it to a null distribution formed of sums of link weights between the same number of random genes in the phenotypic-linkage network. I controlled for both node degree and coding-sequence (CDS) length during the randomizations, by selecting random genes the node degree and CDS length of which matched the node degree and CDS length of the studied genes, respectively. To each of the studied genes in turn I assigned a list of 100 genes with the same or most similar node degree and CDS length, as described in Chapter 3.

4.3 Results

To verify that there was no apparent population stratification which could confound findings, I carried out routine principal component analysis in PLINK. To test for relatives, I calculated identity-by-descent (IBD) values between all possible pairs of individuals with PLINK and removed one individual from each pair with an IBD value of 0.1875 or more, approximately corresponding to a first-cousin relation. 230 cases and 884 controls remained for the subsequent analyses. Not all positions in the exome

were effectively genotyped in all samples, therefore I restricted analyses to variants where genotypes were available in 99% of the samples. In total, 94,365 suitable SNVs were identified. The distribution of these SNVs by mutation category is shown in **Table 4.1**.

Known variants present in *GBA* (N370S and L444P) and *LRRK2* (R1441H and G2019S) were detected through sequencing. However, the *GBA* L444P variant failed on a filtering step (VQSLOD), as this variant had a variant quality score log odds ratio lower than the threshold. To include this variant in the analyses would have required lowering the VQSLOD threshold and including 11,407 additional SNVs. We took a conservative approach and did not lower the VQSLOD threshold.

Table 4.1: Summary statistics for SNVs called in all samples.

Total number of SNVs	94,365
Rare variants[1]	39,566
Polyphen: damaging	6,541
STOP gained or lost	313
Non-synonymous	21,235
Synonymous	17,210
Affected genes	10,252

[1] Minor allele frequency <0.01

4.3.1 Single variant association testing

I started with association testing of single variants for PD risk. Given the lack of power for identifying single variants associated with PD, I restricted single-variant association analysis to non-synonymous (stop gained or lost, splice region and missense) variants in both cases and controls. The p-value considered significant under Bonferroni

correction for the number of the tested non-synonymous SNVs (n=21,235) was $p=2.355 \times 10^{-6}$; no variant passed this threshold (**Figures 4.1** and **4.2**). I tried both Fisher's exact test and logistic regression with between 1–10 covariates in PLINK and obtained the best-fitting Q–Q plot with the use of two covariates in logistic regression (**Figures 4.1**).

Q–Q plots compare the shapes of distributions by plotting their quantiles against each other. The expected p-values (red line) are derived from a uniform distribution expected to be seen in the absence of a statistical signal. The lowest expected p-value is determined by the sample size; a substantial departure from the diagonal towards the top of the plot would reflect significantly associated variants or SNPs. However, a widespread departure from the diagonal could indicate inflated p-values owing to population stratification or other confounders. I tried different subsets of the UK10K controls and obtained the best-fitting Q–Q plot with the set of six control data sets listed in section 4.2.4. In summary, at the single variant level, I found no robust genome-wide significant SNV association with Parkinson's disease.

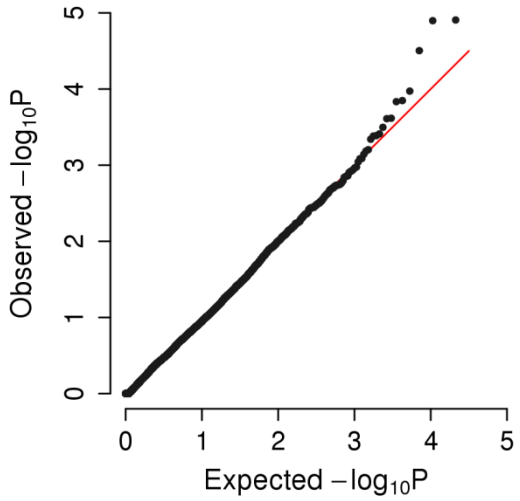


Figure 4.1: Quantile–quantile plot of p-values. I used logistic regression in PLINK with 2 covariates, only considering non-synonymous variants. The top p-values are not unexpected by chance.

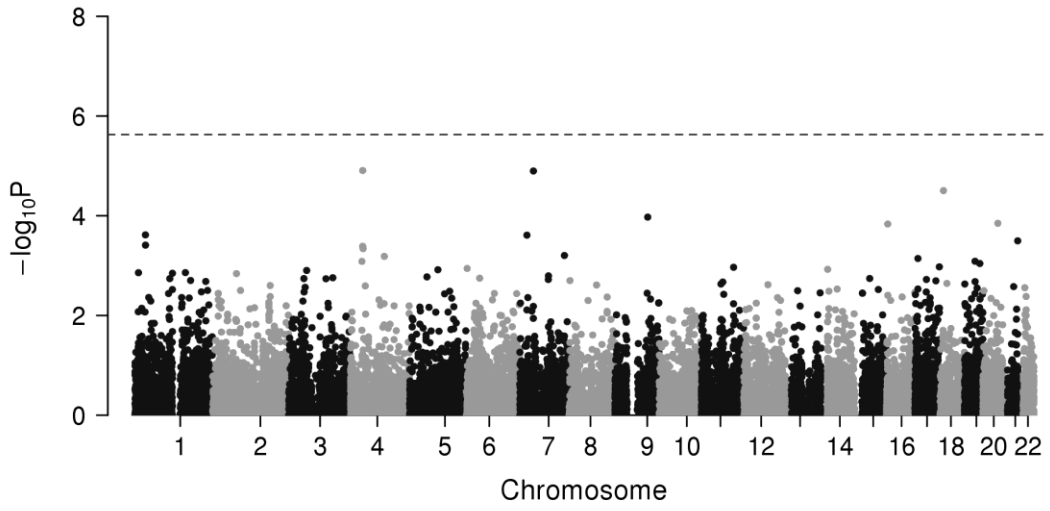


Figure 4.2: Manhattan plot of all non-synonymous variants' p-values. None of the SNVs reached genome-wide significance. The dashed line marks the Bonferroni correction threshold.

4.3.2 Gene-level association testing

Next, I performed association tests at the gene level for PD risk with SKAT (Wu *et al*. 2011). The p-value considered significant under Bonferroni correction based on this number of genes (n=21,235) was $p=2.355 \times 10^{-6}$: no gene passed this threshold for any test so no gene could be considered to have evidence of association with PD.

Table 4.2: Genes with the top/strongest P-values from the burden test.

ENSG ID	HUGO	P-value	#SNVs
ENSG00000168158	*OR2C1*	6.550e-05	3
ENSG00000122194	*PLG*	0.000190	3
ENSG00000121904	*CSMD2*	0.000220	8
ENSG00000205978	*NYNRIN*	0.000332	2
ENSG00000105479	*CCDC114*	0.000658	2
ENSG00000181523	*SGSH*	0.000755	1
ENSG00000162490	*DRAXIN*	0.001044	1
ENSG00000157259	*GATAD1*	0.001060	1
ENSG00000124570	*SERPINB6*	0.001078	1
ENSG00000152670	*DDX4*	0.001127	2

Although none of the case–control association tests produced a genome-wide significant result, the obtained p-values could be ordered and some variants tended to be more frequent in the cases. Likewise, some genes tended to have a higher burden of mutations (**Table 4.2**). I hypothesized that groups of these genes might show a functional enrichment and tested various numbers of genes with the strongest p-values for functional enrichment. However, I found no significant functional signal with any of the gene networks.

4.4 Discussion

I carried out association tests both at the level of single-nucleotide variants and at the gene level but found none of the variants to be significantly associated with Parkinson's disease after multiple-testing correction. Burden tests or collapsing methods, such as SKAT, are expected to have improved power over single-marker tests, but they still require thousands of samples to reach acceptable statistical power (Ladouceur *et al*. 2012; Wu *et al*. 2011). I also tested the functional associations of the genes with the strongest p-values but found no functional signal. The failure to detect a signal may have been the result of an inadequacy of the used functional-enrichment analysis method, powerlessness of the association tests at these sample sizes and a consequence of imperfect study design.

I tried different subsets of the UK10K controls and obtained the best-fitting Q–Q plot with the set of six control data sets described in section 4.2.4. (**Figure 4.1**). Notably, the use of the two largest UK10K cohorts, COHORT_TWINS (~1,700 samples) and COHORT_ALSPAC (740 samples), produced inflated p-values that could not be controlled by filtering or covariates. I found that some of the responsible minor allele frequencies in these two cohorts were very different from the allele frequencies in other controls, suggesting a population stratification problem.

DNA sequencing enables the direct detection of causal mutations and sequencing studies are superseding microarray-based genotyping methods, but some of the limitations of GWAS still apply to these studies. Thousands of cases and controls are required to robustly associate rare variants with a common disease and population stratification has to be controlled for (Foo *et al*. 2012; Kiezun *et al*. 2012). The

heritability of PD is not high, so it can be expected that the identification of causative variants or risk alleles for PD may be more difficult than for some other, common neurological disorders (Wirdefeldt *et al*. 2011). In addition, somatically acquired mutations accumulated in aging tissues may be behind a substantial number of PD cases and this sort of mutations is challenging to identify (De 2011).

Chapter 5: Conclusions

The main advantage of the phenotypic-linkage network may lie in the extraction of extra information from bio-ontologies by means of semantic similarity. The use of semantic similarity with Gene Ontology has been well established (Pesquita *et al*. 2009), and I applied it to other data sources as well, including the Mammalian Phenotype Ontology (Eppig *et al*. 2012). The use of semantic similarity enabled me to extract more information about functional associations between human genes from diverse data types and to assess the reliability of inferred associations.

I could confirm the hypothesis that dysfunctions of genes associated with each other in terms of functional genomic and proteomic data tend to give rise to the same phenotypes on a genomic and phenomic scale (**Figure 2.1**, p. 40). These correlations are consistent with the concept that the products of genes whose mutations are implicated in the same disease function together in the same biological pathways and it is the disruption of these pathways that underlies the disease (Brunner and van Driel 2004). The overlapping nature of pathways is well captured by gene networks, representing functional links among a large number of genes.

Gene networks are now commonly used in functional-enrichment analyses, but I have found that these analyses are often confounded by coding-sequence (CDS) length bias. Gene length bias had been implicated in studies of CNVs (Raychaudhuri *et al*. 2010)

and in RNA-seq analyses (Young *et al*. 2010), but their relevance to DNA sequencing studies has been overlooked. Instead, O'Roak *et al*. considered gene-specific mutation rate estimates while Neale *et al*. controlled for node degree, thus overestimating the functional enrichment in the implicated genes.

Overlaps and convergences of genes implicated in different neurodevelopmental disorders (Hoischen *et al*. 2014) may also be driven by CDS length bias. It is unclear what proportion of the long genes implicated in disease are true positives. The comparable length of genes mutated in cases and unaffected siblings suggests that numerous variants are in fact neutral mutations, which are likely to appear in genes with longer CDS. Thus, it is important to control for CDS length in functional analyses in order to assess the functional signal.

Room for improvement remains concerning the properties of gene networks as well. They represent the functional associations between genes as static states, as a snapshot, or an average state of an otherwise dynamic system (Fraser and Marcotte 2004). Gene networks specific to a cell type, tissue or condition are expected to better represent a biological system of interest and they could provide a substantial improvement over a universal network. Furthermore, the identification of a perturbed pathway does not equal the identification of a disease mechanism, it simply suggests a molecular system for further experimental investigations.

I applied my methods to clinical phenotypes related to intellectual disability (ID) and have found an increased functional coherence in the subgroups, which suggested that the sub-classification of ID-associated genes captured functional clusters or pathways. Thus, functional dissection of heterogeneous disorders into different subphenotypes

could facilitate the uncovering of the underlying biological pathways (Krumm *et al*. 2014). I have also observed an increased predictive potential in identifying genes associated with these phenotypes, but this analysis may have been affected by circular reasoning.

I have also performed case–control association analyses of variants from an exome-sequencing study of Parkinson's disease and tested the functional associations of the mutated genes, but found no signal. The failure to detect a signal may have been caused by limitations of my functional-enrichment analysis method, a powerlessness of the association tests owing to small sample sizes and by the imperfect study design.

I have advanced a framework for the identification of functional enrichments, both by technical improvements and by identifying coding-sequence length bias and pointing out their implications. These findings are hoped to facilitate the identification of biological pathways and disease mechanisms.

Future areas of possible research include the utilisation of new data types, such as genomic binding sites for transcription factors and enhancer regions, which can be used to infer potential links between genes based on shared transcription factors. These could be weighted and assessed in the same manner as the other data types dealt with earlier. Doubtless there will be challenges specific to each new data type during their utilisation, but the methodology to test and optimise their use is in place now.

Further applications of a phenotypic-linkage networks include the identification of clusters of functionally-related genes located next to each other on a chromosome; the disruption of such clusters by a CNV may be a frequent cause of developmental

disorders (Andrews *et al*. 2015). Furthermore, gene networks based on select datasets can be used for improved prediction of haploinsufficient genes and the functional consequences of gene disruption (Steinberg *et al*. 2015). Considering the surge of interest in sequencing studies and disease genomics, gene network approaches are not likely to be forgotten anytime soon.

References

Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, McVean GA (2012) An integrated map of genetic variation from 1,092 human genomes. Nature 491:56-65

Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR (2010) A method and server for predicting damaging missense mutations. Nat Methods 7:248-9

Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, Boldrick JC, Sabet H, Tran T, Yu X, Powell JI, Yang L, Marti GE, Moore T, Hudson J, Jr., Lu L, Lewis DB, Tibshirani R, Sherlock G, Chan WC, Greiner TC, Weisenburger DD, Armitage JO, Warnke R, Levy R, Wilson W, Grever MR, Byrd JC, Botstein D, Brown PO, Staudt LM (2000) Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature 403:503-11

Allen AS, Berkovic SF, Cossette P, Delanty N, Dlugos D, Eichler EE, Epstein MP, Glauser T, Goldstein DB, Han Y, Heinzen EL, Hitomi Y, Howell KB, Johnson MR, Kuzniecky R, Lowenstein DH, Lu YF, Madou MR, Marson AG, Mefford HC, Esmaeeli Nieh S, O'Brien TJ, Ottman R, Petrovski S, Poduri A, Ruzzo EK, Scheffer IE, Sherr EH, Yuskaitis CJ, Abou-Khalil B, Alldredge BK, Bautista JF, Boro A, Cascino GD, Consalvo D, Crumrine P, Devinsky O, Fiol M, Fountain NB, French J, Friedman D, Geller EB, Glynn S, Haut SR, Hayward J, Helmers SL, Joshi S, Kanner A, Kirsch HE, Knowlton RC, Kossoff EH, Kuperman R, McGuire SM, Motika PV, Novotny EJ, Paolicchi JM, Parent JM, Park K, Shellhaas RA, Shih JJ, Singh R, Sirven J, Smith MC, Sullivan J, Lin Thio L, Venkat A, Vining EP, Von Allmen GK, Weisenberg JL, Widdess-Walsh P, Winawer MR (2013) De novo mutations in epileptic encephalopathies. Nature 501:217-21

Alpi AF, Patel KJ (2009) Monoubiquitylation in the Fanconi anemia DNA damage response pathway. DNA Repair (Amst) 8:430-5

Anderson IM, Haddad PM, Scott J (2012) Bipolar disorder. BMJ 345:e8508

Andrews T, Honti F, Pfundt R, de Leeuw N, Hehir-Kwa J, Vulto-van Silfhout A, de Vries B, Webber C (2015) The clustering of functionally related genes contributes to CNV-mediated disease. Genome Res 25:802-13

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25:25-9

Brown SD, Moore MW (2012) The International Mouse Phenotyping Consortium: past and future perspectives on mouse phenotyping. Mamm Genome 23:632-40

Brunner HG, van Driel MA (2004) From syndrome families to functional genomics. Nat Rev Genet 5:545-51

Carter NP (2007) Methods and strategies for analyzing copy number variation using DNA microarrays. Nat Genet 39:S16-21

Caspi R, Altman T, Billington R, Dreher K, Foerster H, Fulcher CA, Holland TA, Keseler IM, Kothari A, Kubo A, Krummenacker M, Latendresse M, Mueller LA, Ong Q, Paley S, Subhraveti P, Weaver DS, Weerasinghe D, Zhang P, Karp PD (2013) The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. Nucleic Acids Res 42:D459-71

Chial H (2008) Gene mapping and disease. Nature Education 1:50

Cooper GM, Stone EA, Asimenos G, Green ED, Batzoglou S, Sidow A (2005) Distribution and intensity of constraint in mammalian genomic sequence. Genome Res 15:901-13

Cordaux R, Batzer MA (2009) The impact of retrotransposons on human genome evolution. Nat Rev Genet 10:691-703

Couto FM, Silva MJ, Coutinho PM (2007) Measuring semantic similarity between Gene Ontology terms. Data & Knowledge Engineering 61:137-152
Croft D, O'Kelly G, Wu G, Haw R, Gillespie M, Matthews L, Caudy M, Garapati P, Gopinath G, Jassal B, Jupe S, Kalatskaya I, Mahajan S, May B, Ndegwa N, Schmidt E, Shamovsky V, Yung C, Birney E, Hermjakob H, D'Eustachio P, Stein L (2011) Reactome: a database of reactions, pathways and biological processes. Nucleic Acids Res 39:D691-7
Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, McVean G, Durbin R (2011) The variant call format and VCFtools. Bioinformatics 27:2156-8
Darwin CR (1859) On the origin of species by means of natural selection, Lond.
de Lau LM, Breteler MM (2006) Epidemiology of Parkinson's disease. Lancet Neurol 5:525-35
de Ligt J, Willemsen MH, van Bon BW, Kleefstra T, Yntema HG, Kroes T, Vulto-van Silfhout AT, Koolen DA, de Vries P, Gilissen C, del Rosario M, Hoischen A, Scheffer H, de Vries BB, Brunner HG, Veltman JA, Vissers LE (2012) Diagnostic exome sequencing in persons with severe intellectual disability. N Engl J Med 367:1921-9
De S (2011) Somatic mosaicism in healthy human tissues. Trends Genet 27:217-23
Deane CM, Salwinski L, Xenarios I, Eisenberg D (2002) Protein interactions: two methods for assessment of the reliability of high throughput observations. Mol Cell Proteomics 1:349-56
Dietmann S, Georgii E, Antonov A, Tsuda K, Mewes HW (2009) The DICS repository: module-assisted analysis of disease-related gene lists. Bioinformatics 25:830-1
Down JLH (1866) Observations on an ethnic classification of idiots. London Hospital Reports 3:259-262
Edvardsen J, Torgersen S, Roysamb E, Lygren S, Skre I, Onstad S, Oien PA (2008) Heritability of bipolar spectrum disorders. Unity or heterogeneity? J Affect Disord 106:229-40
Ellegren H, Smith NG, Webster MT (2003) Mutation rate variation in the mammalian genome. Curr Opin Genet Dev 13:562-8
Eppig JT, Blake JA, Bult CJ, Kadin JA, Richardson JE (2012) The Mouse Genome Database (MGD): comprehensive resource for genetics and genomics of the laboratory mouse. Nucleic Acids Res 40:D881-6
Epstein CJ, Erickson RP, Wynshaw-Boris AJ (2008) Inborn errors of development : the molecular basis of clinical disorders of morphogenesis. Oxford University Press, New York, NY ; Oxford
Faraone SV, Perlis RH, Doyle AE, Smoller JW, Goralnick JJ, Holmgren MA, Sklar P (2005) Molecular genetics of attention-deficit/hyperactivity disorder. Biol Psychiatry 57:1313-23
Foo JN, Liu JJ, Tan EK (2012) Whole-genome and whole-exome sequencing in neurological diseases. Nat Rev Neurol 8:508-17
Francioli LC, Polak PP, Koren A, Menelaou A, Chun S, Renkens I, van Duijn CM, Swertz M, Wijmenga C, van Ommen G, Slagboom PE, Boomsma DI, Ye K, Guryev V, Arndt PF, Kloosterman WP, de Bakker PI, Sunyaev SR (2015) Genome-wide patterns and properties of de novo mutations in humans. Nat Genet 47:822-6
Franke L, van Bakel H, Fokkens L, de Jong ED, Egmont-Petersen M, Wijmenga C (2006) Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes. Am J Hum Genet 78:1011-25
Fraser AG, Marcotte EM (2004) A probabilistic view of gene function. Nat Genet 36:559-64
Frazer KA, Murray SS, Schork NJ, Topol EJ (2009) Human genetic variation and its contribution to complex traits. Nat Rev Genet 10:241-51
Garrod AE (1923) Inborn errors of metabolism. H. Frowde and Hodder & Stoughton, London
Gatz M, Reynolds CA, Fratiglioni L, Johansson B, Mortimer JA, Berg S, Fiske A, Pedersen NL (2006) Role of genes and environments for explaining Alzheimer disease. Arch Gen Psychiatry 63:168-74

Gillis J, Pavlidis P (2011) The impact of multifunctional genes on "guilt by association" analysis. PLoS One 6:e17258
Gilman SR, Iossifov I, Levy D, Ronemus M, Wigler M, Vitkup D (2011) Rare de novo variants associated with autism implicate a large functional network of genes involved in formation and function of synapses. Neuron 70:898-907
Grantham R (1974) Amino acid difference formula to help explain protein evolution. Science 185:862-4
Gratten J, Visscher PM, Mowry BJ, Wray NR (2013) Interpreting the role of de novo protein-coding mutations in neuropsychiatric disease. Nat Genet 45:234-8
Graves DJ (1999) Powerful tools for genetic analysis come of age. Trends Biotechnol 17:127-34
Gulsuner S, Walsh T, Watts AC, Lee MK, Thornton AM, Casadei S, Rippey C, Shahin H, Nimgaonkar VL, Go RC, Savage RM, Swerdlow NR, Gur RE, Braff DL, King MC, McClellan JM (2013) Spatial and temporal mapping of de novo mutations in schizophrenia to a fetal prefrontal cortical network. Cell 154:518-29
Hacia JG, Collins FS (1999) Mutational analysis using oligonucleotide microarrays. J Med Genet 36:730-6
Harris JC (2006) Intellectual disability : understanding its development, causes, classification, evaluation, and treatment. Oxford University Press, Oxford
Hastings PJ, Lupski JR, Rosenberg SM, Ira G (2009) Mechanisms of change in gene copy number. Nat Rev Genet 10:551-64
Hebert LE, Scherr PA, Bienias JL, Bennett DA, Evans DA (2003) Alzheimer disease in the US population: prevalence estimates using the 2000 census. Arch Neurol 60:1119-22
Hoischen A, Krumm N, Eichler EE (2014) Prioritization of neurodevelopmental disease genes by discovery of new mutations. Nat Neurosci 17:764-72
Honti F, Meader S, Webber C (2014) Unbiased functional clustering of gene variants with a phenotypic-linkage network. PLoS Comput Biol 10:e1003815
Hughes AJ, Daniel SE, Kilford L, Lees AJ (1992) Accuracy of the clinical diagnosis of idiopathic Parkinson's disease: a clinico-pathological study of 100 cases. J Neurol Neurosurg Psychiatry 55:181-4
Hunter S, Jones P, Mitchell A, Apweiler R, Attwood TK, Bateman A, Bernard T, Binns D, Bork P, Burge S, de Castro E, Coggill P, Corbett M, Das U, Daugherty L, Duquenne L, Finn RD, Fraser M, Gough J, Haft D, Hulo N, Kahn D, Kelly E, Letunic I, Lonsdale D, Lopez R, Madera M, Maslen J, McAnulla C, McDowall J, McMenamin C, Mi H, Mutowo-Muellenet P, Mulder N, Natale D, Orengo C, Pesseat S, Punta M, Quinn AF, Rivoire C, Sangrador-Vegas A, Selengut JD, Sigrist CJ, Scheremetjew M, Tate J, Thimmajanarthanan M, Thomas PD, Wu CH, Yeats C, Yong SY (2012) InterPro in 2011: new developments in the family and domain prediction database. Nucleic Acids Res 40:D306-12
Jimenez-Sanchez G, Childs B, Valle D (2001) Human disease genes. Nature 409:853-5
Kampmann B, Hemingway C, Stephens A, Davidson R, Goodsall A, Anderson S, Nicol M, Scholvinck E, Relman D, Waddell S, Langford P, Sheehan B, Semple L, Wilkinson KA, Wilkinson RJ, Ress S, Hibberd M, Levin M (2005) Acquired predisposition to mycobacterial disease due to autoantibodies to IFN-gamma. J Clin Invest 115:2480-8
Kanehisa M, Goto S, Furumichi M, Tanabe M, Hirakawa M (2010) KEGG for representation and analysis of molecular networks involving diseases and drugs. Nucleic Acids Res 38:D355-60
Kerrien S, Aranda B, Breuza L, Bridge A, Broackes-Carter F, Chen C, Duesbury M, Dumousseau M, Feuermann M, Hinz U, Jandrasits C, Jimenez RC, Khadake J, Mahadevan U, Masson P, Pedruzzi I, Pfeiffenberger E, Porras P, Raghunath A, Roechert B, Orchard S, Hermjakob H (2012) The IntAct molecular interaction database in 2012. Nucleic Acids Res 40:D841-6

Khatri P, Sirota M, Butte AJ (2012) Ten years of pathway analysis: current approaches and outstanding challenges. PLoS Comput Biol 8:e1002375
Kiezun A, Garimella K, Do R, Stitziel NO, Neale BM, McLaren PJ, Gupta N, Sklar P, Sullivan PF, Moran JL, Hultman CM, Lichtenstein P, Magnusson P, Lehner T, Shugart YY, Price AL, de Bakker PI, Purcell SM, Sunyaev SR (2012) Exome sequencing and the genetic basis of complex traits. Nat Genet 44:623-30
Kjeldsen MJ, Kyvik KO, Christensen K, Friis ML (2001) Genetic and environmental factors in epilepsy: a population-based study of 11900 Danish twin pairs. Epilepsy Res 44:167-78
Krumm N, O'Roak BJ, Shendure J, Eichler EE (2014) A de novo convergence of autism genetics and molecular neuroscience. Trends Neurosci 37:95-105
Kryukov GV, Pennacchio LA, Sunyaev SR (2007) Most rare missense alleles are deleterious in humans: implications for complex disease and association studies. Am J Hum Genet 80:727-39
Kumar P, Radhakrishnan J, Chowdhary MA, Giampietro PF (2001) Prevalence and patterns of presentation of genetic disorders in a pediatric emergency department. Mayo Clin Proc 76:777-83
Ladouceur M, Dastani Z, Aulchenko YS, Greenwood CM, Richards JB (2012) The empirical power of rare variant association methods: results from sanger sequencing in 1,998 individuals. PLoS Genet 8:e1002496
Lage K, Karlberg EO, Storling ZM, Olason PI, Pedersen AG, Rigina O, Hinsby AM, Tumer Z, Pociot F, Tommerup N, Moreau Y, Brunak S (2007) A human phenome-interactome network of protein complexes implicated in genetic disorders. Nat Biotechnol 25:309-16
Lee I, Blom UM, Wang PI, Shim JE, Marcotte EM (2011) Prioritizing candidate disease genes by network-based boosting of genome-wide association data. Genome Res 21:1109-21
Lee I, Date SV, Adai AT, Marcotte EM (2004) A probabilistic functional network of yeast genes. Science 306:1555-8
Li H (2011) A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics 27:2987-93
Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R (2009) The Sequence Alignment/Map format and SAMtools. Bioinformatics 25:2078-9
Liu CC, Kanekiyo T, Xu H, Bu G (2013) Apolipoprotein E and Alzheimer disease: risk, mechanisms and therapy. Nat Rev Neurol 9:106-18
Lonie IM, Hippocrates (1981) The Hippocratic treatises, "On generation," "On the nature of the child," "Diseases IV" : a commentary. De Gruyter, Berlin ; New York
Lukk M, Kapushesky M, Nikkila J, Parkinson H, Goncalves A, Huber W, Ukkonen E, Brazma A (2010) A global map of human gene expression. Nat Biotechnol 28:322-4
Macklon NS, Geraedts JP, Fauser BC (2002) Conception to ongoing pregnancy: the 'black box' of early pregnancy loss. Hum Reprod Update 8:333-43
Mardis ER (2008) The impact of next-generation sequencing technology on genetics. Trends Genet 24:133-41
McGrath J, Saha S, Chant D, Welham J (2008) Schizophrenia: a concise overview of incidence, prevalence, and mortality. Epidemiol Rev 30:67-76
McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 20:1297-303
McKusick VA (1966) Mendelian inheritance in man : catalogs of autosomal dominant, autosomal recessive, and X-linked phenotypes. Heinemann, London
McLaren W, Pritchard B, Rios D, Chen Y, Flicek P, Cunningham F (2010) Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. Bioinformatics 26:2069-70

Michaelson JJ, Shi Y, Gujral M, Zheng H, Malhotra D, Jin X, Jian M, Liu G, Greer D, Bhandari A, Wu W, Corominas R, Peoples A, Koren A, Gore A, Kang S, Lin GN, Estabillo J, Gadomski T, Singh B, Zhang K, Akshoomoff N, Corsello C, McCarroll S, Iakoucheva LM, Li Y, Wang J, Sebat J (2012) Whole-genome sequencing in autism identifies hot spots for de novo germline mutation. Cell 151:1431-42

Morgan TH (1915) The mechanism of Mendelian heredity. H. Holt, New York

Morgan TH (1925) Evolution and genetics. Princeton University Press, Princeton

Nalls MA, Plagnol V, Hernandez DG, Sharma M, Sheerin UM, Saad M, Simon-Sanchez J, Schulte C, Lesage S, Sveinbjornsdottir S, Stefansson K, Martinez M, Hardy J, Heutink P, Brice A, Gasser T, Singleton AB, Wood NW (2011) Imputation of sequence variants for identification of genetic risks for Parkinson's disease: a meta-analysis of genome-wide association studies. Lancet 377:641-9

Nasmyth K, Haering CH (2009) Cohesin: its roles and mechanisms. Annu Rev Genet 43:525-58

Neale BM, Kou Y, Liu L, Ma'ayan A, Samocha KE, Sabo A, Lin CF, Stevens C, Wang LS, Makarov V, Polak P, Yoon S, Maguire J, Crawford EL, Campbell NG, Geller ET, Valladares O, Schafer C, Liu H, Zhao T, Cai G, Lihm J, Dannenfelser R, Jabado O, Peralta Z, Nagaswamy U, Muzny D, Reid JG, Newsham I, Wu Y, Lewis L, Han Y, Voight BF, Lim E, Rossin E, Kirby A, Flannick J, Fromer M, Shakir K, Fennell T, Garimella K, Banks E, Poplin R, Gabriel S, DePristo M, Wimbish JR, Boone BE, Levy SE, Betancur C, Sunyaev S, Boerwinkle E, Buxbaum JD, Cook EH, Jr., Devlin B, Gibbs RA, Roeder K, Schellenberg GD, Sutcliffe JS, Daly MJ (2012) Patterns and rates of exonic de novo mutations in autism spectrum disorders. Nature 485:242-5

Newschaffer CJ, Croen LA, Daniels J, Giarelli E, Grether JK, Levy SE, Mandell DS, Miller LA, Pinto-Martin J, Reaven J, Reynolds AM, Rice CE, Schendel D, Windham GC (2007) The epidemiology of autism spectrum disorders. Annu Rev Public Health 28:235-58

Ng PC, Henikoff S (2003) SIFT: Predicting amino acid changes that affect protein function. Nucleic Acids Res 31:3812-4

Nielsen TO, West RB, Linn SC, Alter O, Knowling MA, O'Connell JX, Zhu S, Fero M, Sherlock G, Pollack JR, Brown PO, Botstein D, van de Rijn M (2002) Molecular characterisation of soft tissue tumours: a gene expression study. Lancet 359:1301-7

O'Roak BJ, Vives L, Fu W, Egertson JD, Stanaway IB, Phelps IG, Carvill G, Kumar A, Lee C, Ankenman K, Munson J, Hiatt JB, Turner EH, Levy R, O'Day DR, Krumm N, Coe BP, Martin BK, Borenstein E, Nickerson DA, Mefford HC, Doherty D, Akey JM, Bernier R, Eichler EE, Shendure J (2012a) Multiplex targeted sequencing identifies recurrently mutated genes in autism spectrum disorders. Science 338:1619-22

O'Roak BJ, Vives L, Girirajan S, Karakoc E, Krumm N, Coe BP, Levy R, Ko A, Lee C, Smith JD, Turner EH, Stanaway IB, Vernot B, Malig M, Baker C, Reilly B, Akey JM, Borenstein E, Rieder MJ, Nickerson DA, Bernier R, Shendure J, Eichler EE (2012b) Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. Nature 485:246-50

Oti M, Brunner HG (2007) The modular nature of genetic diseases. Clin Genet 71:1-11

Pesquita C, Faria D, Bastos H, Ferreira AE, Falcao AO, Couto FM (2008) Metrics for GO based protein semantic similarity: a systematic evaluation. BMC Bioinformatics 9 Suppl 5:S4

Pesquita C, Faria D, Falcao AO, Lord P, Couto FM (2009) Semantic similarity in biomedical ontologies. PLoS Comput Biol 5:e1000443

Plomin R, Pedersen NL, Lichtenstein P, McClearn GE (1994) Variability and stability in cognitive abilities are largely genetic later in life. Behav Genet 24:207-15

Prendergast JG, Semple CA (2011) Widespread signatures of recent selection linked to nucleosome positioning in the human lineage. Genome Res 21:1777-87

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, Sham PC (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81:559-75

Ramanan VK, Shen L, Moore JH, Saykin AJ (2012) Pathway analysis of genomic data: concepts, methods, and prospects for future development. Trends Genet 28:323-32

Rauch A, Wieczorek D, Graf E, Wieland T, Endele S, Schwarzmayr T, Albrecht B, Bartholdi D, Beygo J, Di Donato N, Dufke A, Cremer K, Hempel M, Horn D, Hoyer J, Joset P, Ropke A, Moog U, Riess A, Thiel CT, Tzschach A, Wiesener A, Wohlleber E, Zweier C, Ekici AB, Zink AM, Rump A, Meisinger C, Grallert H, Sticht H, Schenck A, Engels H, Rappold G, Schrock E, Wieacker P, Riess O, Meitinger T, Reis A, Strom TM (2012) Range of genetic mutations associated with severe non-syndromic sporadic intellectual disability: an exome sequencing study. Lancet 380:1674-82

Raychaudhuri S, Korn JM, McCarroll SA, Altshuler D, Sklar P, Purcell S, Daly MJ (2010) Accurately assessing the risk of schizophrenia conferred by rare copy-number variation affecting genes with brain function. PLoS Genet 6:e1001097

Reijns MA, Kemp H, Ding J, de Proce SM, Jackson AP, Taylor MS (2015) Lagging-strand replication shapes the mutational landscape of the genome. Nature 518:502-6

Resnik P (1995) Using information content to evaluate semantic similarity in a taxonomy. Ijcai-95 - Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence, Vols 1 and 2:448-453

Robinson PN, Kohler S, Bauer S, Seelow D, Horn D, Mundlos S (2008) The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease. Am J Hum Genet 83:610-5

Robinson PN, Kohler S, Oellrich A, Wang K, Mungall CJ, Lewis SE, Washington N, Bauer S, Seelow D, Krawitz P, Gilissen C, Haendel M, Smedley D (2014) Improved exome prioritization of disease genes through cross-species phenotype comparison. Genome Res 24:340-8

Ronald A, Happe F, Bolton P, Butcher LM, Price TS, Wheelwright S, Baron-Cohen S, Plomin R (2006) Genetic heterogeneity between the three components of the autism spectrum: a twin study. J Am Acad Child Adolesc Psychiatry 45:691-9

Ropers HH (2007) New perspectives for the elucidation of genetic disorders. Am J Hum Genet 81:199-207

Rossin EJ, Lage K, Raychaudhuri S, Xavier RJ, Tatar D, Benita Y, Cotsapas C, Daly MJ (2011) Proteins encoded in genomic regions associated with immune-mediated disease physically interact and suggest underlying biology. PLoS Genet 7:e1001273

Ruepp A, Waegele B, Lechner M, Brauner B, Dunger-Kaltenbach I, Fobo G, Frishman G, Montrone C, Mewes HW (2010) CORUM: the comprehensive resource of mammalian protein complexes--2009. Nucleic Acids Res 38:D497-501

Samocha KE, Robinson EB, Sanders SJ, Stevens C, Sabo A, McGrath LM, Kosmicki JA, Rehnstrom K, Mallick S, Kirby A, Wall DP, MacArthur DG, Gabriel SB, DePristo M, Purcell SM, Palotie A, Boerwinkle E, Buxbaum JD, Cook EH, Jr., Gibbs RA, Schellenberg GD, Sutcliffe JS, Devlin B, Roeder K, Neale BM, Daly MJ (2014) A framework for the interpretation of de novo mutation in human disease. Nat Genet 46:944-50

Sander JW, Shorvon SD (1988) Epilepsy in the community. J R Coll Gen Pract 38:51-2

Sanders SJ, Murtha MT, Gupta AR, Murdoch JD, Raubeson MJ, Willsey AJ, Ercan-Sencicek AG, DiLullo NM, Parikshak NN, Stein JL, Walker MF, Ober GT, Teran NA, Song Y, El-Fishawy P, Murtha RC, Choi M, Overton JD, Bjornson RD, Carriero NJ, Meyer KA, Bilguvar K, Mane SM, Sestan N, Lifton RP, Gunel M, Roeder K, Geschwind DH, Devlin B, State MW (2012) De novo mutations revealed by whole-exome sequencing are strongly associated with autism. Nature 485:237-41

Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A 74:5463-7

Schaner ME, Davidson B, Skrede M, Reich R, Florenes VA, Risberg B, Berner A, Goldberg I, Givant-Horwitz V, Trope CG, Kristensen GB, Nesland JM, Borresen-Dale AL (2005)

Variation in gene expression patterns in effusions and primary tumors from serous ovarian cancer patients. Mol Cancer 4:26
Schuster-Bockler B, Lehner B (2012) Chromatin organization is a major influence on regional mutation rates in human cancer cells. Nature 488:504-7
Shyamsundar R, Kim YH, Higgins JP, Montgomery K, Jorden M, Sethuraman A, van de Rijn M, Botstein D, Brown PO, Pollack JR (2005) A DNA microarray survey of gene expression in normal human tissues. Genome Biol 6:R22
Stark C, Breitkreutz BJ, Chatr-Aryamontri A, Boucher L, Oughtred R, Livstone MS, Nixon J, Van Auken K, Wang X, Shi X, Reguly T, Rust JM, Winter A, Dolinski K, Tyers M (2011) The BioGRID Interaction Database: 2011 update. Nucleic Acids Res 39:D698-704
Stefl S, Nishi H, Petukh M, Panchenko AR, Alexov E (2013) Molecular mechanisms of disease-causing missense mutations. J Mol Biol 425:3919-36
Steinberg J, Honti F, Meader S, Webber C (2015) Haploinsufficiency predictions without study bias. Nucleic Acids Res
Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, Block D, Zhang J, Soden R, Hayakawa M, Kreiman G, Cooke MP, Walker JR, Hogenesch JB (2004) A gene atlas of the mouse and human protein-encoding transcriptomes. Proc Natl Acad Sci U S A 101:6062-7
Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A 102:15545-50
Sullivan PF, Kendler KS, Neale MC (2003) Schizophrenia as a complex trait: evidence from a meta-analysis of twin studies. Arch Gen Psychiatry 60:1187-92
Supek F, Lehner B (2015) Differential DNA mismatch repair underlies mutation rate variation across the human genome. Nature 521:81-4
Szklarczyk D, Franceschini A, Kuhn M, Simonovic M, Roth A, Minguez P, Doerks T, Stark M, Muller J, Bork P, Jensen LJ, von Mering C (2011) The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. Nucleic Acids Res 39:D561-8
Vanneste E, Voet T, Le Caignec C, Ampe M, Konings P, Melotte C, Debrock S, Amyere M, Vikkula M, Schuit F, Fryns JP, Verbeke G, D'Hooghe T, Moreau Y, Vermeesch JR (2009) Chromosome instability is common in human cleavage-stage embryos. Nat Med 15:577-83
Veltman JA, Brunner HG (2012) De novo mutations in human genetic disease. Nat Rev Genet 13:565-75
Visscher PM, Hill WG, Wray NR (2008) Heritability in the genomics era--concepts and misconceptions. Nat Rev Genet 9:255-66
Vockley J, Rinaldo P, Bennett MJ, Matern D, Vladutiu GD (2000) Synergistic heterozygosity: disease resulting from multiple partial defects in one or more metabolic pathways. Mol Genet Metab 71:10-8
Warde-Farley D, Donaldson SL, Comes O, Zuberi K, Badrawi R, Chao P, Franz M, Grouios C, Kazi F, Lopes CT, Maitland A, Mostafavi S, Montojo J, Shao Q, Wright G, Bader GD, Morris Q (2010) The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. Nucleic Acids Res 38:W214-20
Webber C (2011) Functional enrichment analysis with structural variants: pitfalls and strategies. Cytogenet Genome Res 135:277-85
Willcutt EG (2012) The prevalence of DSM-IV attention-deficit/hyperactivity disorder: a meta-analytic review. Neurotherapeutics 9:490-9
Winter RM (1996) Analysing human developmental abnormalities. Bioessays 18:965-71
Wirdefeldt K, Gatz M, Reynolds CA, Prescott CA, Pedersen NL (2011) Heritability of Parkinson disease in Swedish twins: a longitudinal study. Neurobiol Aging 32:1923 e1-8

Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X (2011) Rare-variant association testing for sequencing data with the sequence kernel association test. Am J Hum Genet 89:82-93

Young MD, Wakefield MJ, Smyth GK, Oshlack A (2010) Gene ontology analysis for RNA-seq: accounting for selection bias. Genome Biol 11:R14

Zapala MA, Hovatta I, Ellison JA, Wodicka L, Del Rio JA, Tennant R, Tynan W, Broide RS, Helton R, Stoveken BS, Winrow C, Lockhart DJ, Reilly JF, Young WG, Bloom FE, Barlow C (2005) Adult mouse brain gene expression patterns bear an embryologic imprint. Proc Natl Acad Sci U S A 102:10357-62

2.6 Appendix

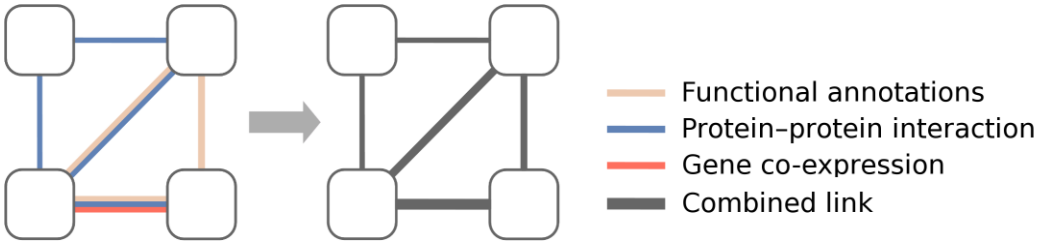


Figure A.1: Integration of different data types linking genes. When multiple data sources suggested functional linkage between the same two genes, I integrated the link weights into one for each gene pair. The rounded rectangles represent genes.

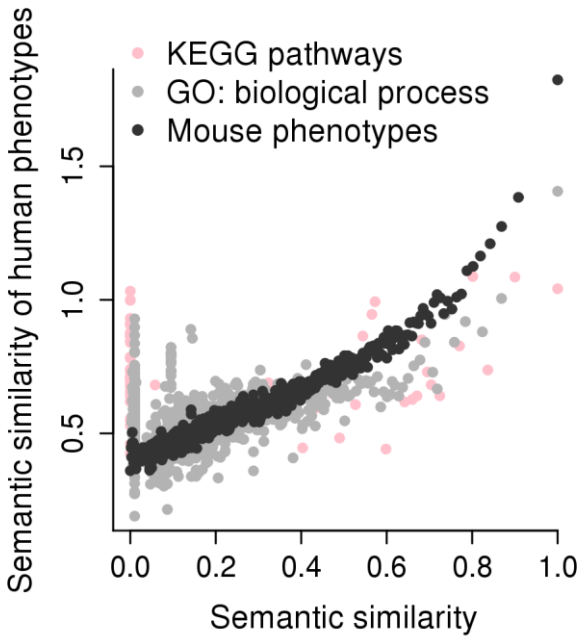


Figure A.2: Correlation between semantic similarities measured with different gene annotations. Gene pairs were ordered by their semantic similarity scores based on either the human Gene Ontology biological process (grey) or mouse phenotype annotations to genes (black dots). The ordered pairs were divided to bins of 1,000 and the median of the semantic similarity scores measured with Human Phenotype Ontology annotations has been calculated for each bin of gene pairs.

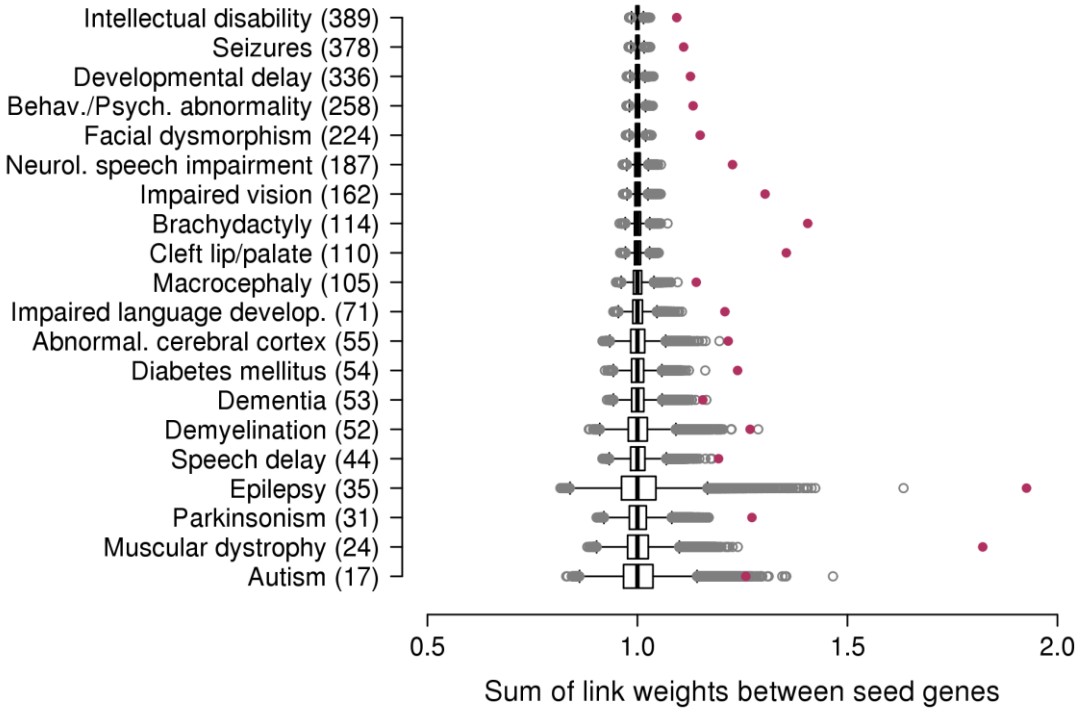


Figure A.3: Clustering of genes for Human Phenotype Ontology (HPO) phenotypes in a gene network built on the semantic similarity of mouse phenotypes. I calculated the sum of link weights among genes annotated with the same symptom and used it to represent the degree of clustering of these sets of genes. The box plots show the distribution of the sums of link weights for 100,000 sets of randomly selected genes with the same node degrees as the seed genes. The sums of link weights are presented as fold changes compared to the median of the specific distribution, set to equal 1 for each term. For each HPO phenotype, I randomly selected the same number of genes as there were annotated with that symptom in the HPO. This number is shown in parentheses; the red marks indicate the sum of link weights among the actual genes annotated with the corresponding HPO term.

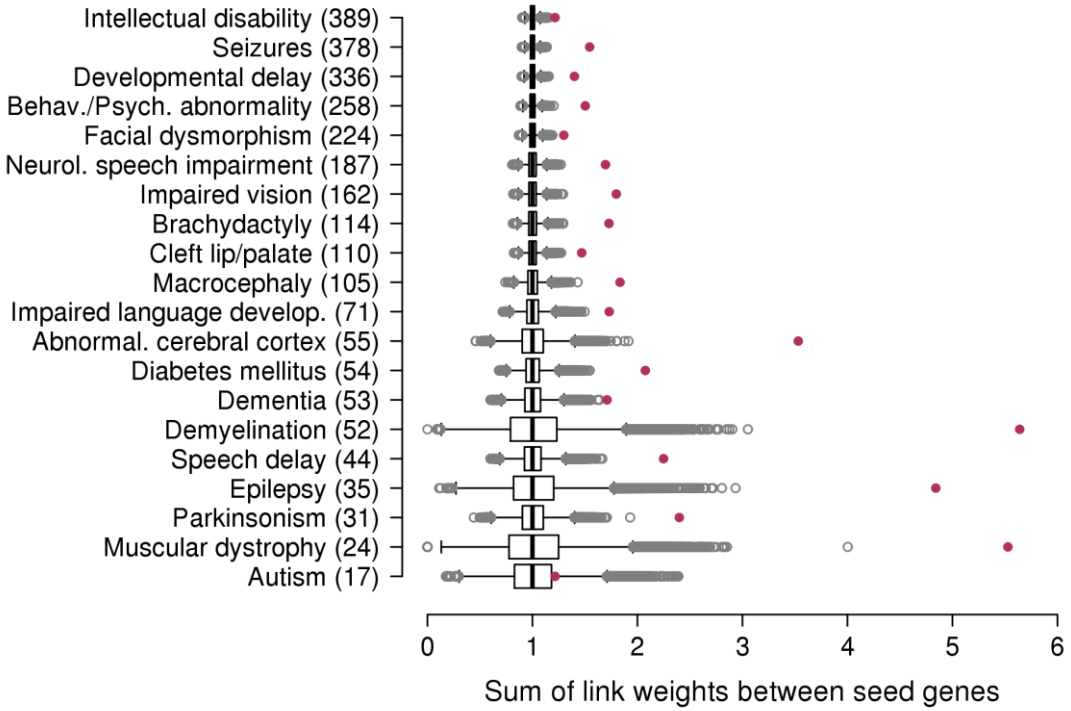


Figure A.4: Clustering of genes for Human Phenotype Ontology (HPO) phenotypes in a gene network built on the semantic similarity of Gene Ontology biological process annotations. I calculated the sum of link weights among genes annotated with the same symptom and used it to represent the degree of clustering of these sets of genes. The box plots show the distribution of the sums of link weights for 100,000 sets of randomly selected genes with the same node degrees as the seed genes. The sums of link weights are presented as fold changes compared to the median of the specific distribution, set to equal 1 for each term. For each HPO phenotype, I randomly selected the same number of genes as there were annotated with that symptom in the HPO. This number is shown in parentheses; the red marks indicate the sum of link weights among the actual genes annotated with the corresponding HPO term.

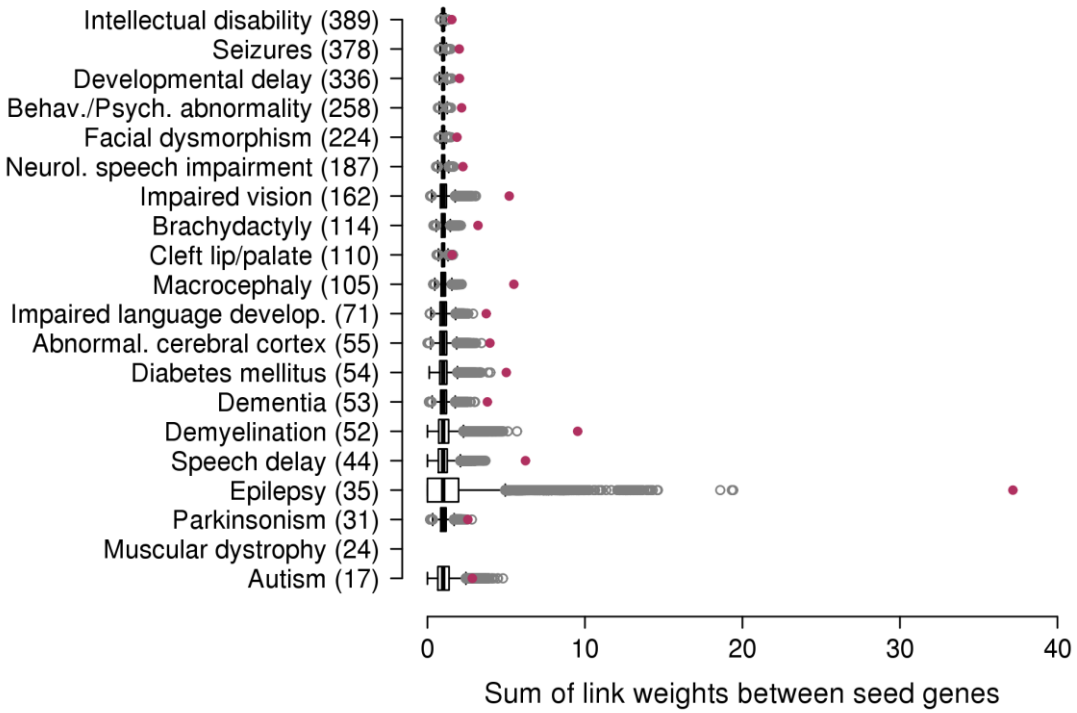


Figure A.5: Clustering of genes for Human Phenotype Ontology (HPO) phenotypes in a gene network based on protein–protein interactions. I calculated the sum of link weights among genes annotated with the same symptom and used it to represent the degree of clustering of these sets of genes. The box plots show the distribution of the sums of link weights for 100,000 sets of randomly selected genes with the same node degrees as the seed genes. The sums of link weights are presented as fold changes compared to the median of the specific distribution, set to equal 1 for each term. For each HPO phenotype, I randomly selected the same number of genes as there were annotated with that symptom in the HPO. This number is shown in parentheses; the red marks indicate the sum of link weights among the actual genes annotated with the corresponding HPO term.

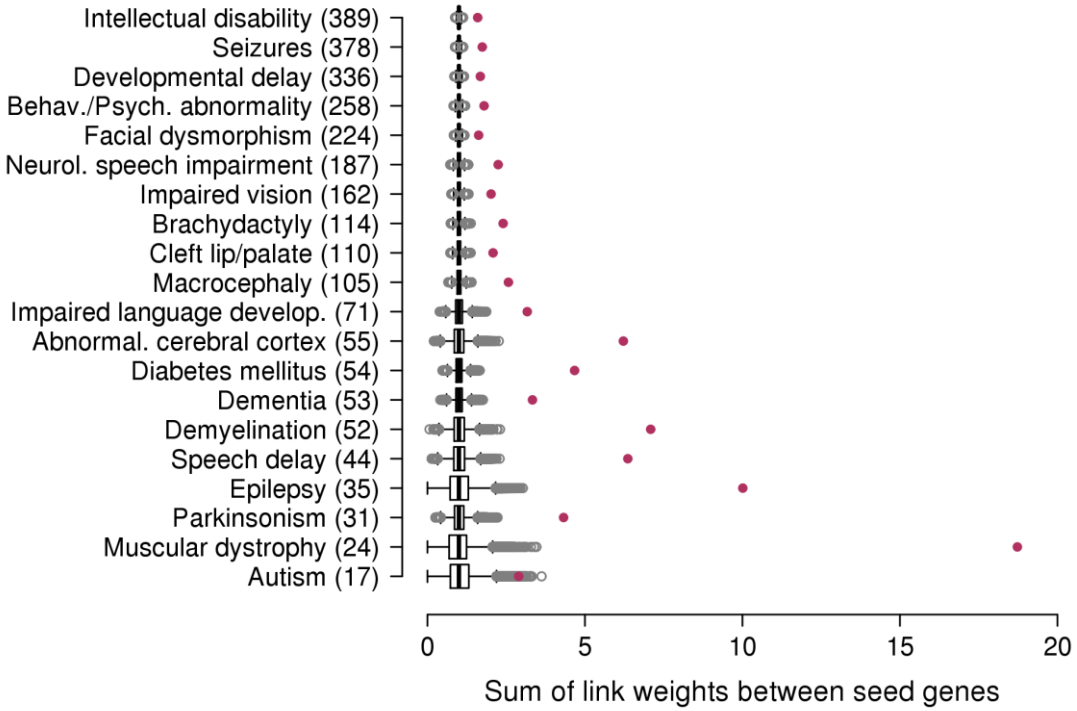


Figure A.6: Clustering of genes for Human Phenotype Ontology (HPO) phenotypes in a gene network built on the co-citation of mouse genes. I calculated the sum of link weights among genes annotated with the same symptom and used it to represent the degree of clustering of these sets of genes. The box plots show the distribution of the sums of link weights for 100,000 sets of randomly selected genes with the same node degrees as the seed genes. The sums of link weights are presented as fold changes compared to the median of the specific distribution, set to equal 1 for each term. For each HPO phenotype, I randomly selected the same number of genes as there were annotated with that symptom in the HPO. This number is shown in parentheses; the red marks indicate the sum of link weights among the actual genes annotated with the corresponding HPO term.

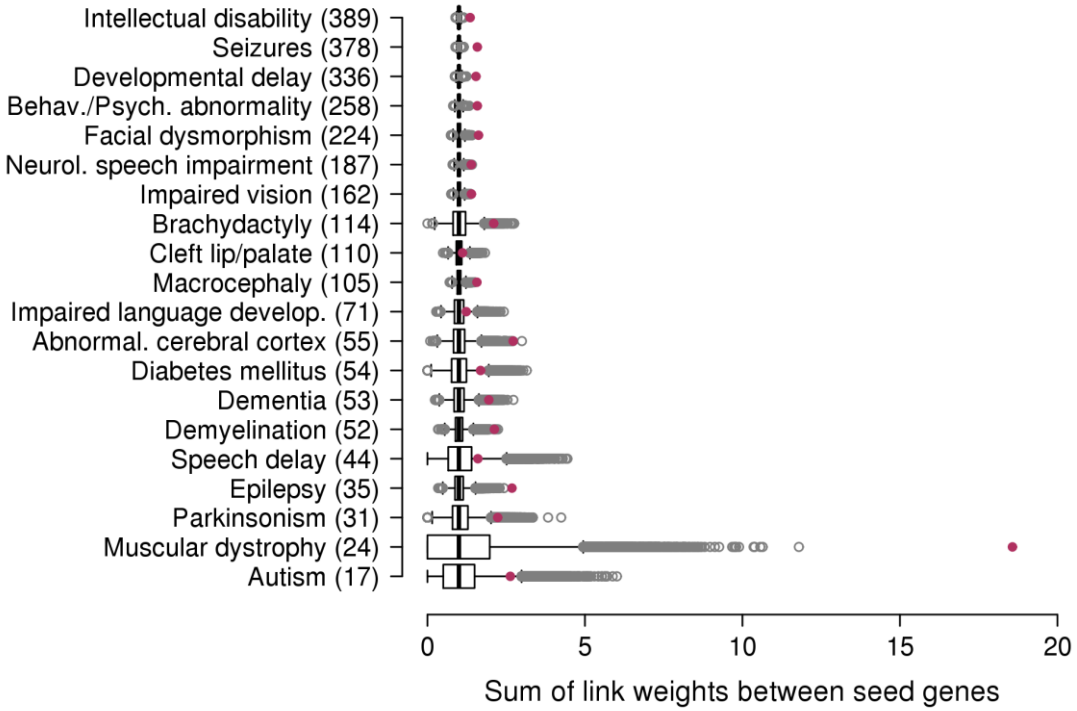


Figure A.7: Clustering of genes for Human Phenotype Ontology (HPO) phenotypes in an integrated co-expression network based on microarrays. I calculated the sum of link weights among genes annotated with the same symptom and used it to represent the degree of clustering of these sets of genes. The box plots show the distribution of the sums of link weights for 100,000 sets of randomly selected genes with the same node degrees as the seed genes. The sums of link weights are presented as fold changes compared to the median of the specific distribution, set to equal 1 for each term. For each HPO phenotype, I randomly selected the same number of genes as there were annotated with that symptom in the HPO. This number is shown in parentheses; the red marks indicate the sum of link weights among the actual genes annotated with the corresponding HPO term.

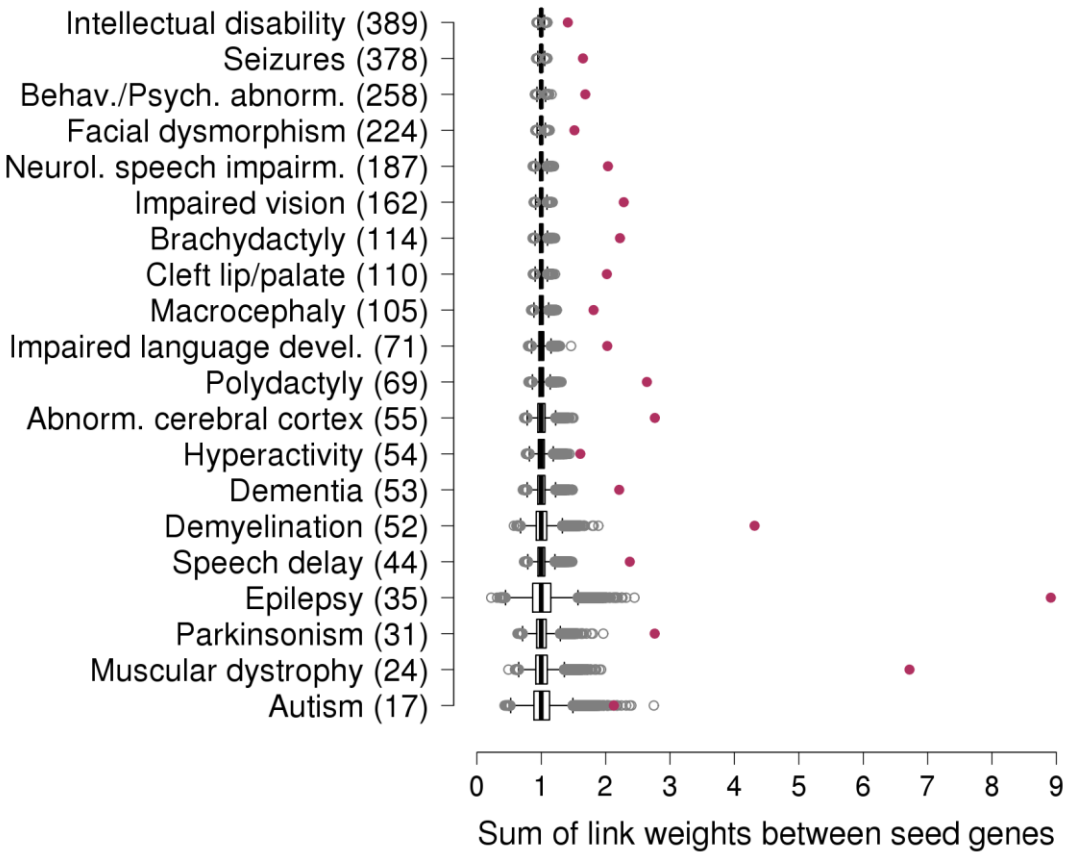


Figure A.8: Clustering of genes for Human Phenotype Ontology (HPO) phenotypes in the integrated phenotypic-linkage network. I calculated the sum of link weights among genes annotated with the same symptom and used it to represent the degree of clustering of these sets of genes. The box plots show the distribution of the sums of link weights for 100,000 sets of randomly selected genes with the same node degrees as the seed genes. The sums of link weights are presented as fold changes compared to the median of the specific distribution, set to equal 1 for each term. For each HPO phenotype, I randomly selected the same number of genes as there were annotated with that symptom in the HPO. This number is shown in parentheses; the red marks indicate the sum of link weights among the actual genes annotated with the corresponding HPO term.

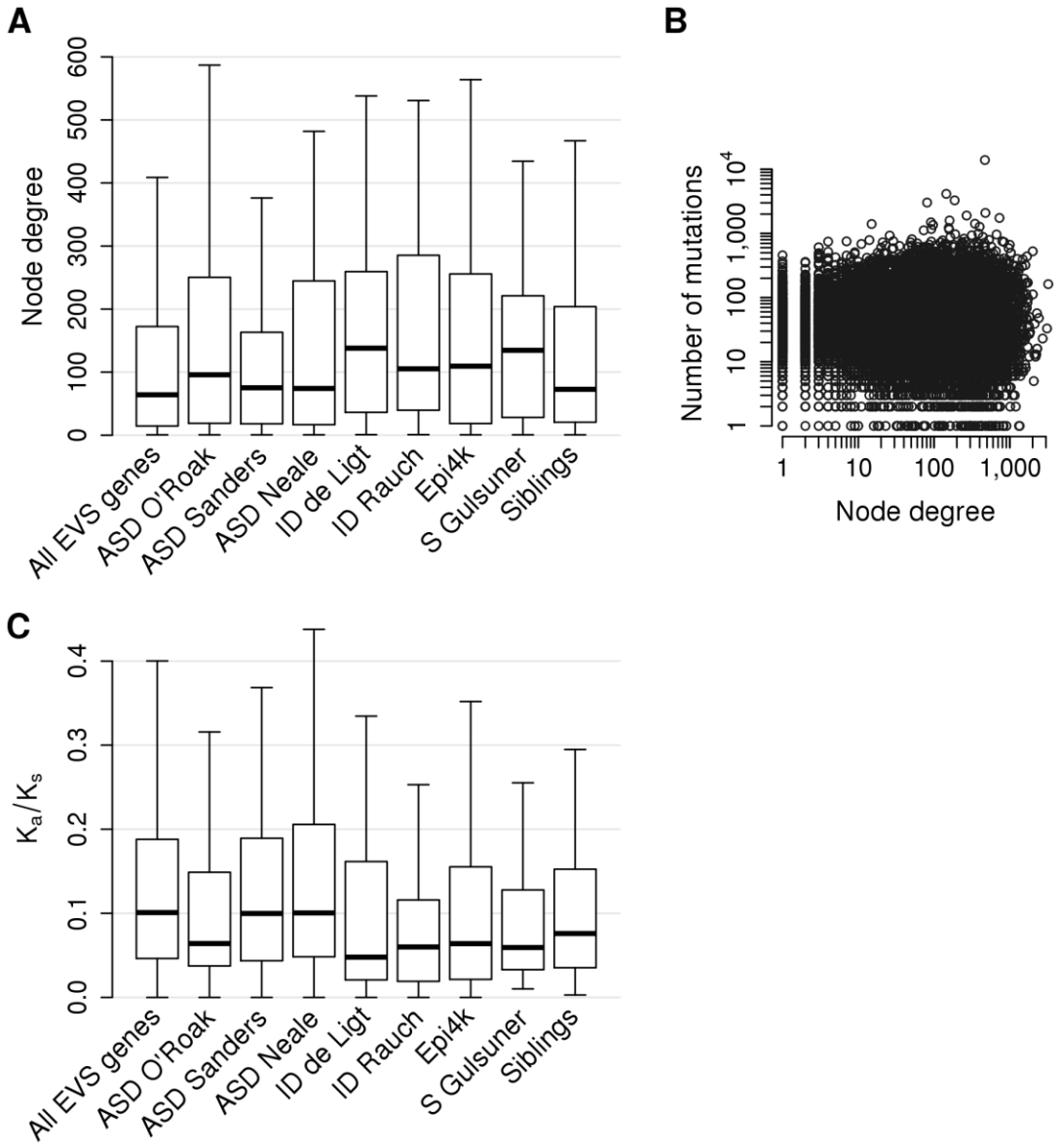


Figure A.9: Node degrees of genes with *de novo* variants (A) 'All EVS genes' denotes all genes in the Exome Variant Server, 'Siblings' denotes genes with *de novo* mutations in non-autistic siblings of ASD cases published by O'Roak *et al*. and Sanders *et al*. The node degrees are significantly higher only in the the O'Roak *et al*., de Ligt *et al*., Rauch *et al*. and Epi4k candidate gene sets. The node degrees represent the number of connections of a gene in the integrated phenotypic-linkage network. (B) Mutational burden does not correlate with node degree in the Exome Variant Server (Spearman's ρ=0.007; http://evs.gs.washington.edu/EVS). All nonsynonymous mutations were considered across all human chromosomes. (C) The same gene sets that have higher node degrees show significantly increased sequence conservation (lower K_a/K_s ratio), indicating that the degree bias could be due to a

functional signal in these gene sets. K_a/K_s is the ratio of the number of nonsynonymous substitutions per nonsynonymous site (K_a) to the number of synonymous substitutions per synonymous site (K_s), based on one-to-one orthologs between human and mouse genes.

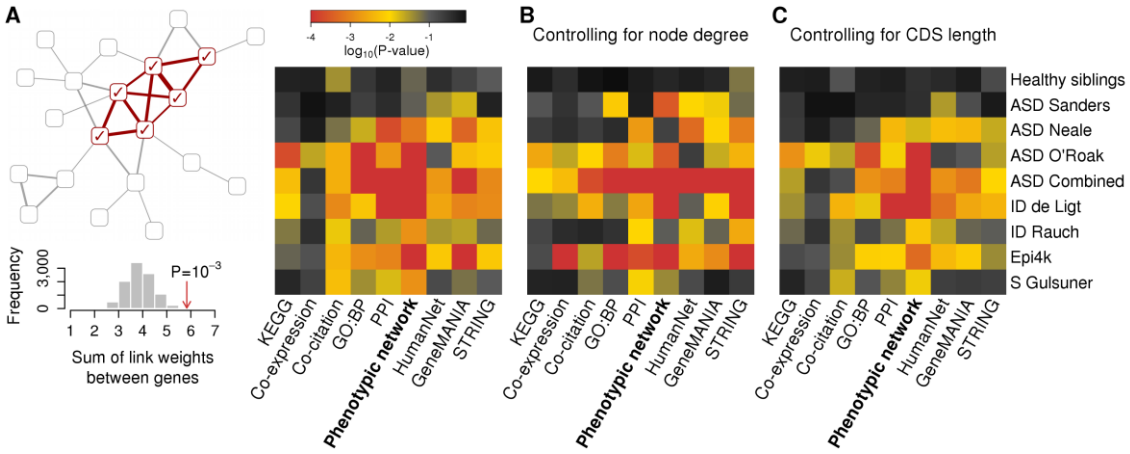


Figure A.10: Clustering of genes hit by *de novo* nonsynonymous substitutions. (A) I have examined the network properties of whole sets of genes with nonsynonymous mutations implicated by recent exome-sequencing studies in autism (ASD), severe intellectual disability (ID), epilepsy or schizophrenia (S). I calculated the sum of link weights among genes from a set and compared this sum to that calculated for randomized gene sets in order to assess the degree of functional clustering. (B and C) The implicated genes are significantly more strongly interconnected with each other by means of functional genomics data than random gene sets of the same size, but controlling for coding-sequence length considerably affects the p-values. The genes mutated in the same disease cluster most significantly in the integrated phenotypic-linkage network, while genes mutated in healthy controls do not cluster.

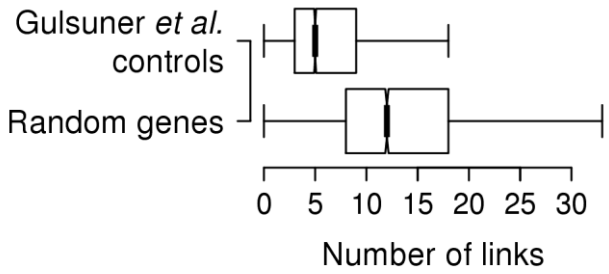


Figure A.11: Interconnectedness of controls used in simulations. I calculated the number of links between 54 randomly selected control genes carrying damaging mutations in unaffected siblings, as in Gulsuner *et al*., in the GeneMania physical interaction data set (http://pages.genemania.org/data). I also calculated the number of links between randomly selected genes matched in CDS length to the genes mutated in the Gulsuner *et al*. probands in the same network (*Random genes*). The box plots show the distribution of the numbers of links for 10,000 sets of randomly selected genes. The null distribution used in controlling for CDS length has a larger spread, indicating that controlling for CDS length in testing for clustering is more conservative.

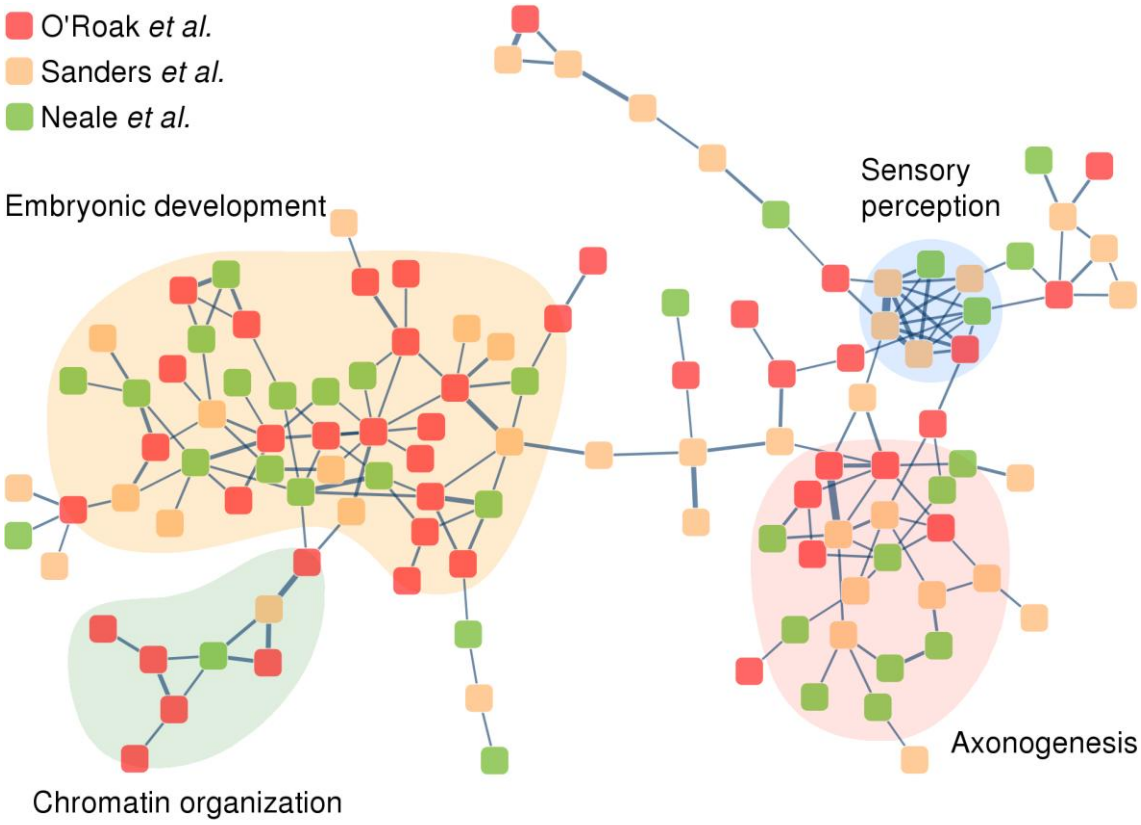


Figure A.12: Functional subclusters of genes implicated in autism within the integrated gene network. Only the strongest 166 links are shown among 115 genes. The terms represent the most significantly enriched GO biological process annotations among the genes forming the subclusters. Links based on the semantic similarity of GO annotations were included in the integrated network, but these enrichments are still useful in characterizing the subclusters and illustrate that the subclusters fit well with recent insights into the etiological variation underlying ASD (Krumm *et al*. 2014).

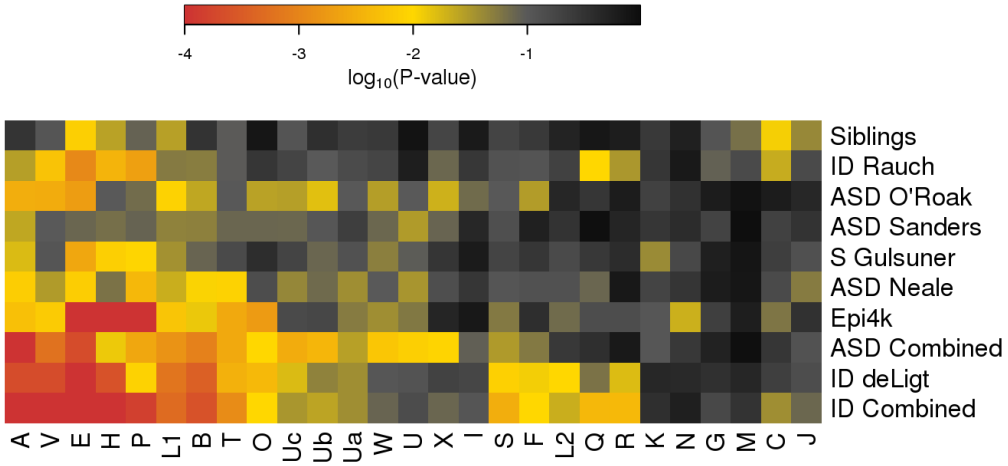


Figure A.13: Co-clustering of ID-accompanying clinical features with *de novo*-mutated genes in the integrated phenotypic-linkage network. I calculated the sum of the weighted links connecting these gene sets and compared this sum to that calculated by randomising each one of the gene sets in turn to assess the degree of functional clustering (Methods). 'Siblings' are unaffected controls from Sanders *et al*. and O'Roak *et al*. ASD: Autism spectrum disorder, Epi4k: epilepsy, S: schizophrenia.

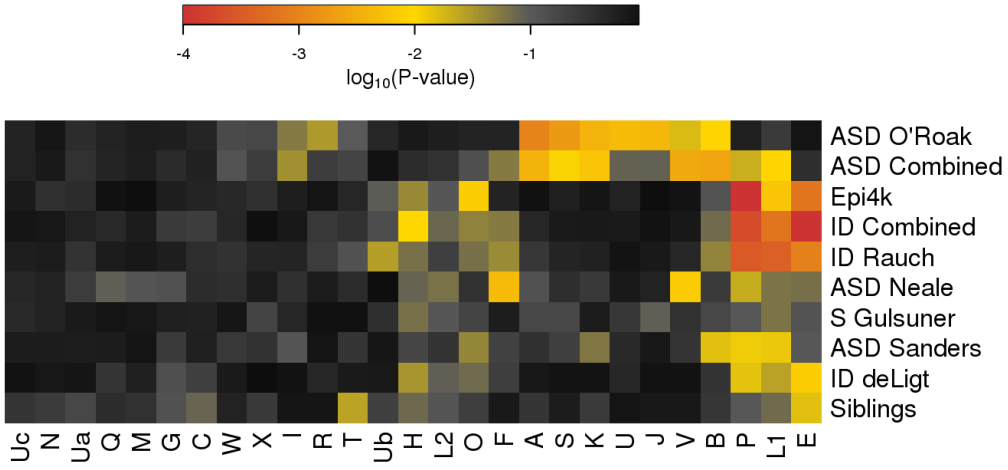


Figure A.14: Co-clustering of ID-accompanying clinical features with *de novo*-mutated genes in a co-expression network based on microarrays.

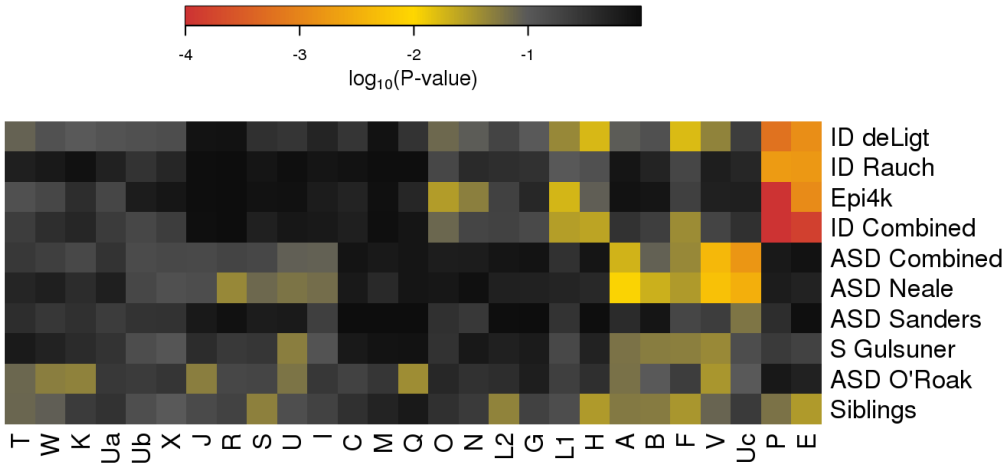


Figure A.15: Co-clustering of ID-accompanying clinical features with *de novo*-mutated genes in a co-expression network based on GTEx.

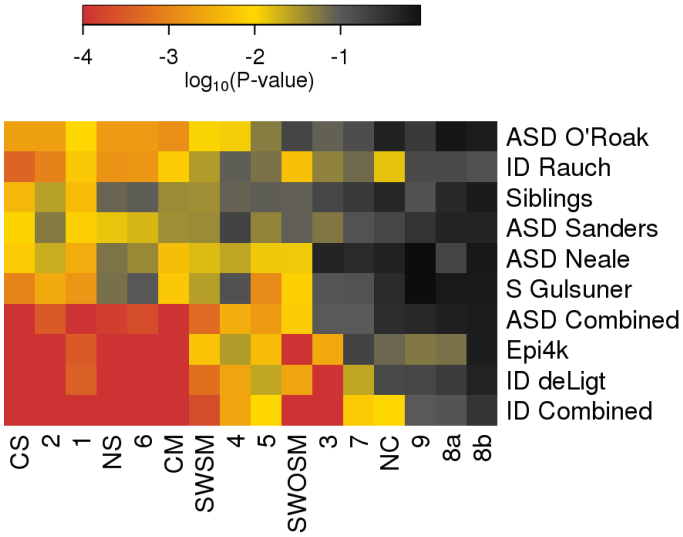


Figure A.16: Co-clustering of the clinical classes with *de novo*-mutated genes in the phenotypic-linkage network.

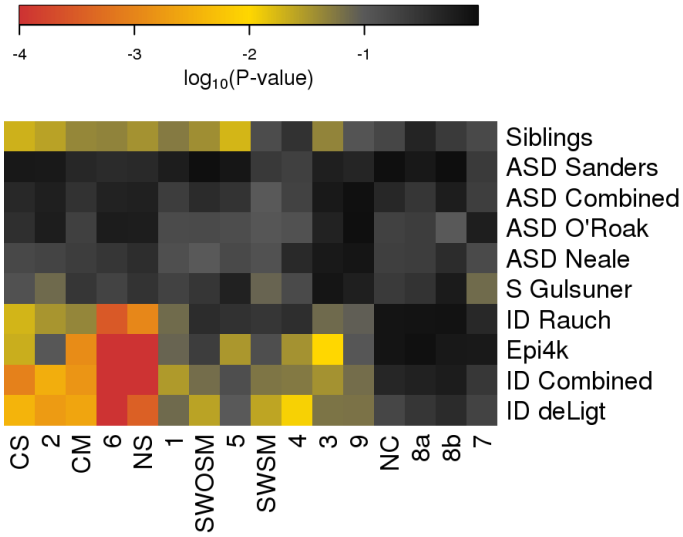


Figure A.17: Co-clustering of the clinical classes with *de novo*-mutated genes in a co-expression network based on GTEx.

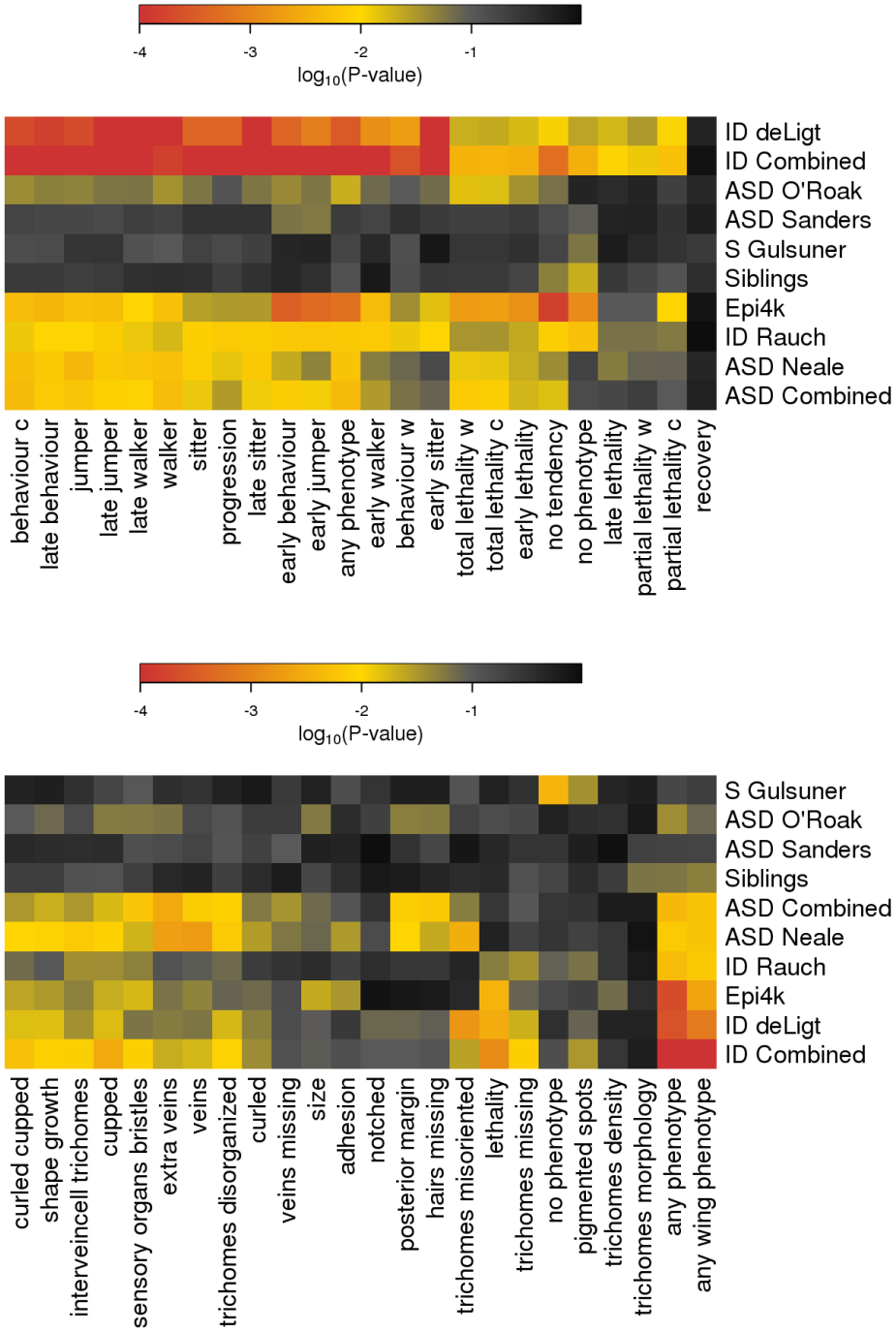


Figure A.18: Co-clustering of neuronal and fly wing phenotypes with *de novo*-mutated genes in the phenotypic-linkage network.

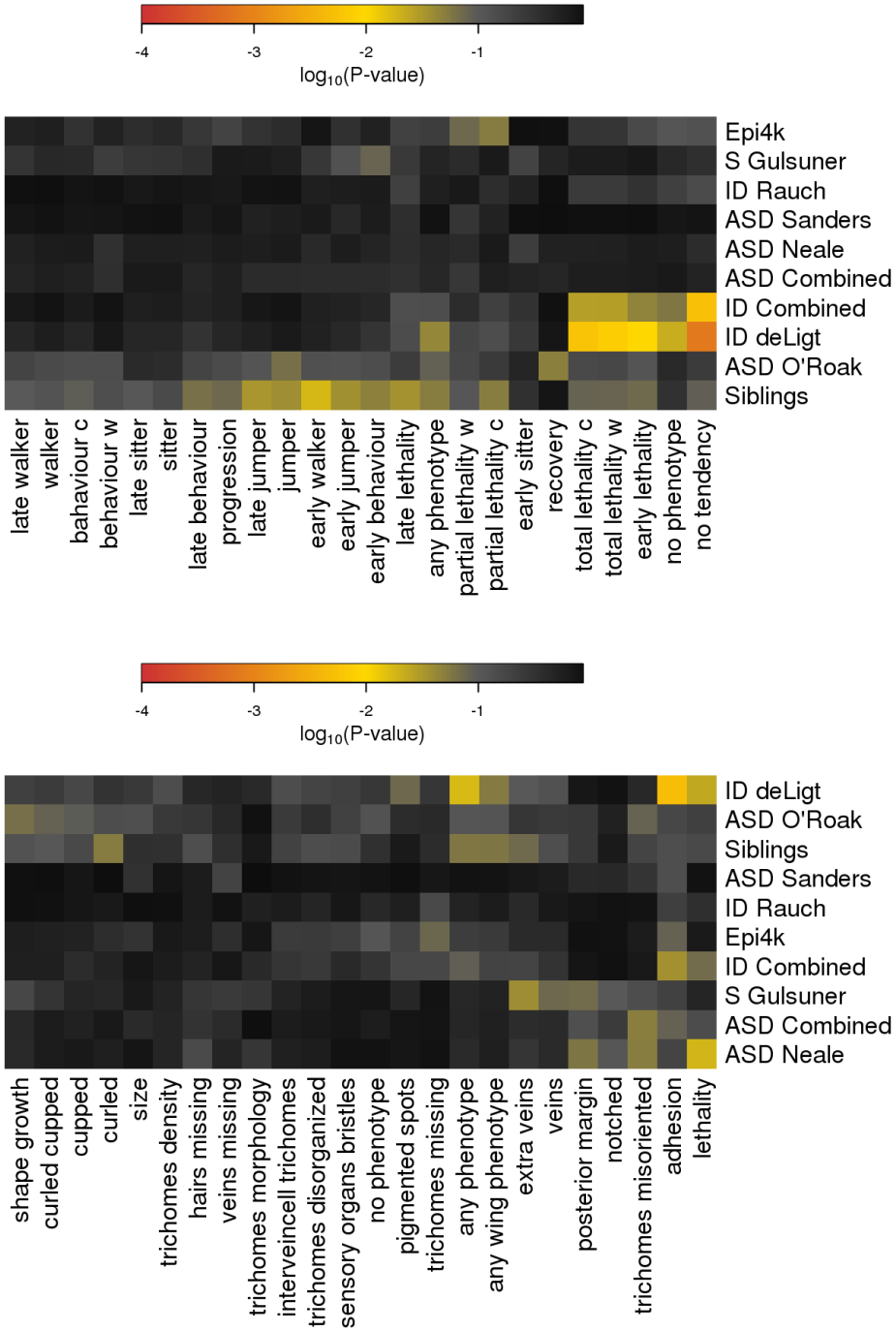


Figure A.19: Co-clustering of neuronal and fly wing phenotypes with *de novo*-mutated genes in a co-expression network based on GTEx.

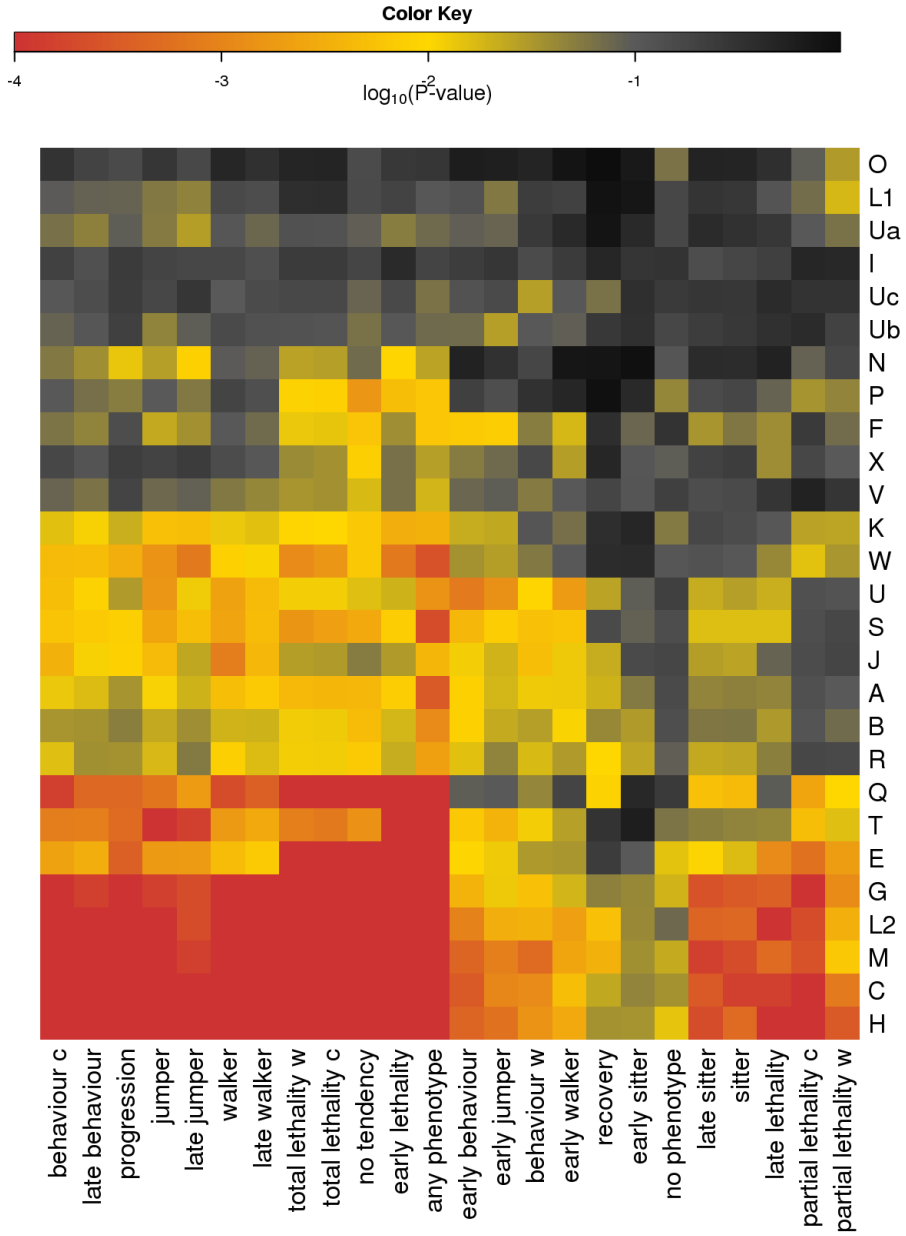


Figure A.20: Co-clustering of fly neuronal phenotypes with ID-accompanying clinical features in a co-expression network based on GTEx.

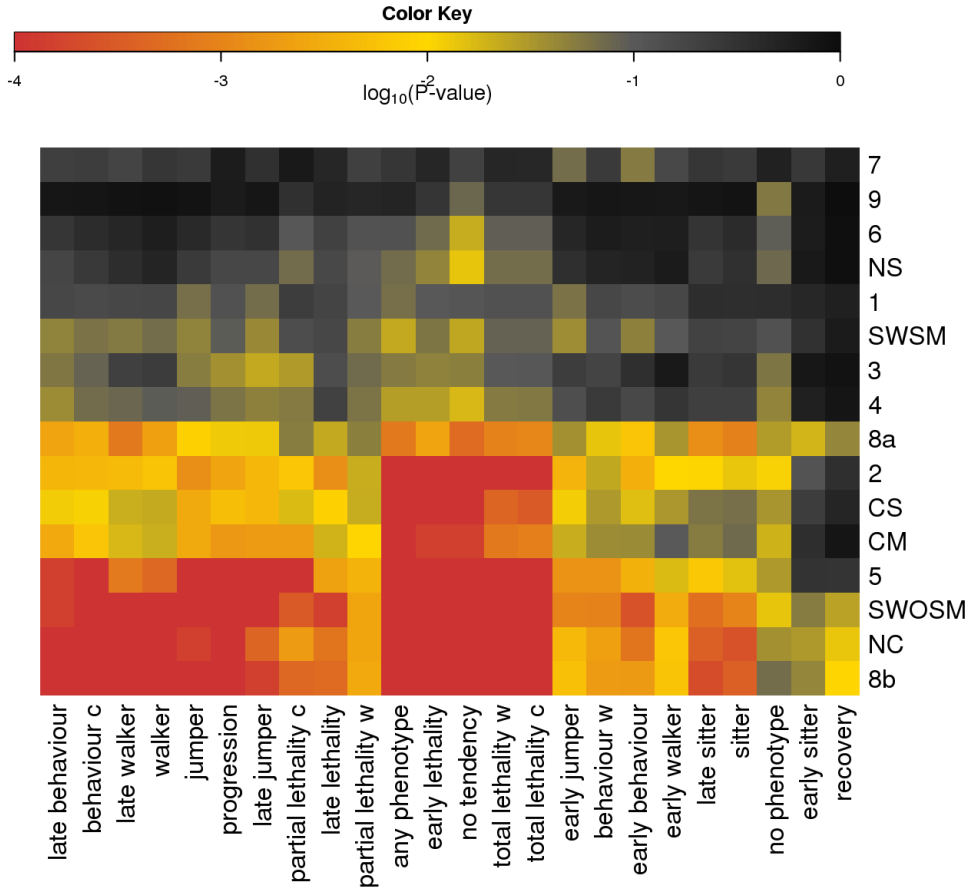


Figure A.21: Co-clustering of fly neuronal phenotypes with the clinical classes in a co-expression network based on GTEx.

Table A.1: Data that were assessed but not included in the integrated network.

Data type	Source	Type
Physical interactions	IntAct	Mouse proteins MI:0029 cosedimentation through density gradient
		Mouse proteins MI:0676 tandem affinity purification
Sequence patterns	InterPro	Mouse proteins
Gene expression	SMD[1]	Abate *et al*. 2004
		Adler *et al*. 2006
		Adler *et al*. 2008
		Alizadeh *et al*. 2010
		Alter *et al*. 2003
		Appari *et al*. 2009
		Baldwin *et al*. 2003
		Bashyam *et al*. 2005
		Bebermeier *et al*. 2006
		Beck *et al*. 2010
		Ben-Chetrit *et al*. 2006
		Benetkiewicz *et al*. 2005
		Bergamaschi *et al*. 2006
		Bergamaschi *et al*. 2009
		Berry *et al*. 2010
		Bhamre *et al*. 2009
		Blader *et al*. 2001
		Bohen *et al*. 2003
		Boldrick *et al*. 2002
		Bredel *et al*. 2005a
		Bredel *et al*. 2005b
		Bredel *et al*. 2005c
		Bredel *et al*. 2006
		Brouard *et al*. 2007
		Brown *et al*. 2006
		Buess *et al*. 2006
		Buess *et al*. 2007
		Bullinger *et al*. 2004
		Bullinger *et al*. 2007
		Bullinger *et al*. 2008
		Cario *et al*. 2005
		Chang *et al*. 2002
		Chang *et al*. 2004
		Chang *et al*. 2006
		Chen *et al*. 2002

Chen *et al*. 2003
Chen *et al*. 2004
Chi *et al*. 2003a
Chi *et al*. 2003b
Chi *et al*. 2006
Chi *et al*. 2007
Chua *et al*. 2007
Clement *et al*. 2002
Cuadras *et al*. 2002
Dairkee *et al*. 2004
DeAvalos *et al*. 2002
DePrimo *et al*. 2002
Detweiler *et al*. 2001
Diehn *et al*. 2002
Diehn *et al*. 2005
Diehn *et al*. 2006
Dumas *et al*. 2007
El-Etr *et al*. 2004
Faherty *et al*. 2010
Fine *et al*. 2004
Fine *et al*. 2005
Fletcher *et al*. 2009
Fortna *et al*. 2004
Fouts *et al*. 2007
Galgano *et al*. 2008
Garber *et al*. 2001
Giacomini *et al*. 2005
Gilks *et al*. 2005
Griffiths *et al*. 2005
Guillemin *et al*. 2002
Gyorffy *et al*. 2005
Hao *et al*. 2006
Hao *et al*. 2007
He *et al*. 2006
Hendrickson *et al*. 2008
Hendrickson *et al*. 2009
Hertel *et al*. 2004
Heuser *et al*. 2005
Higgins *et al*. 2003
Higgins *et al*. 2004
Higgins *et al*. 2007
Holterhus *et al*. 2003

Holterhus *et al*. 2007
Holweg *et al*. 2011
Houshdaran *et al*. 2010
Hurowitz *et al*. 2007
Iacobuzio *et al*. 2003
Iyer *et al*. 2009
Jacobsen *et al*. 2007
Ji *et al*. 2002
Ji *et al*. 2003
Jones *et al*. 2003
Jones *et al*. 2005a
Jones *et al*. 2005b
Juric *et al*. 2005
Juric *et al*. 2007
Kainz *et al*. 2004
Kao *et al*. 2009
Kapp *et al*. 2006
Kasperkovitz *et al*. 2005
Kharas *et al*. 2010
Kim *et al*. 2007
Klapholz *et al*. 2007
Kosinski *et al*. 2007
Kwei *et al*. 2008a
Kwei *et al*. 2008b
Lacayo *et al*. 2004
Langerod *et al*. 2007
Lapointe *et al*. 2004
Lapointe *et al*. 2007
Lee *et al*. 2006
Lee *et al*. 2008
Leung *et al*. 2002
Leung *et al*. 2004
Li *et al*. 2004
Liang *et al*. 2005
Lin *et al*. 2011
Linn *et al*. 2003
Liu *et al*. 2006
Liu *et al*. 2007
Lossos *et al*. 2002
Lowe *et al*. 2007
Mazzucotelli *et al*. 2007
Melk *et al*. 2005

Mueller *et al*. 2004
Munagala *et al*. 2004
Murray *et al*. 2004
Myers *et al*. 2006
Nagarayan *et al*. 2007
Naume *et al*. 2007
Nicolau *et al*. 2007
Nielsen *et al*. 2004
Novoradovskaya *et al*. 2004
Palmer *et al*. 2006
Pathan *et al*. 2004
Patil *et al*. 2005
Perou *et al*. 1999
Perou *et al*. 2000
Pollack *et al*. 2002
Popper *et al*. 2007
Popper *et al*. 2009
Rajski *et al*. 2010
Rinn *et al*. 2006
Roose *et al*. 2003
Rosenwald *et al*. 2001
Ross *et al*. 2000
Ross *et al*. 2001
Rubins *et al*. 2004
Rubins *et al*. 2007
Rubins *et al*. 2008
Rucker *et al*. 2006a
Rucker *et al*. 2006b
Rudnicki *et al*. 2009
Saaf *et al*. 2006
Saaf *et al*. 2007
Sarwal *et al*. 2003
Shaner *et al*. 2003
Schwarze *et al*. 2002
Segal *et al*. 2007
Simmons *et al*. 2007
Sivertsen *et al*. 2006
Sood *et al*. 2006b
Sood *et al*. 2008
Sorlie *et al*. 2001
Sorlie *et al*. 2003
Sperger *et al*. 2003

Sridhar *et al*. 2009
Subramanian *et al*. 2004
Subramanian *et al*. 2005
Storz *et al*. 2003
Thompson *et al*. 2008
Thompson *et al*. 2009
Timmer *et al*. 2007
Townley-Tilson *et al*. 2006
Tsai *et al*. 2006
VanBaarsen *et al*. 2006
VanBaarsen *et al*. 2008
VanBaarsen *et al*. 2010
VanDerPouw *et al*. 2003a
VanDerPouw *et al*. 2003b
VanDerPouw *et al*. 2007
VanDerPouw *et al*. 2008a
VanDerPouw *et al*. 2008b
Wang *et al*. 2006
Waddel *et al*. 2010
West *et al*. 2004
West *et al*. 2005
West *et al*. 2006
Whitfield *et al*. 2002
Whitfield *et al*. 2003
Whitney *et al*. 2003
Wong *et al*. 2008b
Zhang *et al*. 2003
Zhang *et al*. 2004
Zhang *et al*. 2008
Zhao *et al*. 2002
Zhao *et al*. 2004a
Zhao *et al*. 2004b
Zhao *et al*. 2005a
Zhao *et al*. 2005b
Zhao *et al*. 2005c
Zhao *et al*. 2006

[1] Stanford Microarray Database (http://smd.princeton.edu).

Table A.2: Very weak correlation between CDS length and node degree.

Data type	All links	Top 100,000 links
Phenotypic network	0.050	0.057
PPI	-0.001	0.004
GO:BP	0.075	0.051
Co-citation	0.050	0.014
Co-expression	-0.060	-0.055
KEGG	0.046	NA
HumanNet	0.050	-0.028
GeneMANIA	0.075	0.017
STRING	0.050	0.014
InWeb	0.069	NA

The values represent the Spearman's correlation coefficient.

Table A.3: Fractions of gene pairs co-annotated with the same phenotype in the integrated phenotypic-linkage network.

Gene–gene links[1]	Fraction	Phenotype
1 to 10,000	5.60%	Abnormality of the nervous system
	5.17%	Abnormality of metabolism/homeostasis
	4.47%	Abnormality of the musculoskeletal system
	3.42%	Abnormality of the eye
	3.22%	Abnormality of the abdomen
	3.10%	Abnormality of the head and neck
	3.01%	Abnormality of the integument
	2.54%	Abnormality of musculature
	2.51%	Abnormality of the genitourinary system
	2.25%	Neoplasia
10,001 to 20,000	3.56%	Abnormality of the nervous system
	3.44%	Abnormality of metabolism/homeostasis
	3.18%	Abnormality of the musculoskeletal system
	2.06%	Abnormality of the abdomen
	1.98%	Abnormality of the head and neck
	1.92%	Abnormality of the eye
	1.87%	Abnormality of musculature
	1.81%	Abnormality of the integument
	1.65%	Abnormality of the cardiovascular system
	1.55%	Abnormality of the genitourinary system

20,001 to 30,000	3.92%	Abnormality of the nervous system
	3.32%	Abnormality of the musculoskeletal system
	3.18%	Abnormality of metabolism/homeostasis
	2.43%	Abnormality of the eye
	2.17%	Abnormality of the head and neck
	1.95%	Abnormality of the integument
	1.88%	Abnormality of the abdomen
	1.66%	Abnormality of musculature
	1.63%	Abnormality of the genitourinary system
	1.41%	Abnormality of the ear
30,001 to 40,000	3.70%	Abnormality of the nervous system
	3.19%	Abnormality of the musculoskeletal system
	2.89%	Abnormality of metabolism/homeostasis
	2.15%	Abnormality of the eye
	1.93%	Abnormality of the head and neck
	1.88%	Abnormality of the abdomen
	1.82%	Abnormality of the integument
	1.63%	Abnormality of musculature
	1.50%	Abnormality of the genitourinary system
	1.31%	Growth abnormality
40,001 to 50,000	3.11%	Abnormality of the nervous system
	2.72%	Abnormality of the musculoskeletal system
	2.50%	Abnormality of metabolism/homeostasis
	1.94%	Abnormality of the head and neck
	1.80%	Abnormality of the eye
	1.76%	Abnormality of the integument
	1.56%	Abnormality of the abdomen
	1.33%	Abnormality of musculature
	1.26%	Abnormality of the genitourinary system
	1.11%	Abnormality of the cardiovascular system

[1] Ordered by strength.

Table A.4: Clustering of genes associated with clinical classes of ID.

Class	#genes$_{GTEx}$	P$_{microarrays}$	P$_{GTEx}$
1	54	0.0222978	0.1228877
2	105	0.1989801	0.0042996
3	13	0.6062394	0.0706929
4	77	0.4123588	2e-04
5	159	1e-04	1e-04
6	62	1e-04	1e-04
7	27	0.2571743	0.0019998
8a	53	0.3411659	0.0020998
8b	77	1e-04	1e-04
9	1	1	1
SWSM	148	0.010499	1e-04
SWOSM	365	1e-04	1e-04
NS	75	1e-04	1e-04
CS	168	3e-04	0.0013999
CM	290	1e-04	1e-04
NC	152	1e-04	1e-04

Table A.5: Clustering of genes associated with clinical phenotypes.

Letter	Feature	#genes$_{GTEx}$	P$_{microarrays}$	P$_{GTEx}$
A	Short stature	127	0.0225977	1e-04
B	Microcephaly	133	2e-04	1e-04
C	Lethality	120	1e-04	1e-04
E	Epilepsy	193	1e-04	1e-04
F	Overgrowth/macrocephaly	23	0.6945305	0.2768723
G	Progression/regression	113	1e-04	1e-04
H	Neurological symptoms	265	1e-04	1e-04
I	Malignancies	21	0.3790621	0.1056894
J	Immunological anomalies	28	0.9225077	0.0010999
K	Endocrine anomalies	43	0.1121888	4e-04
L1	Brain malformations	82	0.0211979	0.0015998

L2	Non-structural MRI anomalies	123	1e-04	1e-04
M	Metabolic/mitochondrial anomalies	159	1e-04	1e-04
N	Obesity	22	0.6568343	2e-04
O	Vegetative anomalies	21	1	0.0326967
P	Behavioural anomalies	106	1e-04	1e-04
Q	Myopathy or muscular anomalies	56	1e-04	1e-04
R	Blood cell anomalies	25	0.0072993	1e-04
S	Ectodermal anomalies	78	0.0376962	0.0037996
T	Eye anomalies	121	0.0127987	1e-04
U	Skeletal anomalies	46	0.2044796	0.0039996
Ua	Limb anomalies	49	0.0145985	0.0030997
Ub	Vertebral/skull anomalies	27	0.2213779	0.3493651
Uc	Clefts	20	1	0.0286971
V	Cardiac malformations	52	0.0527947	2e-04
W	Urogenital or renal anomalies	62	0.0083992	1e-04
X	Other malformations	29	0.0947905	0.0323968

Table A.6: Clustering of genes associated with fly neuronal phenotypes.

Neuronal phenotype	#genes	$P_{microarrays}$	P_{GTEx}
any_phenotype	197	1e-04	1e-04
no_phenotype	49	0.0859914	0.1960804
early_lethality	126	1e-04	1e-04
late_lethality	35	0.2832717	0.2128787
early_behavior_phenotype	48	0.079592	0.0194981
late_behavior_phenotype	101	0.0015998	0.0045995
phenotype_no_tendency	141	5e-04	1e-04
phenotype_progression	96	5e-04	0.0031997
phenotype_recovery	8	0.290471	0.1379862
total_lethality_worst_case	115	1e-04	1e-04
partial_lethality_worst_case	29	0.8742126	0.2708729
behavioral_phenotype_worst_case	53	0.0535946	0.0170983
total_lethality_collective	115	1e-04	1e-04
partial_lethality_collective	46	0.3611639	0.1180882
behavioral_phenotype_collective	109	0.0018998	0.0040996
walker	99	0.0012999	0.0046995
sitter	74	0.0157984	0.0807919
jumper	83	0.070193	0.0015998
early_walker	28	0.1533847	0.0750925
early_sitter	13	0.3715628	0.2947705
early_jumper	40	0.1478852	0.0544946
late_walker	93	0.0005999	0.0076992
late_sitter	72	0.0036996	0.0918908
late_jumper	74	0.0966903	0.0045995

Table A.7: Clustering of genes associated with fly wing phenotypes.

Wing phenotype	#genes	$P_{microarrays}$	P_{GTEx}
no_phenotype	10	1	0.5155484
any_phenotype	233	1e-04	1e-04
any_wing_phenotype	211	3e-04	1e-04
lethality	46	0.0017998	0.0018998
shape_growth_overview	127	0.0112989	1e-04
curled_cupped	122	0.0268973	2e-04
curled	88	0.0333967	1e-04
cupped	74	0.0706929	0.0163984
size	33	0.0267973	0.0819918
adhesion	26	0.0830917	0.0361964
posterior_margin_overview	11	0.2607739	0.0851915
notched	2	1	0.0987901
hairs_missing	13	0.4257574	0.5362464
interveincell_trichomes_overview	69	0.0243976	0.010299
trichomes_missing	20	0.2613739	0.140386
trichomes_misoriented	3	1	0.0751925
trichomes_disorganized	48	0.0564944	0.0065993
trichomes_density	14	1	0.3326667
trichomes_morphology	7	1	0.5007499
pigmented_spots	35	0.5144486	0.0150985
veins_overview	98	0.1216878	0.0020998
veins_missing	29	0.6109389	0.2875712
extra_veins	83	0.0957904	0.0016998
sensory_organs_bristles	104	0.009799	0.0009999