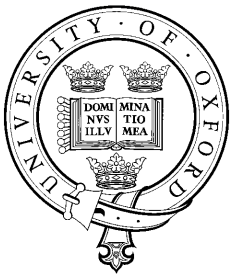
CLUSTERING GENES BY FUNCTION TO UNDERSTAND DISEASE PHENOTYPES



TALLULAH ANDREWS

Somerville College

MRC Functional Genomics Unit

Department of Physiology, Anatomy, and Genetics

University of Oxford

A thesis submitted in partial fulfilment of the requirements for the degree of

Doctor of Philosophy

24 April 2015

Abstract

Developmental disorders including: autism, intellectual disability, and congenital abnormalities are present in 3-8% of live births and display a huge amount of phenotypic and genetic heterogeneity. Traditionally, geneticists have identified individual monogenic diseases among these patients but a majority of patients fail to receive a clinical diagnosis. However, the genomes of these patients frequently harbour large copy-number variants (CNVs) but their interpretation remains challenging. Using pathway analysis I found significant functional associations for 329 individual phenotypes and show that 39% of these could explain the patients' multiple co-morbid phenotypes; and multiple associated genes clustered within individual CNVs. I showed there was significantly more such clustering than expected by chance. In addition, the presence of a multiple functionally-related genes is a significant predictor of CNV pathogenicity beyond the presence of known disease genes and size of the CNV. This clustering of functionally-related genes was part of a broader pattern of functional clusters across the human genome. These genome-wide functional clusters showed tissue-specific expression and some evidence of chromatin-domain level regulation. Furthermore, many genome-wide functional clusters were enriched in segmental duplications making them prone to CNV-causing mutations and were frequently seen disrupted in healthy individuals. However, the majority of the time a pathogenic CNV affected the entire functional cluster, where as benign CNVs tended to affect only one or two genes. I also showed that patients with CNVs affecting the same functional cluster are significantly more phenotypically similar to each other than expected even if their CNVs do not affect any of the same genes. Lastly, I considered one of the major limitations in pathway analysis, namely ascertainment biases in functional information due to the prioritization of genes linked to human disease, and show how the modular nature of gene-networks can be used to identify and prioritize understudied genes.

Acknowledgements

I'd like to thank Steve Meader, Avigail Taylor, Julia Steinberg for their contributions to this work as well as my supervisors Caleb Webber and Chris Ponting. Thanks to other members of the Webber lab and the department for various helpful comments and suggestions. Finally, thanks to all the friends and family members who provided moral support and the occasional distraction during this project.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

1.1.1 Developmental disorders

Developmental disorders (aka genetic disorders) are a group of diseases which include autism, intellectual disability, and congenital abnormalities and are present in 3-8% of live births (1; 2). These disorders occur early in life (before the age of 25) with many evident before the age of two (1). Despite this high frequency, each of the approximately 7,000 known syndromes is individually extremely rare with a frequency on the order of 1:10,000 (3). Common phenotypes exhibited by patients with these disorders include various neurological/behavioural defects (70%), heart defects (30%), facial dysmorphisms (30-40%), limb defects (20%) and genital/urinary defects (20%) (3; 4; 5). These phenotypes are highly heritable, with heritability estimates in the range of 70-90% (6); and often have severely debilitating effect, 2.5% die within the first week of life (5). Furthermore these disorders display a large amount of phenotypic and genetic heterogeneity (5; 6; 7; 8), resulting in up to 80% of patients failing to receive a clinical or genetic diagnosis(3; 5). Furthermore estimates of the number of genes which can cause individual phenotypes frequently range in the hundreds, eg. 150-300 different loci are estimated to contribute to autism (9; 10), and more than 400 genes have been linked to intellectual disability which is likely to be only half of all contributing genes(11).

Types of variants

Many different genetic variants are overrepresented in the genomes of these patients, including large scale chromosomal aberrations, sub-microscopic deletions and duplications and *de novo* single nucleotide variants (8; 12; 13; 14; 15). Large-scale chromosomal aberrations, where a large section (10Mb and larger) of a chromosome is deleted or duplicated, contribute to diseases such as Down's syndrome and are detectable using common karyotyping methods (8). Sub-microscopic deletions and duplications are those variants that are smaller than chromosomal aberrations, thus not visible in a karyotyping test, but are still greater than 1 kb in size. They are detected using the relative signal strength from either array comparative genomic hybridization (aGCH) or SNP-genotyping chips. Collectively both these classes of variants are included in under the term 'copy-number variant' (CNV), which are defined as all DNA deletions or duplications greater than 1 kb in size (12; 14). Whereas single nucleotide variants (SNVs) are most commonly a single base-pair change in the genome, but may also include small indels (1-10bp), and typically refer to rare variants (<1% frequency) as opposed to single nucleotide polymorphism (SNPs) which are typically common variants (frequency >5%) and only include single-base substitutions. SNVs are identified using whole exome or whole genome sequencing of patient's DNA.

Genome-wide aCGH analysis is commonly used as a first-tier diagnosis procedure for patients with developmental disorders (2; 16); and after various quality control measures provides a potential genetic diagnosis in the form of a likely pathogenic CNV for 10-20% of patients (2; 8; 12; 15). As a result several large clinical quality datasets have been made available to researchers (eg.(17; 18)). Newer methods employing whole genome or whole exome sequencing identify potentially deleterious single nucleotide variants (SNVs) in another 10-20% of patients (11; 12). Few common single nucleotide polymorphisms have been associated with these disorders, rather it

is the rare and *de novo* variation which is significantly associated with these disorders (4; 6; 10; 15). Furthermore the high selection against these variants due to the effect on survival and fecundity is expected to keep any causal variants in the population at low frequency(2; 19).

Function of Variants

Despite this success in identifying associated variants, interpreting the functional consequences of these variants remains challenging. In many cases it is unclear whether there is a single gene responsible ('monogenic') for a particular disorder or many genes contributing. It has been argued that CNVs affect a single dosage sensitive gene which is responsible for the respective disorder (8) or that deletion CNVs may reveal a null allele inherited from the other parent (13). However, functional analyses of CNVs consistently reveals multiple candidate genes within each variant (20; 21; 22; 23). In addition, there is a strong correlation between CNV size and the severity or penetrance of the affected patient's phenotype (4; 24). However CNV size alone is insufficient to explain the disorder as similarly sized CNVs can be found in 5-10% of healthy individuals(25). Some studies have even suggested a two-hit model where a secondary variant increases the penetrance of a primary CNV (7; 14). Further evidence that many developmental disorders have a more complex genetic cause, as opposed to a monogenic disease, comes from the observation that many variants found in these patients are also found at low frequency in healthy individuals indicating the variant alone is not sufficient to cause the disease (12).

Many phenotypes have been associated with multiple different genes in different patients, eg. Coffin-Siris syndrome which is due to a mutation in any of at least six different genes(14). These observations have lead to the rise of pathway analysis approaches for elucidating common biological functions disrupted by mutations in different genes (26; 27). Pathway analyses rely on the observation that disruptions to different genes

which function in the same pathway often produce similar phenotypes (28). For instance the six different genes known to cause Coffin-Siris syndrome all participate in the same protein complex (14). Others have identified multiple genes involved in GTPase/Ras signalling (29) or the MAPK pathway (21) disrupted in patients with ASD. Pathway analysis differs from the traditional approach, in which clinician-researchers identify variants in a single gene as the cause of a particular syndromes by examining unrelated individuals with strikingly similar phenotypes (30; 31), by not relying on finding patients with identical genetic variants; rather pathway analysis approaches get power from different but related genes affected by mutations across multiple patients with related phenotypes. This can be extremely advantageous when dealing with developmental disorders since many genetic variants are so rare (< 1%) or the prevalence of the syndrome is so low (< 1:100,000) that they are often seen in only a single individual in a study (6; 10; 25; 32).

1.1.2 Gene functional networks and pathways

Genes can work together in many different ways. Genes can physically interact in protein complexes, chemically modify each other, regulate each others expression, or catalyse sequential metabolic reactions. Several resources exist which compile the genes involved in various reactions and pathways from the scientific literature, eg. the Kyoto Encyclopaedia of Genes and Genomes (KEGG, (33)) and Reactome, but these are quite limited in scope. The Gene Ontology (GO) augments gene functions reported in the literature with computational inferences(34). Pathways can also be inferred from observing similar phenotypes (in humans or model organisms) which are available from various public databases, eg. the Human Phenotype Ontology (HPO, (35)) or the Mouse Genome Informatics (MGI, (36; 37)). Most of these resources organize the functional descriptions into an ontology, a set of controlled terms organized into a hierarchy such that specific terms (eg. “cleft palate”) are grouped under more general parental terms (eg. “abnormality of the mouth”). This ensures reliable comparisons since a discrete number of terms are used to describe functions according to consistent

definitions. In addition, by moving up or down the hierarchy of terms analyses can be performed at various levels of detail. Variants associated with a particular disease can be tested for an enrichment of genes with a particular functional annotation (38). This approach has found significant associations in many studies (20; 22; 39; 40; 41). However, functional annotation datasets vary both in the quality of the annotations (false-positive rate) and the quantity of annotations (false-negative rate). High-throughput experiments and computational-inferences, eg. yeast-2-hybrid protein-protein interactions, typically contribute a large number of annotations (lower false-negative rate) but these tend include more false positives (lower quality) than low-throughput experiments, eg. co-purification of interacting proteins (42; 43; 44; 45)

The annotations described above are all qualitative descriptions of gene function; the main quantitative measure of gene function is gene expression, which is typically measured using microarrays or RNA sequencing, which can infer tissue and temporally specific gene functions or be used to infer groups of co-operative genes on the basis of the similarity of their expression patterns. This quantitative information is frequently represented as a co-expression network, where genes are represented as nodes and the correlation coefficient between their expression patterns is represented as edges between them (46; 47; 48; 49; 50). Network representation is extremely flexible and a large number of gene functional networks have been constructed with edges representing different measure of functional similarity between genes. Common types of functional networks are the previously mentioned co-expression networks (eg. COXPRESdb (46)), protein-protein interaction networks (eg. STRING (51), iRefIndex (52)), and integrated networks (eg. HumanNet (53), FunCoup (54)). Integrated networks are constructed by using statistical methods to combined multiple functional datasets including the qualitative annotations discussed above into a single measure of functional similarity between genes. In recent years the advantage of integrated networks in their ability to combine data which reflect different ways genes can interact has been recognized and a multitude of networks tailored to different model organisms(eg. *S. cerevisiae* (55; 56), *C. elegans* (57), *D. melanogaster* (58)) have been created and have demon-

strated a greater ability to predict gene function than networks based on single data source (56; 59; 60; 61; 62).

Another advantage of representing the functional relationships between genes as a network is the large number of existing algorithms and measures which have been developed for analysing networks in other fields which can be applied to analyse biological networks (63). Degree, the number of edges or the sum of the weights of all edges attached to a given node[gene] (64), can be a potential confounder in network-based pathway analyses if the method used to identify deleterious variants is biased towards genes with a high degree; for instance exome sequencing is biased towards long genes and long genes are more likely to have multiple functions(65), and RNASeq differential gene expression is biased toward long and highly expressed genes which tend to be well-studied genes thus tend to have a high degree in functional networks(66; 67). These biases towards high degree genes can result in a high number of connections between the selected genes (suggesting they participate in the same pathway) which is purely due to the large number of other genes connected to each selected and not a consequence of a tendency of selected genes to be connected to each other more than to other genes. Shortest-paths, the fewest number of links or the least total distance (inverted similarity) needed to travel between two links (68), can be used to calculate functional similarity between two genes which do not have a direct edge between them. Modelling a random-walk on a network, a 'walker' traverses the network from node to node following edges at random where the probability of travelling across an edge is proportional to its weight, provides a good approximation for the flow of information through a network and can be used to find bottlenecks or relatively independent pathways (69). Edge density, the number of edges between a set of nodes divided by the number of nodes (local edge density of a region of a network)(70) or divided by the total number of possible edges between the nodes (network edge density)(71), can be used as a measure of the strength of connection within a group of genes. Finally, examining the clustering (aka 'modularity' or 'community structure') of a network refers to identifying regions with unusually high amounts of interconnection between genes,

for instance where the local edge density is higher that expected (72). Clustering a functional network has been shown to identify groups of genes which participate in the same biological pathway (28; 63; 73; 74; 75), thus is extremely useful for pathway analysis since these groups (often called 'functional modules') typically include far more genes than are covered by the literature-based annotations.

1.1.3 Functional clustering

A finding that has emerged from availability of genome-wide functional information is that functionally-related genes are located close together in eukaryotic genomes. This clustering of functionally-related genes in the linear genome has been observed in yeast (76; 77; 78), mouse(79; 80), fly(81; 82; 83), worm (84), zebrafish (85) and human (80; 86; 87; 88; 89; 90; 91). Furthermore, it has been detected using many different sources of functional information, including: protein-protein interactions (76; 92), KEGG pathways (89), GO functional annotations (86), and phenotypes exhibited following gene knock-downs (84). Gene expression has been used most often and clusters of broadly expressed housekeeping genes (80; 81; 93), highly expressed genes (90; 93) and co-expressed or tissue-specific genes (78; 79; 81; 83; 85; 88; 90) have been reported. However, many different methods at many different scales, from just the closest gene pairs to windows of 39 genes, have been used often with conflicting results on the extent and biological importance of the functional clusters (81; 87; 93). Furthermore all these studies corrected for the presence of paralogs thus the functional clusters are not due to tandem duplications.

The most common approach to identifying functional clustering has been a sliding window, where the genome is examined as a sequence of windows and the distribution of gene function across or within the windows is compared to that expected from a randomized genome. However the size of the window will affect the results as clusters bigger than the window or much smaller than the window are unlikely to be detected

or be estimated to be much larger or smaller than they ultimately are. For example the human genome was found to contain 30 clusters of 30-200 housekeeping genes and co-expressed clusters (after removing housekeeping genes) were identified when using a window of 10 genes(80) but when using a window of 300 kb only clusters of highly expressed housekeeping genes were detected (93). In D. melanogaster, when a 10 gene window was used 20% of the genome appeared to participate in co-expressed clusters containing 10-30 genes (82). In yeast, adjacent gene pairs (window size of 2) are significantly co-expressed, more likely to engage in protein-protein interactions, and have the same functional annotations (76; 78). Another study in zebrafish showed that as the size of the window increased from 3 to 20 adjacent genes the degree of co-expressiong decreases (85). Yet even when using a very large window of 39 genes (90) or 5 Mb (94) there is still significant functional clustering in the human genome. This is another of the major issues with sliding window approaches, that any or all of the genes in a particular window could be contributing its significance.

Another common approach is to identify pairs of genes which are close together in the genome (either adjacent to each other or within a particular linear distance along the sequence) then determining whether they are functionally similar to each other. Binning gene pairs based on the distance between them reveals a significant enrichment of protein-protein interactions and significantly high co-expression among the genes closest together (88; 91). Others used a growing algorithm where genes where grouped with their neighbours as long as with were within a particular distance of the next gene; and the resulting groups were then tested for an enrichment of various gene functions (79; 85; 89). In zebrafish, a threshold of 25kb identified groups of co-expressed genes but they were not enriched in particular GO functions (85). Whereas in mice using various thresholds from 25kb-100kb identified groups of 2-5 genes enriched in testis-specific genes (79). In humans grouping together genes in the top 1% closest gene-pairs revealed 50% of KEGG pathways were significantly enriched among these tightly clustered groups (89). While these approaches partially relieved the issue of resolution limit of the sliding window by allowing groups of clusters with a much

broader range of size to be detected, they still relied on arbitrarily chosen distance thresholds. In addition, they identify potential clusters using only genomic distance between genes thus it is possible for only a couple genes in a particular cluster which have a related function to make the whole cluster appear functionally similar.

Only a single study incorporated both genomic distance and functional similarity to define the functional clusters in a genome. Weber et al. (81) identified genes which were either highly co-expressed (top 5% most co-expressed) or were broadly expressed (housekeeping genes) and which had fewer than four other genes located between them in the fly genome. This revealed that the previously reported large functional clusters (10-30 genes) based on a sliding window (82) were in fact composed of multiple small clusters of 2-3 highly co-expressed genes, which were separate from the 512 clusters of housekeeping genes.

The problem of determining the appropriate resolution for functional clustering is the lack of knowledge on the biological causes or consequences of functional clustering. Almost all studies I have cited corrected for the presence of paralogous genes thus it is unlikely that tandem duplications is sufficient to explain the observed functional clustering. Highly or broadly expressed functional clusters may be explained by the presence of dosage-sensitive essential genes as selection may act on chromatin level factors, such as nucleosome occupancy, to promote the maintenance of open chromatin regions and lower transcriptional noise (95; 96). Another reason for the clustering of essential genes may be to protect them from disruption since they are frequently found in regions of low recombination (77; 84). Co-expression clusters may be subject to chromatin level regulation as suggested by a relatively high level of nucleosome occupancy (97). However, it is unclear whether the clustering of functionally-related genes is selectively advantageous since individual clusters are poorly conserved across species (83; 86), though others dispute this finding (91). Since many of the studies have been based on expression data, another plausible explanation is that they are a result of leaky gene expression rather than selection to locate co-expressed genes together

in the genome (92). However, this cannot explain the significant clustering reported when only protein-protein interactions or functional annotations were considered (eg. (76; 86; 89; 91)). A final possible explanation is selection to keep genes which genetically interact (ie. epistasis) linked together to reduce the frequency of recombination between them (98).

Functional clustering may be particularly relevant to diseases arising from CNVs since a single *de novo* CNV frequently affects as many as a dozen different genes present at the same locus in the genome. If those genes genetically interact or participate in the same biological process because they belong to a functional cluster, then a single variant affecting multiple genes in the cluster may have more deleterious effects or have increased penetrance due to the simultaneous loss of multiple compensatory mechanisms. However, there has yet been no systematic examination of the relevance of functional clustering to diseases linked to CNV mutations.

1.2 Thesis Structure

In this thesis I will be investigating the role of biological pathways and functional clustering to developmental disorders. In the following chapter (Chapter 2) I will describe the various patient and control CNV datasets, the many functional resources, and the different network clustering algorithms used throughout my work. The first results chapter (Chapter 3) addresses the question of the usefulness of biological pathways in understanding the phenotypes of patients with developmental disorders. Multiple standard pathway approaches are applied to a large, well characterized cohort of patients. I introduce a novel way to evaluate the performance of the different techniques by comparing the patterns of co-morbid phenotypes among patients with mutations affecting the pathway to those without. And I identify the phenotypes that show significant changes in prevalence depending on the pathway affected by variants in the patient. In addition, I examine multiple biases which could lead to the large number of

false-positives in pathway analyses. In the second results chapter (Chapter 4), I examine the extent and significance of functional clustering in CNVs found in patients with developmental disorders using a novel integrated functional network. Furthermore the relationship between functional clusters and known disease genes including their ability to distinguish pathogenic from benign CNVs is examined; and the presence of an association with functional clusters and specific phenotypes or phenotype severity is considered. The third results chapter (Chapter 5) uses the results from the previous chapter to look for disease-relevant functional clustering in the human genome. Various genomic properties and context of the clusters is examined. Furthermore the deleteriousness and association with disease of different genome-wide clusters is examined and various functional annotations associated with pathogenicity (with respect to developmental disorders) are identified. In addition, the concordance of patient phenotypes whose CNVs affect the same function cluster is evaluated. The final results chapter (Chapter 6) changes gear and looks at how biological networks can be used to identify understudied genes in order to fill the gaps in our understanding of gene function, in particular the prioritization of genes for mouse-knockout phenotyping experiments is considered. I identify biases in existing mouse phenotype resources both with respect to gene function and various genomic properties that could affect the results of human-genetics studies. Then I demonstrate how biological networks could be used to correct these biases by prioritizing genes of unknown function. Finally I discuss the significance of my results in a broader scientific context in Chapter 7. Extra figures and long tables which did not fit in the main text of my chapters can be found organized by chapter in the Appendices (Figures in Appendix A, Tables in Appendix B) following the bibliography.

Chapter 2

Materials and Methods

2.1 CNV Datasets

Throughout this thesis I examined copy number variants (CNVs), defined as stretches of DNA at least 1 kb in length duplicated or deleted relative to the reference genome. These CNVs were obtained from various published sources and publicly accessible databases. CNVs were identified either within patients diagnosed with one or more developmental phenotypes such as intellectual disability, autism, and/or congenital malformations or within apparently healthy individuals. In two of these datasets (DECIPHER and GENCODYS) attempts were made to examine the parents of the affected individuals to determine the inheritance of the variants. Most analyses focuses on the *de novo* variants since these consistently show the greatest deleterious effects and are the most penetrant (10; 12; 14). Furthermore, frequently I filtered these *de novo* variants to remove very large CNVs (>5 Mb) since the large number of genes they contain can obscure functional associations (99), and/or remove very small CNVs (<100 kb) due to high error rates when calling small CNVs because of the limited resolution of arrays used to detect them (4; 100; 101). Furthermore at larger sizes (>500kb) differences in calling CNVs due to differences in array platform disappear (25).

2.1.1 DECIPHER

CNVs were obtained from the Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources (17) (DECIPHER) in November 2009. These variants were contributed by over 200 different centers around the world and represent over 10,000 cases. Participating centers were mainly located in the USA and UK. Each case was examined using a high resolution array and CNVs were called separately by each centre; in addition the patient's phenotype was described using controlled terms of the London Dysmorphology Database (LDDB) by a clinical geneticist. The finalized data were uploaded and stored by the DECIPHER project at the Wellcome Trust Sanger Institute.

At the time of download, DECIPHER contained 4,614 CNVs with end points recorded relative to the hg18 human genome build, of which 626 were *de novo* variants, 2,464 were inherited variants, and 1,524 were of unknown inheritance (Table 2.1). As previously reported *de novo* CNVs were larger and affected a greater number of genes than inherited CNVs (4). CNVs of unknown inheritance were generally between these two extremes. In contrast to *de novo* CNVs, inherited CNVs were mostly duplications (60%) whereas *de novo* CNVs were mostly deletions (74%). This may be a result of deletions being more deleterious or more penetrant (4) thus less likely to be passed from parent to child due to reduced fecundity among affected individuals (19).

Syndromes

In addition to these patient CNVs, I obtained a list of regions associated with particular developmental syndromes from DECIPHER on 7 March 2013. This list included 66 regions, five of which were located on the Y chromosome and were excluded, leaving 61 regions associated with 58 different syndromes (Table 2.1, B.1). Nine regions were included twice since the deletion and duplication of the same region was associated with different syndromes. These syndrome regions were of comparable size and gene content to the *de novo* CNVs. However there were fewer phenotypes annotated to

Table 2.1: DECIPHER CNV data as of November 2009. Ancestral phenotype terms have been imputed.

	inherited	unknown	de novo	de novo <5Mb	de novo >100kb & <5Mb	Syndromes
Median size (bp)	185,557.5	344,058	2,229,944	1,278,074	1,483,416	1,398,217
No. CNVs	2464	1524	626	471	427	61
No. Losses	1010	775	464	348	317	42
No. Gains	1454	749	162	123	110	19
No. CNVs with genes	1843	1289	582	427	406	60
Median genes per CNV	4	7	20	11	13	15.5
No. Patients	359	785	516	384	372	NA
Phenotypes per patient	11	15	15	14	14.5	9

these syndromes likely due to the high level of heterogeneity among patients with developmental disorders.

2.1.2 GENCODYS

Collaborators at the Radbound University Medical Centre, Nijmegen, provided me with data on 4,240 patients with intellectual disability, developmental delay, and/or multiple congenital abnormalities (these have since been published in (18)). These data are part of the Genetic and Epigenetic Networks in Cognitive Dysfunction (GENCODYS) project. Each patient was phenotyped by clinicians using a uniform and standardised clinical form with phenotypes described using the Human Phenotype Ontology (HPO) terms (35). Of 10,000 possible HPO phenotypic terms covering the full spectrum of human phenotypic abnormalities, 1350 terms were assigned to one or more patients within the cohort. DNA samples were mainly taken via peripheral blood and analysed using the Affymetrix 250k Nspl SNP array platform (262,264 SNPs, 200Kb resolution). This array was based on hg17, thus end points were mapped to hg18 using liftOver prior to mapping genes (102). CNVs were called where there were at least five or seven consecutive aberrant SNPs, for losses and gains, respectively. Where CNVs were observed, parental DNA was considered in order to determine whether CNVs were *de novo*, or to determine the mode of inheritance. In all, 426 *de novo* CNVs, 636 inherited CNVs and 597 CNVs of unknown inheritance were detected, ranging in size from 5kb to 158Mb (Table 2.2). Again I find that *de novo* CNVs were larger and affected more genes than inherited CNVs. The respective patients exhibited a greater number of phenotypes for *de novo* variants than inherited ones. Again I find that inherited CNVs were more likely to be duplications (63%), whereas *de novo* CNVs were more likely to be deletions (60%).

Notably, GENCODYS CNVs of each type were larger than their DECIPHER counterparts, likely due to the lower resolution of the arrays employed. However, they were more consistent since all patients were analysed using the same array and the same

Table 2.2: GENCODYS CNV data. Ancestral phenotype terms have not been imputed.

	inherited	unknown	de novo	de novo <5Mb	de novo >100kb & <5Mb
Median size (bp)	518,884	1,210,311	2,670,612	1,483,415	1,525,370
No. CNVs	636	597	426	288	277
No. Losses	238	331	254	172	164
No. Gains	398	266	172	116	113
No. CNVs with genes	536	545	412	274	268
Median genes per CNV	4	10	24	16	17
No.Patients	495	352	296	222	216
Phenotypes per patient	28	23	39	38	37

clinical form unlike those of DECIPHER. GENCODYS patients had many more phenotypes annotated to each patient than DECIPHER but this was likely due to the higher level of detail included in HPO compared to the LDDB.

2.1.3 Benign CNVs

Benign CNV data was obtained from (103). All variants present in at least one of the Caucasian, Asian, or African populations were included. These benign CNVs were much smaller and affected many fewer genes than the CNVs identified in either DECIPHER or GENCODYS patient cohorts (Table 2.3). When all CNVs were included there was a large bias towards deletions (74%) however when large CNVs were considered (between 100kb and 5Mb in length) the majority were duplications (59%) as seen in the inherited CNVs in both DECIPHER and GENCODYS datasets. This suggests the issue isn't one of detection but that large deletions are more likely to be pathogenic and thus less likely to be inherited than similarly sized duplications.

Table 2.3: Benign CNV data.

	All	100kb< and <5Mb	100kb< and <5Mb affecting at least 2 genes
Median size (bp)	12,555	182,427.5	208,334.5
Number of CNVs	19,295	2478	1082
Number of Losses	14,191	1018	463
Number of Gains	5,104	1460	619
CNVs with genes	5,678	1628	1082
Median genes/CNV	1	2	3

2.1.4 Case-Control CNVs

A further set of CNV data from another cohort of patients with various neurodevelopmental abnormalities was obtained from (4). These data were downloaded from dbVar (study nstd54) on January 24th 2013. Patient phenotypes were described using descriptions from the HPO ontology which I mapped back to their HPO terms and imputed parental terms. However due to the more minimal phenotype descriptions, on average 7 terms/patient vs >20 terms/patient using the same ontology in the GENCODYS dataset, I did not use them in the analyses involving the specifics of patients' phenotypes (Table 2.4). The case CNVs in this dataset most closely resemble the DECIPHER inherited CNVs in terms of size (Case CNVs 100kb, DECIPHER inherited 200kb) and distribution of losses/gains (Case CNVs 63% gains, DECIPHER inherited 60% gains). Only a minority of CNVs were *de novo* variants in DECIPHER and GENCODYS thus it is likely these Case CNVs are predominantly inherited variants. In contrast to the benign CNV dataset above, the Control CNVs in this dataset were very small (2kb) and predominantly deletions (89%) may be due to the difficulty in identifying small duplications or the greater mutation rate for small deletions than small duplications (25; 104).

Table 2.4: Case-Control CNV data. After imputation.

	Case	Control
Median size (bp)	106,693	1,853
Number of CNVs	58,012	886,671
Number of Losses	21,236	790,371
Number of Gains	36,776	96,300
CNVs with genes	34,633	101,693
Median genes/CNV	2	1
Number of Patients	9,526	not available
Median phenotypes/patient	7	NA

2.1.5 Mapping & Imputing Phenotypes

DECIPHER patient phenotypes were recorded using LDDB and GENCODYS patient phenotypes were recorded using HPO. I converted between these ontologies as needed using the file provided on the HPO website (35), `http://www.human-phenotype-ontology.org/contao/index.php/downloads.html`. All provided mappings of any quality were considered. For those LDDB terms mapped to more than one HPO term the most general HPO term was used.

In addition, unless otherwise stated I used the hierarchical structure of each ontology (LDDB or HPO) to recursively assign more general ancestral terms to each patient based on the specific clinical assignments provided in each data set, eg. a patient with the specific term "Hypertelorism" would be assign the more general term "abnormality of the eye". LDDB imputations were done with November 2009 version of the LDDB ontology. HPO imputations were done using the October 2012 version of the Human Phenotype Ontology.

2.1.6 Assigning Genes

Genes were assigned to each copy-number variant (CNV) or syndrome region using the gene models in Ensembl (version 54 (105); human genome build hg18). Regions were first mapped to this human genome build as necessary using liftOver (102). All genes where at least one exon from every transcript was within the boundaries of the

genomic region (CNV or syndrome) was deemed affected by that region. This method has been shown to reduce biases towards either long or short genes (106); long genes are more likely to be assigned if any extent of overlap between the region and the genes deemed sufficient for assignment and short genes are more likely if only genes completely within the region are assigned.

2.2 Calculating Phenotypic Similarity

Since the patient phenotypes were described using standardized terms from the LDDB or HPO ontology they can be readily compared and an overall similarity between the patients could be calculated. Similarly phenotype annotations to genes either from mouse knock-outs (using the Mammalian Phenotype Ontology (107)) or from curated human disease descriptions from HPO (35) could be used to calculated phenotypic similarity between genes with these annotations. However, in general only a small number phenotype-gene relationships have been discovered thus there is a high false negative rate in these datasets due to unexamined or undiscovered associations. In contrast patients with developmental disorders typically have very well described phenotypes. Thus different measures of phenotypic similarity were appropriate for each of these cases.

2.2.1 Definitions of Measures

Semantic Similarity

Semantic similarity is a way of measuring the similarity between pairs of terms which belong to a hierarchical ontology. For genes or patients with multiple terms annotated to them these pairwise term-to-term similarities are then combined to produce an overall similarity between the two genes. Semantic similarity uses both the structure of the ontology and the overall frequency of each term in a given population (ie. all genes in the genome with any annotations) in the calculation of the similarity.

To calculate the semantic similarity between two terms, all the common ancestors between the two terms are identified using the hierarchical ontology. The information content of each of these common ancestors is calculated as: $-\log p_i$ where p_i is the proportion of genes with that term (after imputing ancestral terms) (108). The term-to-term semantic similarity was defined as the average of the most informative disjoint, one is not a parental term of the other, common ancestors (Figure 2.1) (108; 109).

These pairwise term similarities were then combined for the whole set of terms annotated to each of a pair of genes as follows: for each term annotated to Gene A the most similar term of those annotated to Gene B was identified (best-match). This isn't necessarily symmetrical so this must be repeated swapping Gene A and Gene B. Next the most similar pair of the best-matches was identified (maximum best-match) and the average of all the best-matches was calculated (average best-match). Finally the average of these two values was taken as the final gene-to-gene semantic similarity (109).

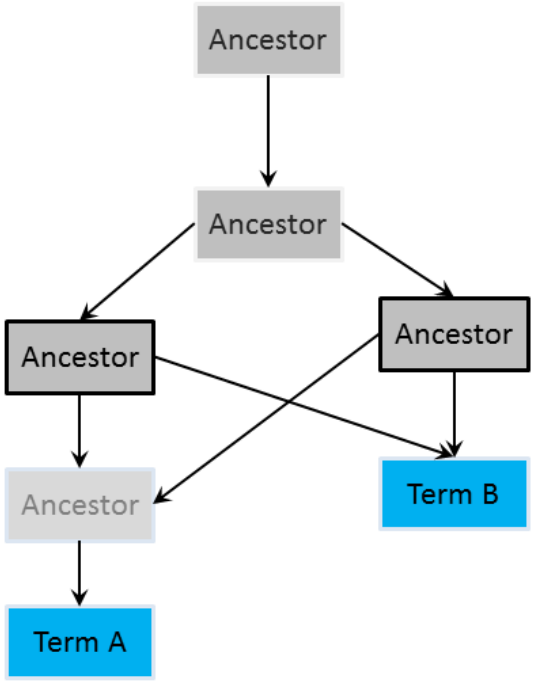


Figure 2.1: Calculating term to term semantic similarity. Dark boxes are common ancestors of Term A and Term B, black outline indicates the disjoint common ancestors. Arrows point from parent to child terms.

Goodall3

Phenotypic similarity between patients was calculated using the Goodall3 measure (110). The Goodall3 measure considers all phenotypes in the patient population and gives a high weight to the shared presence of rare phenotypes and the shared absence of common phenotypes for each pair of patients (Figure 2.2). For each pair of patients, the phenotypic similarity was calculated as the sum of the weighted similarity (G) of the presence/absence of each of all the phenotypes annotated to any of the patients considered, where G is weight by the frequency (f_i) of the phenotype in the respective patient population:

$$G_i = \begin{cases} 1 - f_i^2 & \text{if } i \text{ is present in both patients} \\ 1 - (1 - f_i)^2 & \text{if } i \text{ is present in neither patient} \\ 0 & \text{if } i \text{ is present in exactly one patient} \end{cases} \tag{2.1}$$

The resulting sum was divided by the total number of phenotypes considered, which is identical for all patient pairs in a given population, in order to ensure the resulting score was always between 0 and 1.

2.2.2 Comparing Measures

Both semantic similarity (SS) and Goodall3 are weighted by the frequency of the phenotype term in the population thus sharing a more specific rare term results in a higher value than sharing a more general common term. However, the absence of a common term within two patients will only result in a high score when using the Goodall3 index, not when using SS (Figure 2.2). SS only uses information from phenotypes which are present in the patient(s) to obtain a similarity score. Thus if a dataset is likely to contain many false negatives, for instance because the phenotypes were not examined due to ascertainment biases, SS score can only be lower than they should be; whereas such a situation could result in higher or lower Goodall3 scores. Thus SS is the preferred measure if a dataset is suspected of having a high false negative rate (eg. among

mouse phenotype data). However, when phenotypes have been thoroughly examined, as is the case for the GENCODYS dataset, Goodall3 is a better measure.

Furthermore, SS weights rare phenotype terms much more highly than Goodall3. Depending on the quality of the annotations of these specific rare phenotypes this may or may not be desirable. For instance if phenotypes of moderate frequency are quite general (eg. the MPO term *abnormal nervous system physiology* which is annotated to 23% of genes with any MPO term) it may be desirable to give greater emphasis to more specific terms. However, rare phenotypes may be difficult to measure/detect thus be more prone to ascertainment bias (eg. *abnormal cochlear microphonics* which is annotated to only 0.2% of genes).

SS explicitly accounts for the structure of the phenotype ontology during its calculation thus can be calculated on a set of terms either before or after imputing the ancestral terms; imputing the ancestral terms has relatively little effect on the absolute value of the measure since they will have the same or lower term-to-term similarity as their child terms. In contrast, it was necessary to impute all ancestral phenotypes prior to the calculation of the Goodall3 similarity measure in order to account for the structure of the phenotype ontology since its calculation is based on the presence of absence of identical terms in each pair of patients.

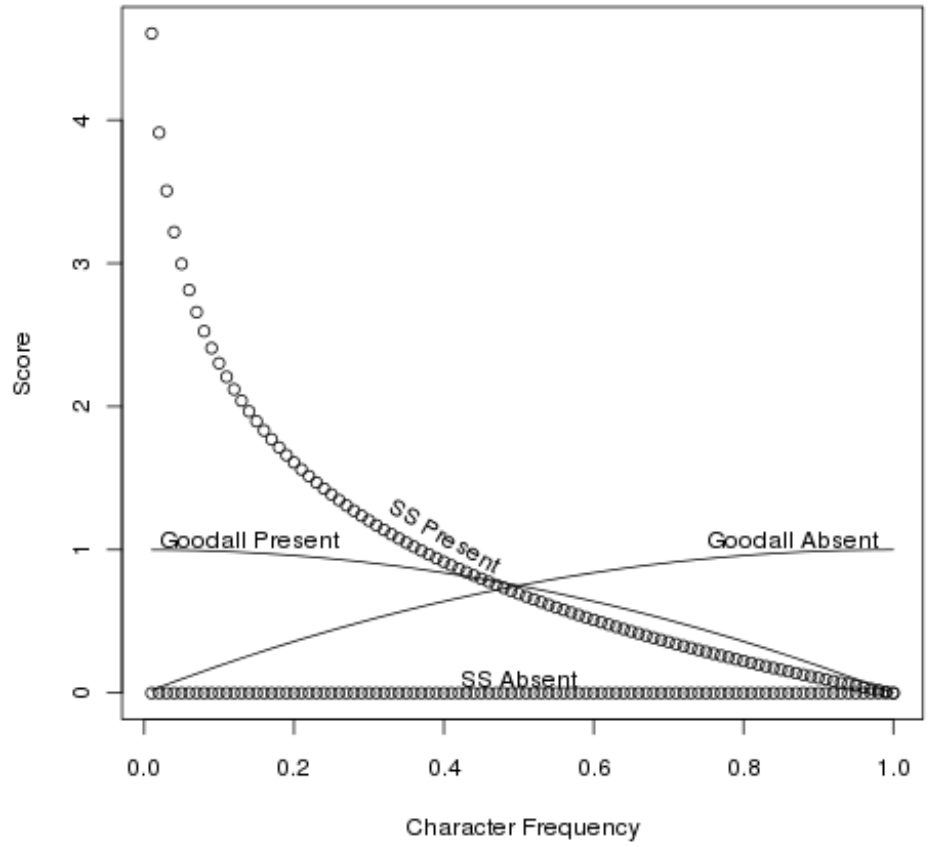


Figure 2.2: Comparing Goodall similarity metric to semantic similarity(SS) for patient phenotype comparisons. While semantic similarity strongly weights the presence of a shared rare character (in this case a phenotype term), nothing is learned from the shared absence of a character. By comparison, the Goodall metric considers both the presence of shared rare character and the absence of a common shared character towards the overall similarity. The Goodall metric is thus more suitable where both the presence and absence of phenotypes are known, as is the case with the systematically phenotyped GENCODYS cohort. Where as semantic similarity is more suitable when there is a high number of false-negatives due to incomplete information, as is the case with the mouse phenotype dataset.

2.3 Determining Gene Function

The function of a gene or a protein can be described or defined in a multitude of ways. Proteins often have more than a single function: eg. Rac1 catalyzes the hydrolysis of GTP but also binds to a variety of proteins and participates in many signalling pathways to regulate the cell growth and the cytoskeleton (111). Many sources of data can be used to identify or infer the function of a particular protein/gene including: sequence homology, studies in the scientific literature, protein-protein interactions, gene expression patterns, phenotypes when mutated in a model organism, protein 3D struc-

ture, and genetic interactions. In addition, these data sources can be combined to improve functional inferences (55; 56; 59).

2.3.1 Functional Annotations

The most common method for determining the function represented by a set of genes is to identify an enrichment of genes with a particular functional annotation among the set using a hypergeometric test (equivalent to a Fisher's exact test on a 2 by 2 contingency table). Any terms used to categorize genes (or their protein products) can be used but in this thesis I used the following: the Gene Ontology terms (GO), Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways, and mouse-knockout phenotypes from the Mouse Genome Informatics Database (MGI). I also used human disease annotations from the Human Phenotype Ontology (HPO) and the Online Mendelian Inheritance in Man (OMIM).

Gene Ontology (GO)

Gene Ontology annotations were obtained from Ensembl (version 62, (105)) for each of the three ontologies: cellular component (CC), molecular function (MF), and biological process (BP). The CC ontology contained 1,200 unique terms annotated to 17,415 genes. The MF ontology contained 3,578 unique terms annotated to 17,373 genes. The BP ontology contained 9,792 unique terms annotated to 15,685 genes. A list of GOSlims terms were downloaded on 2nd January 2013 from the GO website (`http://www.geneontology.org/GO.slims.shtml`). These 148 terms are chosen to broadly represent all three ontologies.

Kyoto Encyclopedia of Genes and Genomes (KEGG)

KEGG data was obtained from an existing copy of the database prior to it ceasing to be publicly accessible (dating from 2011). These annotations included all 241 curated pathways and the classes and subclasses they are organized into. 5,264 Entrez genes

were annotated to on average 1.9 different pathways, and for a total of 5.7 different KEGG terms, (using the id-mapping file included in the download). Ensembl gene IDs were mapped to Entrez gene ID using Ensembl 54 (105) as needed to calculate enrichments.

Mouse Phenotypes (MGI)

Knockout-mouse phenotype information was obtained from the Mouse Genome Informatics (MGI, (36; 37)) website (`ftp://ftp.informatics.jax.org/pub/reports/index.html#pheno`). Unless stated otherwise the data used was downloaded on February 7th 2012. However for Chapter 6 a second version was obtained on January 8th 2014. These were mapped to their human 1-1 orthologs as defined by the MGI database. The provided gene symbols were mapped to Ensembl gene IDs (version 54). Results from heterozygous knockouts (haplo-insufficiency) were obtained from this same source. Overall there were 7,706 unique Mammalian Phenotype Ontology (MPO) terms organized into 29 overarching categories annotated to 6,434 human genes.

Human Phenotypes & Disease Genes

The Human Phenotype Ontology (HPO, (35)) uses rigorously defined terms to describe the exhibited phenotypes of various diseases/traits and links these descriptions to the causal genes based on information available in the scientific literature as well as the Online Mendelian Inheritance in Man database. These annotations were downloaded on 17th October 2012 from the HPO website (`www.human-phenotype-ontology.org/`). Gene IDs were mapped from Entrez to Ensembl IDs using EnsemblMart 54 (105).

In addition, a list of known disease genes was obtained from the Online Mendelian Inheritance in Man database (OMIM, (112)) on March 20th 2012. This list was filtered to remove genes with uncertain disease associations, leaving only genes described as:

"the molecular basis for the disorder is known; a mutation has been found in the gene" or "a continguous gene deletion or duplication syndrome, multiple genes are deleted or duplicated causing the phenotype" and where these results have been confirmed in at least two laboratories or in several families. This filter removed 11,502 of the 13,250 gene-disease annotations, leaving 1,648 high-confidence disease genes after mapping to Ensembl 54 gene IDs.

Imputation

More general terms were recursively annotated, traversing up the relevant ontology, to each gene according to the specific terms present in the annotation file for each dataset.

2.3.2 Phenotypic Linkage Network (PLN)

An alternative to examining multiple functional datasets individually is to combine them together into an integrated functional network (55; 56; 59). This network represented the functional relationships between genes as weighted edges (functional similarity/relatedness) connecting nodes (genes/proteins). In creating our phenotypic linkage network (PLN), we (Frank Honti) integrated co-expression, protein-protein interactions, and various functional annotations into a single network (Table 2.5). Functional information identified in mice was transferred to human-genes using one-to-one orthology relationships. Each dataset was evaluated using the semantic similarity between mouse phenotypes (36; 37). The portion of the data with a positive correlation with the semantic similarity between mouse phenotypes was selected and re-scored to mouse-phenotype equivalents using a fitted regression then combined using a weighted sum where those datasets with the strongest correlation with the mouse phenotypes receiving the highest weight as described in (65) and (113):

$$G_i = \left\{ \; S_{AB} = L_0 + \sum_{i=1}^{n} \frac{L_i}{D \times i} \right. \tag{2.2}$$

Table 2.5: Datasets included in the phenotypic linkage network (PLN). All datasets except the phenotype data were first integrated based on the strength of their relationship with the mouse phenotype data. The resulting integrated network was combined with the mouse phenotype data based on the relationship with the similarity of human phenotypes annotated to the genes to create the final PLN.

Datatype	Source
Biological Process	Gene Ontology
Cellular Location	Gene Ontology
Molecular Function	Gene Ontology
Protein-Protein Interactions	BioGRID, IntAct, Corum, DICS, Reactome (Mouse & Human)
Gene Coexpression	GNF2, GSE3594 (Gene Expression Omnibus), MTAB-62 (Gene Expression Atlas) , 5 SMD sets (Stanford Microarray Database) from (114; 115; 116; 117; 118)
Protein Domains	InterPro
Pathways, Reactions	KEGG & Reactome
Genetic Interactions	Yeast
Co-citation (Literature PPIs)	STRING

where L_i is the re-scored weight from dataset i ordered in decreasing size, ie. L_0 is the largest link weight among all the links between the two genes, L_i is the second largest etc... and D whose optimized value based on (65) was $D = 5$. This initial integrated network contained 1,396,596 similarities between 17,011 genes.

The resulting integrated network (65) was then combined with the mouse phenotype information using the same method training on human phenotypes obtained from HPO. Again the qualitative phenotype information from HPO was quantified using semantic similarity (see above) then a polynomial regression between these pairwise gene similarities and the gene similarity obtained by calculating the semantic similarity of the mouse phenotypes or the initial integrated network. The resulting regression was used to convert both the initial integrated network and the mouse phenotypes to the expected human phenotype-based similarity score and the weighted sum of these two scores was used to construct the final Phenotypic Linkage Network (PLN). This network both contained a similarity score for a larger number of genes-pairs (10,792,987), and these similarity scores were a better reflection of the phenotypic

effects of these genes in human than either the mouse phenotypes information or the initial integrated network alone (Figure 2.3).

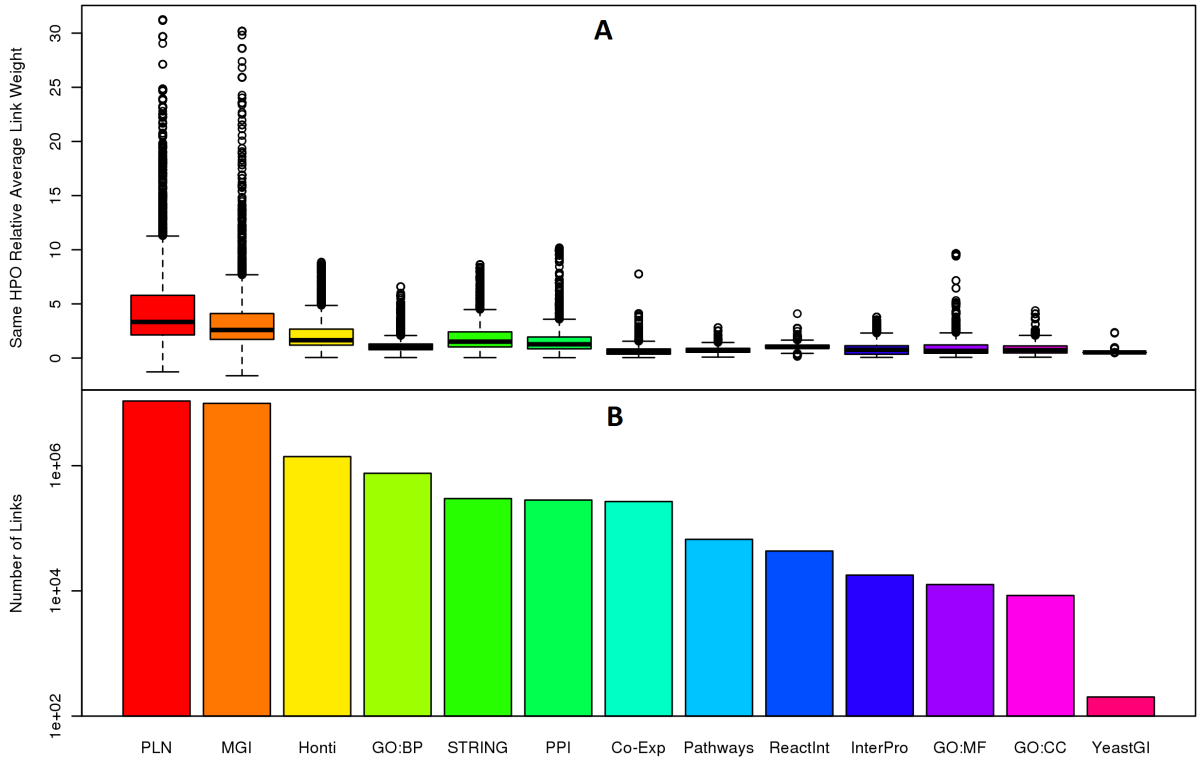


Figure 2.3: Incorporating the mouse knockout phenotypes increases the specificity (A) and coverage (B) of the Phenotypic Linkage Network (PLN). (A) the average link weight between genes annotated with a given Human Phenotype Ontology (HPO) term for each network divided by the average link weight between all genes with any HPO term. All HPO terms with frequency < 0.5 were considered. (B) Total number of links in the whole network. PLN = the Phenotypic Linkage Network. MGI = semantic similarity between mouse phenotypes only network. Honti = the integrated functional network from (65) which combines all datasets except the mouse phenotypes. GO:BP = Gene Ontology Biological Process. PPI = combined protein-protein interaction network from all databases considered Table 2.5. Co-Exp = gene co-expression. Pathways = Reactome + KEGG. ReactInt = Reactome Interactions. GO:MF = Gene Ontology Molecular Function. GO:CC = Gene Ontology Cellular Component. YeastGI = Yeast genetic interactions.

2.3.3 Other Networks

In addition to the PLN, many major results were replicated using two publicly available networks. First was HumanNet (version 1, downloaded from `http://www.functionalnet.org/humannet/about.html` on February 18th 2013) an integrated network which combined gene expression, protein-protein interactions and co-citation of human genes and their orthologs in worm and fly(53). This was a relatively sparse network with under half a million edges linking 16,000 genes (Table 2.6). The second network was a human-specific co-expression network from COXPRESdb (version 6, downloaded from `coxpresdb.jp/download.shtml` in April 2012)(46). This network was derived from 123 experiments analyzed on the Affymetrix Human Genome U133 Plus 2.0 Array (4,401 GeneChips) by calculating a weighted correlation to account for redundancies in the data. Due to the noise in gene expression data, I filtered this network to remove all links where the Pearson correlation was less than 0.5 (Table 2.6, (119; 120; 121; 122)). This left 1.7 million edges between 13,236 different genes.

Finally, other networks and functional resources were used in individual chapters, these have been described in detail in the respective chapter's Specific Methods section. Briefly these included additional co-expression networks derived from BrainSpan (123), or GTEx (124); protein-protein interactions from DAPPLE (125), iRefIndex (52) or STRING (51); and Co-citation from Pubmed (126).

Table 2.6: Networks used to defined functional similarity.

	PLN	HumanNet	COXPRESdb
Genes	17,039	16,243	13,236
Direct Edges	10,792,987	476,399	1,771,841
Filtering	None	None	r >= 0.5
Shortest Paths	142,864,287 (98.4%)	129,146,568 (97.9%)	81,046,264 (92.45%)
1% Threshold	0.1	0.6	0.5
Data Types	PPI, co-expression, pathways, co-citation, domains, functional annotations, mouse phenotypes	PPI, co-citation, co-expression from human, worm, fly, and yeast	>100 expression experiments
Reference	(65)	(53)	(46)

2.3.4 Tissue-specific Expression

In several chapters I also examined the particular tissues genes were expressed in or the broadness of the expression of the gene. For this I used the Illumina Body Map 2.0 (127) or the Gene Expression Barcodes (128). The Illumina Body Map RNASeq expression data was downloaded on January 21st 2014 from the Gene Expression Atlas (ID: E-MTAB-513) (129; 130). This dataset contained FPKM results for 18,680 genes across 16 different tissues, a gene was considered expressed in a tissue if it had an FPKM > 1. The Gene Expression Barcodes were downloaded on 28th August 2013 from `http://barcode.luhs.org/`. I used the data for the Affymetrix Human Genome U133A Plus 2.0 (obtained from a large number of studies from the Gene Expression Omnibus) since after mapping the probes to gene models using Ensembl54 this contained data for 18,374 genes where as the U133A array data was mapped to only 12,599 genes. For each gene only the most specific probes (with the least cross-hybridization) were considered. A gene was considered expressed in every tissues where at least one of the respective probes was expressed with a probability > 0.5 based on their mixed-model as suggested in (128). This data was filtered to retain only the 106 human normal tissues or cell-types (cancerous and non-human primate data was excluded).

2.4 Clustering Methods

Each results chapter involves identifying groups of related genes based on one or more of the above functional networks. These groups are typically called 'clusters' and the process of identifying them is referred to as 'clustering'; however it may also be referred to as 'community detection' when applied to partitioning a full network into unusually dense sub-graphs ('communities')(72). However, since there is no single optimal algorithm to perform this clustering I used different algorithms to suit the situation addressed in each chapter. Here I will discuss the differences between these algorithms and the reasons each one was chosen for each particular situation featured in the different chapters. Particulars on the parameters used and the specific networks

they were applied to are addressed in the Specific Methods section of each respective chapter.

2.4.1 Clustering Gene Sets

A common question addressed using clustering is whether a given set of genes belong to the same cluster, for instance whether sets of genes mutated in patients with particular phenotypes participate in the same pathway as seen in Chapter 3. There are two common approaches to take to address this question: i) identify clusters in the whole network then look for an enrichment of genes belonging to the set of interest among a particular cluster (eg. (131)), or ii) determining if genes in the set are more strongly connected to each other than expected (eg. (21; 22)). The first approach has some difficulties due to the large number of different algorithms and possible parameters which could affect the resulting clusters. Chiefly among these is the problem of the appropriate resolution at which to identify the clusters. Using a coarse resolution results in large clusters which will dilute the signal if there are a small number of genes in the given set. Whereas a fine resolution which gives small clusters may result in a large functionally connected gene set spreading across many clusters. In contrast, measuring the strength of connections between genes in a set is equally effective for large and small gene sets. Calculating the strength of connections is as simple as summing the weights of the links between genes (or counting the number of links if they are not weighted, eg. protein-protein interactions) in the given set then comparing that to the expectation given random genes in the network or after controlling for the total strength of connections for each gene ('degree') in the network (avoids potential issues due to biases towards more well studied genes). Thus I used this latter method (summing link weights and controlling for degree) to identify clustering of genes sets associated with particular phenotypes.

2.4.2 Hierarchical Clustering

Hierarchical clustering is another commonly used algorithm. Starting with each gene in a separate cluster, iteratively the two clusters most similar to each other are combined until only a single cluster remains. This results in a dendrogram showing the clusters at each stage of the algorithm. The dendrogram can then be cut at a desired height to produce a set of clusters. The advantages of this algorithm is it's simplicity, it requires little pre-processing of the data and run-time is proportional to the square of the number of items to be clustered ($O(n^2)$); this makes it suitable for clustering from two to a few thousand genes. Furthermore it is completely deterministic, always producing the same final dendrogram, and can be adapted to many different problems by changing the measure of similarity between clusters.

Thus various versions were used throughout this thesis. Complete-linkage hierarchical clustering, the similarity between clusters is the most different pair of genes with one gene from each cluster, was used to organize various bits of data in plots (n = 60, or 1,000) in Chapter 3 and Chapter 5. Average-linkage hierarchical clustering, where the similarity between clusters is the average similarity between pairs of genes between then, along with Ward's method hierarchical clustering (similarity between clusters is calculated to reflect the variance within the combined cluster) were used to group roughly 1,000 of genome-wide functional clusters affected by different types of copy-number variants (Chapter 5).

In addition, functional clusters were identified in copy-number variants (Chapter 4) using single-linkage hierarchical clustering, the similarity between two clusters is the most similar pair of genes with one gene from each cluster, with a cut-height chosen prior to the clustering. Thus all genes more similar than the cut-height to at least one other gene in the cluster were added to the cluster. An extended version of this algorithm was used to identify functional clusters across the genome (Chapter 5). Each of these copy number variants contained fewer than 1,000 genes.

2.4.3 Clustering a Network

In Chapter 6 I identify functional modules, groups of genes which participate in the same biological process or pathway, within various networks. Hierarchical clustering is too slow to apply to networks containing tens of thousands of genes and millions of edges, thus a faster clustering algorithm was need to address this problem. A recent study (132) found the two best algorithms for identifying clusters of a range of sizes within large networks were the Blondel algorithm for maximizing modularity (133) and the Infomap algorithm (69). Modularity is an aggregated measure of the density of links within each cluster compared to what is expected by chance; due to recent concerns that this measure fails to correctly identify relatively small clusters (134) I elected to not use the Blondel algorithm. The Infomap algorithm is another optimization algorithm, however rather than maximizing modularity it aims to find a minimal mapping for random walks in the network. The idea is that by partitioning the network into neighbourhoods where a random walker will tend to stay within the neighbourhood, and assigning these neighbourhoods a short code enables the compression of codes describing random walks on the network (69). An advantage of the Infomap algorithm is its basis in random walks through the network which is often used as a proxy for information flows through a signalling pathway or the flow of molecules through metabolic pathways; thus is able to capture patterns such as cycles or feedback loops which may not be evident from just the density of links.

Consensus Clustering

However, one of the disadvantages of the Infomap algorithm is the stochastic optimization algorithm which is necessary to reduce the runtime complexity ($O(n)$)) but this also introduces the issue of determining the stability of the achieved optimal clustering. One way to address the instability introduced by the stochasticity of the algorithm is to find a consensus clustering which is typically very stable (135). Consensus clusterings are ways to combine the outcome from multiple runs of a stochastic clustering algorithm into a single set of clusters. For simplicity, I used the consensus cluster-

ing method where the outcome of multiple runs (in this case 100 runs) of the Infomap algorithm were used to create an association matrix, which represents the frequency with which each pair of genes are placed into the same cluster by the algorithm, and the Infomap algorithm is simply applied to this association matrix to obtain the final consensus clustering (135).

Chapter 3

Pathways in patients with developmental disorders

3.1 Introduction

Developmental disorders and congenital abnormalities affect 3% of births, and represent an extremely heterogeneous group of disorders including intellectual disability, autism, developmental delay, often with a diverse range of structural and morphological defects (5). These disorders are highly heritable and considered primarily genetic in origin (29; 136; 137; 138). Patients have been found to possess an increased burden of copy number variants (CNVs; regions of the genome > 1Kb that are deleted or duplicated) (14; 22; 29; 139). CNV screens now routinely included in primary diagnostics (16; 139). Clinicians have identified numerous genetic syndromes on the bases of phenotypic similarities across a broad range of features in a group of patients sharing a particular genetic cause (eg. (30)). However, the heterogeneity of patient phenotypes is reflected in the underlying genetics, with many patients possessing unique complements of phenotypes and/or unique CNVs, making the identification of the particular genes contributing to the phenotype difficult. Pathway analysis can circumvent these problems since patients with distinct CNVs affecting functionally related genes or presenting similar phenotypes can be pooled together, thus removing the issue of pheno-

typically or genetically unique individuals while still being enriched for a particular aetiology(26; 27; 28).

Large-scale pathway analysis has been facilitated by the use of phenotype ontologies, such as the Human Phenotype Ontology (HPO)(35), which facilitate the identification of groups of patients with various levels of phenotypic similarity by structuring phenotype descriptions using a hierarchy of terms with more specific child terms organized beneath broader parent terms. In addition, large databases of functional information about genes, such as the Gene Ontology (34), permit the systematic examination of multi-genic disease through the use of functional enrichments (38). This approach assumes that genes affected by multiple variants associated with a particular phenotype act within a common pathway thus will share other functional characteristics such as being annotated to the same biological process (39; 40), follow a similar expression pattern (131), whose protein products physically interact (21), and/or whose disruption in model organisms results in the same phenotype (41). Sources of functional information can vary with respect to error rates (false positives & false negatives) particularly when derived from high-throughput experiments or computational inferences (42; 43; 44; 45; 140). Combining functional information from different sources may increase confidence that genes participate in the same biological pathway (61; 62)

In this chapter, we identified functional enrichments associated with particular phenotypes in a large cohort of patients with CNVs and developmental disorders. We combined multiple sources of functional information and verified they reflect molecular pathways using protein-protein interactions (PPI). I introduced a novel validation of these pathways based on phenotypic similarity between patients, and show that most pathways were associated with multiple distinct phenotypic characteristics (pleiotropy). I used this phenotype-based validation to evaluate the efficacy of the different sources of functional information as well as examine several potential sources of error which could lead to false-positive functional enrichments. Finally, I identified the phenotypes significantly associated with the various pathways disrupted by variants

in the respective patients to better inform future phenotypic descriptions of patients.

3.2 Specific Methods

3.2.1 Copy number variants

For this study, I exploited the consistent and comprehensive phenotyping of the patients in the GENCODYS dataset(18). This cohort contains data on 4,240 patients with intellectual disability, developmental delay, and/or multiple congenital abnormalities in which 1,659 CNVs were identified using an Affymetrix 250k SNP array (200Kb resolution). Of this cohort, 296 patients possessed *de novo* CNVs and 197 possessed *de novo* CNVs of <5Mb. We focused on this subset of 197 patients with small *de novo* CNVs since *de novo* CNVs are likely to be causative and very large CNVs introduce a lot of noise due to extraneous genes found in the region.

As a replication cohort we selected patients from the DatabasE of Chromosomal Imbalance and Phenotype in Human using Ensembl Resources (DECIPHER) who had a phenotype with a "Good" mapping (using `http://compbio.charite.de/svn/hpo/trunk/src/mappings/`) to the HPO phenotypes we examined in the GENCODYS cohort, (Table 3.1). Only Microcephaly (London Dysmorphology Database code: 32.08.05) had a sufficiently large protein-protein interaction (PPI) network (≥ 10) and an adequate number of DECIPHER *de novo* CNVs (≥ 5) to complete the analysis. Seventy-six CNV regions (CNVRs) were obtained by merging overlapping (by at least 1bp) or bookended CNVs (losses and gains combined). Eight of which affected genes participating in the Microcephaly network based on the GENCODYS CNVs and were removed, and 12 CNVRs did not affect any genes annotated within the PPI database could not be considered. Thus the final replication cohort contained 55 DECIPHER CNVRs from Microcephaly patients overlapping 606 genes in the PPI database (Table 3.1).

Genomic co-ordinates of CNV regions were mapped from hg17 to hg18 using the UCSC liftOver tool (102) through the bedTools application (`https:\code.google.com/p/bedtools/`) . Genes were assigned to each CNV if the CNV intersected with at least one exon in every transcript in the Ensmart 54 database (hg18) database. This method ensures that all coding transcripts of a gene are affected, and has been demonstrated to reduce length biases associated with genes that show brain-specific expression patterns (106). As we were primarily interested in genes whose *de novo* copy change would be highly penetrant, we removed 5,768 genes that had been observed as copy number changed in the same direction within a control cohort that represents deleted and duplicated regions in individuals with no overt abnormalities from (103).

Table 3.1: Mappings between HPO and the LDDM ontologies for the PPI network phenotypes. HPO: Human Phenotype Ontology id. LDDB: London Dysmorphology Database code phenotype with a "Good" mapping in `http://compbio.charite.de/svn/hpo/trunk/src/mappings/`, Explained : No. of DECIPHER patients whose CNVs overlap a gene already in the PPI networks identified among GENCODYS patients with this phenotype (Table B.6), GENCODYS PPI genes: Genes with the PPI network identified amongst the GENCODYS patients (Table B.6). Edges: Number of interactions between the GENCODYS PPI network genes and gene copy number variant in DECIPHER patients with the same phenotype. P-value: significance of the number of interactions observed compared to randomized CNVRs.

Term	HPO	LDDB	DECIPHER CNVs	Explained	Remain with PPI genes	GENCODYS Net Size	Edges	Pval
Microcephaly	HP:0000252	32.08.05	93	8	73	12	64	0.03796
Cryptorchidism	HP:0000028	20.03.12	14	0	14	2	na	na
Abnormality of the teeth	HP:0000164	13.01.00	6	1	4	5	na	na
Thick lower lip vermillion	HP:0000179	11.02.05	5	0	5	2	na	na
Thin upper lip vermillion	HP:0000219	11.04.05	27	0	26	2	na	na
Abnormality of the face	HP:0000271	10.01.00	19	0	15	3	na	na
Downward slanting palpebral fissures	HP:0000494	08.05.05	16	0	14	2	na	na
Abnormality of the nervous system	HP:0000707	32.31.00	2	0	2	6	na	na
Behavioural/Psychiatric abnormality	HP:0000708	32.05.00	12	0	11	6	na	na
Tapered fingers	HP:0001182	23.04.17	30	0	23	4	na	na
Abnormality of the toes	HP:0001780	26.05.00	4	0	4	2	na	na
Exaggerated cupid's bow	HP:0002263	11.01.03	2	0	2	2	na	na
Abnormality of the musculature	HP:0003011	32.30.00	2	0	1	2	na	na
Abnormality of the palpebral fissures	HP:0008050	08.05.00	2	0	2	94	na	na
Abnormality of the male genitalia	HP:0010461	20.03.00	2	0	2	2	na	na

3.2.2 Functional Enrichments

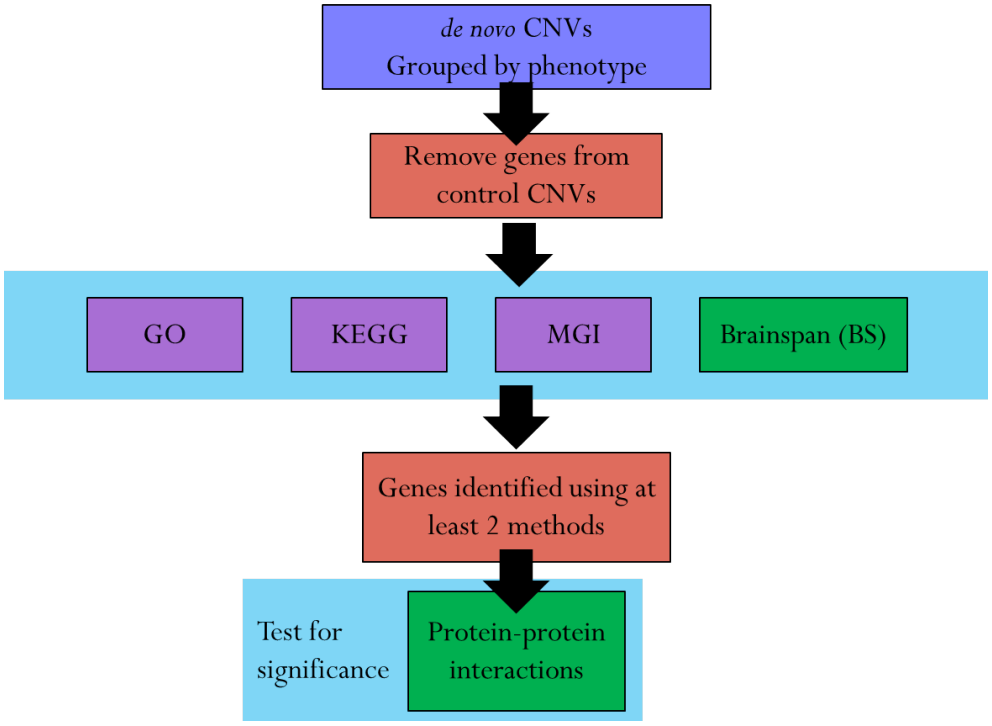


Figure 3.1: Functional enrichment analysis workflow. Qualitative descriptions of gene function (purple) were tested using a hypergeometric test. Network connectivity (green) was tested against random samples of genes controlling for the degree of the observed genes in the respective network. Tests for enrichment had a 5% FDR multiple testing correction applied.

The cohort of 197 patients were assigned to non-exclusive groups based on the presence of a specific HPO term. Functional enrichments were identified amongst genes disrupted by CNVs possessed by patients with each specific HPO phenotype using four different dataset (Figure 3.1): Mouse phenotypes (MGI,(36; 37)), Gene Ontology terms (GO, (34)), KEGG pathways (KEGG, (33)), and co-expression in the developing human brain (BrainSpan, (123)). The first three (MGI, GO and KEGG) were recorded as qualitative annotations for which significance was evaluated by comparing the frequency of the annotation with that expected given the whole genome background using a hypergeometric test, a False Discovery Rate (FDR) of less than 5% was applied to each resource (141). I replicated the significant results by testing against a background of all genes assigned to the *de novo* CNVs.

Functional Annotations

The phenotypes identified during published genetic mouse model experiments were obtained from the Mouse Genome Informatics (MGI; http://www.informatics.jax.org) (36; 37) and mapped to human genes using 1:1 human:mouse gene orthology relationships as defined by MGI. Higher level terms were imputed and assigned using the Mammalian Phenotype Ontology (MPO, (107)) . For each set of patients annotated with a specific HPO term, we identified the most relevant of 33 overarching categories within the MPO (Table B.2 in Appendix B). Enrichments were evaluated for each of the mouse phenotypic terms within the relevant overarching category, excluding phenotypes populated by less than 1% of the genes in the overarching category to reduce underpowered results.

All GO and KEGG annotations were examined for enrichments. Gene Ontology (GO) annotations and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway annotations were obtained from their respective websites ((34), http://www.geneontology.org/ and (33), http://www.genome.jp/kegg/). Higher level terms were imputed and enrichments were evaluated for each term.

Gene Expression Networks

Normalized RNAseq gene expression data from was downloaded from BrainSpan ((123), http://www.brainspan.org; 16 brain regions, 41 individuals aged from 8 weeks post-conception to 40 years). Genes with RPKM < 1 in $> 95\%$ of the samples were excluded and the expression correlation between each pair of remaining genes was calculated. A network was built with genes as nodes and edges between two genes weighted with their correlation coefficient r, considering only edges with weight $r \geq 0.7$ gave 13,953 unique genes with at least one edge. I confirmed our findings at $r \geq 0.6$ and $r \geq 0.8$, but found inconsistent and diminished results when $r \geq 0.9$ (Figure 3.3) which may result from a small number of links remaining at this threshold (151,636 links vs 5,680,000 links with $r \geq 0.7$ threshold). We compared the strength of

connections between genes in our test set (i.e. the sum of the correlation coefficients between the genes), as compared to genes randomly sampled from the co-expression network. We controlled for the number of edges associated with each gene (i.e. degree) by randomly sampling without replacement a maximum of 10,000 sets equal in gene number from the 100 genes with the most similar degree to each original gene to determine significance.

Given our findings of pleiotropic effects associated with the perturbations of the inferred pathways we report here (see Results), I repeated the analysis using a recently-released body-wide expression data, GTEx (124), instead of the brain-specific dataset. A co-expression network was derived from the GTEx data using the weighted correlation method used in (46) which accounts for missing data. However, I found that BrainSpan-derived molecular associations outperformed GTEx-derived association in the phenotype comparisons on which this decision would have been based, and thus retained the BrainSpan-derived co-expression approach (Figure 3.2).

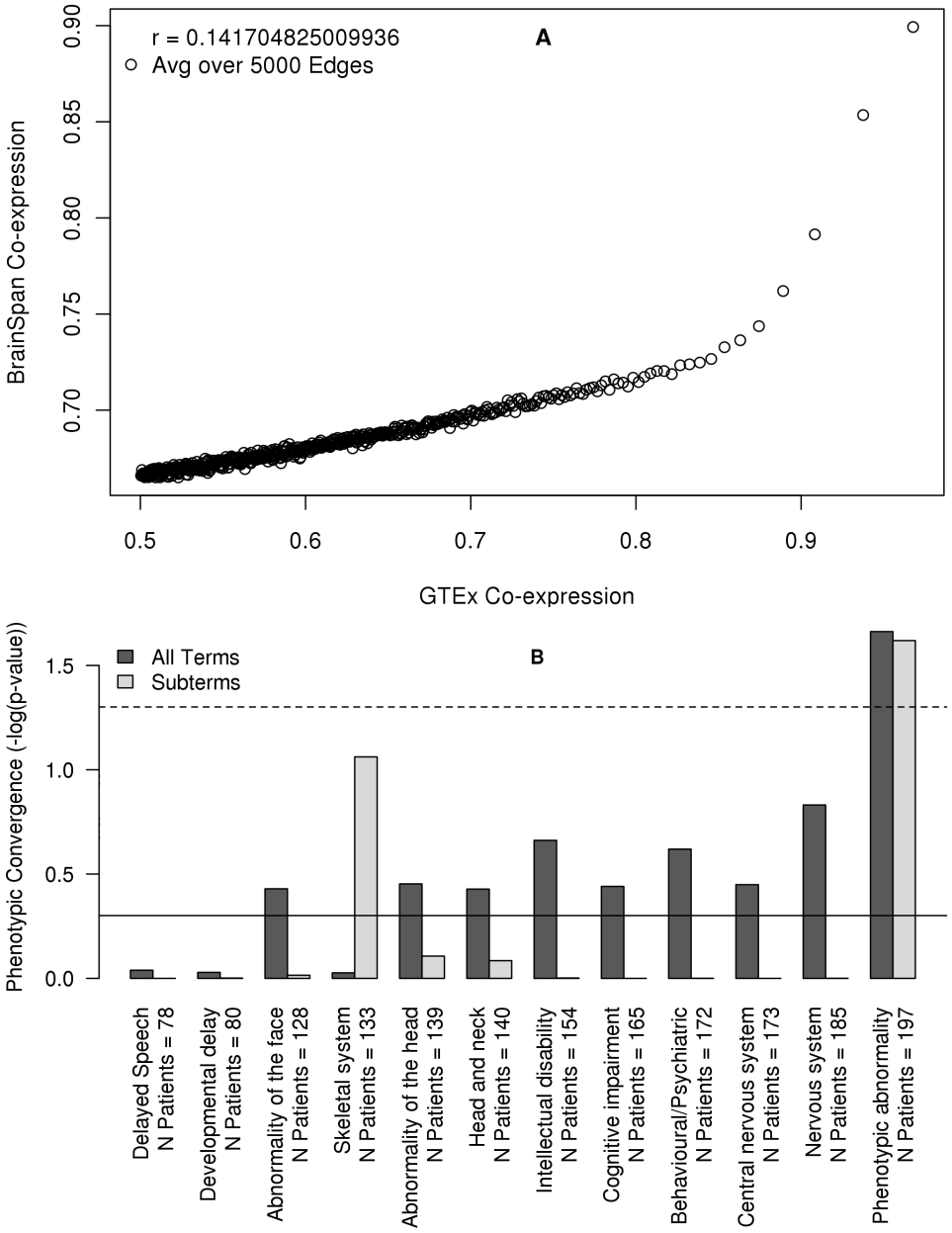


Figure 3.2: The GTEx co-expression networks did not perform as well as the BrainSpan co-expression network. (A) Pearson correlation between GTEx and BrainSpan co-expression networks (all edges with $r > 0.5$ in both networks). Each point is the average taken over 5000 edges. The pearson correlation coefficient on the unbinned data is noted in the corner of the plot ($r = 0.14$); (B) Phenotypic similarity of subset of patients contributing to a significant GTEx co-expression network. Solid line is p = 0.5, dashed line is p = 0.05. Dark bars are using all phenotypes to calculate phenotypic similarity, light bars are using only the child phenotypes of the original human phenotype to calculate phenotypic similarity.

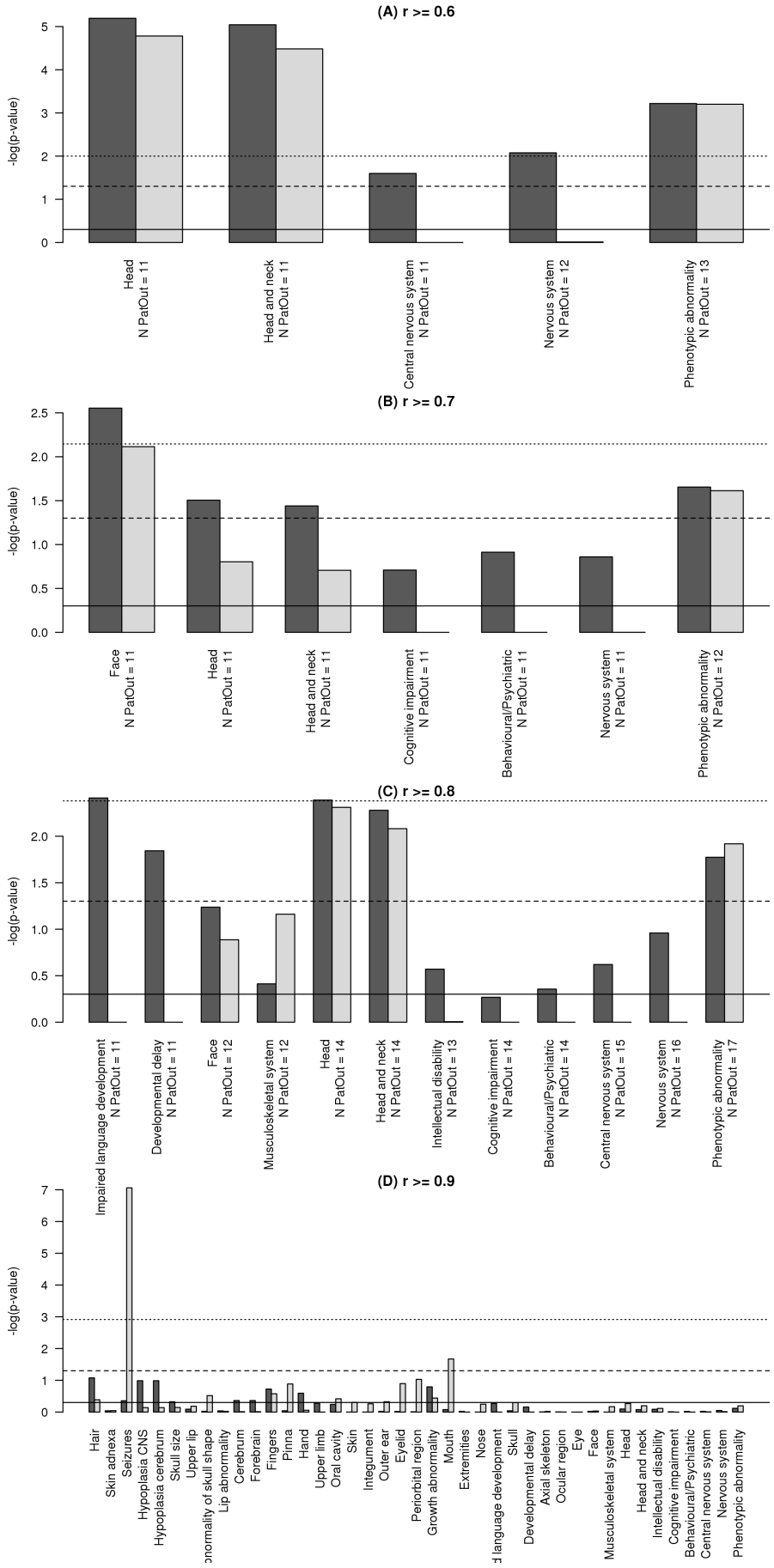


Figure 3.3: Phenotypic convergence for the BrainSpan co-expression network significant functional enrichements using four different thresholds on the Pearson correlation for links to be maintained in the network (A $r \geq 0.6$, B $r \geq 0.7$, C $r \geq 0.8$, D $r \geq 0.9$). BrainSpan consistently shows phenotypic convergence for a range of thresholds, but fails at $r \geq 0.9$ likely as a result of the paucity of links retained at this threshold (only 151,636 links vs 5,680,000 links at $r \geq 0.7$). "N PatOut" is the number of patients without genes participating in the co-expressed clusters. Solid line is p = 0.5, dashed line is p = 0.05, dotted line is significant after a Bonferroni correction. Dark bars are using all phenotypes to calculate phenotypic similarity, light bars are using only the child phenotypes of the original human phenotype to calculate phenotypic similarity.

Protein-Protein Interactions

All protein-protein interaction (PPI) data were obtained from the Dapple website (`http://www.broadinstitute.org/mpg/dapple/dapple.php`). Genes identified using at least two of GO, KEGG, MGI and BrainSpan enrichments (multi-method genes) for the same patient-phenotype groups set were clustered within the protein-protein interaction network by comparing the number of interactions between multi-method genes to randomly sampled gene sets, controlling for the number of degrees in the same manner as performed with the gene-expression network.

To investigate the connections between the genes within the GENCODYS Microcephaly (HP:0000252) PPI network and the genes variant in DECIPHER patients presenting with Microcephaly (Figure 3.5 B), we took the set of 12 genes participating in the GENCODYS-derived microcephaly PPI network (Figure 3.5 A) and asked whether they were more connected to the 606 genes (in the PPI database) affected by 55 CNVRs identified amongst the DECIPHER microcephaly patients (Table 3.1), than to genes hit by 500 randomised sets of 55 CNVRs, matched in the number of contiguous genes that were present in the PPI database as the original set. Randomised regions were prohibited from containing any of the 12 genes participating in the NIJMEGEN-derived microcephaly PPI network.

3.2.3 Phenotypic Similarity

Phenotypic similarity between patients was calculated using the Goodall3 measure ((110), see: Chapter 2 Section 2.2.1 for details). The Goodall3 measure gives a high weight to the shared presence of rare phenotypes and the shared absence of common phenotypes. Since the GENCODYS cohort was systematically phenotyped we are confident both in the presence of recorded phenotypes and in the absence of unrecorded phenotypes thus Goodall3 was deemed more appropriate than other measures such as semantic similarity (Figure 2.2). For each of the significant functional enrichments, the group of patients sharing the respective HPO term was divided into those patients

with variant genes participating in the enrichment (”contributing patients”) and those without (”non-contributing patients”). The population frequencies of each phenotype were derived from the 197 patients with *de novo* CNVs. The significance of the difference between the phenotypic similarity amongst contributing patients and between contributing patients and non-contributing patients was evaluated using a two-sided Wilcox-rank-sum test. To ensure the test was well-powered, only those cases where there were at least 10 contributing patients and at least 10 non-contributing patients were considered. he extent of phenotypic convergence for pathways identified using different resources was evaluated by comparing the number of pathways showing nominally significant convergence compared to 10,000 permutations of the functional resource labels.

Testing Errors and Biases

We considered several possible sources of error which could result in functional enrichments failing to identify phenotypically similar subpopulations within the cohort. Spearman and Pearson correlations were calculated between the observed phenotypic convergence among the 197 patients with *de novo* CNVs (<5Mb in size) and various measures reflecting research ascertainment bias, dispersion of contributing genes among patients, statistical power, and the effect of using a whole genome background for enrichments. Citation frequency, a measure of ascertainment bias in the scientific literature, was calculated from the number of abstracts which are associated with a given gene in Pubmed (using `ftp://ftp.ncbi.nih.gov/gene/DATA/gene2pubmed.gz`). Dispersion of contributing genes was measured using Shannon entropy (aka Shannon diversity) on the proportion of genes with the annotation found in each patient (142) which is calculated as:

$$Sh = -\sum_{i=1}^{n} p_i \log(p_i)$$

where n is the total number of contributing patients and p_i is the proportion of all

contributing genes which were affected by variants in patient i. Statistical power was measured as the number of patients with genes contributing to the enrichment, except for BrainSpan where almost all patients contributed genes in which case the number of patients without contributing genes was used. Finally, to examine potential biases resulting from using the whole genome background the replication p-value from a hypergeometric test comparing against the expected frequency given all genes affected by *de novo* CNVs in the 197 patients.

Replicating Phenotypic Convergence

We replicated the phenotype analysis described above using inherited CNVs amongst those 3,871 patients who did not possess a *de novo* CNV (adjusting the population fequencies of the phenotype term appropriately). We reasoned that if the pathways identified above are indeed responsible for those patients' phenotype then these same pathways could be used to identify the potentially pathogenic CNVs from the likely benign CNVs among the 1,043 inherited or unknown inheritance CNVs identified in these patients. We placed patients with the same specific phenotypic abnormality into three mutually exclusive groups: (1) "candidate pathways", defined as those patients whose CNVs affect genes identified previously in the *de novo* CNV analyses above that contributed to significant functional enrichments; (2) "Extended pathways", defined as those patients whose CNVs affect none of the genes included in the candidate pathways but that are nonetheless annotated with the same GO, KEGG or MGI function as previously associated functional enrichments or, for BrainSpan and the PPI, genes with a direct functional link in the respective network to one of the genes in the functional enrichment; (3) "No pathway", defined as those patients that possess either no CNVs or whose CNVs do not affect genes within previously associated pathways or extended pathways. Firstly, phenotypic similarity was calculated within the group of patients affecting candidate pathways as compared to the phenotypic similarity observed between the patients in this group and those within the combined group of patients affecting extended pathways or no pathway. Secondly, the phenotypic similarity was calculated within the group of patients whose CNVs affected the extended path-

ways as compared to the phenotypic similarity between patients within the extended pathway group and patients affecting no pathway.

3.2.4 Identifying Important Phenotypes

I identified phenotypes important for the observed phenotypic convergence amongst patients with CNVs affecting the same pathway by reversing the typical functional enrichment analysis. Each phenotype was tested for a significant association with each of the pathways we identified above by comparing its frequency among patients with CNVs affecting genes in that pathway, regardless of whether that patient possessed the phenotype the pathways was significantly enriched in, compared to its frequency in the whole cohort using a two-sided binomial test. A Bonferroni multiple testing correction was applied to conservatively eliminate potentially spurious results. This analysis was replicated three-fold: in the 197 patients with small *de novo* CNVs, in all 4,240 patients , and using the extended pathways with all 4,240 patients.

In addition the specificity of the convergent phenotypes was determined by calculating the intersection/union of all the genes (candidate or extended as appropriate) belonging to each pathway that were significantly associated with the phenotype using the reverse-functional-enrichments. This would only result in a high value if all the different pathways a given phenotype was associated with defined the same set of genes, and thus were effectively synonymous. Phenotypes only associated with one pathway received a specificity of 1.

3.3 Results

3.3.1 Summary of Dataset

We sought to identify and validate pathways underlying the heterogeneous phenotypes exhibited by patients with developmental disorders. For this we obtained copy-

number variant (CNV) data as well as detailed phenotype information from 4,240 patients diagnosed with intellectual disability, developmental delay and/or congenital abnormalities (18). Wherever possible trios were examined to determine the inheritance pattern of the identified CNVs. There were 1,659 CNVs observed within 1,388 different patients within the cohort. Only 426 CNVs could be identified as *de novo*, 636 were identified as inherited, and 597 were of unknown inheritance. We focused our analysis on sporadic patients possessing *de novo* CNV events because it is likely that the *de novo* CNV mutation is pathogenic, thus the disruption of at least one of the affected genes is responsible for the phenotype (16; 30). Patients with CNVs larger than 5Mb were excluded since the large number of genes affected by such CNVs introduce considerable noise which obscures the presence of functional associations (99). The remaining 197 patients possessed a total of 219 *de novo* CNVs (82 duplications and 137 deletions) with a median size of 1.37 Mb. A gene was considered affected by one of these CNVs if at least one exon of each transcript was overlapped by the CNV (106). The 219 *de novo* CNVs affected 2907 unique genes, with a range of 0-190 genes affected per patient (median = 95).

One of the advantages of this cohort was the standardized approach taken to describing phenotypes presented by the patients. All patients were evaluated for a comprehensive set of phenotypes described using terms from the Human Phenotype Ontology (HPO). The HPO organizes over 10,000 terms describing clinical phenotypic abnormalities into a hierarchical structure, grouping specific phenotypic terms together under more general parent terms. All parent terms were imputed from the each patient's clinically-assigned phenotype and assigned to the respective patient. The 197 patients with *de novo* CNVs possessed 826 distinct HPO phenotypes, with individuals possessing 2-182 phenotypes (median = 31). 154 patients (78%) had intellectual disability of varying severity, 80 were diagnosed with developmental delay, 54 with growth abnormalities, and 37 with autistic spectrum disorder (ASD).

3.3.2 Inferring molecular pathways

Biological pathways can be inferred from multiple different data-sources. We (Steve Meader) employed four methods commonly used in pathway analysis, namely, i) testing for enrichment of Gene Ontology (GO) terms among affected gene (34), ii) testing for enrichment of pathway annotations from the Kyoto Encyclopedia of Genes and Genomes (KEGG) (33), iii) examining abnormal phenotypes observed in mouse models (21; 143) , and iv) (Julia Steinberg) examining the level of co-expression between affected genes using a co-expression network. Since these patients had predominantly neurological phenotypes, we calculated gene co-expression from the BrainSpan dataset (123), which measured gene expression across 16 brain regions at 6 developmental time points (see: 3.2.2).

All four methods were applied to the genes affected by *de novo* CNVs in patients sharing a specific HPO phenotype where there were at least 3 patients with the phenotype (408 patient-phenotype groups). Only two patient-phenotype groups possessed significant functional associations using all four methods, namely *Seizure* (HP:0001250) and *Intellectual disability, Severe* (HP:0010864). A further 64 patient-phenotype groups shows significant enrichment using three of the methods; 120 groups had enrichments using two methods,and 143 groups had enrichments with only one method. Comparing the results from different methods we (Steve Meader) find mouse-ortholog mutant phenotypes (MGI) gave the fewest significant results with significant associations in only 12 patient-phenotype groups (Table B.7 in Appendix B). However, the enriched terms were most obviously relevant to the respective human phenotype, eg. patients with *Seizures* were enriched in genes whose mouse orthologue when disrupted results in *Absence seizures* (MP:0003216; 6.2-fold enrichment; $p = 3 \times 10^{-4}$) and patients with *Intellectual Disability, Severe* were enriched in genes resulting in synaptic phenotypes in mouse such as *Abnormal synaptic transmission* (MP:0003635; 3.3-fold enrichment; $p = 2.0 \times 10^{-5}$). Other methods resulted in a larger number of significant associations but the magnitude of the enrichment was smaller and the gene categories involved were typically less specific and more difficult to relate to the human phenotype (Tables

B.3, B.4, B.5 in Appendix B). GO term enrichments were identified for 121 human phenotypes (Table B.3), KEGG enrichments for 189 phenotypes (Table B.4), and significant clustering in the BrainSpan network for 262 phenotypes (Table B.5), In all, 186 of the 408 phenotypes had significant functional association using more than one method. In 177 phenotypes (95%) the different methods identified the same genes, with a range of 1-355 gene identified using multiple methods per phenotype (Figure 3.4).

We (Steve Meader) confirmed that the functional enrichments described above represent molecular networks using protein-protein interactions (PPIs). Since many protein functions require direct interactions between different proteins (eg. G-protein coupled receptors), we expect genes participating in the same biological pathway to be more likely to produce proteins who physically interact. We (Steve Meader) counted the number of PPIs in the DAPPLE PPI network (125) which were between genes in each of the 177 sets of genes contributing to significant functional associations with a respective human phenotype using more than one method. Sixty-five (37%) of the phenotype-specific candidate gene sets were significantly clustered within the PPI network after controlling for the total number of interactions recorded for each gene (Figure 3.4, Table B.6 in Appendix B).

Furthermore, we (Avigail Agam, randomizations for empirical p-values by Tallulah Andrews) attempted to replicate the association between the 65 specific phenotypes and their respective molecular pathways identified using the PPI network in an independent cohort of patients with developmental disorders. Patients with *de novo* CNVs were extracted from the DatabasE of Genomic variants and Phenotype in Humans Using Ensembl Resources (DECIPHER, (17)). Unlike the GENCODYS cohort, these patients have not been systematically phenotyped and their phenotypes are described using the London Dysmorphology Database (LDDB) dysmophology terms. Fifteen of the 65 respective HPO terms could be mapped to synonymous LDDB terms (Table 3.1) using the mappings provided on the HPO website (`http://compbio.charite.de/svn/hpo/trunk/src/mappings/`). Of these only *Microcephaly* was associated with

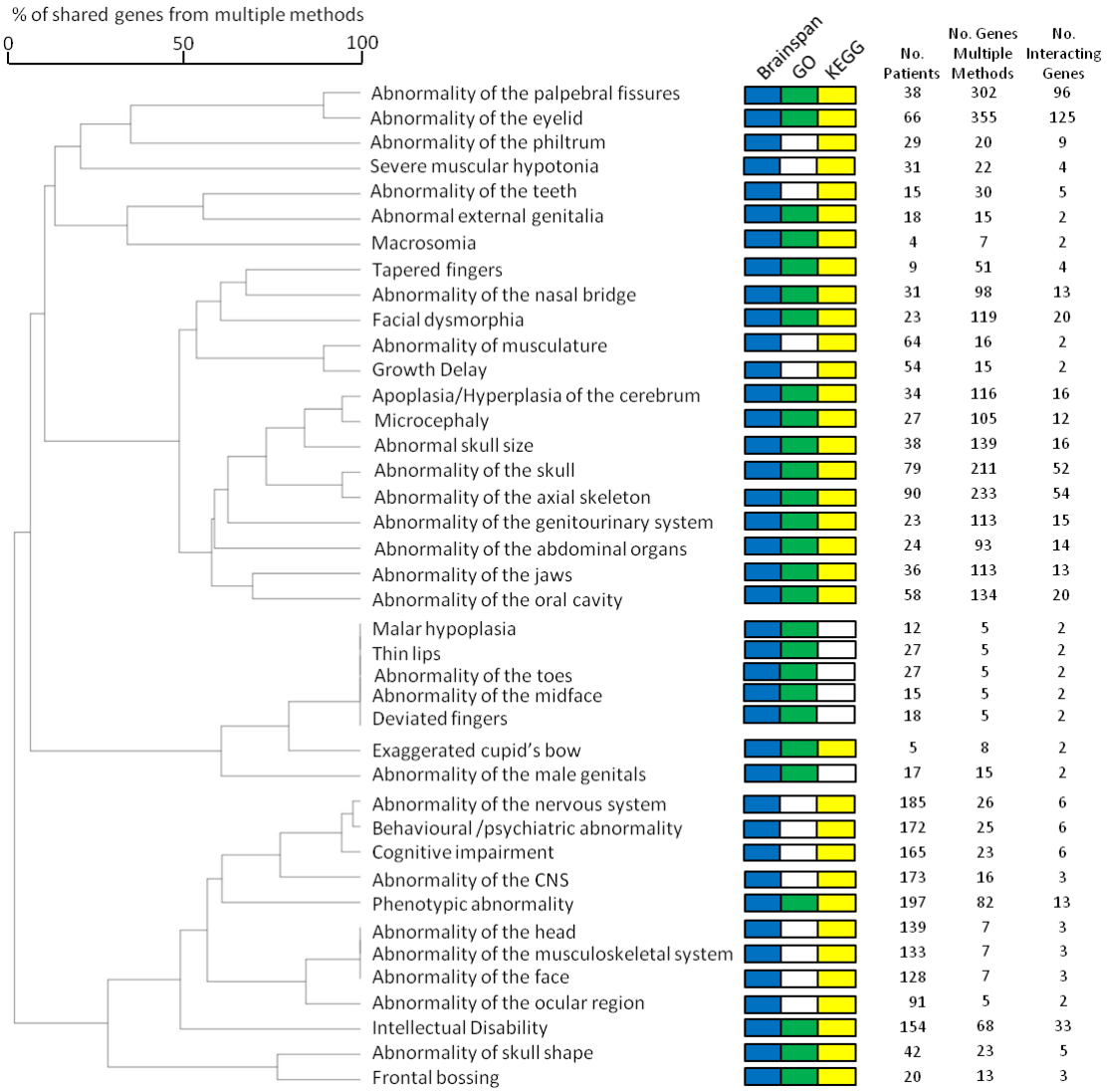


Figure 3.4: Forty non-exclusive phenotype-patient groups, each group's patients share the same HPO term, amongst whom individual copy number variant genes were identified by multiple functional genomics methods and whose recurrently-identified candidate genes possess a significant number of protein-protein interactions (PPI). The dendrogram displays the relationships between phenotype-categories based upon the number of candidate genes identified by multiple methods that are shared between the phenotype-group patients. Categories are marked if there were significant enrichments using clustering in a gene expression network (Blue), GO (Green) or KEGG (yellow), none were associated with significant enrichments of mouse phenotypes (MGI). (Created by: Steve Meader, reprinted with permission)

a sufficient number of genes in both the GENCODYS PPI network and the DECIPHER patient's *de novo* CNVs to test for interactions. Thirteen of the 71 patients in the DECIPHER Microcephaly cohort possessed *de novo* CNVs affecting genes in the original GENCODYS PPI and an additional 30 possessed CNVs affecting genes with a direct interaction with the GENCODYS PPI. The genes affected by CNVs in the DECIPHER Microcephaly cohort were significantly more connect to the GENCODYS Microcephaly

PPI network than expected by chance even after removing the 13 patients with CNVs affecting genes in the GENCODYS PPI ($p = 0.04$, Figure 3.5). Importantly, the GENCODYS Microcephaly network is split into four disconnected parts but the DECIPHER Microcephaly genes connect these four parts together into a single coherent pathway (Figure 3.5 B). In all, this coherent Microcephaly network is disrupted in 14/27 (52%) of GENCODYS patients and 43/71 (60%) of DECIPHER patients, thus potentially explaining the phenotype of 57/98 (58%) of all microcephaly patients considered.

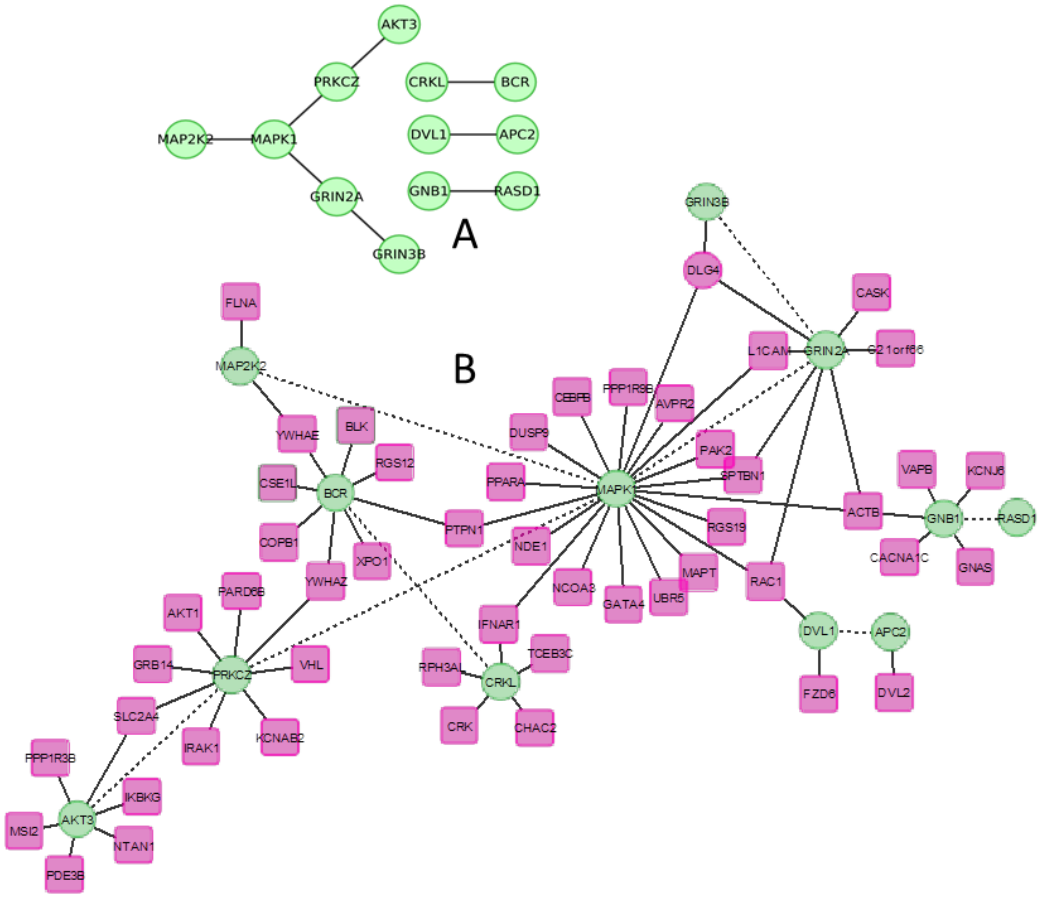


Figure 3.5: Molecular pathways identified among patients with *Microcephaly* in two large cohorts. (A) 12 copy variant genes drawn from 14 of 27 GENCODYS patients with *Microcephaly* that were identified using multiple functional resources (KEGG, GO, and BrainSpan co-expression) and cluster strongly ($p = 0.04$) in the DAPPLE protein-protein interaction (PPI) network. (B) Genes (n=51; Red)) that were copy number variant in 30 of 71 DECIPHER patients with *Microcephaly* were found to possess a significant number of interactions with the genes from panel A (Green) ($p = 0.04$), forming a single coherent microcephaly PPI network. (Created by: Avigail Agam, reprinted with permission)

3.3.3 Phenotypic similarity between patients contributing to enrichments

Developmental syndromes, such as Prader-Willi syndrome, are frequently identified based upon a set of shared phenotypic characteristics; for Prader-willi syndrome these include hypotonia, obesity, short stature, hypogonadism and small hands and feet (OMIM #176270). The large numbers of phenotypic expressions may reflect the pleiotropic effects of a recurrently mutated gene (144; 145; 146; 147; 148). Beyond a single gene or a single locus, if the functional enrichments we have identified above represent shared biological pathways repeatedly disrupted in the respective patients then disruption of the same pathway could result in similar pleiotropic outcomes. To test this hypothesis, I subdivided each patient-phenotype group by each of the significant functional enrichments identified within it. Patients with the specific phenotype and a CNV affecting genes with the significantly associated functional annotation (or within the significant PPI or co-expression network) were termed 'contributing patients' and those patients whose CNV genes do not contribute were termed 'non-contributing patients'. For each significant functional association I compared the distribution of pairwise phenotypic similarity amongst contributing patients to the pairwise similarity between contributing and non-contributing patients. Phenotypic similarity is a fair test in this cohort because of the consistent and standardized phenotyping procedure applied (18). I calculated phenotypic similarity using a simplified version of the Goodall3 index (110), which is based on calculating the probability of seeing the same or more similarity just by chance given the frequencies of each characteristic in the cohort (see: Chapter 2 Section 2.2.1 for details on this metric).

Comparison of different functional resources

Considering the four different sources of functional enrichment (GO, KEGG, MGI and BrainSpan), nominal significance ($p < 0.05$) was reached in ten of the 106 instances where we had sufficient power to conduct a test (at least ten contributing and ten non-contributing patients), which was significantly more than expected ($p = 4 \times 10^{-4}$,

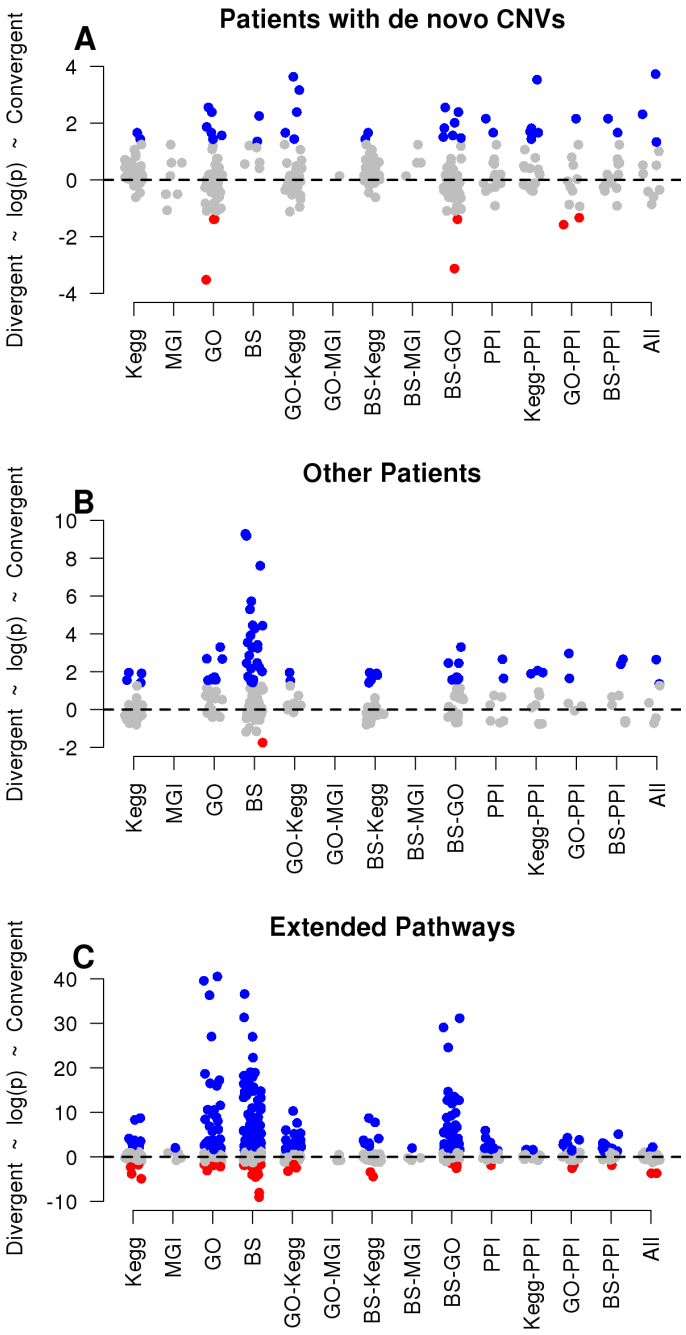


Figure 3.6: Patients contributing to the same functional association are phenotypically similar ($p < 1 \times 10^{-4}$). The the significance of the phenotypic similarity amongst patients contributing to a functional association was compared to non-contributing patients in the phenotype group. Each point represents a single significant functional association, grouped by enrichment method: KEGG, MGI (mouse phenotypes), GO, BS (BrainSpan co-expression), and genes identified by more than one method (e.g. GO-KEGG). PPI is the protein-protein interactions (Figure 3.4). Dots coloured blue or red indicate nominally significant phenotypic similarity (Convergent) or dissimilarity (Divergent) respectively (two-sided Wilcox test, $p < 0.05$). (A) Using the 197 patients with small *de novo* CNVs, (B) Using the 3,871 patients without *de novo* CNVs and the same pathway genes as in A, (C) Using the patients without *de novo* CNVs and only extended pathway genes, other genes with the same functional annotation as the initial pathway (GO, KEGG, MGI) or with a direct edge to the genes belonging to the initial pathway (PPI, BS).

one-sided binomial test, Figure 3.6 A). However, there was high variability between functional genomics resources and between patient-phenotype groups with even a couple functional enrichments having nominally significant trends in the opposite direction. Only pathways identified using the BrainSpan co-expression data identified subsets of patients that were consistently more similar to each other as compared to patients without genes in the inferred pathway. BrainSpan enrichments consistently shows phenotypic convergence when the correlation threshold was varied from 0.6-0.8 (Figure 3.3). I also compared the extent of phenotypic convergence seen for the

BrainSpan enrichments with those from a body-wide co-expression network derived from the GTEx data ((124), Figure 3.2). Despite a significant correlation between the two datasets ($r = 0.14$, $p < 1 \times 10^{-10}$), GTEx associations resulted in less phenotypic convergence than BrainSpan. However, only a small number of BrainSpan functional associations could be tested due to a lack of patients without contributing genes; the six I was able test had fewer than 15 patients without genes participating in the significantly co-expressed cluster (Figure 3.3).

To address the issue of insufficient patients in either group, I replicated the phenotypic similarity analysis using the remaining 3,871 patients without *de novo* CNVs. While I expected many of the inherited CNVs to be benign, I reasoned if the pathways are responsible for the patient phenotypes then they should also identify those pathogenic CNVs among all the remaining CNVs. Employing only those candidate genes identified in the 197 *de novo* CNVs for each functional association, I was able to test 66 of the 262 significant BrainSpan functional associations of which 25 exhibited nominally significant phenotypic convergence among patients possessing CNVs in contributing genes ($p = 3.4 \times 10^{-23}$) and 15 of which were significant after applying a Bonferroni correction for multiple tests (Figure 3.6 B). While for GO and KEGG, respectively, only 8/29 (28%, $p = 0.49$) and 4/30 (13%, $p = 0.020$) of the functional enrichments with sufficient power to be tested resulted in nominally significant phenotypic convergence (38%, $p = 0.025$ for BrainSpan). I further replicated the phenotypic analysis after extending the pathways, by including all genes with the significantly enriched annotation or all genes with a direct link to the associated expression or PPI network. To ensure this was a truly independent replication all patients with CNVs affecting genes in the original pathway were excluded(Figure 3.6 C). Once again BrainSpan networks exhibited a large excess of nominally significant phenotypic convergence (88/188, $p = 1.3 \times 10^{-87}$). However, the increase in power revealed that GO enrichments also frequently exhibited nominally significant phenotypic convergence(47/119, $p = 3.0 \times 10^{-43}$).

Importantly, combining information from more than one resource did not increase the proportion of instances where contributing patients were more similar to each other than to non-contributing patients (Figure 3.6). Considering only those candidate genes identified by two different functional resources in patients with *de novo* CNVs, 14/128 (11%, $p = 0.48$) of pathways identified using multiple resources and 10/63(16%, $p = 0.15$) of significant protein-protein interaction pathways (in any combination with other resources) showed nominally significant phenotypic convergence compared to 10/106 (9%, $p = 0.26$) of single-resource pathways. This was also true when the phenotypic analysis was replicated in the 3,871 patients without *de novo* CNVs whether considering only the original candidate (Figure 3.6 B; one method = 37/125 30%, two methods = 15/72 21%, PPI = 9/31 29%, $p > 0.1$) or considering the extended pathways (Figure 3.6 C; one method = 132/373 35%, two methods = 65/209 31%, PPI = 39/97 40%, $p > 0.1$). Two plausible explanations for the failure of intersecting multiple methods to identify the most phenotypically informative pathways are: (i) all the methods (except Brainspan) suffer from ascertainment bias so intersecting them simply selects for the most well studied genes, and/or (ii) intersecting multiple functional associations reduces the number of genes in the pathway thus the number of contributing patients and thus the power of the phenotypic similarity test.

Replication of Specific Signals

While the overall patterns of phenotypic convergence replicated well in the inherited and unknown CNVs using both the pathway genes from the *de novo* CNVs and the genes in the extended pathways, individual signals of phenotypic convergence were more inconsistent (Figure 3.7). Of the 27 pathway-phenotype associations, which showed nominally significant phenotypic convergence in the initial 197 patients with *de novo* CNVs, and that I had sufficient power to test in all three analyses, 18 (67%) showed nominal significance it at least one of the replication tests, but only three (11%) reached nominal significance in both replications. In addition, the three doubly-replicating signals all involved the PPI network associated with *Abnormality of the eyelid*, and thus represent very similar sets of genes.

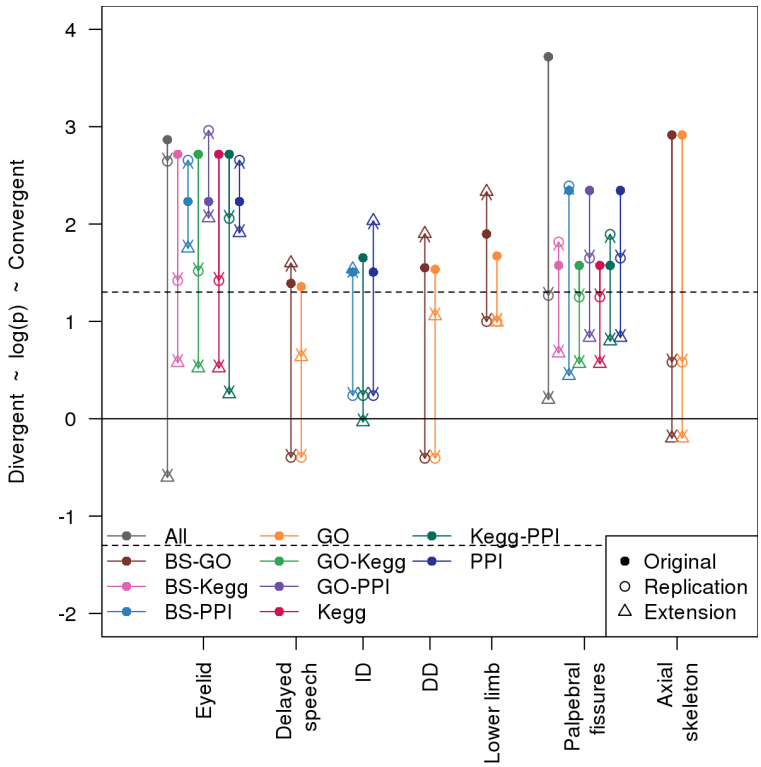


Figure 3.7: Replication of those pathways showing nominally significant phenotypic convergence in the patients with small *de novo* CNVs (Figure 3.6 A). Results are grouped by phenotype (condensed names appear along x-axis) and different ways of identifying pathways are colour coded. Original (solid dots) refers to results from considering only patients with small *de novo* CNVs, Replication (open dots) refers to the same pathway genes that were identified in the *de novo* CNVs tested in the 3,871 patients without *de novo* CNVs, Extension (open triangles) refers to testing the extended pathways in the patients without *de novo* CNVs (see: Methods 3.2.3)

Co-morbidity vs specific phenotypes

The next logical question to address is whether this observed phenotypic convergence is due to the pathway delineating a subset of patients exhibiting specific sub-phenotype(s) of the phenotype significantly associated with the pathway or whether the patients with mutations in the pathway exhibit similar patterns of co-morbidity of different phenotypes. I considered those significant functional associations giving rise to nominally significant phenotypic convergence ($p < 0.05$, two-sided Wilcox-rank-sum test; Figure 3.6, blue points) and repeated the phenotypic convergence analysis considering only those child phenotypes (terms found below the associated-phenotype in the HPO). As with the original phenotypic-similarity analysis this was replicated using the 197 patients with small *de novo* CNVs (Figure 3.8 A), the 3,875 patients without *de novo* CNVs (Figure 3.8 B), and considering just the extension of the original pathways in the patients without *de novo* CNVs (Figure 3.9). In a large majority of tests (62%)

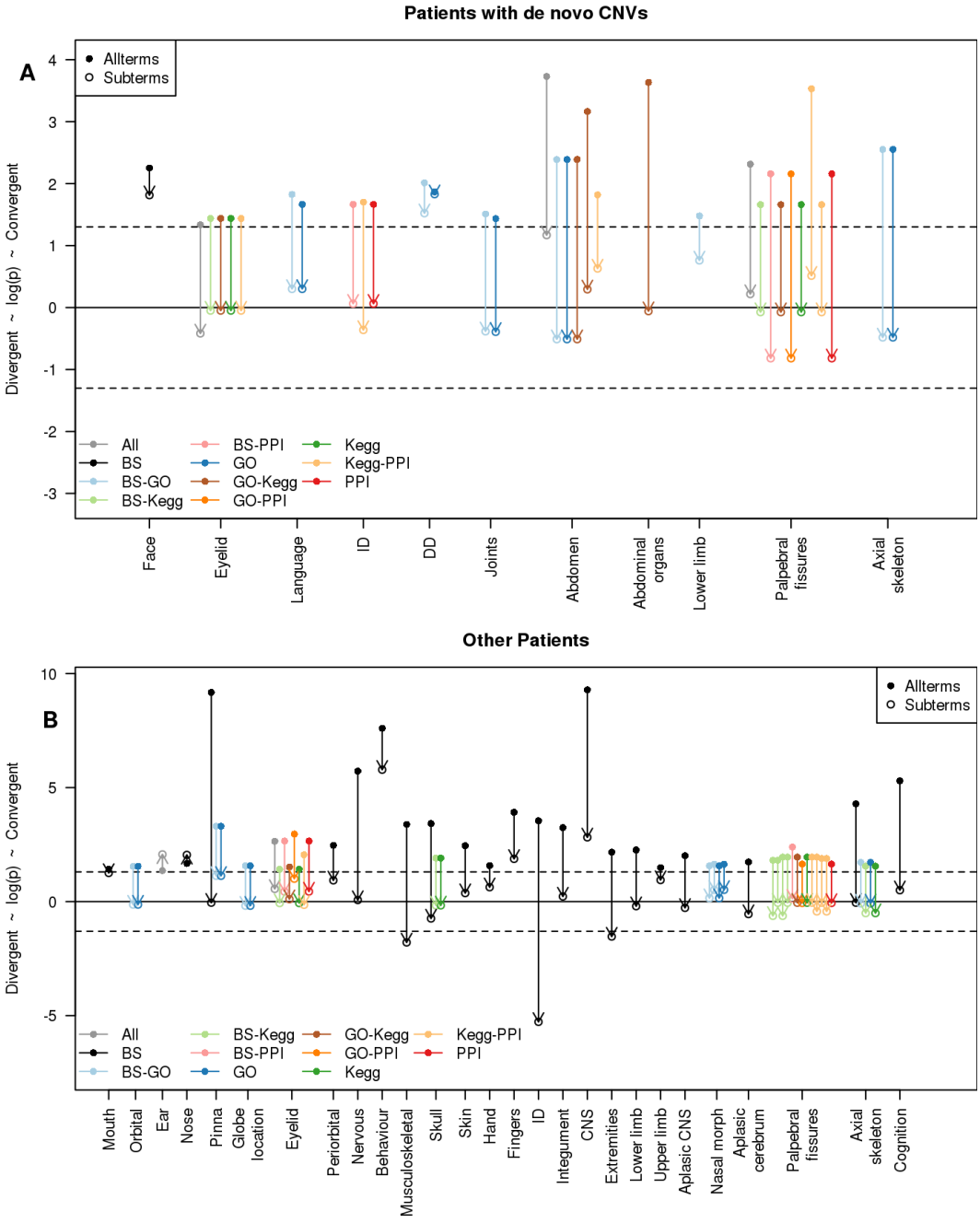


Figure 3.8: The significant phenotypic similarities amongst patients who contribute to the same functional association are not derived from these patients presenting more specific subphenotypes of the original phenotype. The Y-axis gives the significance of the overall phenotypic similarity amongst patients within a patient-phenotype group whose variant genes contribute to a functional association as compared to those patients in the same phenotype group who do not contribute (Convergent indicates contributing patients were more similar to each other than others, Divergent is the reverse). For all nominally significant enrichments (solid points) we recalculated the patient phenotypic similarities considering only child terms of the original HPO phenotype (open points connected to their respective solid point by an arrow). Points are grouped horizontally by HPO and coloured by enrichment-type. Solid line: p = 0.5, dashed line: p = 0.05. (A) Considering the 197 patients with small *de novo* CNVs, (B) Considering the 3,871 patients without *de novo* CNVs using the same pathway genes as in A. (Results for extended pathway can be found in Figure 3.9).

the distribution of child phenotypes was indistinguishable between patients contributing to the pathway enrichment and non-contributing patients, demonstrating that the phenotypic convergence between patients whose variant genes contribute to the same functional association was produced by the sharing of co-morbid phenotypes distinct from the associated phenotype. This may in part be due to insufficiently detailed phenotypes being available (eg. *Abnormality of the palpebral fissues* has only 8 child phenotypes annotated to patients in the cohort), but even relatively well described phenotypes (eg. *Abnormality of the axial skeleton* with over 50 child phenotypes present in the cohort) did not exhibit phenotypic convergence when only the child phenotypes were considered (Table B.8 in Appdendix B).

The few phenotypes where the child phenotypes show similar levels of phenotypic convergence as when all terms were included were relatively general terms associated with the head and face, the nervous system, and behaviour (Figure 3.8 3.9). Terms such as *Abnormality of the head, mouth, ear* or *nose* were found to have similar levels of phenotypic convergence when considering all terms or just their respective child phenotypes replicating the known importance of cranio-facial features in diagnosing patients with developmental syndromes (3; 149). However, terms describing the eye, such as *Abnormality of the orbital region*, *eye*, or *eyelid*, showed minimal phenotypic convergence among their child phenotypes despite having 11, 166, and 25 examined child phenotypes within the full cohort respectively, suggesting that detailed eye phenotypes are not as relevant to the underlying biological processes as other facial features. However, this could also be a result of difficulty in diagnosing or describing eye-related phenotypes compared to other facial features.

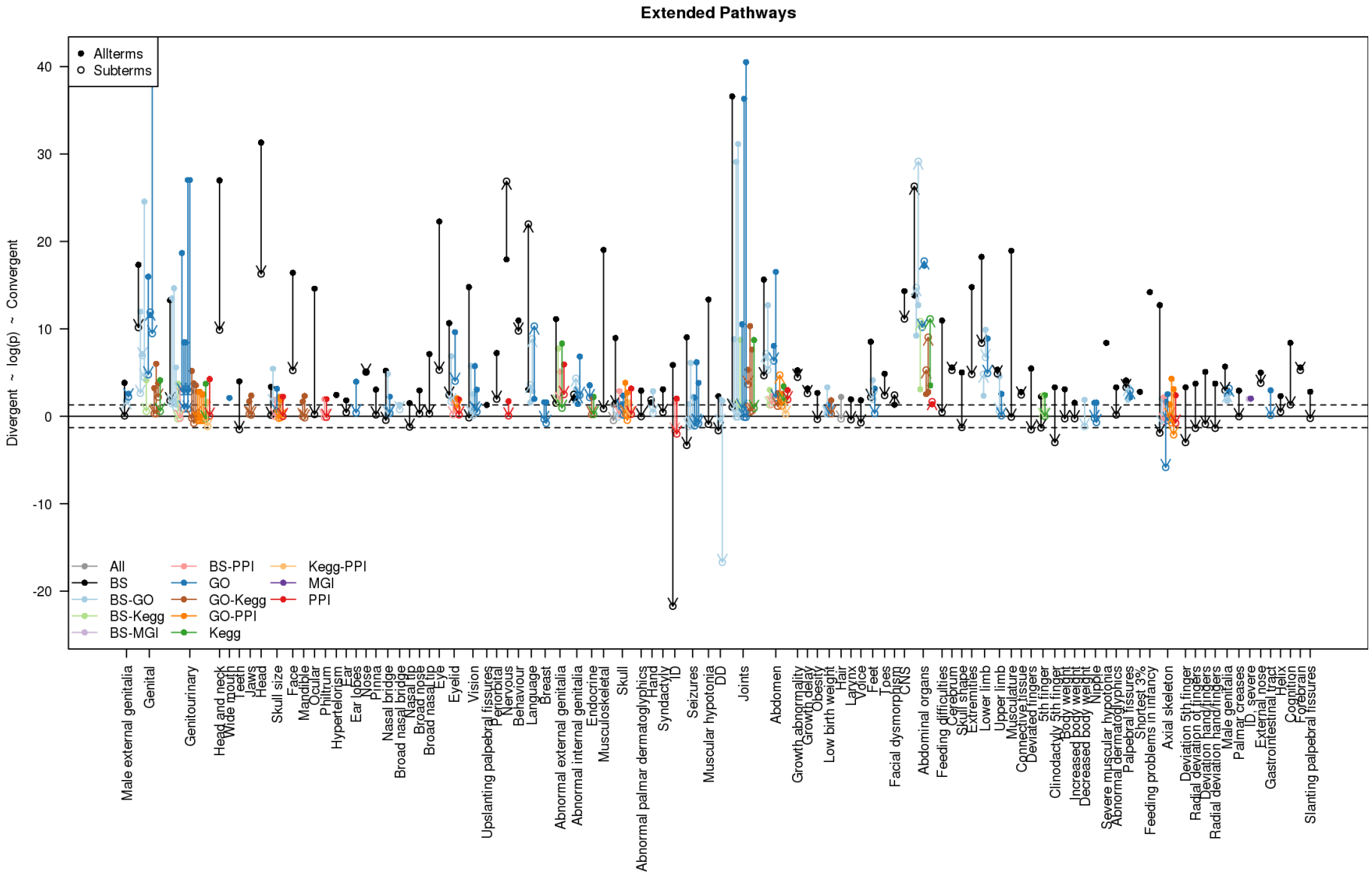


Figure 3.9: This is the same as 3.8 but for the extended pathways in patients without *de novo* CNVs. Y-axis is the significance of convergence/divergence of phenotypes. For all nominally significant enrichments (solid points) we recalculated the patient phenotypic similarities considering only child terms of the original HPO phenotype (open points connected to their respective solid point by an arrow). Points are grouped horizontally by HPO and coloured by enrichment-type. Solid line: p = 0.5, dashed line: p = 0.05.

3.3.4 Errors and Biases in Pathway Approaches

The occurrence of CNVs in the genome is significantly biased towards locations with long, low-copy number repeats (104). Many of the KEGG and GO annotations were significantly associated with multiple human phenotypes, and those annotations associated with more than 30 different phenotypes did not exhibit any phenotypic convergence (Figure 3.10 A). This suggested the most frequently associated annotations may be false positives resulting from the use of the whole genome as the background for the initial functional enrichment tests (see: Specific Methods 3.2). Thus I replicated the functional enrichments using a background of just the genes affected by at least one of the small (<5Mb) *de novo* CNVs. There was a high Spearman correlation ($\rho = 0.71$) between the significance of the enrichment against the whole genome and against only *de novo* CNV genes (Figure 3.10 B). Furthermore, 94% of the whole genome significant functional enrichments reached at least nominal significance ($p < 0.05$) when testing against only the genes found in *de novo* CNVs. Finally I found no correlation between the significance of the enrichment against the *de novo* CNVs background and the significance of the observed phenotypic convergence (Figure 3.10 C). Thus the biased occurrence of CNVs in the genome does little to explain the lack of phenotypic convergence for some functional annotations.

In addition to frequently enriched terms, KEGG terms related to highly studied processes (eg. Glycolysis, and Cancer) did not exhibit phenotypic convergence (Figure 3.10 A). Since GO, KEGG and MGI terms are frequently derived from small scale experiments they may suffer from research bias, the tendency to study the same set of genes/proteins or functions for which there is already existing information (140). Thus some terms may give significant results simply because the larger number of genes known to participate in the pathway increases the power to detect those terms whereas other more relevant pathways may not have been studied as extensively. Thus I considered the average number of papers associated with each gene contributing to each of the observed functional enrichments in Pubmed (from `ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/gene2pubmed.gz`) and compared this to the significance of

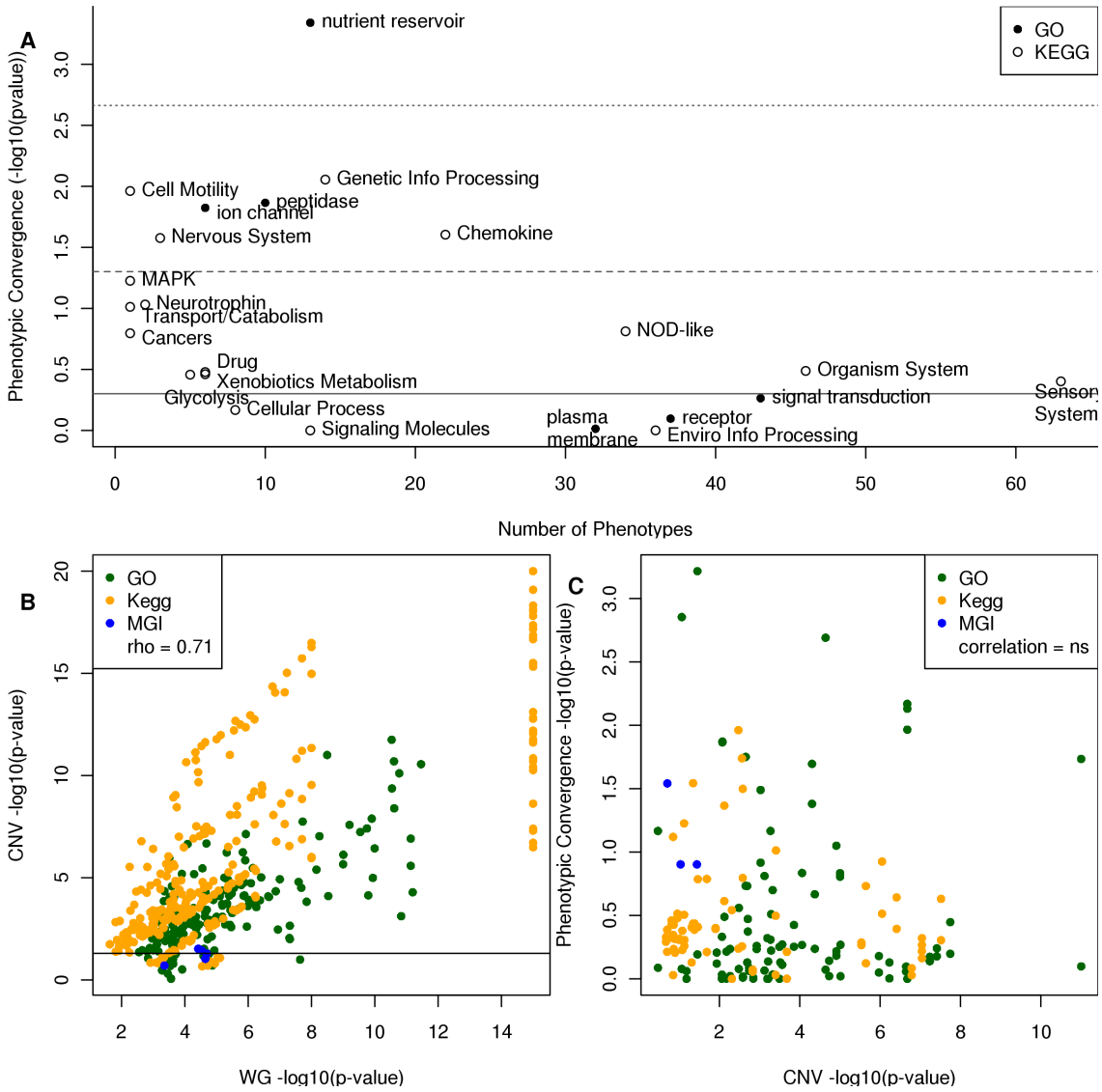


Figure 3.10: Systematic biases in CNV occurrence and functional enrichments. (A) GO or KEGG pathways significantly enriched in more than 30 patient phenotype groups do not show phenotypic convergence. Patient-phenotype groups associated with the same KEGG pathway or GO term were combined and for each association the phenotypic similarity amongst those patients whose variant genes contributed to the given association were compared to those who did not contribute. Y-axis is the observed significance of the phenotypic convergence (one-sided Wilcox rank-sum test). solid line is at $p = 0.5$, dashed line is at $p = 0.05$, and dotted line indicates significance after a Bonferroni correction. (B) Replication of the enrichments identified using the whole genome as a background (WG) when just genes found in *de novo* CNVs in our cohort are used as the background (CNV). Solid line indicates $p = 0.05$. (C) There was no significant relationship between how well the functional enrichment replicated using the CNV background and the significance of the phenotypic convergence observed.

the phenotypic convergence observed amongst contributing patients (Figure 3.11 A). There was no significant correlation between the number of associated papers and the phenotypic convergence. However, different resources exhibited different amounts of research bias, KEGG and MGI functional enrichments were due to much more highly studied genes (average 122 and 148 associated papers/gene respectively) than GO functional enrichments (average 38 associated papers, $p < 0.002$ two-sided t-test). Furthermore, the PPI networks built from genes significantly associated using more than one method contained even more highly studied genes (278 associated papers/gene on average, $p < 0.0008$). As noted above, intersecting multiple methods and adding the PPI information did not increase the proportion of pathways exhibiting phenotypic convergence, a possible explanation of this result is that intersecting the results from multiple methods simply selects for the most studied genes rather than identifying a more cohesive pathway.

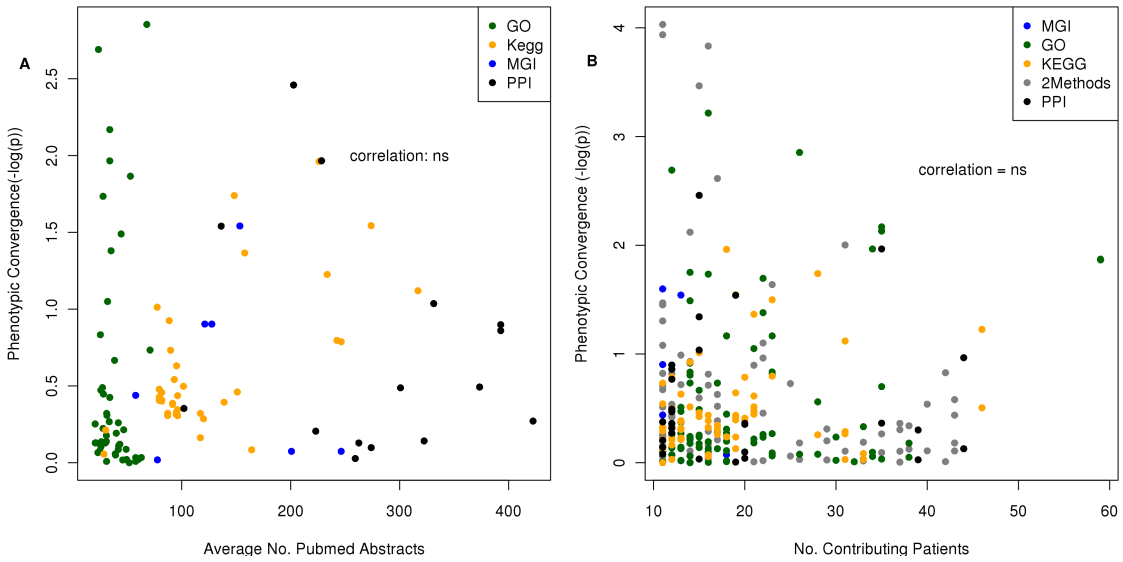


Figure 3.11: Ascertainment bias and power of phenotypic convergence test. (A) Ascertainment bias was examined using the average number of Pubmed abstracts associated with the gene. (B) Power was examined by determining the number of patients contributing to the functional enrichment. No significant correlation (Pearson or Spearman) was identified between either ascertainment bias or power and the significance of phenotypic convergence.

Another issue which could affect the paucity of functional enrichments showing phenotypic convergence is a lack of power. Only those functional enrichments with at least 10 contributing and 10 non-conbributing patients were tested to try and ensure suffi-

cient power but this threshold may have been too low. In addition, intersecting multiple methods could result in fewer contributing patients, thus decreasing the power to detect phenotypic convergence. I investigated this by considering the number of patients contributing to the functional enrichment compared to the significance of observed phenotypic convergence (Figure 3.11 B). Note that unlike GO, KEGG, and MGI, BrainSpan functional associations tended to include genes from nearly all patients thus they were excluded from this analysis. There was no correlation between the number of patients contributing to the functional enrichment and the observed phenotypic convergence. The intersections of different annotations or the significant PPI networks did not have particularly small numbers of contributing patients(average of 21 contributing patients vs 23 patients for single resource enrichements). Thus while the number of patients with or without *de novo* CNVs in contributing genes severely limited the number of enrichments that could be tested (Figure 3.6) the applied threshold of at least 10 of each was sufficient to ensure a well powered test for phenotypic convergence.

A final bias that could result in false positive functional associations is the non-random organization of genes in the genome. Paralogous genes tend to be located in close proximity in the genome as a result of tandem duplications. In addition the clustering of functionally related genes which are not paralogs has been repeatedly observed (84; 86). The clustering of functionally related genes in the genome could result in a single *de novo* CNV contributing many genes to a single functional association if it occurs at one of the clusters, which would inflate the significance of that functional association and introduce spurious results. At least 25% of patients possessed four or more genes contributing to the same KEGG, GO, or MGI functional enrichment and/or at least two genes to a respective PPI network (Figure 3.12 A). However, BrainSpan enrichments were far more extreme with 50% of contributing patients possessing at least 9 significantly co-expressed genes. To determine whether this clustering of functional genes was likely to be resulting in spurious functional associations, I compared the spread of genes amongst contributing patients (as measured by the Shannon Diversity Index (142)) to the significance of observed phenotypic convergence (Figure 3.12

B). There was a significant positive correlation between the degree of spread of contributing genes amongst patients and the extent of phenotypic convergence ($r = 0.34$, $p = 1.6 \times 10^{-4}$). Functional enrichments where a small number of patients possessed CNVs in a large number of contributing genes showed much less significant phenotypic convergence than enrichments where contributing genes were more evenly spread among patients. Tight BrainSpan co-expression associations had both the most consistent phenotypic convergence as well as the most patient diversity resulting from the large number of genes involved in these associations, however if they are excluded from the analysis there remains a significant positive correlation between the spread of contributing genes and phenotypic convergence ($r = 0.22$, $p = 0.02$). Thus the clustering of functionally related genes due to paralogs resulting from tandem duplication or the clustering of unrelated functionally-similar genes may be causing spurious functional associations (I will investigate this phenomenon further in Chapter 4).

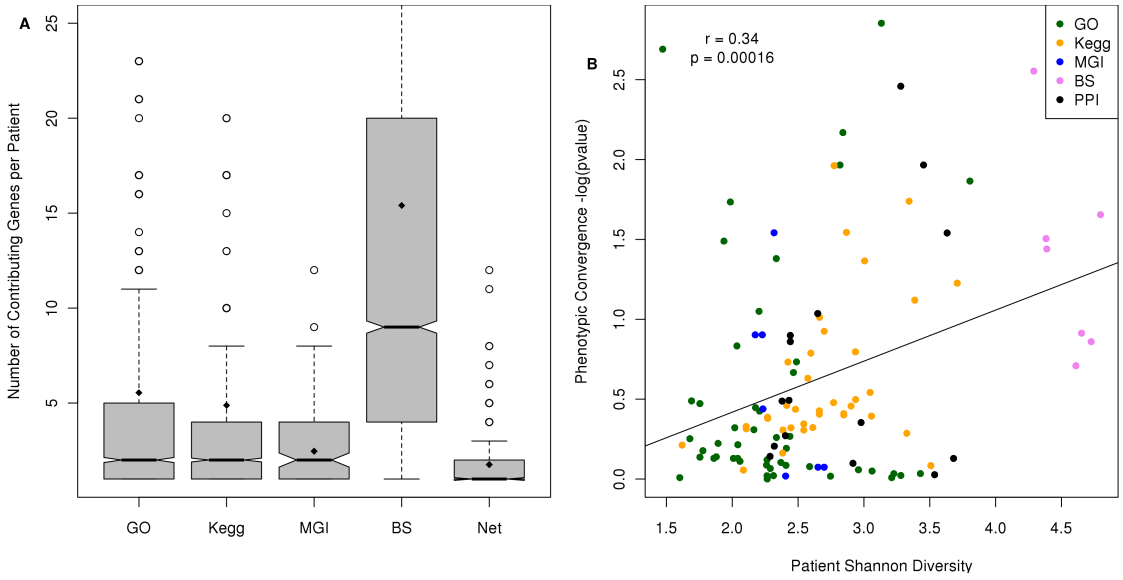


Figure 3.12: Multiple functionally-similar within the same patient may contribute to false positive functional enrichments. (A) Distribution of number of genes within each functional pathway found in each contributing patient for each resource. In most cases at least 25% of the time a single patient is contributing at least 4 genes to a single enrichment. (B) Shannon diversity of the distribution of pathway genes among patients (Patient Shannon Diversity) was significantly positively correlated with the significance of phenotypic convergence.

3.3.5 Important Phenotypes

The next thing I investigated was which phenotypes were important for identifying patients with mutations in particular pathways. This was done by reversing the typical functional enrichment methodology. Instead of testing the genes affected by mutations in patients with each phenotype for an enrichment in of genes associated with a pathway, I tested each group of patients with mutations in genes belonging to a particular pathway for an enrichment of each phenotype. This analysis was performed on just the 197 patients with *de novo* CNVs, all patients, and using the extended pathways with all patients. Eighty-seven (out of 1,868) different phenotypes were significantly enriched among patients with mutations in genes annotated with one of the pathways and extended pathways we identified above in all three versions of the analysis(Figure 3.13, Table B.9 in Appendix B).

Despite only identifying functional enrichments in 186 patient-phenotype groups, using just the patients with *de novo* CNVs we found 313 phenotypes with significant enrichment (none were depletions) in patients with genes contributing to a particular pathway. This reflects the phenotypic convergence of comorbid phenotypes observed above but may also be partially explained by the lower number of tests performed resulting in a less harsh multiple testing correction. However, when I replicated the analysis using all 4,240 patients in the cohort for either the *de novo* CNV pathway genes or the extended pathways only a subset (223 and 91 respectively) of these phenotypes continued to be enriched in patients carrying mutations in particular pathways. Interestingly, only 14 additional phenotypes were significant when all 4,240 patients were considered that were not significant using just the patients with *de novo* CNVs despite the increase in power from the increased number of patients. This suggests there is much more phenotypic noise among patients with inherited CNVs. My results corroborate the previous finding that particularly heterogeneous developmental disorders tend to be associated with multiple inherited CNVs rather than a single *de novo* mutation (7).

Across the three versions of the analysis, none of the 87 consistently were significantly enriched among the patients with mutations in a single pathway, however many of the pathways overlapped because they describe similar or related processes (eg. the KEGG term "Signalling Molecules" and the GO term "signal transduction"). Thus, the specificity of the phenotype to the pathways, among whose patients the phenotype was significantly enriched, was determined by calculating the number of genes belonging to all the pathways the phenotype was associated with divided by the number of genes belonging to any of the pathways (intersection/union). Thirty of the phenotypes were consistently associated with multiple completely independent pathways (intersection/union = 0). Only seven phenotypes were consistently associated with overlapping pathways: *Upslanting palpebral fissures*, *Aplasia/Hypoplasia of the nails*, *Epicanthus*, *Abnormality of the toes*, *Cardiac malformation*, *Joint hypermobility*, and *Abnormality of body weight* (Table B.9). This agrees with the existing knowledge of large overlaps between phenotypic presentations between different disorders (150; 151).

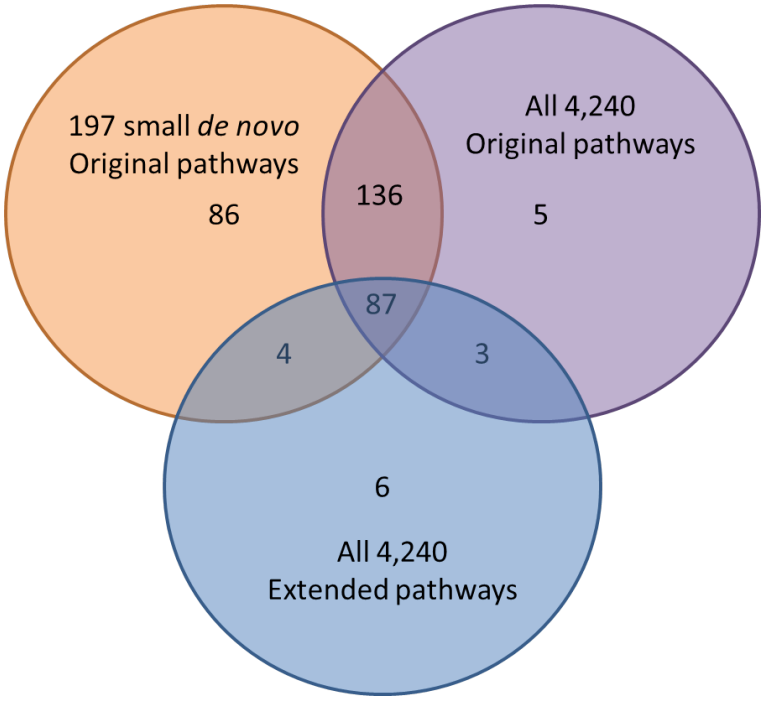


Figure 3.13: Phenotypes significantly enriched among patients contributing to one of the pathways using each of the three different versions: 197 patients with *de novo* CNVs, all 4,240 patients and the pathways identified in the *de novo* CNVs, and all 4,240 patients with the extended pathways (share all genes with the significantly associated term or with a direct link to one of the pathway genes).

3.4 Conclusion

In this chapter, we identified biological pathways enriched in groups of patients with each of 329 different phenotypes amongst 197 patients with developmental disorders and *de novo* CNVs (Tables B.5, B.3, B.4, B.7). Of these 186 had significant associations using more than one of the four different methods employed. Of the multiply identified pathways we identified significant protein-protein interactions amongst 65 pathways (Figure 3.4, Table B.6). Taking advantage of the systematic phenotyping performed on this cohort, I devised a novel method to identify phenotypic convergence amongst patients with copy-number variant genes within the pathway to further validate the phenotypic importance of the identified pathways. Using this method I showed that 39 % of the identified pathways (which could be tested) exhibited phenotypic convergence primarily a consequence of shared co-morbid phenotypes among patients with CNVs in the same pathway (Figure 3.8, 3.9). I examined various sources of error that could explain the many pathways which failed to exhibit phenotypic convergence, including: ascertainment bias, power, biases in CNV occurrence and clustering of functionally related genes. Ascertainment bias provided an explanation for the fact that combining multiple functional resources did not increase the proportion of pathways showing phenotypic convergence (Figure 3.11). In addition, clustering of functionally related genes explained many false positive pathway enrichments showing non-significant phenotypic convergence (Figure 3.12). Finally by inverting the functional enrichment analysis, I identified 87 phenotypes significantly enriched among patients with mutations in particular pathways (Figure 3.13).

Chapter 4

CNVs and Functional Clustering

4.1 Introduction

Several recent studies of copy number variants (CNVs; deletions or duplications > 1kb in size) have found individual CNVs often contain multiple functionally-related genes any one of which might be responsible for the associated disease phenotype (20; 22; 23). Boulding and Webber (20) and Dolken *et al* (22) each found multiple genes within single CNVs that when individually knocked out in a model organism (mouse and zebrafish respectively) cause phenotypes related to those observed in the patient in whom the CNV was observed. Furthermore, Golzio *et al* (23) identified genetic interactions between *KCTD13*, *MVP*, and *MAPK3*, all of which are present within the 16p11.2 CNV locus, that produce microcephaly or macrocephaly when their orthologs were simultaneously over- or under-expressed in zebrafish. Furthermore in Chapter 3 we found that the CNVs in individual patients often contained more than four genes associated with the same biological pathway. However, none of these studies have systematically examined the extent of this phenomenon and whether it is a result of chance due to the large size of many of these variants, largely a result of tandem gene duplications producing regions containing many paralogous genes, or represents a significant contribution to the pathogenic effects of the CNVs. This chapter will examine these three different hypotheses in the context of two large datasets of *de novo* CNVs

identified in patients with developmental disorders.

Genes and the proteins they produce interact in many different ways: proteins physically interact, regulate gene expression, modify the activity of other proteins, and catalyse sequential metabolic reactions. Many studies have found that non-paralogous genes encoding functionally-related proteins tend to be located close together (in terms of linear distance) in the genomes of humans(80; 86; 87; 88; 89; 90; 91; 152), mouse (79; 80), zebrafish(85), worm(84), fly (81; 82; 83), and yeast (76; 77; 78). Many of these studies have been based on gene expression, since it is available for nearly every gene, and have identified clusters of broadly expressed housekeeping genes (80; 81; 87; 93) and clusters of co-expressed/tissue-specific genes (78; 79; 81; 83; 85; 88; 90). However, other types of functional information have also been used to identify significant clustering of functionally-related genes in the genome (I will refer to this as 'functional clustering') including: protein-protein interactions (76; 91), Gene Ontology terms (86), KEGG pathways (89), and phenotypes exhibited from gene knock-downs (84). CNVs that affect one of these functional clusters could confer deleterious effects due to the compounding effects on the same biological process.

In this chapter I examine the presence of functional clusters in *de novo* CNVs found in patients with developmental disorders. I employ three different functional networks for defining functional similarity and control for the presence of tandem duplications. I show that (i) the *de novo* CNVs contain significantly large clusters of functionally-related genes, (ii) that the presence of functional clusters is robust to the network used (iii) that the clusters are not driven by co-expression (iv) that these clusters contain disease genes but are better indicators of CNV pathogenicity than known disease genes, (v) that functional clusters were not specific to any particular phenotype, rather that they are broadly associated with disease, and (vi) that the various clusters found in CNVs belonging to patients with hypotonia form interconnected networks with molecular functions plausibly connected to the disease phenotype.

4.2 Specific Methods

4.2.1 Copy Number Variants

For this chapter, I employed two large datasets of CNVs from patients with developmental disorders for which inheritance patterns for a large proportion of CNVs has been determined. I obtained 626 *de novo* CNVs and the respective patient phenotypes (described using London Dysmophology Database terms) from the Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources (DECIPHER, (17)). The second independent set of 426 *de novo* CNVs along with the respective patient phenotypes (described using the Human Phenotype Ontology) was obtained from the Genetic and Epigenetic Networks in COgnitive DYSfunction project (GENCODYS, (18)). The most common phenotypes amongst these cohorts were nervous system abnormalities (>90% of patients), intellectual disability (80%), Eye abnormalities (43%) and facial abnormalities (39%). I included only *de novo* CNVs since they are considered likely to be pathogenic(6; 12; 153). I also test the 61 known contiguous gene syndrome regions described in the DECIPHER database the pathogenicity of which has been well established (Table B.1).

I further filtered these CNV sets by removing all CNVs less than 100kb in length since many arrays have limited resolution leading to a high false positive rate among small CNVs (8; 25; 101; 139; 154), whereas CNVs >100kb are likely to pass validations (4). Furthermore, the pathogenicity of small *de novo* CNVs is less clear (155). In addition to removing small CNVs, I also excluded very large CNVs (>5Mb) since they may contain many extraneous genes which act as noise and can obscure the any signal within the cohort (99). After this filtering 427 and 237 *de novo* CNVs remained in the DECIPHER and GENCODYS cohorts respectively (See Materials and Methods, Table 2.1, 2.2).

Coordinates of CNVs were mapped to hg18 using LiftOver (102). Genes were mapped to these CNV regions using Ensembl build 54 gene models such that only genes for

which some exonic sequence from every transcript was within the CNV region were included, a method that has been shown to reduce length bias of mapped genes (106).

CNV Randomizations

The patient *de novo* CNVs contained significantly more genes than random sequences of equal length (Figure 4.3), recapitulating what has been seen in autistic patients (9). To our knowledge there are no datasets of *de novo* CNVs from healthy individuals, thus the genes affected by each *de novo* CNV were compared to 10,000 randomly chosen, equally-sized sets of genes that were contiguous on a single chromosomal arm, (termed "gene-number matched randomizations"), to determine the expected functional similarity between the genes affected by these CNVs. Genes not present in the relevant gene network were excluded and paralogs collapsed such that randomizations had the same number of genes remaining as the original CNV.

Collapsing Paralgous Genes

Human paralogs were identified using both Ensembl Compara (version 54, (156)) and the OPTIC database (157). Both use phylogenetic methods to identify paralogs, I selected all paralogs identified using zebrafish as the out-group. All paralogous relationships identified in either resource were included when identifying instances of paralogy. Within each gene set (a CNV or a gene-number-matched randomization) paralogous genes were collapsed such that the first member of the family encountered is retained and all other members are removed.

4.2.2 Finding Functional Clusters

Functional Gene Networks

We were interested in the functional similarity between genes found within each *de novo* CNV. Interactions between genes can be obtained or inferred from genome-wide

databases of known pathways & functions, expression patterns, protein-protein interaction (PPI) experiments, sequence information and knock-out phenotypes displayed by model organisms. Each of these data-types has errors and covers a subset of genes. We augmented an existing integrated functional network (65) with mouse phenotype information using the same method they describe. Briefly, the network combined many biological datasets (Table 2.5) by re-scoring them according to the regression of the dataset against the similarity of mouse phenotypes annotated to the 1-1 orthologs and then summed after weighting each dataset according to the strength of its relationship with phenotypic similarity. To combine the previous network with the mouse knockout phenotypes we calculated the semantic similarity between the Mammalian Phenotype Ontology (MPO, (107)) phenotypes annotated to each gene derived from the Mouse Genome Informatics database (36; 37). Term to term similarity was calculated using the average information content of the most informative disjoint common ancestors (108). The pairwise similarities between terms annotated to each pair of genes was combined by taking the average of the similarity between the single most similar pair of terms (maximum best-match) and the average of the similarity between all best-matching term pairs (average best-match) as described by (109). We then used a polynomial regression between the semantic similarity between genes based on the MPO terms from mouse models and their semantic similarity based on phenotype annotations in the human phenotype ontology (HPO, (35)) to re-score all genes with MPO annotations (regardless of whether they also have HPO annotations). This was combined with the already integrated molecular datasets by weighting each according the strength of their relationship with the semantic similarity between genes based on their HPO terms (65). We name this final integrated functional network the Phenotypic Linkage Network (PLN).

To ensure my results were not an artefact due to the particulars of the construction of the PLN, I replicated my results using two other networks: HumanNet (53), another publicly available integrated functional network, and COXPRESdb (46), a co-expression only network. However, I later focused on results obtained using the PLN

due to its greater coverage of genes compared to HumanNet (PLN: 17,039 genes; HumanNet: 16,243 genes; see Materials and Methods Table 2.6) and the improvement of integrated functional networks over co-expression or protein-protein interaction only networks when predicting gene function (55; 56; 59).

Clustering Algorithm

The PLN contained roughly 11 million direct edges connecting 17,039 genes which is on 7.4% of all possible pairwise-similarities. To increase the coverage we calculated the shortest-paths between genes through this network which gave a similarity value for 142,864,287 gene pairs (>98% of all possible pairwise comparison). Shortest-paths were calculated by converting original network similarity edges into distances using: $dist = 1/(1 + sim)$ and applying Dijkstra's shortest-path algorithm (68). The resulting shortest-paths were converted back to similarities using the inverse function: $sim = 1/dist - 1$ (shortest-path similarities). This method was applied to all three networks considered (Table 2.6). Clusters of functionally-related genes were identified using single-linkage hierarchical clustering using a height threshold equal to the top 1% shortest-paths in the network. Results were replicated using 5% and 0.1% shortest-paths thresholds.

4.2.3 Disease Genes

Known disease genes were obtained from the Online Mendelian Inheritance in Man (OMIM, (112)). Only OMIM disease genes classed as confirmed and where the molecular basis or mutation in the gene is known or where the gene is part of a known contiguous gene syndrome were considered known disease genes; these were mapped to 1,648 Ensembl genes. In addition, 297 curated human haplo-insufficient genes (HIS) were obtained from the literature (158). Significance of the enrichment of these disease genes in clusters of functionally-related genes (vs all CNV genes) was calculated using a one-sided hypergeometric test. The phenotypic consequences of HIS genes and OMIM disease genes were not recorded in rigorously defined terms so they could not

be easily compared to the patients' phenotypes as recorded in DECIPHER and GENCODYS. However, gene-phenotype annotations from the Human Phenotype Ontology (HPO, (35)) could be easily compared to the patients' phenotypes in GENCODYS, which were also recorded using HPO terms, and to the patients' phenotypes in DECIPHER, which were recorded using London Dysmorphology Database (LDDB) terms which could be mapped to HPO (see below). Each gene had all terms ancestral to the terms found in the HPO database assigned to them by imputing on the hierarchy of HPO terms. The patients' phenotypes were not imputed to avoid matches between distinct but related phenotypes (eg. epilepsy and autism). Significance of the enrichment of these phenotype-specific genes in clusters of functionally-related genes (vs all CNV genes) was calculated using a one-sided binomial test. In addition, a previous study found significant associations between LDDB phenotypes and mouse phenotypes (20). I took these associations and mouse knockout phenotypes from the Mouse Genome Informatics (MGI) database (36; 37) to identify candidate genes specific for each patient's phenotype in both datasets (HPO terms were mapped to LDDB for GENCODYS patients) enrichments were tested using a one-sided binomial test.

I performed a logistic regression to compare the predictive ability of functional clusters versus known disease genes in identifying pathogenic CNVs. Both apparently benign (from (103)) and *de novo* CNVs were filtered to remove CNVs >5Mb or <100kb in length. The presence/absence of a functional cluster (at least 2 functionally-related genes), at least one known disease gene, or at least one known HIS gene were each treated as a binary predictor (1 or 0) with or without also including the number of genes in the CNV or CNV length (in units of 100kb) as a predictor. The regression was performed using the `glm` function in the R statistical package (159).

Mapping Phenotypes

DECIPHER patient phenotypes were recorded using LDDB while GENCODYS patient phenotypes were recorded using HPO. I converted between these ontologies as needed using the file provided on the HPO website (35), `http://www.human-phenotype-ontology.`

org/contao/index.php/downloads.html. All provided mappings of any quality were considered. For those LDDB terms mapped to more than one HPO term the most general HPO term was used.

4.3 Results

4.3.1 Copy Number Variants

I investigated the extent and frequency of groups of multiple functionally-related genes within copy number variants (CNVs) from two independent cohorts of patients with developmental disorders (Tables 2.1, 2.2). Patients were diagnosed with heterogeneous developmental disorders with 1-79 distinct clinically assigned phenotypes such as autism, seizures, specific craniofacial malformations, cardiac or other morphological, physiological or behavioural abnormalities. 626 *de novo* CNVs were obtained from the Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources (DECIPHER, (17)). A second set of 426 *de novo* CNVs were obtained from collaborators at the Genetic and Epigenetic Networks in Cognitive Dysfunction Consortium (GENCODYS, (18)). In contrast with DECIPHER, GENCODYS CNVs were identified using lower resolution arrays (only 250K probes) but the same array and a standard diagnostic quality control criterion was used for all CNVs as well as a standardized phenotyping procedure.

Due to the differences in the density of array probes across the genome and different algorithms and/or quality control criteria used by DECIPHER and GENCODYS, GENCODYS CNVs tended to be larger than DECIPHER CNVs (Figure 4.1). I removed CNVs less than 100kb in size to reduce false positive calls resulting from the limited resolution of many arrays (8; 25; 101; 139; 154). In addition, I removed CNVs over 5Mb in size to reduce noise from the large number of extraneous genes in large CNVs. The remaining 427 and 237 CNVs from DECIPHER and GENCODYS respectively were much more comparable in terms of size and number of genes affected (Tables 2.1, 2.2).

However, GENCODYS CNVs still tend to contain more genes than DECIPHER CNVs this may be due to differences in the quality control procedures applied during the identification of the CNVs, or due to differences in the distribution of probes on the arrays used to identify CNVs. Major results were replicated using the full datasets to ensure robustness to this filtering. These datasets were considered separately due to differences in CNV calling/arrays as well as the differences in phenotyping. Both datasets show a predominance of loss CNVs (deletions) compared to gain CNVs (duplications or triplications) since losses are easier to detect and more deleterious than gains (4; 19; 104).

Differences in the number of phenotypes per patient (14.5 for DECIPHER, and 37 for GENCODYS) are largely due to differences in the specificity of the two different ontologies used. DECIPHER patient phenotypes were recorded using terms from the London Dysmorphology Database (LDDB) which are less specific than the Human Phenotype Ontology (HPO) used to record the patient phenotypes in GENCODYS.

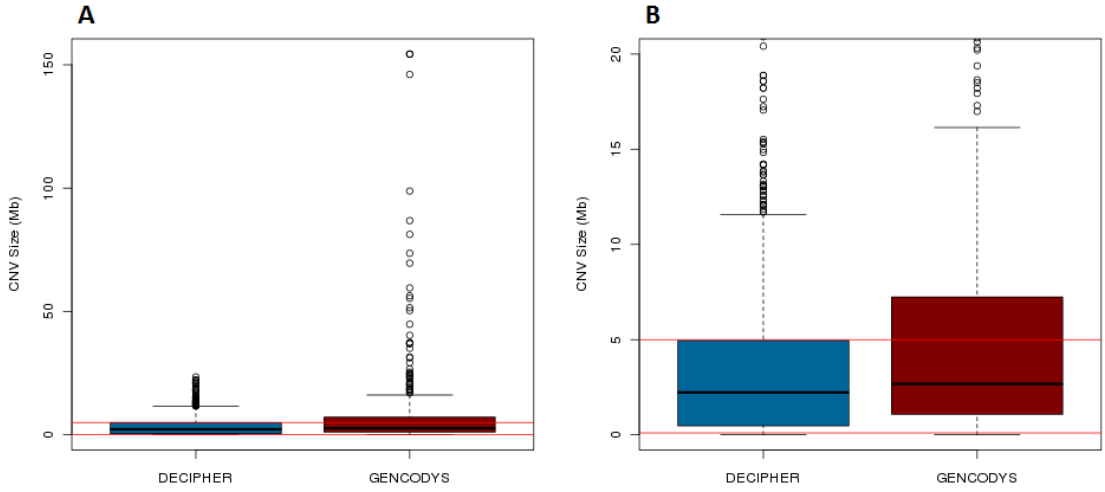


Figure 4.1: Overall GENCODYS *de novo* CNVs were larger than DECIPHER *de novo* CNVs. Red lines indicate 100kb and 5Mb used to filter the datasets. B is a zoomed in view of A.

4.3.2 Defining Functional Similarity

In order to investigate the presence of groups of functionally similar genes within individual CNVs, I assessed functional similarity between genes using an integrated network created by combining multiple large functional genomics datasets including physical interactions between proteins, correlations between RNA expression levels, molecular function annotations, and protein domains (Table 2.5). This network was created by treating each gene as a node and creating weighted edges, representing the functional similarity, between them. The functional similarity values for each edge were obtained by augmenting an existing integrated functional network (65) with information from mouse-knockout phenotypes. Combining many datasets in this way has been shown to outperform any single dataset on its own (56; 59; 160). In addition, it increases the number of genes which can be considered as each individual dataset has information for just a subset of genes. However, even after integrating all the datasets the resulting phenotypic linkage network (PLN) only contained 10,792,987 functional similarity values (direct edges) out of a possible 145,155,241 gene pair comparisons (<10%). Thus, to calculate the functional similarity for additional gene pairs I converted the similarities (s) into distances (d) as $d = 1/(1+s)$ and calculated the smallest sum of distances to travel, possibly by way of additional genes, from one gene to another (shortest-paths). This sum was converted back into a similarity inverting the equation $s = 1/d - 1$. As a result of this calculation, I determined a similarity value for 142,864,287 gene pairs; the remaining missing values result from a few genes being disconnected from the rest of the network thus no paths were possible.

To ensure my findings were not particular to the datasets or methods used to define functional similarity between genes I used the same procedure on two other publicly available gene networks. Applying the above procedure to HumanNet, a publicly available integrated functional network which uses very similar statistical methods to combine molecular and co-citation in the same articles in scientific literature for human genes and their orthologs in worm, fly, and yeast (53) , yielded similarity values for 129,146,568 gene pairs based on shortest paths (Table 2.6). COXPRESdb (46),

which calculates Pearson correlation coefficients between human genes using over 100 expression experiments from the Gene Expression Omnibus database, was treated similarly after all correlation coefficients less than 0.5 were removed due to the high levels of noise in expression datasets (119; 120; 121; 122). The resulting functional similarity values for 81,046,264 gene pairs (Table 2.6) were used along with HumanNet to replicate results.

Metabolic or signalling pathways are frequently chains of connected proteins rather than dense balls of proteins all interacting with each other (Figure 4.2). Most clustering algorithms are designed to identify groups with many connections between them which is typical of protein complexes (Figure 4.2 A). However, I was also interested in signalling or metabolic pathways which tend to have a less dense chain-like structure (Figure 4.2 B). Thus, groups of functionally similar genes (functional clusters) were identified by only requiring a gene to be more similar than a given threshold to at least one gene within a group to become a member of that group (single-linkage clustering). This method also has the advantage of being simple and efficient to implement by simply iteratively adding genes to the appropriate cluster and will always result in the same functionally similar groups regardless of which gene is used as the starting point. I used a threshold equivalent to the top 1% shortest-paths in each network. Alternative thresholds equivalent to the top 0.1% and top 5% were also considered to check the sensitivity of results.

De novo CNVs are typically much larger than common CNVs (14; 161). Unfortunately, to our knowledge there is not a substantial set of *de novo* CNVs from healthy controls to which we could compare our patient-derived *de novo* CNVs. In addition, the observed *de novo* CNVs contained many more genes than expected based on their length (Figure 4.3). Thus, I determined significance of observed functional similarity by comparing the observed *de novo* CNVs from DECIPHER and GENCODYS to randomly chosen sets of adjacent genes from the genome such that the total number of genes was identical to the original CNVs, termed 'gene-number matched randomizations'.

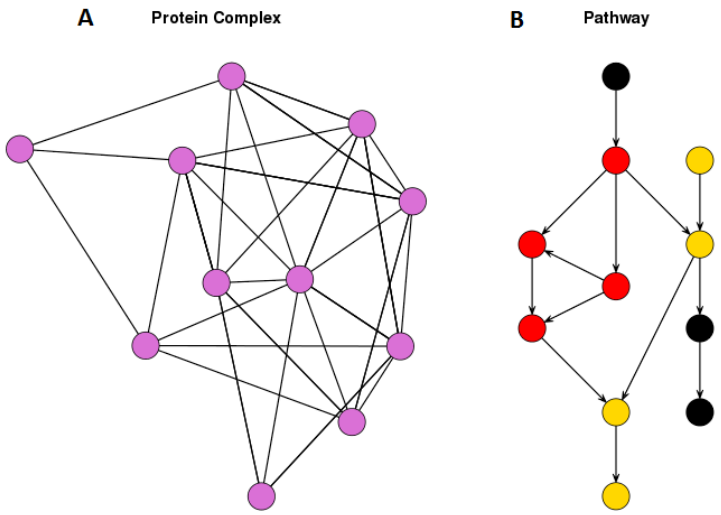


Figure 4.2: Protein complexes form globular clusters in functional networks (A) whereas pathways may not (B). (A) Typical structure of a protein complex with many interactions between genes. (B) Chain like structure typical of metabolic or signalling pathways with many fewer interactions between genes. Red indicates the type of cluster identified by most clustering algorithms. However, the clustering algorithm used in this work is also capable of identifying clusters such as the yellow circles.

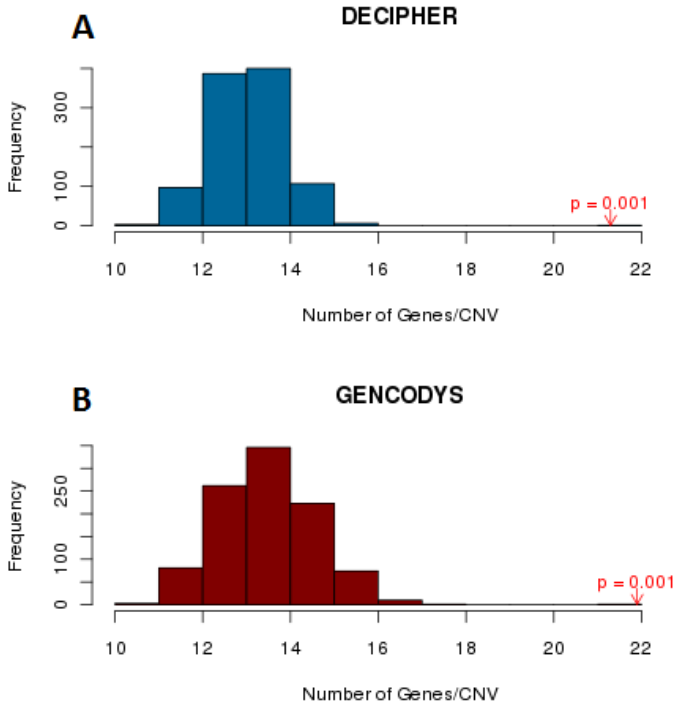


Figure 4.3: *De novo* CNVs contained more genes than expected. DECIPHER & GENCODYS *de novo* CNVs between 100kb and 5Mb in size compared to 1,000 random genomic segments of equal length. Red arrow indicates observed value and p-value.

4.3.3 Functional Clusters in CNVs

Both DECIPHER and GENCODYS *de novo* CNVs had significantly larger functional clusters (average of 3.56 genes/cluster and 3.63 genes/cluster respectively) compared to 10,000 gene-number matched CNV randomizations ($p = 0.0034$ and $p = 0.0015$ respectively) using single-linkage clustering of the PLN (Figure 4.4 A,B). To ensure this was not due to tandem gene duplications, paralogous genes, identified in OPTIC (157) or Ensembl (156) using zebrafish as the out-group, within each CNV/randomization were collapsed to a single copy. When paralogs were not collapsed to a single copy, clusters were less significantly large and on average were slightly smaller in both DECIPHER (3.52 genes/cluster, $p = 0.0119$) and GENCODYS (3.63 genes/cluster, $p = 0.0015$). The single largest cluster in each CNV was by far the largest containing on average 4.83 and 4.92 genes for DECIPHER & GENCODYS respectively and the most significant ($p = 0.0002$,$p = 0.0002$) vs the second and third largest clusters which are only slightly larger than the minimum size of 2 genes in both datasets (2.39, $p = 0.6735$, and 2.30,$p = 0.0112$ for DECIPHER; and 2.41, $p = 0.3490$, and 2.38,$p = 0.0129$, for GENCODYS) (Figure 4.4 C). 131 and 74 CNVs contained larger clusters than expected from the 10,000 gene number matched randomizations (ie. $p < 0.5$) in DECIPHER and GENCODYS respectively (Figure 4.4 D). Both deletions and duplications contain functional clusters with DECIPHER gains (duplications) and GENCODYS losses (deletions) most prominent (Figure 4.6 C). The patients, whose CNVs contained unusually large functional clusters, did not have any patient phenotype term significantly over-represented compared to all patients with *de novo* CNVs for either dataset (hypergeometric test, Bonferoni multiple testing correction).

DECIPHER *de novo* CNVs were also significantly more likely to contain at least one functional cluster compared to 10,000 gene-number matched randomizations ($p < 0.001$, Figure 4.4 E). GENCODYS *de novo* CNVs were more likely to contain a functional cluster than expected but did not reach significance ($p = 0.07$, Figure 4.4 F). The lower number of probes on GENCODYS arrays results in lower resolution CNVs which are larger and likely to contain extraneous genes (Figure 4.1). These extraneous

genes will increase the probability of hitting a functional cluster just by chance thus resulting in the higher null distribution seen for GENCODYS (GENCODYS median = 50%, 119/237; DECIPHER median = 44%, 188/427). In contrast we see the proportion of CNVs containing a functional cluster is more similar between the two datasets (49.4% DECIPHER, 54% GENCODYS expected). Since clusters were significantly large, within cluster similarity was significantly low compared to the 10,000 randomizations (DECIPHER: 16% lower, p = 0.031; GENCODYS: 20% lower, p = 0.1780). Since each gene in a cluster needs to be highly similar to just one other gene in the cluster to be included, adding more genes increases the number of potentially low weight edges thus reducing the average within cluster similarity). In addition, the number of functional clusters per CNV was not significantly different from chance in either dataset (DECIPHER: 1% more numerous, p = 0.3561; GENCODYS: 5% fewer, p = 0.2089).

Functional clusters within a patient

Next, I considered whether functional clusters within the same CNV or within different CNVs in the same patient are similar to each other, thus representing two parts of a larger pathway of genes, or are distinct representing two independent pathways. 110 DECIPHER CNVs and 68 GENCODYS CNVs contained more than one functional cluster. I defined the similarity between functional clusters as the similarity between the most similar pair of genes where one gene belongs to one cluster and the other gene belongs to the other cluster, this definition has the advantage of being reciprocal (similarity from A to B = similarity from B to A). Average between cluster similarity for clusters within the same CNV had very different patterns depending on the clustering threshold and CNV dataset examined. When compared to 10,000 randomizations average between cluster similarity was significantly low in GENCODYS ($p = 0.0001$) when a clustering threshold of the top 0.1% was used but slightly high in DECIPHER ($p = 0.1605$); when the top 1% threshold was used neither CNV dataset showed much of a trend in either direction (GENCODYS high $p = 0.3978$, DECIPHER low $p = 0.4563$). Only 59 DECIPHER patients and 67 GENCODYS patients had more than one *de novo* CNV and in only 22 and 34 respectively did more than one CNV

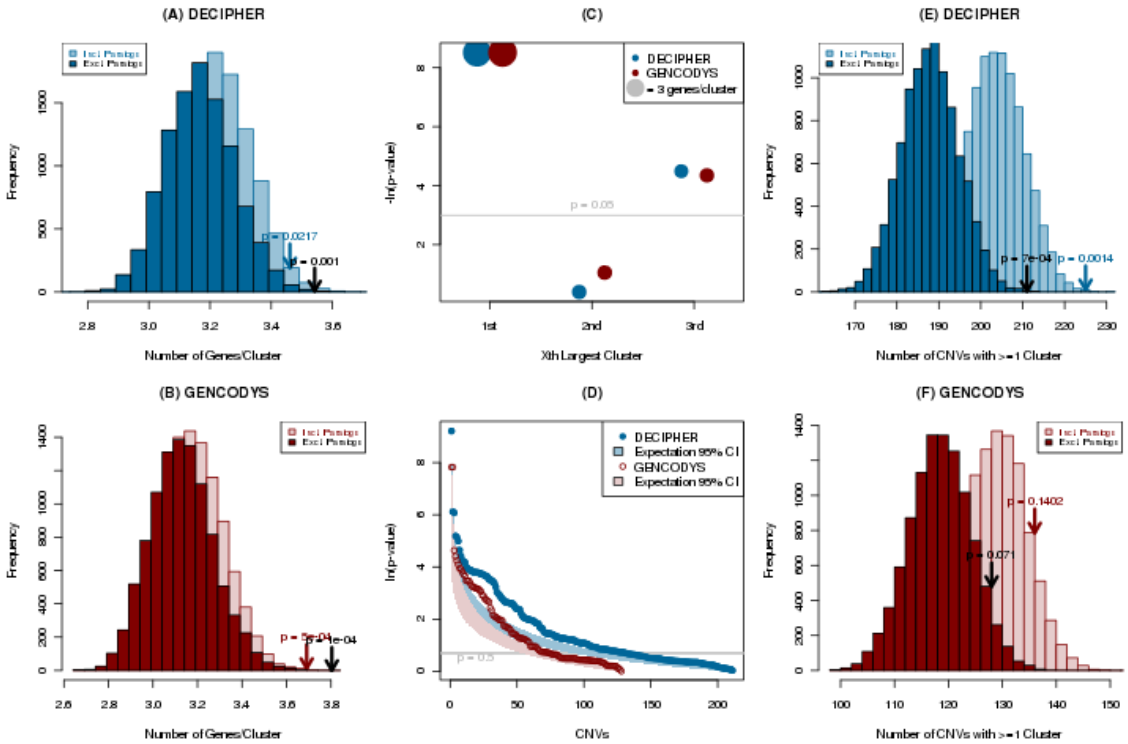


Figure 4.4: *De novo* CNVs from patients with developmental disorders contain significantly large numbers of functionally similar genes, as defined by proximity in the phenotypic linkage network (PLN). Blue = DECIPHER, Red = GENCODYS. (A & B) DECIPHER (and respectively GENCODYS) *de novo* CNVs contain significantly large clusters of functionally similar genes compared to 10,000 gene-number matched randomizations, the significance of which increases when paralogous genes within the same CNV are collapsed to a single copy. Arrows indicate observed value and p-value. (C) the largest group of functionally similar genes is most significant in both datasets. Size of the circle indicates the average cluster size, light grey line indicates p = 0.05, datasets are offset due to high overlap. (D) Many *de novo* CNVs contain more functionally similar genes than expected (points, grey line indicates p = 0.5), shaded areas indicate 95% confidence intervals given a uniform distribution of p-values. (E & F) More DECIPHER (and GENCODYS) *de novo* CNVs contain clusters of functionally similar genes compared to 10,000 gene-number matched randomizations. Only DECIPHER was significantly different. Arrows indicate observed value and p-value.

contain a functional cluster resulting in too little power to reliably determine the significance of any analysis.

Functional clusters in contiguous gene syndromes

In addition to the *de novo* CNVs, I obtained a list of 61 regions associated with known contiguous gene syndromes (CGSs, Table B.1), a set of consistent phenotypes associated with a deletion (or duplications) of a region of the genome containing multiple protein coding genes. 36 out of 61 of these syndrome regions contained at least two functional cluster genes which was not significantly ($p = 0.0736$) more than expected

compared to 10,000 gene-number matched randomizations. However on average 4.23 genes per functional cluster were affected by the CGS regions which was significantly higher than expected ($p = 0.0081$). Thus, functional clustering could contribute to more than half known CGSs (Figure 4.5). Many of these CGSs resist attempts to narrow down to specific genes (24; 162), but it is unlikely every gene in the region contributes significantly to the phenotype. If functional clusters are used to identify candidate genes to explain the phenotype of affected patients within the region, this would reduce the pool of genes by 70% (Figure 4.5).

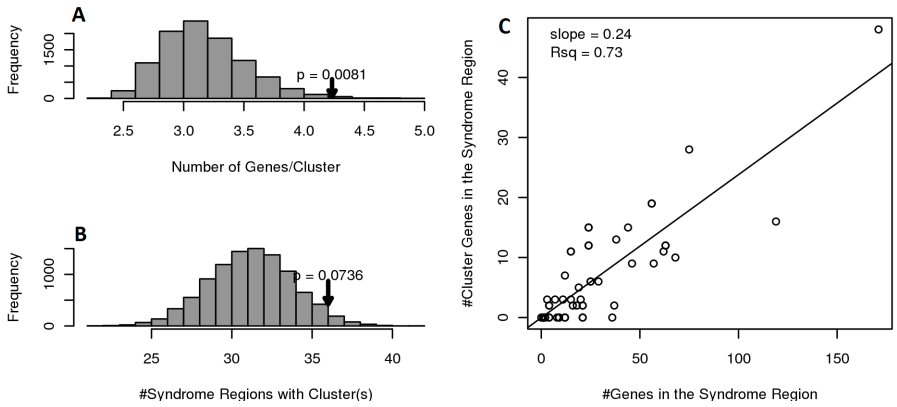


Figure 4.5: Contiguous Gene Syndrome regions contain functional clusters. The average number of genes belonging to each functional cluster within a CGS region was significantly high compared to 10,000 gene-number matched randomizations (A), the number of CGS regions which include at least 2 genes from the same functional cluster was elevated but not significant (B). Functional cluster genes are only 30% of all the genes in the CGS region (C).

4.3.4 Robustness of functional clustering

Both the HumanNet integrated functional network and the COXPRESdb co-expression network revealed large functional clusters in both DECIPHER and GENCODYS CNVs (Figure 4.6 A). DECIPHER *de novo* CNVs contained significantly ($p = 0.0492$) large clusters using the top 1% shortest paths in HumanNet and using the top 1% shortest paths in COXPRESdb ($p = 0.0227$). GENCODYS *de novo* CNVs contained highly significantly large clusters using COXPRESdb and larger than expected (but not significant) clusters using HumanNet ($p < 0.0001$ and $p = 0.3742$ respectively). Considering only the direct edges of the network prior to calculation of shortest paths in the PLN using the same similarity value as when using shortest paths (top 1% shortest paths) was

almost identical to using the shortest paths for both datasets (DECIPHER $p = 0.0035$, GENCODYS $p = 0.0015$ average cluster size). When I used all 626 DECIPHER and 426 GENCODYS *de novo* CNVs without filtering out very small and large CNVs, both datasets showed the same trends as they did using only the smaller CNVs but clusters are only significantly large in DECIPHER ($p = 0.0137$) not in GENCODYS ($p = 0.0817$), which suggests the very large and very small CNVs introduce considerable noise. In addition, I considered a lower clustering threshold equivalent to the top 5% shortest paths in the PLN. DECIPHER showed the same trend as using the higher threshold but not significant ($p = 0.0936$); whereas GENCODYS contained slightly smaller clusters than expected but not significantly so ($p = 0.3925$). However, using the lower clustering threshold more than 80% of all genes, whose mouse orthologs had been phenotyped within the Mouse Genome Informatics database, within the CNVs were found in the same cluster suggesting the clusters identified using the lower threshold may contain genes with quite different functions simply because those genes are well studied. Whereas less than half the genes, whose mouse orthologs had phenotype information available within the Mouse Genome Informatics database, that are located within the CNVs were found in the same cluster using the higher (top 1%) threshold. Thus the higher threshold requires genes to have similar functions not just be well studied, whereas the lower threshold may not.

The significance and trend of the number of CNVs containing functional clusters compared to 10,000 gene-number matched randomizations was more sensitive to the network used to define functional similarity (Figure 4.6 B). COXPRESdb maintained the same trend, of more CNVs with clusters than expected, for both DECIPHER and GENCODYS but neither reached significance. However HumanNet showed the reverse trend, of fewer CNVs with functional clusters than expected, in both datasets, but was not significant. Using only the direct edges, as opposed to shortest paths, and the lower clustering threshold showed roughly the same pattern as the original with significantly more DECIPHER CNVs containing clusters but not GENCODYS. Thus it is unclear whether *de novo* CNVs are more likely to contain a functional cluster than

expected by chance but they consistently contain larger functional clusters.

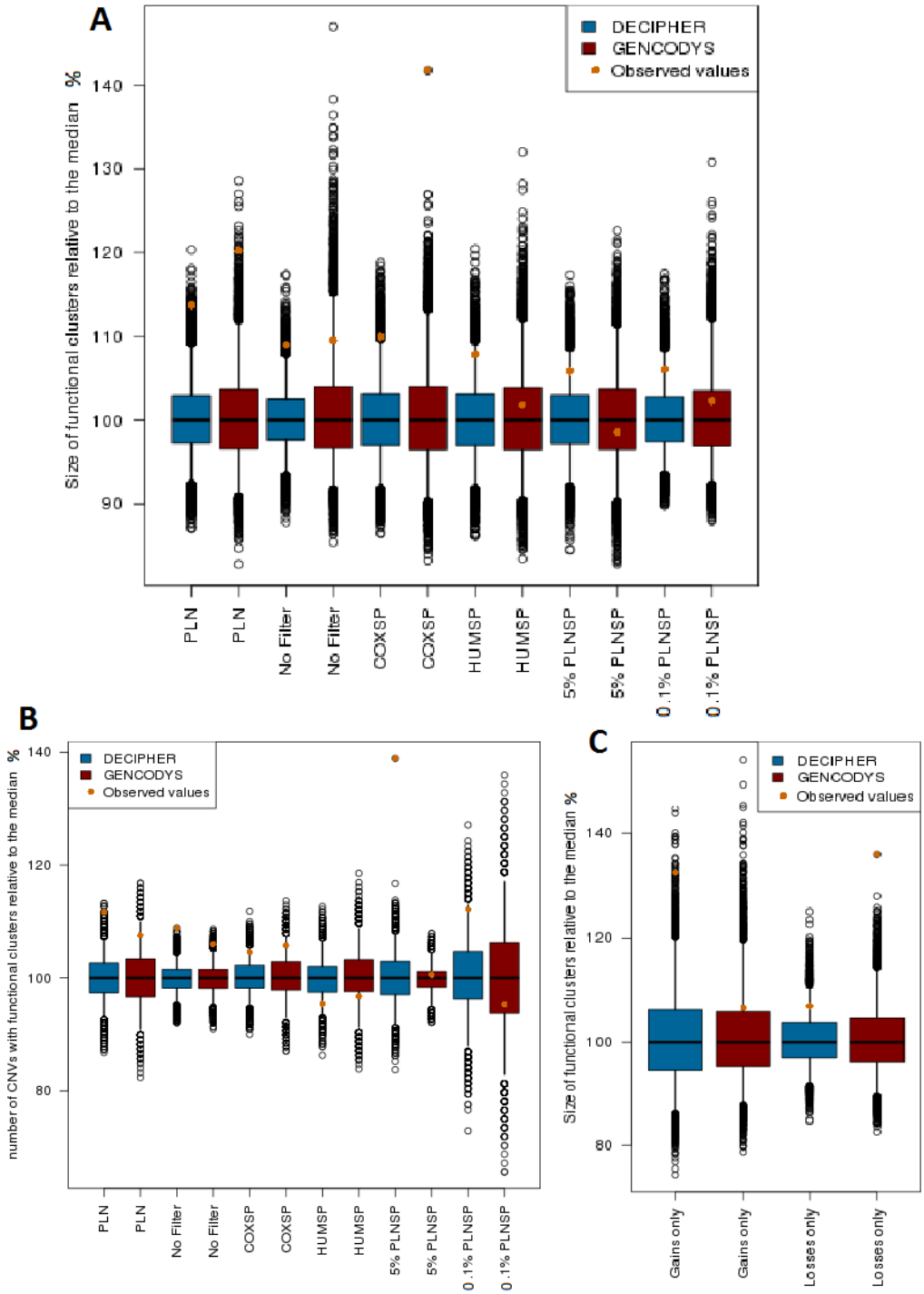


Figure 4.6: Robustness of functional clustering across multiple different networks and parameters. Since average size of functional clusters (A & C) and number of CNVs with a functional cluster (B) depend on the number of genes in the network and is highly sensitive to the threshold used, observed and randomizations were divided by the median of all randomizations to facilitated comparisons. (A & B) COXSP = shortest paths in COXPRESdb (46), HUMSP = shortest paths in HumanNet (53), PLN = direct edges only in the integrated functional network, 5%[0.1%] PLNSP = top 5%[0.1%] shortest paths threshold using the integrated functional network. (C) CNV datasets were divided based on the direction of the copy-number change. Whiskers contain 95% of randomizations. Orange dots indicate observed values

Many data sources contribute to the identification of functional clusters.

The phenotypic linkage network used to identify functional clusters combines many sources of biological information about genes into a single similarity value (Table 2.5). I examined which of these data sources were most important to defining functional clusters by determining the proportion of the similarity values used to include genes in the clusters derived from each of the data sets (Figure 4.7). During the construction of the network it was noticed that similarity between phenotypes produced by mutating the orthologues gene in mouse (from the Mouse Genome Informatics database (MGI), (36; 37)) was an excellent measure of functional similarity and was given more influence over the final similarity value than the other datasets (Figure 2.3). A consequence of this is that all gene similarity values have a high contribution from MGI phenotypes (Figure 4.7). However, gene similarity values which were used to identify functional clusters (cluster-links) had a smaller contribution from MGI phenotype information than the overall network or than other links among the top 1% most similar genes outside of the same CNV. Complementarily, literature-based protein-protein interactions (co-citation) contributed relatively more to the cluster-links. Similarity values between co-CNV genes which were belong to top 1% threshold showed an even more extreme pattern to cluster-links consistent with previous reports that genes which physically interact are located close together in the genome (91). The remaining datasets have a much lower overall contribution to similarity values and this contribution shows only slight deviations in cluster-links. This recapitulates the observations by Doelken et al (22) where they found genes within the same CNV with similar phenotypes in model organisms were also located close together in protein-protein interaction networks.

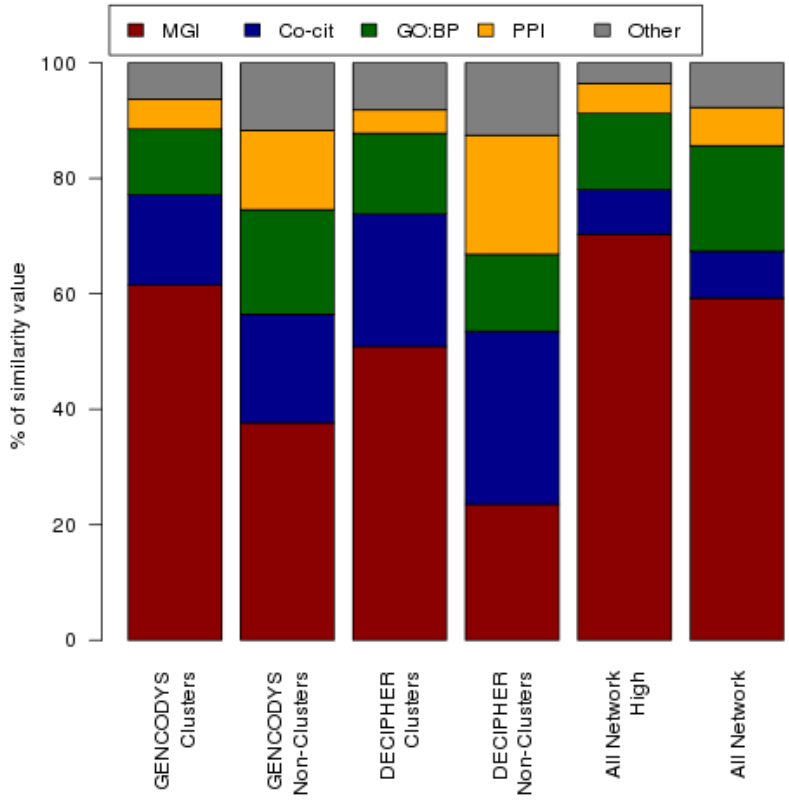


Figure 4.7: Data sources used to determine pairwise gene similarities in the integrated functional network. Mouse phenotypes (MGI, red), literature-based protein-protein interactions (Co-cit, blue), Gene Ontology biological process annotations (GO:BP, green) and physical interactions between proteins (PPI, yellow) supply over 90% of the functional similarity values in the network. All other data sources (Table 2.5) were combined together (grey). Clusters = similarities between genes in the same CNV above the similarity clustering threshold, Non-Clusters = similarities between genes in the same CNV below the similarity clustering threshold, All Network High = all pairwise genetic similarities above the similarity clustering threshold, All Network = all pairwise genetic similarities.

4.3.5 Functional clusters vs known disease genes

Functional clusters contain disease genes

I next considered whether the functional clusters contained known disease genes (Figure 4.8). I compared the proportion of cluster genes with each disease annotation to the proportion of all CNV genes with that annotation using a hypergeometric test. Disease genes from the Online Mendelian Inheritance in Man (OMIM,(112)), curated human haplo-insufficient genes from the literature (158), and genes whose ortholog is haplo-insufficient in mouse (MGI) were used to assess disease potential since their annotated phenotypes were not available in a form that is easily comparable to either LND or HPO phenotype terms. All three disease annotations were highly enriched in

functional clusters in both DECIPHER and GENCODYS (Figure 4.8).

To obtain gene sets which could be related to specific human-phenotypes I used mouse-phenotype enrichments from a previous study based on another independent set of CNVs from patients with developmental disorders which used the LDDB phenotype ontology (20), and genes annotated with Human Phenotype Ontology terms from the HPO database (35). To convert LDDB phenotypes in DECIPHER to HPO phenotypes and HPO phenotypes in GENCODYS to LDDB phenotypes, I used the mapping file available on the HPO website `http://compbio.charite.de/svn/hpo/trunk/src/mappings/`; where multiple terms were matched to a single term only the most general term was used. Since different genes were considered candidates in different patients and some CNVs overlap I used a binomial test to calculate significance for both sets of candidate genes. Both sets of candidate genes were enriched in functional clusters in both DECIPHER and GENCODYS CNVs (Figure 4.8 A & B).

All of the above annotations suffer from ascertainment bias, since well-studied genes are more likely to be annotated and tend to have more functional information available as well. Thus I also examined whether functional cluster genes were more likely to be affected in multiple patients than all CNV genes. Functional cluster genes were significantly more likely to be affected in multiple patients compared to all CNV genes and this enrichment increased with the number of patients who contain CNVs affecting the region (Figure 4.8 C). Thus, the more often a region was seen affected by CNVs in patients with developmental disorders the larger the proportion of genes belonging to functional clusters within the region.

Genes which have been well studied are more likely to have disease annotations and are also more likely to be highly functionally similar to many other genes (resulting in a high node degree : the sum of all similarities to other genes), since more functional information is available, resulting in a possible bias in the enrichments above. Only the enrichment in genes affected by CNVs in multiple patients would not be affected

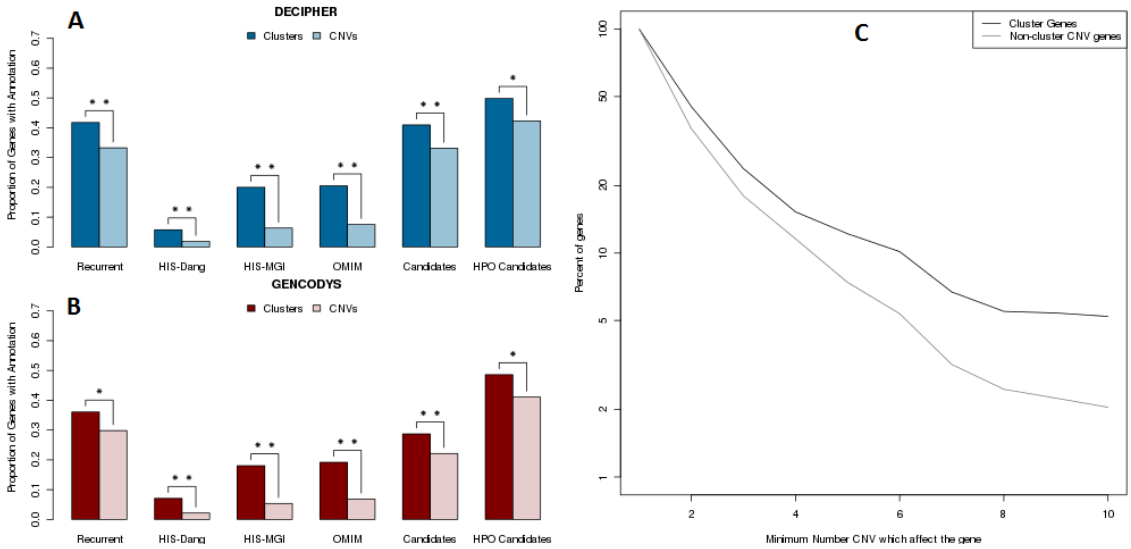


Figure 4.8: Enrichment of various disease relevant annotations in functionally similar genes in DECIPHER and GENCODYS respectively compared to all genes in their CNVs. Recurrent indicates genes found in more than one *de novo* CNV in the same dataset, Dang-HIS are haploinsufficient genes identified in (158), MGI-HIS are genes for which the mouse ortholog is annotated with at least one phenotype when heterozygous for the mutation (36; 37), OMIM are genes causally related to a disease in the Online Mendelian Inheritance in Man database (112), Candidates are those genes annotated with mouse phenotypes that were associated with the respective patient's symptoms in a previous study (20). HPO Candidates are those genes annotated with at least one of the respective patient's symptoms in the Human Phenotype Ontology database (35). One star indicates $p < 0.05$, two stars indicates $p < 0.0005$. (C) Functional clusters are found in highly recurrently affected regions of the genome.

by this bias. When I controlled for the degree distribution of cluster genes, by comparing against a background composed of a random set of genes with similar degree to the cluster genes from the genome, these enrichments lose significance with the exception of HPO Candidate genes which are less common among the degree corrected gene set than CNV genes (Figure 4.9). However, there is an established relationship between degree in various molecular networks, including co-expression, pathways, and protein-protein interactions, and essentiality in yeast which suggests high degree genes may be more pathogenic when disrupted due to their greater influence on cellular networks/function thus may be more likely to be disease genes (163).

Furthermore, I considered the robustness of these disease-gene enrichments across different networks and clustering thresholds (Figure A.1 in Appdendix A). Disease gene enrichments remained significant in most variations tested, with the exception of the COXPRESdb network which did not show an enrichment for general disease genes,

likely reflecting the absence of ascertainment bias in that network. However, COX-PRESdb clusters were still significantly enriched in candidate genes associated with the specific phenotypes displayed by the patient suggesting the functionally clustered genes may still be explaining the patient's phenotype.

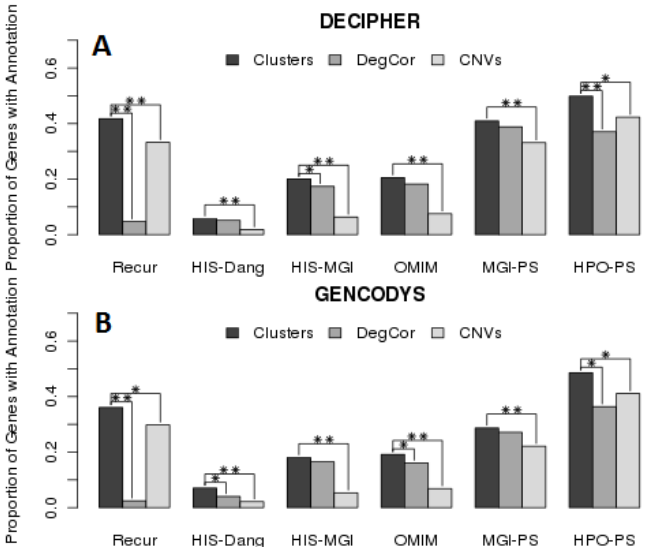


Figure 4.9: Enrichment of various disease relevant annotations in functionally similar genes in DECIPHER and GENCODYS respectively compared to all genes in their CNVs and to genes with similar degree in the PLN. Annotations are the same as Figure 4.8. Medium grey bars (DegCor) are a set of genes chosen at random from genes with the 100 most similar degrees to each cluster gene. One star indicates $p < 0.05$, two stars indicates $p < 0.0005$.

Clusters explain pathogenicity beyond disease genes

I used logistic regression to distinguish the ability of functional clusters from that of disease and HIS genes to differentiate the combined set of 664 *de novo* CNVs from a set of 2,478 CNVs identified in healthy individuals (103). When functional clusters, disease genes and HIS genes were included in the model, they were each significant ($p < 1 \times 10^{-20}$, $p = 4.4 \times 10^{-17}$, $p = 3.0 \times 10^{-11}$ respectively) with the presence of a functional cluster within a CNV having the greatest effect (Odds Ratios: cluster = 9.0, OMIM gene = 3.0 and HIS gene = 3.3). Clusters of functionally-related genes were more specific to pathogenic CNVs than either OMIM or HIS genes: half of pathogenic CNVs affected a cluster of functionally-related genes but only 4% of benign CNVs affected one (Figure 4.10). In contrast, known disease genes were present in 13% of benign CNV; and HIS genes were present in only a third of pathogenic CNVs.

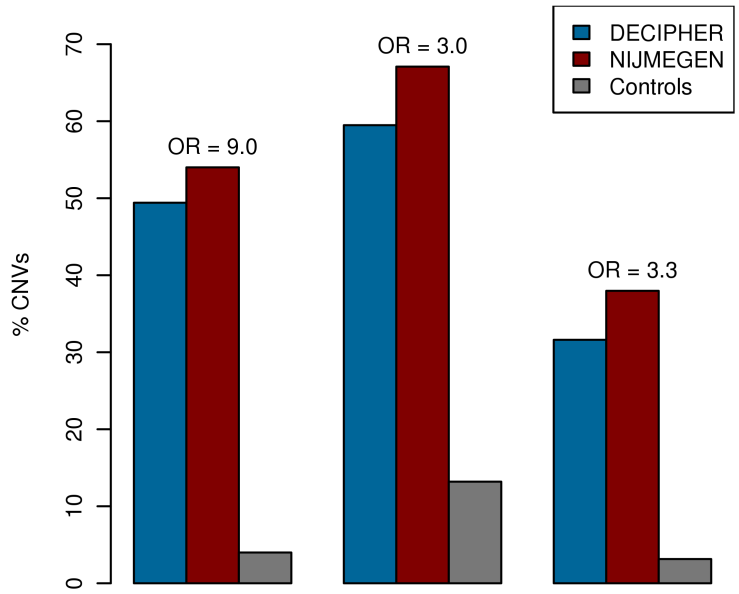


Figure 4.10: The presence of clusters of functionally-related genes in a CNV is a more specific or more sensitive predictor of pathogenicity than the presence of OMIM or HIS genes. OR = Odds Ratio from combined logistic regression

Next, I added the number of genes in the CNV to the logistic regression. I further restricted the dataset to those CNVs affecting at least two genes, thus have the opportunity to affect a functional cluster (Table 4.1). The presence of a functional cluster remains a significant predictor of CNV pathogenicity but is only equally or more powerful than haploinsufficent genes (HIS) or known disease genes (OMIM) when very large CNVs are considered ($\geq$ 10 genes). This result was further replicated in another published set of case-control CNVs from patients with developmental disorders (4). For this set we restricted the set to CNVs which affect at least three genes due to a very large number of CNVs with affecting exactly 2 genes: 15,000 control CNVs (60% all control CNVs which affect no more than 5 genes). Again functional clusters remain significant predictors of CNV pathogenicity and surpass HIS and OMIM genes (Clusters OR = 2.1, HIS OR = 1.6, OMIM OR = 0.9) for CNVs affecting at least 10 genes (Table 4.2). When CNV length was substituted for number of genes affected by the CNV this slightly increased the estimated predictive power of functional clusters. This suggests non-coding sequences have minimal impact on pathogenicity.

Table 4.1: Logistic regression including number of genes in a CNV as a predictor. HIS = presence of a haploinsufficient gene, OMIM presence of a know disease gene, Cluster = presence of a functional cluster

CNV size	Predictor	Odds Ratio [95% CI]	p-value	No. CNVs (% Pathogenic)
>=2 genes	Cluster	2.2 [1.1,4.4]	0.021	1672
	HIS	2.1 [1.3,3.3]	0.0015	(35%)
	OMIM	1.9 [1.4,2.7]	0.00015	
	CNV Length	1.3 [1.2,1.3]	$< 1 \times 10^{-15}$	
>=2 genes	Cluster	1.9 [1.0,3.6]	0.042	1672
	HIS	3.0 [2.1,4.4]	5.3×10^{-9}	(35%)
	OMIM	2.2 [1.7,2.9]	1.5×10^{-8}	
	No. Genes	1.1 [1.1,1.1]	$< 1 \times 10^{-15}$	
>=5 genes	Cluster	2.4 [1.3,4.5]	0.0056	904
	HIS	2.8 [1.8,4.3]	7.7×10^{-6}	(56%)
	OMIM	2.4 [1.7,3.4]	2.8×10^{-7}	
	No. Genes	1.0 [1.0,1.1]	3.4×10^{-7}	
>=10 genes	Cluster	2.5 [1.3,4.6]	0.004	539
	HIS	2.7 [1.5,4.6]	0.0005	(72%)
	OMIM	1.9 [1.2,3.1]	0.0045	
	No. Genes	1.0 [1.0,1.0]	0.0029	
>=15 genes	Cluster	3.1 [1.5,6.2]	0.0021	393
	HIS	2.1 [1.1,4.2]	0.03	(81%)
	OMIM	2.4 [1.3,4.4]	0.0073	
	No. Genes	1.0 [1.0,1.0]	0.22	

Table 4.2: Replication of logistic regression including number of genes in a CNV as a predictor using case-control CNVs from (4). HIS = presence of a haploinsufficient gene, OMIM presence of a know disease gene, Cluster = presence of a functional cluster

CNV size	Predictor	Odds Ratio [95% CI]	p-value	No. CNVs (% Pathogenic)
>=3 genes	Cluster	1.4 [1.1,1.7]	0.0056	30755
	HIS	1.6 [1.5,1.8]	$< 1 \times 10^{-15}$	(56%)
	OMIM	1.3 [1.3,1.4]	$< 1 \times 10^{-15}$	
	No. Genes	1.1 [1.1,1.1]	$< 1 \times 10^{-15}$	
>=5 genes	Cluster	1.4 [1.2,1.8]	8.6×10^{-4}	16273
	HIS	1.6 [1.4,1.8]	1.6×10^{-13}	(69%)
	OMIM	1.3 [1.2,1.4]	1.4×10^{-9}	
	No. Genes	1.1 [1.1,1.1]	$< 1 \times 10^{-15}$	
>=10 genes	Cluster	2.1 [1.7,2.8]	7.5×10^{-9}	8096
	HIS	1.6 [1.4,1.9]	5.2×10^{-9}	(86%)
	OMIM	0.9 [0.8,1.1]	0.50	
	No. Genes	1.1 [1.1,1.1]	$< 1 \times 10^{-15}$	
>=15 genes	Cluster	2.1[1.6,2.9]	2.4×10^{-6}	5573
	HIS	1.5 [1.2,1.9]	0.0019	(93%)
	OMIM	1.2 [1.0,1.6]	0.10	
	No. Genes	1.0 [1.0,1.0]	$< 1 \times 10^{-15}$	

4.3.6 Functional clustering is broadly associated with disease

As reported above, CNVs containing larger clusters than expected are not enriched for any particular patient phenotypes after correcting for multiple testing. I also considered the significance of functional clustering in CNVs from patients with a common phenotype compared to 10,000 randomizations directly (Figure 4.11). I performed the same enrichment tests as above on non-exclusive subsets of CNVs whose respective patients all exhibited a particular phenotype. To deal with the issue of limited power, due to the small size of the CNV datasets, I reduced the number of human phenotypes to test by eliminating those with too few CNVs to draw significance (< 5% of the CNVs in either dataset after filtering) and those with so many CNVs that it is impractical to distinguish between them (>25% of the CNVs in either dataset after filtering). To further increase power only phenotypes which could be tested in both DECIPHER and GENCODYS were considered and parental phenotypes whose child term(s) could be tested were removed. The remaining twenty-one different phenotypes were selected for examination. Only Hypotonia consistently contained large functional clusters across both datasets at $p < 0.05$ (Figure 4.11 left). The 88 and 60 hypotonia CNVs from DECIPHER and GENCODYS contain clusters with an average of 3.66 and 3.51 genes/cluster respectively compared to 3.26 and 3.02 expected by chance ($p = 0.0472$, $p = 0.0462$). Genes from all clusters in patients with Hypotonia in either DECIPHER or GENCODYS were combined into a single set of 385 unique candidate genes to investigate the functions represented by the clusters. The enrichment in Hypotonia cluster genes of a variety of functional annotations was calculated using a one-sided hypergeometric test.

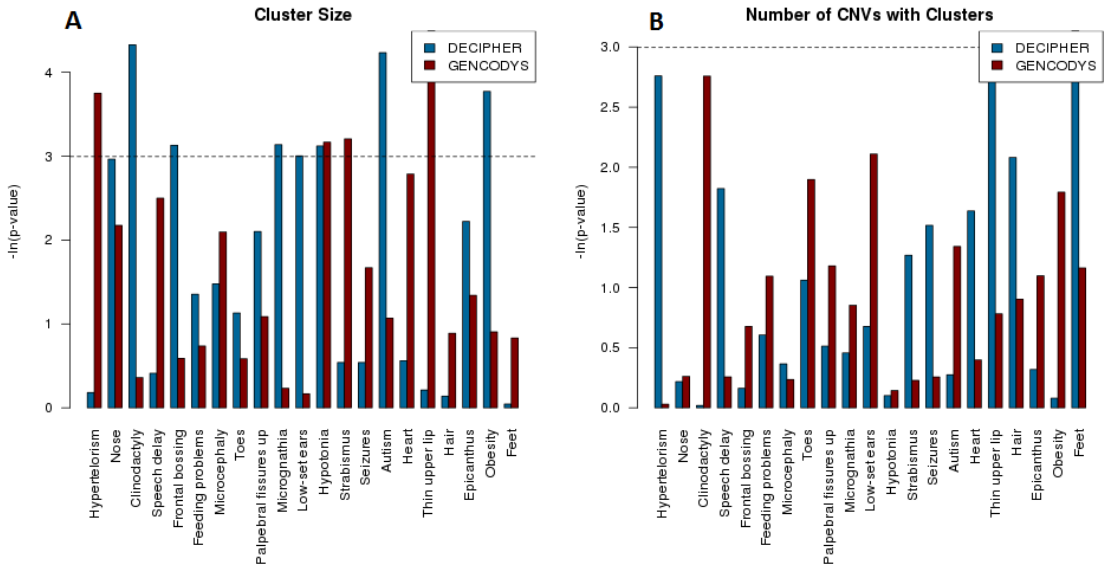


Figure 4.11: Groups of functionally similar genes were not strongly associated with any particular symptom. Significance of functionally similar gene enrichments in CNVs from patients with particular symptoms (dashed line indicates $p = 0.05$) compared to 10,000 gene-number matched randomizations. (A) average cluster size, (B) number of CNVs with clusters. Only Hypotonia reaches nominally significantly high average cluster size in both datasets, none are significant after correcting for multiple tests.

Hypotonia-associated functional clusters

First I determined whether the Hypotonia clusters were enriched in genes known to cause Hypotonia in humans from HPO. 36 out of 95 cluster genes with any HPO annotation were annotated with *Muscular Hypotonia* which was significant when compared to all genes with HPO annotations ($p = 0.0011$), all genes with HPO annotations in the CNVs from Hypotonia patients ($p = 0.0055$), or all genes in functional clusters from all patients ($p = 0.0006$). Since it is unclear which mouse phenotypes would be most relevant to Hypotonia, I tested all Mammalian Phenotype Ontology (MPO) terms assigned to mouse genes with human 1-1 orthologs from the Mouse Genome Informatics database(36; 37). Four terms were significantly enriched in Hypotonia clusters after Bonferroni correction for multiple tests compared to all Hypotonia CNV genes. These terms were: *abnormal behaviour* ($p = 3.8 \times 10^{-6}$), *abnormal learning/memory/conditioning* ($p = 3.1 \times 10^{-6}$), *abnormal nervous system morphology* ($p = 1.3 \times 10^{-6}$) and *nervous system phenotype* ($p = 3.3 \times 10^{-9}$) and accounted for 157 out of 290 Hypotonia cluster genes with mouse phenotype annotations. All four are related to the nervous system which supports their candidacy for the causal genes.

I also examined molecular functions annotated to the Hypotonia cluster genes using Gene Ontology (GO, (34)). Many terms were significantly enriched in clusters compared to all CNV genes (Table B.10 in Appendix B). Overall these terms represent a variety of functions related to neuronal development and nerve function. Many terms suggest cluster genes are involved in signal transduction and tend to be located at synapses. Finally, using DAPPLE (125) I examined physical interactions between Hypotonia cluster genes in the observed PPI network compared to 1,000 permuted networks. Hypotonia cluster genes had 217 interactions between them (Figure 4.12) which was significantly more than expected ($p = 0.003$). In addition the number of indirect links to other genes which interact with at least two Hypotonia candidate genes was significantly higher than the randomizations ($p = 0.006$).

Hypotonia cluster genes tend to function in the nervous system and neural development and form a dense network of physically interacting proteins. When disrupted, these genes result in relevant phenotypes in mouse and 36 are known to cause Hypotonia when disrupted in human. Thus the functional cluster genes found in patients with Hypotonia are likely to be contributing to their phenotype.

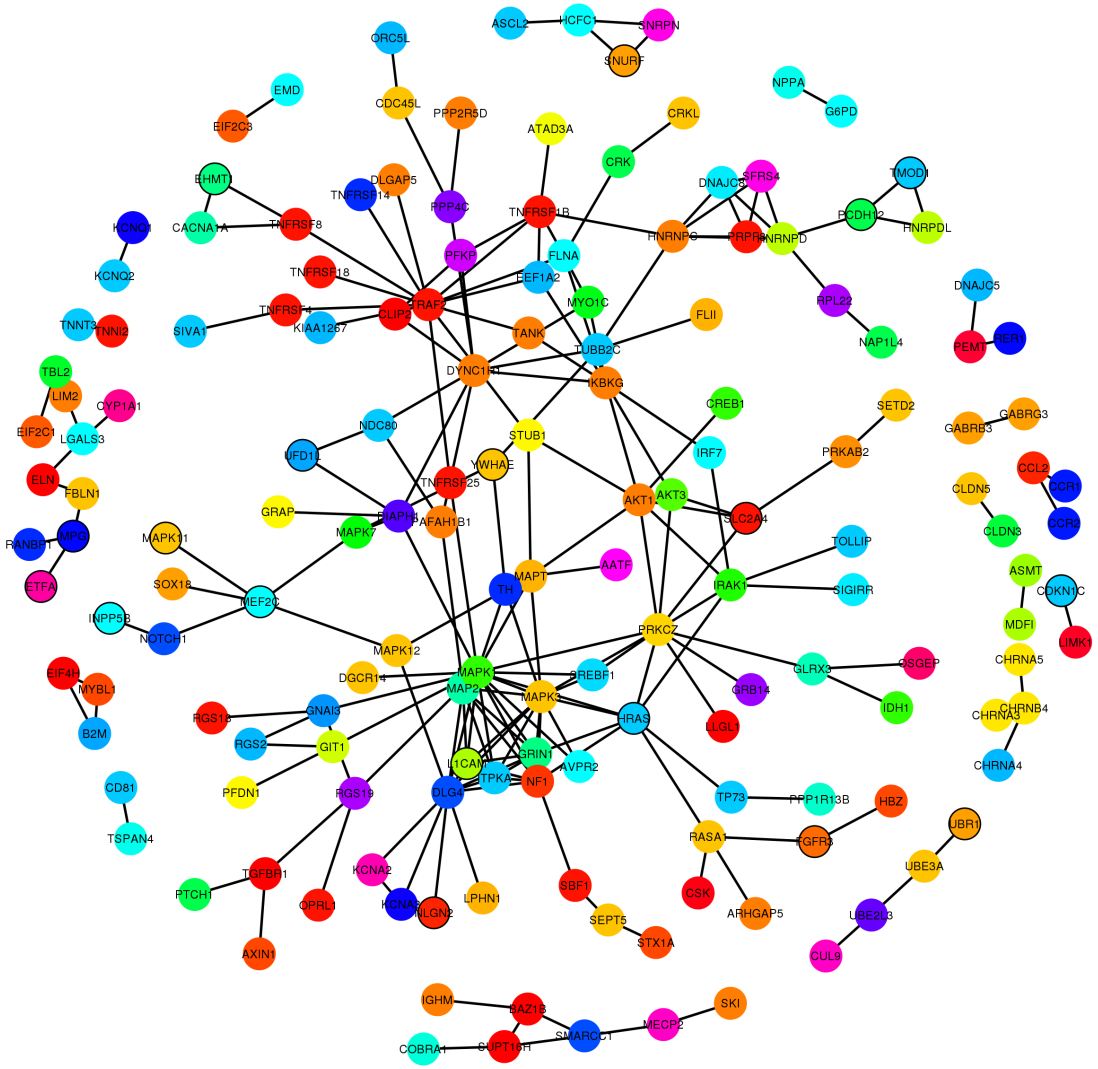


Figure 4.12: Protein-Protein Interaction network for Hypotonia functional cluster genes. Circles represent genes and their respective proteins, lines indicate physical interactions between them. The network contains significantly more direct physical interactions between Hypotonia functional cluster genes than expected ($p = 0.003$). Colours represent genes from different functional clusters from different CNVs, those genes previously associated with *Muscular hyptonia* in HPO are outlined in black.

Functional Clustering and Phenotype Severity

It has been previously reported that more severe symptoms (such as heart malformations, or craniofacial dysmorphologies) are more strongly associated with large CNVs than more subtle symptoms (such as ID, or Autism) (4). To determine if functional clustering explains this observed pattern, I chose the LDDB and HPO terms most similar to the categories used by (4) to group the *de novo* CNVs after imputing more general phenotypes using the respective ontologies (Table 4.3). To improve power, DECIPHER

Table 4.3: LDDB and HPO terms most similar to (4) symptom categories.

Eichler	HPO	HPO Name	LDDB	LDDB Name
DD/ID	HP:0001249	Intellectual disability	32.04.02	Mental retardation /developmental delay
Cardiovascular	HP:0001626	Abnormality of the cardiovascular system	17.05.00	Heart, general abnormalities
Autism	HP:0000729	Autism spectrum disorder	32.05.01	Autism / autistic behaviour
Epilepsy	HP:0001250	Seizures	32.06.00	SEIZURES, general abnormalities
Craniofacial	HP:0000271	Abnormality of the face	10.00.00	FACE

and GENCODYS *de novo* CNVs were combined for each phenotype category. Similar to the technique used in (4), I examined the proportion of CNVs containing a functional cluster larger than a given size, and the average size of functional clusters found in CNVs larger than a minimal size (Figure 4.13). In agreement with previous findings I showed that CNVs with more severe symptoms tended to contain larger functional clusters and that functional cluster size had a positive relationship with CNV size (Figure 4.13). I found different symptom classes were less distinct than previously reported, however my dataset was much smaller thus expected to be more noisy. However, I found that the patient *de novo* CNVs contained much larger clusters than the control CNVs and this difference was much more pronounced than the difference in size found in (4) with almost none of the control CNVs containing functional clusters suggesting functional clustering is a much better marker of pathogenicity than just CNV size (Figure 4.13).

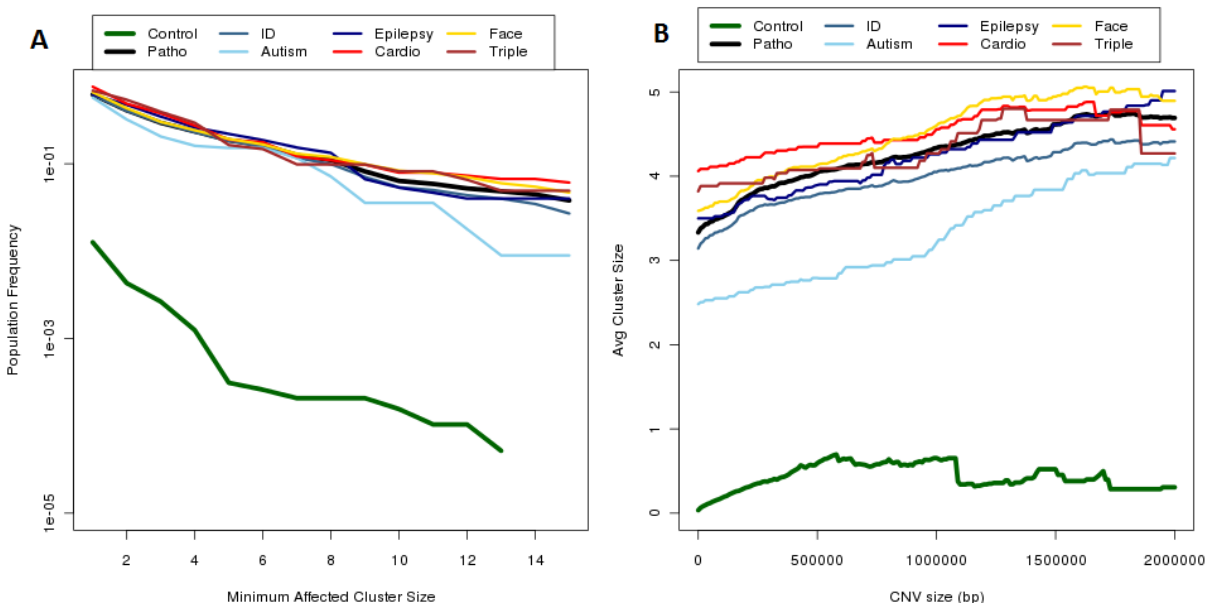


Figure 4.13: Relationship between CNV size, functional cluster size and patient phenotypes. (A) CNVs from patients contain larger clusters than controls. (B) Average functional cluster size increases with CNV size. Blue are neurological (ID, Epilepsy, Autism), red is cardiovascular and yellow is craniofacial. Brown are patients with at least one neurological, cardiovascular and craniofacial phenotype. Dark green are CNVs from healthy individuals from (103) and black is all CNVs from patients.

4.4 Conclusion

I found *de novo* CNVs, and contiguous gene syndrome regions, contain significantly large functional clusters. These clusters were robust and were not symptom specific. These clusters contained many known disease genes and candidate genes for the respective patient's specific phenotype. The presence of a functional cluster was a better indicator of the pathogenicity of a CNV than the presence of known disease genes. The many genes found in functional clusters affected by CNVs in patients with *Hypotonia* form a densely connected molecular network and perform functions related to neural transmission. Thus functional clusters may explain the pathogenicity of half of *de novo* CNVs found in patients with developmental disorders.

Chapter 5

Genome-wide Functional Clustering

5.1 Introduction

Proteins and genes often interact with each other in order to perform their function. This can occur through various mechanisms including: physical interactions, regulating each other's expression, chemical modifications, or catalysing sequential reactions. Many studies have shown that genes encoding functionally-related proteins tend to be located in close proximity along the genome even after accounting for tandem duplications (81; 82; 86; 164).

In Chapter 4, I showed that copy-number variants (CNVs) in patients with developmental disorders contain significantly more functionally-related genes than expected and that these clusters were associated with pathogenicity. This clustering of functionally-related genes has been seen in the genome of multiple eukaryotes including humans (86; 87; 88; 89; 91; 152), mouse (79; 80), zebrafish (85), worm (84), fly (81; 82; 83), and yeast (76; 77; 78). However, these studies are limited in two ways: i) reliance on a single source of functional information to identify clusters, eg. many rely solely on expression information (78; 79; 81; 83; 85; 88; 90) while others rely exclusively on protein-protein interactions (76; 91), or Gene Ontology terms (86); ii) lack of information about the appropriate resolution at which to look for functional clustering, eg. (82) identifies

clusters using a ten gene sliding window and found clusters of 10-30 genes while (81) looked at the same genome using a growing algorithm and found most functional clusters contain fewer than five genes.

In this chapter, I overcome both these limitations. Firstly, to overcome the reliance on a single functional data source I will use the phenotypic linkage network (PLN) which combines the network of (65) with mouse phenotype data as well as HumanNet (53) since integrating functional information together has been shown to improve functional predictions (55; 56; 59). Secondly, to determine the appropriate resolution to search for functional clustering I will use the functional clusters identified in patient CNVs in the previous chapter. I will use these tools to identify disease-relevant functional clustering in the human genome. I will examine various genomic properties which may explain the evolution of these functional clusters. I will examine the differences between functional clusters affected by CNVs in patients with developmental disorders and those affected in controls. Finally I will show that patients with mutations affecting the same functional cluster exhibit similar phenotypes.

5.2 Specific Methods

5.2.1 Identifying Genome-wide functional clusters

In contrast to most previous studies on functional clustering in the genome which relied on only one data-type to define functional similarity, eg. co-expression, or KEGG pathways, or Gene Ontology annotations (76; 81; 85; 86; 88; 91), I used the same integrated functional network as in Chapter 4, which was created by combining all of the large functional genomics datasets in Table 2.5. The functional similarity between pairs of genes, represented as nodes, was represented by weighted edges. The phenotypic linkage network (PLN) covers more genes and provides a better measure of functional similarity than any dataset alone (Figure 2.3, see Chapter 2). As in Chapter 4, to calculate the functional similarity for the 90% of gene pair comparisons without a di-

rect edge in the PLN, I converted the similarities (s) into distances (d) by taking $d = \frac{1}{1+s}$ and calculated the smallest sum of distances to travel, possibly by way of additional genes, from one gene to another (shortest paths), which was converted back into a similarity by taking $s = \frac{1}{d} - 1$, termed shortest-path similarities. The final shortest-path similarities cover 142,864,287 gene pairs (>98% of all possible pairwise comparisons), with the remaining missing values resulting from some genes being disconnected from the rest of the network.

To ensure my findings were not particular to the datasets or methods used to define functional similarity between genes I employed the other two functional networks used in Chapter 4: HumanNet(53), a publicly available integrated functional network , and COXPRESdb (46), which calculates Pearson correlation coefficients between human genes using over 100 expression experiments. Due to high levels of noise in expression data, COXPRESdb was filtered to remove all edges with Pearson correlation < 0.5 (119; 120; 121; 122). The same procedure described above was applied to the filtered COXPRESdb and HumanNet networks to produce 81,046,264 and 129,146,568 shortest-path similarities between gene pairs (Table 2.6) respectively.

Functional clusters were identified across the genome by combining the single-linkage clustering algorithm described in Chapter 4 with a growing algorithm similar to that used in previous studies (79; 81; 85). The resulting algorithm walks along the genome adding a gene to a cluster if it is within a distance threshold, D, and above a functional similarity threshold, T, of another gene (Figure 5.1). This will identify both globular and chain-like functional pathways as well as allow the detection of functional clusters with a large range of sizes while maintaining a minimum density of genes in the functional cluster. As in Chapter 4, I used a functional similarity threshold (T) equivalent to the top 1% shortest paths in each network. The distance threshold (D) was set at 2.1 Mb since 99% of neighbouring gene pairs within the functional clusters identified in *de novo* CNVs in Chapter 4 were within this distance. Major results were replicated varying both T, top 5% or 0.1% most similar genes, and D, 95th percentile (1.3 Mb) or

100th percentile (5 Mb), to ensure results were not sensitive to these choices (Figure 5.2).

To eliminate the effect of tandem duplications contributing to the genome-wide functional clusters, I removed all edges between paralogous genes from each functional network. Within each identified functional cluster sets of paralogous genes were collapsed to a single copy. In addition, I replicated results after eliminating all genes that had any paralogs from the genome ('Strict NP') as well as after removing the entire short arm of chromosome 6 which contains the MHC region ('No MHC') which is know to exhibit high levels of tandemly duplicated genes(165). Paralogous gene were identified using zebra-fish as and outgroup in either OPTIC (157) or Ensembl Compara (156).

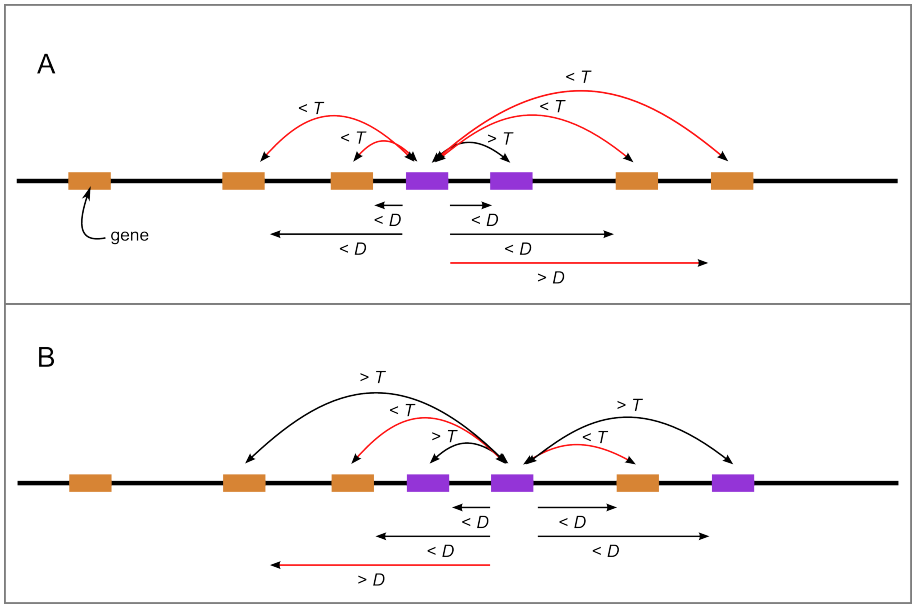


Figure 5.1: Genome-wide functional clustering algorithm. Functional clusters were identified as a chain of genes where each gene was within the distance threshold (D) and above the functional similarity threshold (T). D = 2.1Mb and T = top 1% shortest paths unless otherwise stated. The algorithm then walks along the genome grouping genes according to these rules as depicted in A and the subsequent step in B.

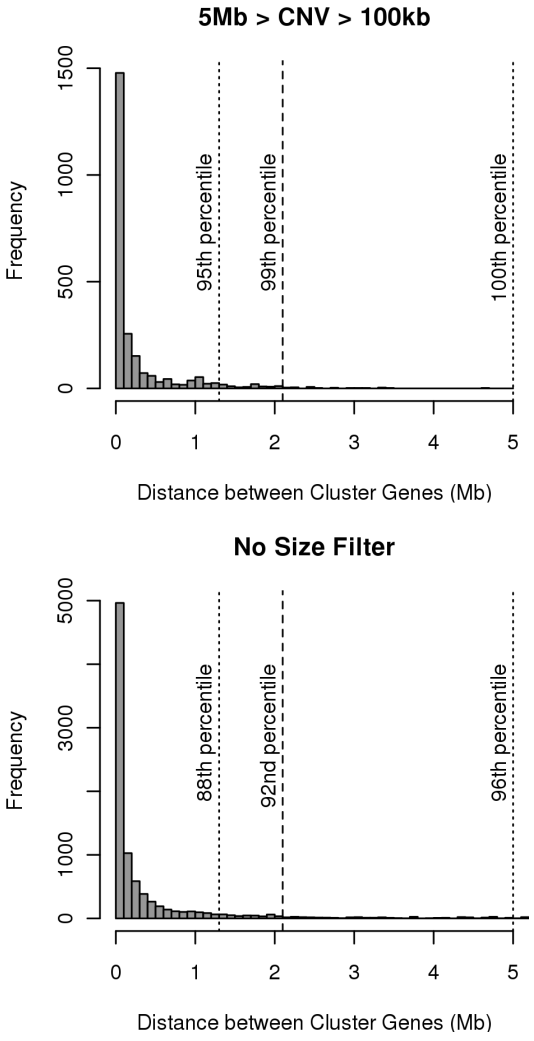


Figure 5.2: Distance thresholds used for genome-wide functional clustering. (top) Distribution of the distances between cluster genes and their nearest neighbour in the same cluster from functional clusters identified in either DECIPHER of GENCODYS *de novo* CNVs between 100kb and 5Mb in size. (bottom) Same as above but considering clusters from all DECIPHER and GENCODYS *de novo* CNVs. Dashed line indicates 2.1 Mb used in most analyses, dotted lines indicate 1.3Mb and 5Mb which were used to check the sensitivity of results to this threshold.

5.2.2 Permutation Clusters

The significance of results was tested by comparing against random genome permutations. These permutations were generated by randomizing the gene labels in the respective functional network, thus permuting the functional relationships between genes. This is equivalent to permuting the gene names across the genome while maintaining the distribution of genes. The clustering algorithm was applied to these permuted networks while retaining the irregular patterns of gene density across the genome. Analyses were applied to the resulting clusters and compared to those obtained from the original network.

5.2.3 Genomic Context

Gene expression data across 106 normal human tissues/cell-lines was obtained from the Gene Expression Barcode database (128). This database models gene expression records for the Affy HGU133A plus 2 array to determine whether each gene is significantly expressed beyond background noise. Housekeeping genes were defined as those expressed in >70% of these tissues based on the bi-modal distribution of gene expression broadness (Figure 5.7).

Cluster-gene sequence included the coding and intronic sequence of each gene belonging to a given functional cluster as defined in the Ensembl database version 54 (105). Inter cluster-gene sequence was defined as all sequence, both genic and non-genic, in between the genes belonging to a cluster according to Ensembl version 54 (hg18).

Lamina-associated Domains (LAD) were obtained from the UCSC Genome Browser track *NKI LADs (Tig3)*, which is derived from LaminB1-DNA interaction data (166). I examined both the proportion of the sequence and the number of LAD boundaries relative to the total length of sequence of cluster genes (or between cluster genes). This was downloaded on 13th Jun 2013 for human genome build hg18.

Segmental duplications were obtained from the UCSC Genome Browser track *genomicSuperDups*, which was derived from the method presented in (167; 168) and defined as large (>= 1kb), nearly-identical (>=90% sequence identity) regions of the genome after excluding high-copy repeats. The August 3rd 2006 edition (hg18) of the data was used. Frequency of Segmental duplications was calculated as the total sequence divided by the number of segmental duplication regions within the region.

Recombination rate was obtained from deCODE (169), the sex-averaged map for human genome build hg18 was obtained on Jun 13th 2013. The average recombination rate relative to the whole genome background was calculated for the cluster-gene se-

quence as well as the inter cluster-gene sequence.

Chromatin interaction data was obtained from the Gene Expression Omnibus accession: GSE18199 (downloaded on 12th Jun 2013). This data was derived from a Hi-C experiment to obtain a map of 3D interactions between chromatin regions in a genome-wide fashion (170). I used the observed divided by expected values over each 100 kb window for the karyotypically normal human lymphoblastoid (gm06690) cell-line to avoid biases resulting from differences in gene length or gene density that might exist between the observed clusters and the genome-permutation clusters. Linear imputation was used for windows only partially overlapped by each sequence type (eg. if 50% of window1 is overlapped by a region and 30% of window2 is overlapped by the other region then the value was the reported interaction score for window1-to-window2 multiplied by $0.5x0.3$). For each cluster the summed interaction score for each cluster-gene to all genes in the cluster (including itself) was divided by the total length of all the genes in the cluster squared (to correct for any differences in gene length between observed and genome-permutation derived clusters). In addition I considered the summed interaction score for each cluster-gene to each inter cluster-gene region to examine possible interactions with enhancers or other elements spanned by the cluster; this was divided by the total length of cluster-genes multiplied by the total-length of inter cluster-gene regions.

5.2.4 Pathogenicity of Functional Clusters

CNV datasets

I identified pathogenic functional clusters by examining the extent to which each genome-wide functional cluster was affected by *de novo* copy-number variants (CNVs) found in patients with developmental disorders or by CNVs found in healthy controls. I used the 626 *de novo* CNVs were from the Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources (DECIPHER, (17)) and 426 *de novo* CNVs from the Genetic and Epigenetic Networks in Cognitive Dysfunction Consor-

tium (GENCODYS, (18)) and 19,295 CNVs from healthy controls from (103) (see: Tables 2.1, 2.2, 2.3).

Another source of pathogenicity could be the presence of haplo-insufficient prenatally lethal genes, however mutations affecting these clusters would not be observed in the population. Since there is minimal knowledge about prenatal-lethality in humans, I used genes whose mouse one-to-one ortholog was completely prenatally-lethal when heterozygously knocked out in mouse (HISPNL genes) as recorded in the Mouse Genome Informatics (MGI) database (36; 37). The MGI database included 78 such genes with 1-1 human orthologs; and 81 clusters included or spanned at least one of these genes.

In addition, I examined the degree to which functional clusters were affected by CNVs identified in patients or controls from a larger CNV dataset obtained from dbVar (study ID: nstd54, (4)). However, since the inheritance patterns of the CNVs identified in patients was not examined in this study many CNVs identified in patients may be benign, thus they could not be used to reliably distinguish pathogenic vs benign clusters. Instead these CNVs were used to validate the pathogenic clusters and the pathogenicity score identified using the *de novo* CNVs.

Pathogenic Clusters

Pathogenicity of clusters was determined by grouping clusters together based on the extent they were hit by each of these CNV datasets and the extent to which they span HISPNL genes. Hierarchical clustering was used to form the groups since it is a simple, standard clustering method.

A cluster was deemed to have been 'hit' by a CNV if it encompassed the genes which formed the largest cluster in that CNV since this was what I found to be significantly different from chance previously (Figure 4.4 C, Chapter 4). This will exclude smaller

clusters which are likely to be incidentally affected by very large CNVs. I compated the proportion of genes participating in each cluster hit by CNVs from each dataset.

The lethality of each cluster was measured on the same scale as CNV hits, since measures on different scales will bias the distance measures used in hierarchical clustering toward the data with the largest values, by using the ratio of the number of HISPNL genes spanned by the cluster divided by the total number of cluster genes. In only one instance was this ratio greater than 1 (it was 1.5) thus to ensure this measure was on the same scale as the proportion of the cluster hit by each type of CNV this one instance was rounded down to 1.

Two different versions of hierarchical clustering were used to ensure robustness of the resulting groups to the details of the clustering method. First I used average-linkage hierarchical clustering which iteratively merges group with the smallest average Euclidean distance between the constituent members of each group starting with each item (in this case each genome-wide functional cluster) in its own group. I identified the appropriate number of clusters using the longest branch in the resulting dendrogram (which reflects the biggest drop in within group similarity as a result of merging clusters) which corresponded to the eight clusters intuitively evident from the data: the seven different possible combinations of CNV datasets and a final group of those containing HISPNL genes. These eight clusters were replicated using Ward's minimum variance method (171) which uses the squared Euclidean distance to merge groups such that the variance within each group is minimized. Both of these methods were performed using the `hclust` method in R (159). The difference between the results from these two methods was calculated as the van Dongen score divided by two which calculates the number of items change group (172).

Pathogenicity Score

I identified functional annotations which were unevenly distributed between clusters assigned to different groups by average-linkage hierarchical clustering (Figure 5.10). I considered both Gene Ontology (GO, (34)) and mouse phenotypes (MGI, (36; 37)) annotations , however to reduce the harshness of the multiple-testing correction applied since there was limited power, only 398 genome-wide functional clusters containing 2,473 unique genes participated in the groups, only terms present in GOSlims (34) and the 29 overarching mouse phenotypes were tested. All genes from all genome-wide functional clusters assigned to the same group were combined to create unique gene lists for each group. One group (G+CTRL) was excluded due to its small size (only three functional clusters with a total of nine unique genes) which is insufficient to detect enrichments. First I considered how terms were distributed between the seven remaining groups using a Fisher's Test on the two by seven contingency table followed by a Bonferroni multiple testing correction (Table 5.2).

Since the three groups containing clusters hit predominantly by patient CNVs (called: Patho, Gen, Dec in Figure 5.10) have a similar proportion of genes with the relevant annotations, as do the three groups hit predominantly by benign CNVs (called: CNV, Ctrl, D+Ctrl in Figure 5.10), these groups were combined together into 'pathogenic' and 'benign' genome-wide functional clusters; clusters belonging to the Lethal group were removed as these clusters may simply span a HISPNL gene rather than the cluster genes themselves being the cause of lethality. Using Fisher's exact test on the resulting 2 by 2 contigency table, I identified which functional annotations were significantly enriched in 'pathogenic' functional clusters after a Bonferroni multiple testing correction.

The 'pathogenicity score' was defined as the average proportion of cluster genes annotated with each of the three GO and six MGI functional annotations found significant between both 'pathogenic' and 'benign' functional clusters as well as between the 7 groups identified using average-linkage hierarchical clustering (Table 5.3). The identi-

fication of functional enrichments across both the seven groups and between the combined 'pathogenic' and 'benign' groups was repeated using all genes spanned by the clusters, which included the genes participating in the functional cluster as well as all genes located in between genes participating in the same functional cluster (thus are located in the same regions of the genome). These were used to define an alternative 'pathogenicity score', defined as the average proportion of all spanned genes with each of the respective enriched functional terms, which should be considerably noisier thus less informative that the one produced by considering just the gene belonging to the functional cluster.

These scores were tested by considering an independent set of case-control CNVs from (4). The cluster with the largest number of genes (and at least two) affected by each CNVs was considered 'hit' and the pathogenicity score for that cluster was contributed to the respective category (Case or Control) and the difference between these distributions was determined using a Wilcox-rank-sum test.

5.2.5 Functional Clusters and Phenotype

Mapping Phenotypes

DECIPHER patient phenotypes were recorded using LDDB and GENCODYS patient phenotypes were recorded using HPO. I converted between these ontologies as needed using the file provided on the HPO website (35), `http://www.human-phenotype-ontology.org/contao/index.php/downloads.html`. All provided mappings of any quality were considered. For those LDDB terms which mapped to more than one HPO term the most general HPO term was used. HPO and LDDB terms without a mapping were excluded from the analysis.

5.3 Results

5.3.1 Functional clusters are present in the genome

Disease-relevant functional clusters in the human genome were identified by combining the single-linkage clustering using in Chapter 4 with a growing algorithm similar to those used in previous studies (79; 81; 85). Using this algorithm to be included in a cluster a gene must share sufficiently high functional similarity (T) with a gene within a distance (D) of another gene (Figure 5.1). As in Chapter 4, the similarity threshold was set to be the top 1% shortest-path similarities in the Phenotypic Linkage Network (PLN). The distance threshold was set at 2.1 Mb corresponding to 99% of all genes found in functional clusters within the *de novo* CNVs examined in Chapter 4 (Figure 5.2).

The human genome (hg18) contained 942 functional clusters located on every chromosome examined (Figure 5.3). To ensure the functional clusters were not simply due to tandem gene duplications, paralogous genes, identified in OPTIC (157) or Ensembl (156) using zebra-fish as the out-group, within each genome-wide cluster were collapsed to a single copy. After collapsing paralogs a total of 3,420 genes were present in functional clusters (including paralogs 3,953 genes participate in 1,071 clusters). I compared the functional clusters identified in the human genome to 1000 genome-permutations, which scrambled the functional relationships between genes while preserving the patterns of gene density and arrays of paralogous genes across the genome. Both the number of clusters (942) and the total number of genes in clusters (3,420) were significantly high ($p < 0.001$) but the average size of clusters was not significantly different ($p = 0.114$) from the randomizations (Figure 5.3). As expected, when paralogs are not collapsed to a single copy there were many more functional clusters identified in the human genome and many more total genes involved in functional clusters (Figure 5.3). Tandem duplications will result in many paralogs to being near each other in the genome thus likely to be grouped together into a functional cluster. However, per-

muting the gene-labels across the genome effectively eliminates this bias, so the null distribution when paralogs are included is almost identical to when paralogs were collapsed.

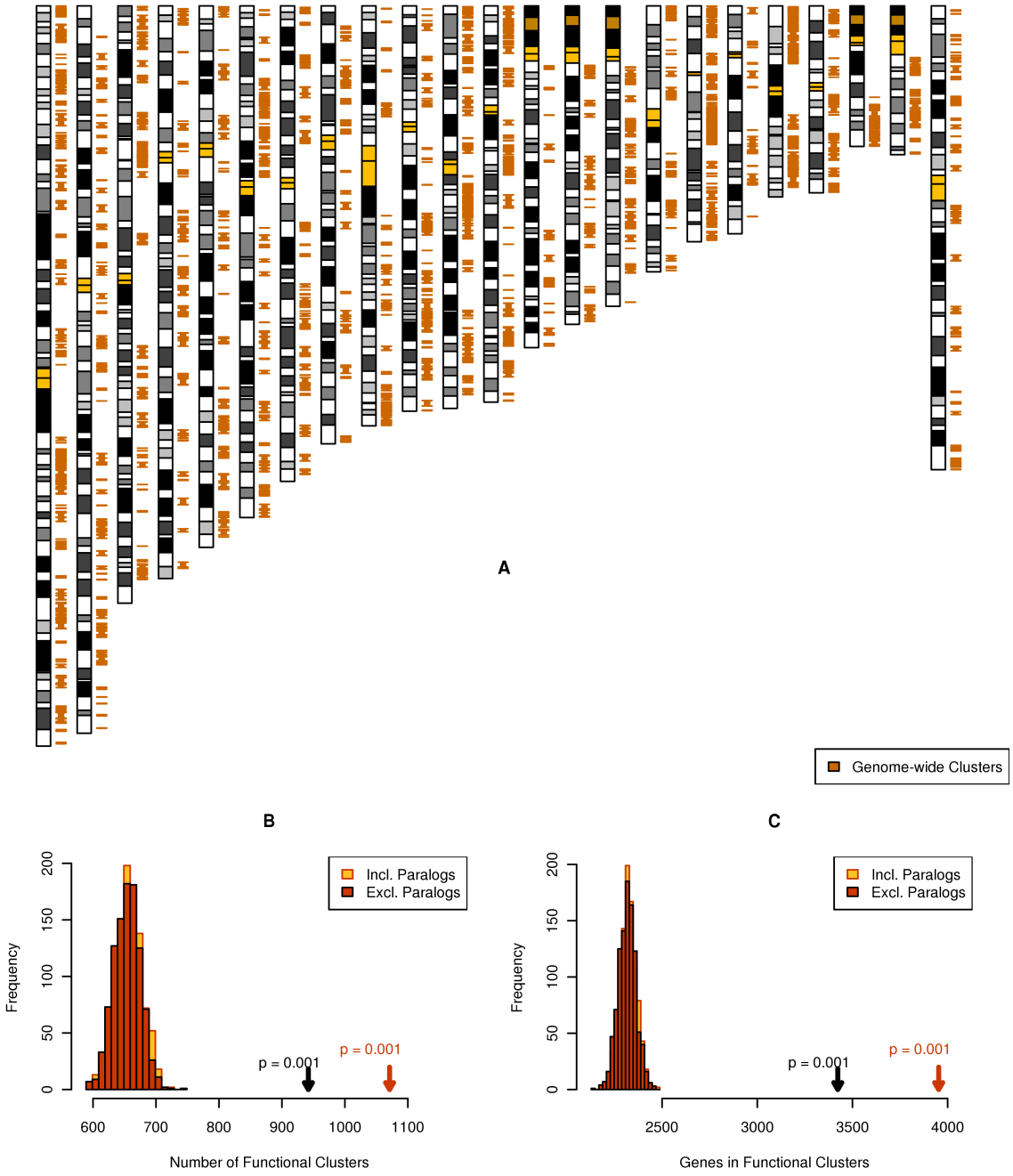


Figure 5.3: Genome-wide functional clustering (A) distribution of functional clusters along chromosomes. (B) The genome contains significantly more functional clusters than seen in any of the 1,000 network node-label permutations. (C) The genome contains significantly more genes in functional clusters than seen in any of the 1,000 network node-label permutations.

To check the sensitivity of genome-wide functional clusters to the network and clustering parameters used, I replicated the identification of functional clusters and collapsing

paralogs to a single copy with two different functional networks, two different clustering thresholds, and two different distance thresholds (Figure 5.4, 5.5). Using either the HumanNet integrated functional network or the COXPRESdb coxpression network, using a similarity threshold of the top 1% shortest-path similarities for each respective network a distance threshold of 2.1 Mb, did not qualitatively change the results with the number of clusters and total genes in functional clusters remaining significantly more than expected. In addition, to ensure my results were not due to paralogous genes I replicated the analysis after excluding any gene with known paralogs (Strict NP) as well as after removing the MHC region (No MHC).

The total number of genes involved in functional clusters was highly significant with $p < 0.001$ in all cases (Figure 5.4A), as was the number of clusters (Figure 5.4B) but not the average number of genes per cluster (Figure 5.5). Changing the similarity threshold (T) had the biggest effect with 8,630 genes participating in functional clustering using the lowest threshold (top 5%) and only 1,110 genes using the highest threshold (top 0.1%); however both thresholds result in a small number of functional clusters being identified, 807 and 440 for $T = 5\%$ and $T = 0.1\%$ respectively. Changing the distance threshold had a smaller effect with the lower threshold ($D = 1.3$Mb) identifying 2,920 genes in 910 clusters and the higher threshold ($D = 5$Mb) identifying 4,460 genes in 866 clusters. I found the initial choice of threshold resulted in the highest number, 942, of distinct functional clusters identified.

The COXPRESdb network identified the most genome-wide functional clustering with 5,340 genes participating in 1,090 clusters, which was twice as many genes and 50% more clusters than the median of the gene-label permutations. Furthermore on average each COXPRESdb functional cluster contain 40% more genes than expected which was a much larger difference than seen in any other case I examined (Figure 5.5). This is consistent with the fact that most previous studies have detected functional clustering using gene expression(78; 79; 81; 83; 85; 88; 90). Whereas, HumanNet which was another integrated functional network identified a similar level of functional clustering

as our PLN (3,990 genes in 902 clusters).

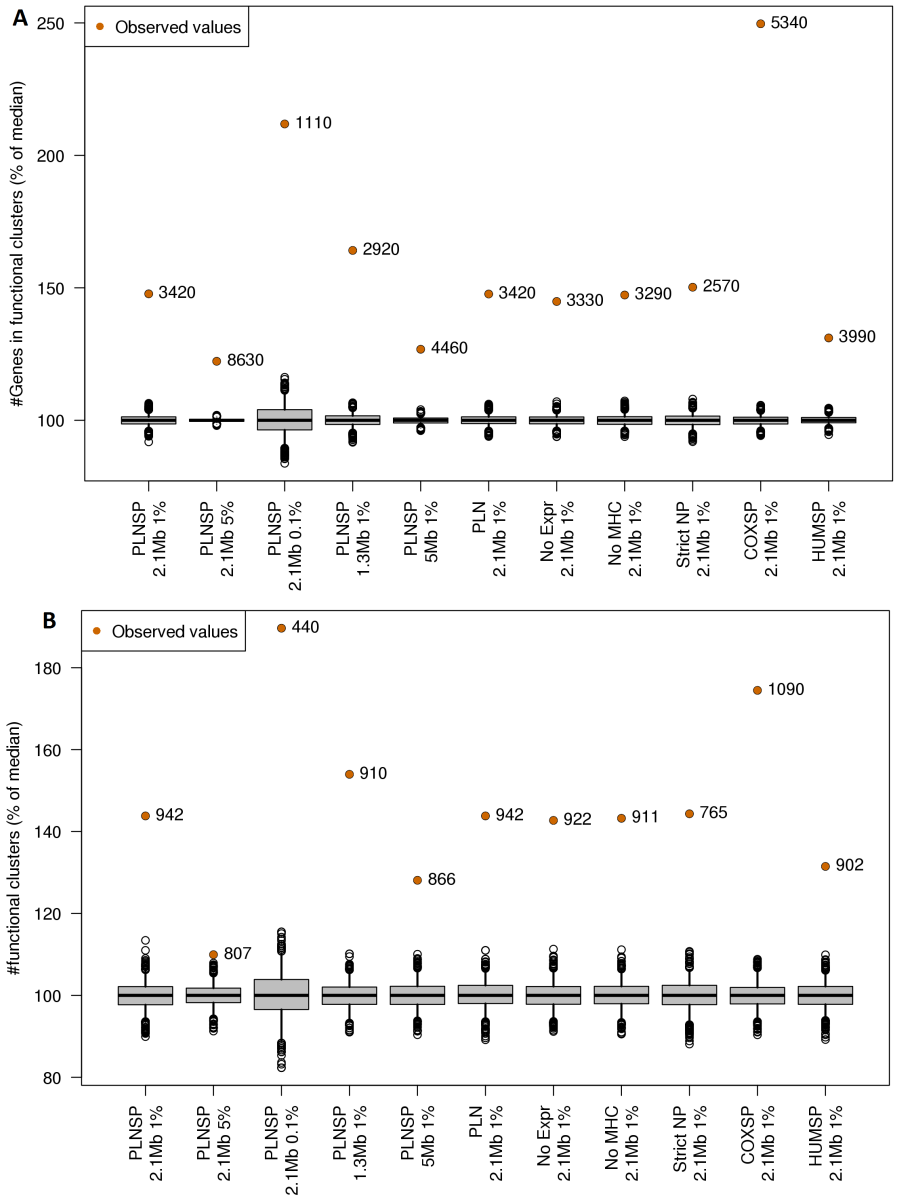


Figure 5.4: (A) Number of genes participating in genome-wide functional clustering is robustly significantly high regardless of the parameters & network used. (B)Number of genome-wide functional clusters is robustly significantly high regardless of the parameters & network used. PLNSP = shortest paths in the Phenotypic Linkage Network. PLN = original edges in the Phenotypic Linkage Network. COXSP = COXPRESdb Shortest Paths. HUMSP = HumanNet Shortest Paths. X% = top X% of similarities was used as the similarity threshold. Y Mb = Distance threshold used. Strict NP = removing all genes with any paralogs from consideration, No MHC = excluding the entire short arm of chromosome 6 including the major histocampatiblity region. No Expr = Shortests paths in the PLN excluding all expression information from its construction. Grey boxplots represent the result from 1,000 node-label permutations of the functional network. Normalized by dividing by the median of the permutations.

Given the ability of gene expression to identify functional clustering observed for the COXPRESdb network, I examined whether the gene expression information contained

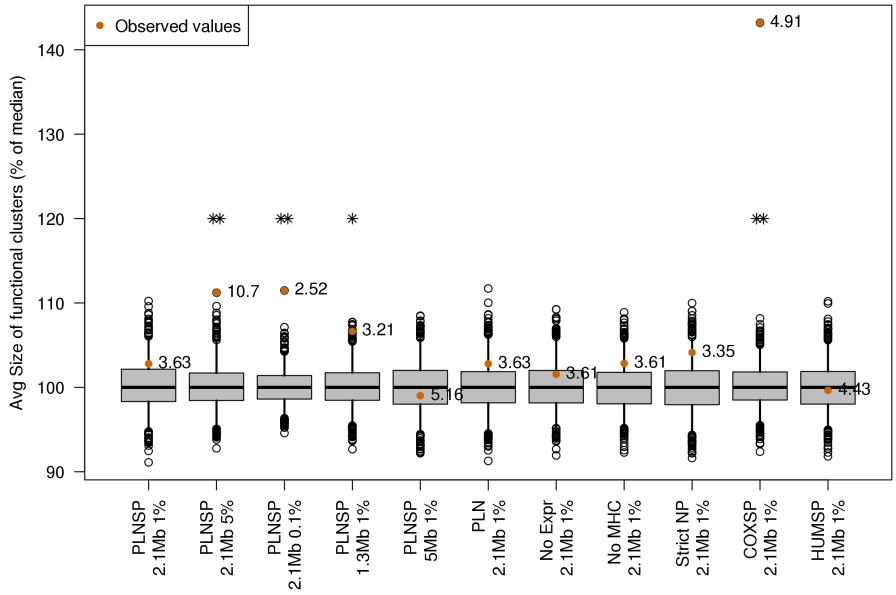


Figure 5.5: Size of genome-wide functional clusters is not significantly high for the majority of the parameters & network used. PLNSP = shortest paths in the Phenotypic Linkage Network. PLN = original edges in the Phenotypic Linkage Network. COXSP = COXPRESdb Shortest Paths. HUMSP = HumanNet Shortest Paths. X% = top X% of similarities was used as the similarity threshold. Y Mb = Distance threshold used. Strict NP = removing all genes with any paralogs from consideration, No MHC = excluding the entire short arm of chromosome 6 including the major histocampatiblity region. No Expr = Shortests paths in the PLN excluding all expression information from its construction. Grey boxplots represent the result from 1,000 node-label permutations of the functional network. Normalized by dividing by the median of the permutations.

within the PLN was responsible for the significance of the functional clustering identified using it. Removing all direct gene-expression information from the construction of the PLN ('No Expr') had little effect on the level or significance of the observed functional clustering with 3,330 genes and 922 clusters still identified (Figure 5.4). To explain this I examined the contributions of the different data-sources (see: Table 2.5) to the PLN edges important in forming the genome-wide functional clusters ($T = 1\%, D = 2.1$Mb). The four most important data-sources (mouse phenotypes, co-citation in the literature, Gene Ontology biological process annotations, and protein-protein interactions) are responsible for > 90% of all similarity values in the network, thus the remaining data-sources (including gene-expression) were grouped together. Genome-wide functional clusters are consistent with *de novo* CNV functional clusters by having very similar datasets contributing to the edges (see Chapter 4 Figure 4.7). As before co-citation was over-represented and mouse phenotypes under-represented in edges used to define genome-wide functional clusters compared to the overall net-

work (Figure 5.6). This difference is partly explained by functional cluster genes being near each other in the genome, which is expected to result in more co-citations due to papers examining genome regions or mapping genetic elements across the genome. Indeed, I find that across the genome genes within 2.1 Mb but not in the same cluster have co-citation over-represented in their links in the PLN. "Other" data-sources which includes co-expression is also more important to genome-wide functional clusters compared to the whole network or just high scoring edges. Edges between genes within 2.1 Mb of each other show a similar pattern to those contributing to clusters but with a larger contribution from protein-protein interactions and less from the mouse phenotypes.

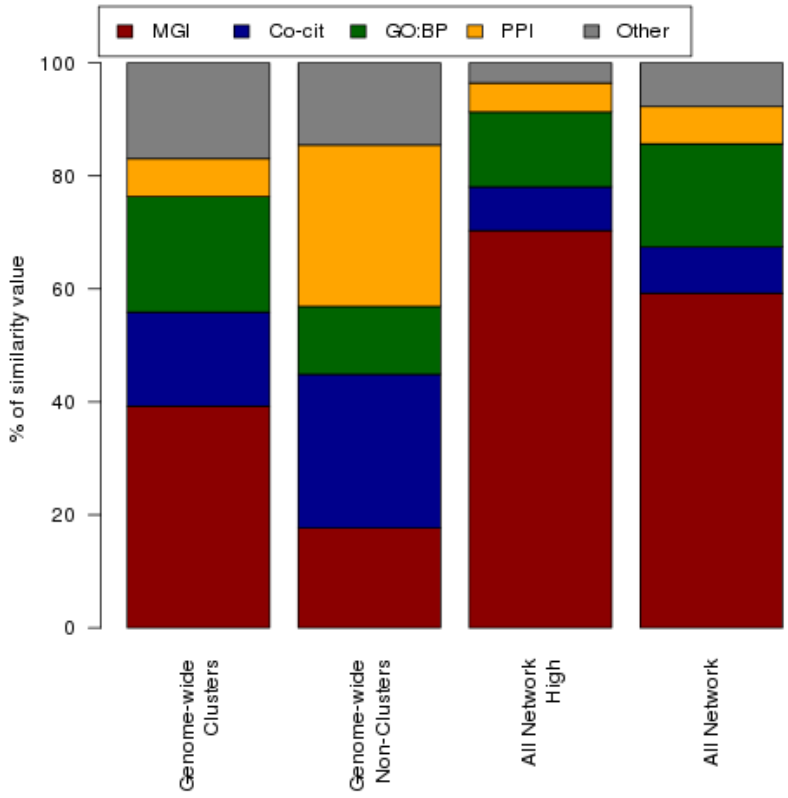


Figure 5.6: Data sources used to determine pairwise gene similarities in the integrated functional network. Mouse phenotypes (MGI, red), co-citation (Co-cit, blue), Gene Ontology biological process annotations (GO:BP, green) and physical interactions between proteins (PPI, yellow) supply over 90% of the functional similarity values in the network. All other data sources (Table 2.5) were combined together (grey). Genome-wide Clusters = similarities between genes in the same genome-wide functional cluster above the similarity clustering threshold, Genome-wide Non-Clusters = similarities between genes within 2.1Mb of each other below the similarity clustering threshold, All Network High = all pairwise genetic similarities above the similarity clustering threshold, All Network = all pairwise genetic similarities.

5.3.2 Gene Expression Patterns of Functional Clusters

Clusters of house-keeping genes have been reported for humans (80; 81; 90; 93) but the existence of tissue-specific clusters is still disputed (93; 94; 173). I employed the gene-expression profiles from the Gene Expression Barcode database (128) to determine whether the functional clusters identified here were composed of house-keeping genes. Genomic functional cluster genes, identified using the PLN, were expressed in a significantly higher proportion of tissues/cell-types than non-cluster genes (two-sided Wilcox rank-sum test, $p = 4.5 \times 10^{-18}$). Housekeeping genes were defined as those genes expressed in > 70% of examined tissues/cell-types. 12.8% of cluster genes and 13.1% of non-cluster gene were considered housekeeping genes ($p = 0.6841$, proportion test). The previously reported clusters of house-keeping genes were large, 30-200 genes, compared to functionally specific gene clusters (80; 81). However, I found that large clusters (>10 genes) contained only 11.7% housekeeping genes while small clusters (<= 10 genes) contained 13.2% (proportion test, $p = 0.2952$) Genes belonging to large and small functional clusters had similar distributions of broadness of expression, overall higher than genes not belonging to functional clusters but not enriched in house-keeping genes.

However, when I examined clusters identified using COXPRESdb or HumanNet both were significantly enriched in housekeeping genes ($p = 3.6 \times 10^{-32}$, $p = 2.4 \times 10^{-52}$, respectively, Figure 5.7 C,D). COXPRESdb functional clusters contained 17.8% housekeeping genes (vs 11.4% of non-cluster genes); and HumanNet clusters contained 20.5% housekeeping genes (vs 11.3% of non-cluster genes). Furthermore large clusters (>10 genes) were more enriched than small clusters in using COXPRESdb (19.0% vs 16.9%, $p = 0.062$) and significantly more enriched for clusters identified using HumanNet (23.7% vs 18.3%, $p = 5.1 \times 10^{-5}$). The major difference which might explain these discrepancies is the very low weight given to gene expression when building the PLN (65) as demonstrated by the negligible change when expression data was excluded from the PLN (Figure 5.7 B). Whereas COXPRESdb and previous studies (80; 81; 90; 93; 94; 173) use exclusively gene expression data and HumanNet combines expression data from

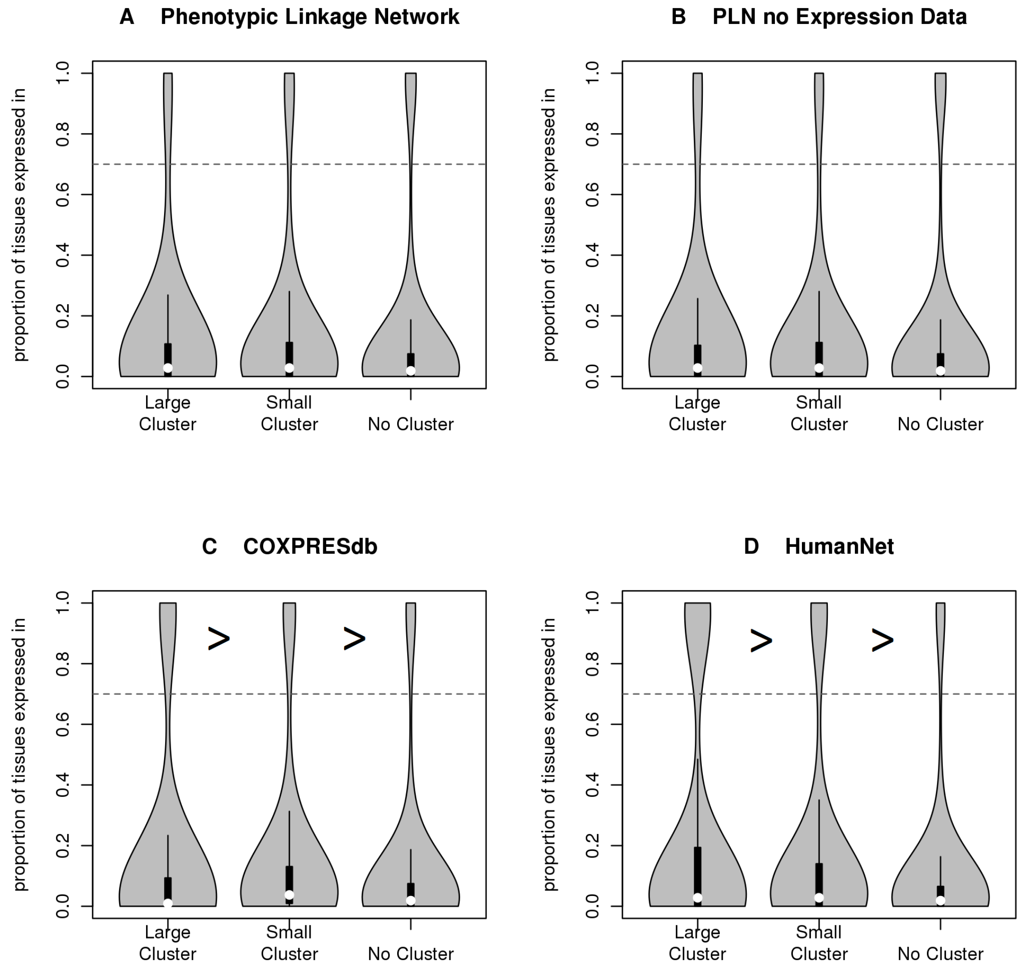


Figure 5.7: Functional clusters are not biased towards housekeeping genes. Expression broadness for genes participating in clusters that were identified using (A) the Phenotypic Linkage Network (PLN), (B) the PLN with excluding the expression data, (C) COXPRESdb co-expression network or (D) the HumanNet integrated network. ">" indicates a significant difference in the proportion of genes which are housekeeping genes (expressed in at least 70% of normal tissues in (128)).

multiple model organisms as well as humans in the construction of the network.

Having found no bias towards housekeeping genes amongst the clusters identified using the PLN, I next considered the tissue-specificity of these clusters. I calculated the proportion of cluster genes expressed in each of the 106 different tissues/cell-types for which expression data was available (Figure 5.8). This revealed only 43 out of 942 (5%) of functional clusters are completely composed of house-keeping genes (all genes were expressed in at least 70% of tissues) furthermore 164 (17%) contained at least one housekeeping gene. Of the remaining 778 clusters without housekeeping genes, 150

(19%) of the functional clusters were tissue specific with all of the genes expressed in the same 1-5 tissues. This rose to 206 (26%) with all of the genes expressed in the same 1-10 tissues (<10% of all tissues/cell-types examined). Finally 173 clusters (18% of all clusters) contained exclusively genes with no detectable expression in any of the tissues. Thus, the PLN identifies clusters of genes sharing similar expression patterns and is not biased towards housekeeping genes.

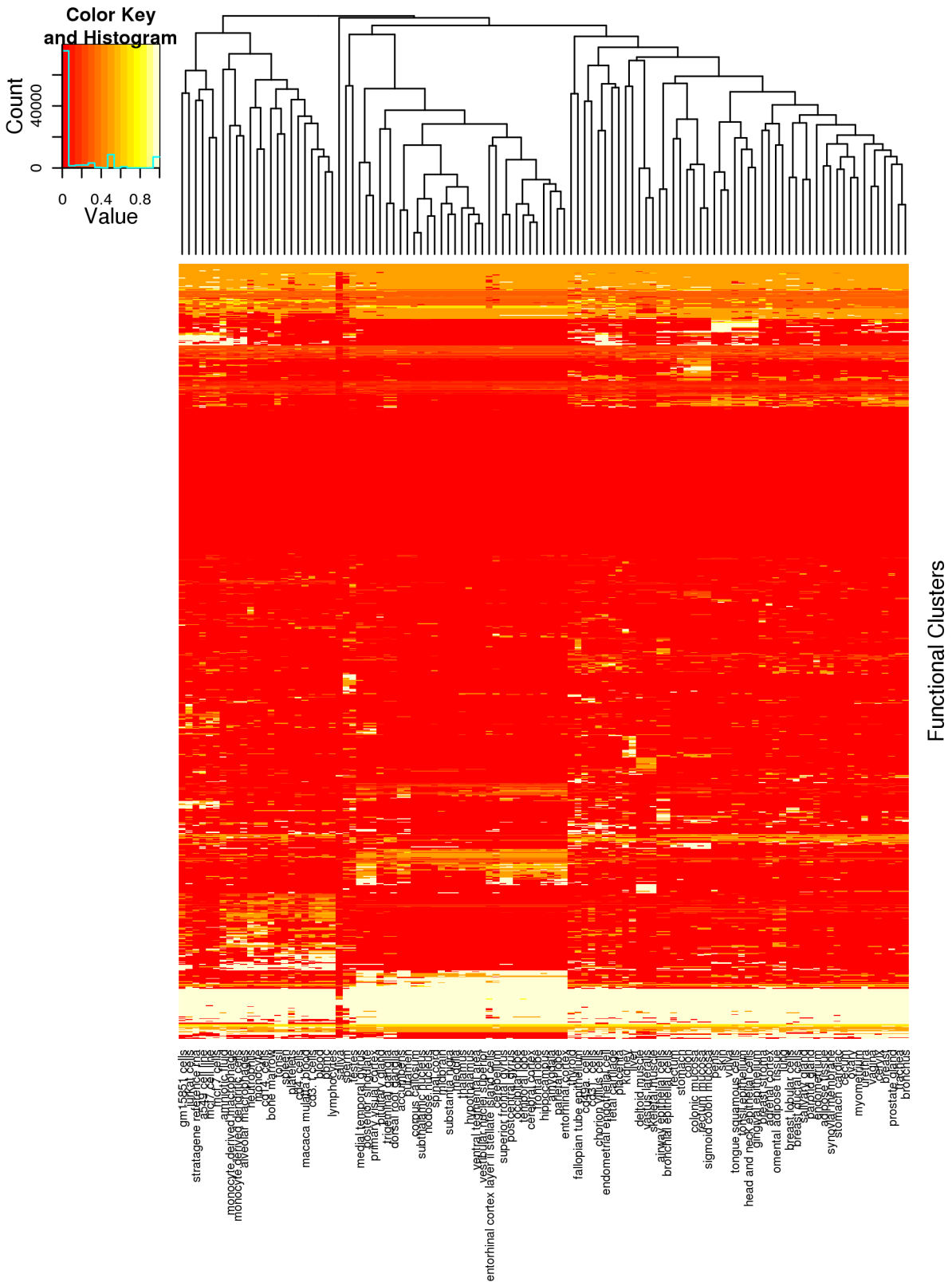


Figure 5.8: Gene expression patterns of functional clusters. I examined whether genes belonging to the same functional cluster showed similar expression tissue-specificity. Each row is a single genome-wide functional cluster. Each column is one of the 106 normal tissues/cell-types with data available. Cells are coloured based on the proportion of cluster genes expressed the respective tissue according to data from (128), white being all genes in the cluster and red being none of the genes.

5.3.3 Genomic context of Genome-wide Functional Clusters

The location of a genomic region within the three dimensional environment of the cell's nucleus can influence the transcription of genes in the region (166; 174). The DNA of a cell is held in a three dimensional globule within the nucleus (175). Some regions are present near the centre of the globule, in the centre of the nucleus, whereas other regions are located at the outside of this globule where it can interact with the protein complexes found on the inside of the nuclear membrane known as the nuclear lamina (166). I examined two recent datasets of the organization of chromatin within the nucleus: i) over 1,000 lamina associated domains (LADs) detected by identifying regions of DNA which associate with Lamin B1, a major constituent of the nuclear lamina (166), and ii) chromatin-chromatin interactions detected using Hi-C, a genome-wide version of chromosome conformation capture (3C) (170) (Figure 5.9) .

Considering the 942 functional clusters identified using the PLN, 30% of the sequence of the genes participating in the clusters was found in LADs; which was significantly higher than the 1,000 sets of genome-wide clusters identified after permuting the node labels in the network ($p = 0.005$, median = 24%, Table 5.1). The number of LAD boundaries, where chromatin transitions from central to lamina-associated, were significantly depleted (one every 1.8Mb versus a median one every 1.4Mb, $p = 0.021$) in the genes in the observed genome-wide functional clusters than those from the permutations. Similarly the genomic regions found between functional cluster genes was more likely to be present in LADs (26% vs 24%, $p = 0.011$) but did not have an unusual number of LAD boundaries compared to inter-cluster-gene regions from network permutation clusters (every 1.1Mb vs every 1.2Mb, $p = 0.137$). Both observed and permutation clusters were less likely to occur in LADs than the genome as a whole (roughly 45%) likely due to their high gene-density. LADs tend to be repressive chromatin environments and it has been shown that chromatin can move in and out of LADs as part of gene activity regulation, which suggests functional clusters' expression is regulated at the chromatin level (176; 177). Since these functional clusters stretch over many megabases, such chromatin level regulation could be an efficient way to coordinate ex-

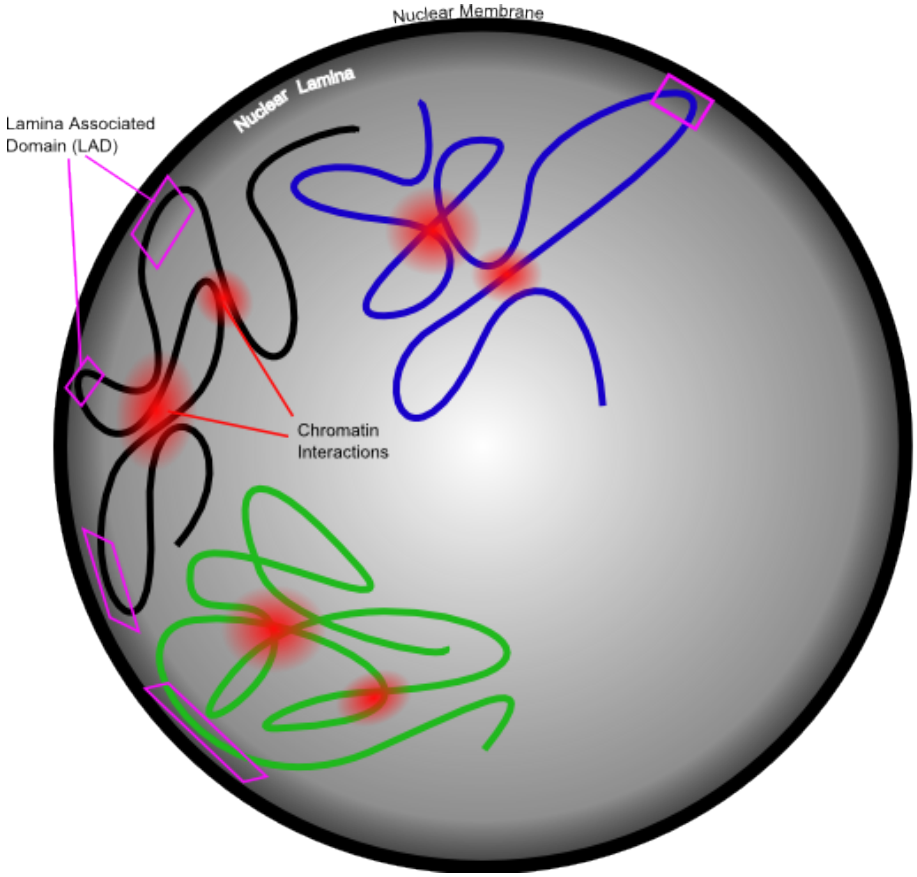


Figure 5.9: Chromatin organization. I tested whether genome-wide functional clusters occur in chromatin regions associated with the nuclear lamina (purple areas), and whether the genes within the same functional cluster are close together in the 3D space in the nucleus such that the chromatin can interact with itself (red). Each coloured line represents a DNA molecule with its associated nucleosomes. Dark grey shaded area is the nuclear lamina composed of many proteins and protein complexes on the inside of the inner nuclear membrane (black circle).

pression of the functionally-related genes.

However, functional clusters identified using the co-expression network (COXPRESdb) were significantly depleted in LADs ($p < 0.05$, Table 5.1). Whereas those identified using HumanNet were not significantly different from the clusters derived from network permutations. A possible explanation for these contradicting results is the bias towards housekeeping genes, which would be less likely to be found in the repressive LADs, exhibited by the COXPRESdb and HumanNet clusters but absent in the PLN clusters (Figure 5.7). However, HumanNet was the most biased towards housekeeping genes yet does not have the lowest proportion of LAD sequence.

Several recent studies have revealed more details of the three dimensional structure

of the genome within the cell using high-throughput chromatin conformation capture (HiC) (170; 178). I obtained the chromatin-chromatin interactions within each chromosome summed over 100kb bins from (170). Using a linear imputation along bins, I found a significantly higher level of chromatin interactions per base between the observed functional cluster genes than between network-permutation cluster genes ($p = 0.025$, Table 5.1). In addition, there were significantly stronger interactions between functional cluster genes and the sequence between them than in the permutations ($p = 0.029$). This again suggests some kind of co-regulation may be occurring since subnuclear localization has been associated with transcription (174) and may be important for the functioning of enhancers (179). However, there was no significant interactions for clusters identified using HumanNet or COXPRESdb (Table 5.1).

Epistasis between functional cluster genes could result in higher levels of linkage disequilibrium and lower levels of recombination between genes belonging to the same functional cluster due to selection to keep epistatically interacting alleles together (180). Using the sex-averaged recombination map from deCODE(169), I determined the average recombination rate per base between functional cluster genes is significantly higher (46% of genome average vs 44% of genome average $p = 0.03$) than that of the network permutation functional clusters but still lower than genome average. This does not support the existence of epistasis between functional cluster genes, however, recombination rate is highly dependent on DNA sequence properties such as GC (181). Sequence length randomizations, random segments of roughly equal length, produced using GAT (182) while controlling for GC content had a similar recombination rate to the original genome-wide functional clusters ($p = 0.314$).

Finally I considered the presence of segmental duplications, low-copy repeats >1kb in size and >90% sequence identity, within the sequence between the genes belonging to each functional cluster. Segmental duplications increase the chance of non-homologous recombination which can result in large duplications and deletions (CNVs) which may cause disease (158; 167; 168). The sequence between genes belonging to

each functional cluster was significantly enriched in segmental duplications compared to clusters resulting from permuted networks, with on average one segmental duplication every 1.5-1.8 kb vs the expectation of one every 2.2-2.7kb (Table 5.1). This suggests functional clusters are particularly prone to CNV mutations thus increasing their potential contribution to disease.

Table 5.1: Genomic Context of Functional Clusters. Proportion of sequence falling within a Lamina Associated Domain (LAD), Fequency of Segmental Duplication (SegDup), average recombination rate (Recomb), Strength of chromatin interactions (Observed/expected) between genes and between genes and the spanned sequence. Expected values from median of permutations in parentheses.

Network	Sequence	LAD	P	SegDup	P	Recomb	P	Chromatin Interactions
COXPRESdb	Cluster Genes	18.6% (22.3%)	0.037	every 255 bp (245 bp)	0.300	0.31 (0.37)	0.001	low (p = 0.152)
	Spanned	26.4% (21.1%)	0.001	every 1,566 bp (2,737 bp)	0.001	0.47 (0.43)	0.003	low (p = 0.226)
HumanNet	Cluster Genes	21.4% (24.4%)	0.069	every 238 bp (268 bp)	0.009	0.40 (0.41)	0.262	high (p = 0.266)
	Spanned	25.1% (24.7%)	0.337	every 1,540 bp (2,214 bp)	0.001	0.44 (0.44)	0.351	high (p = 0.096)
Phenotypic Linkage Network	Cluster Genes	30.0% (23.9%)	0.005	every 278 bp (252 bp)	0.037	0.46 (0.41)	0.01	high (p = 0.025)
	Spanned	26.3% (23.5%)	0.005	every 1,807 bp (2,590 bp)	0.001	0.46 (0.44)	0.03	high (p = 0.029)

5.3.4 Pathogenicity of Genome-wide Functional Clusters

Previously, I showed that *de novo* CNVs from patients with developmental disorders tended to contain large functional clusters (Chapter 4). I have now shown that such clusters are common in the genome (Section 5.3.1. I next investigated the pathogenicity of these genome-wide functional clusters by considering to what extent they are affected by CNVs from patients or by benign CNVs from healthy controls.

For pathogenic CNVs, I obtained 626 *de novo* CNVs were from the Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources (DECIPHER) and 426 *de novo* CNVs from collaborators at the Genetic and Epigenetic Networks in Cognitive Dysfunction Consortium (GENCODYS). Patient phenotypes were recorded using the controlled language of the London Dysmorphology Database (LDDB) for DECIPHER and the Human Phenotype Ontology (HPO) for GENCODYS to ensure patients are easily comparable, these could be mapped from one ontology to the other using a file provided on the HPO website (`http://www.human-phenotype-ontology.org/contao/index.php/downloads.html`). For benign CNVs, I obtained 19,295 CNVs from healthy controls from (103).

In addition, some genome-wide functional clusters may contain or span prenatally lethal genes which would lead to mutations of these clusters being unobserved in either healthy or patient populations. To my knowledge there are no lists of human prenatally lethal genes, thus I used genes whose mouse one-to-one orthologs are completely prenatally-lethal when knocked out in mice. Since CNVs are usually haploid, only mouse orthologs which were prenatally lethal when a single copy was knock-out (haplo-insufficient) were included in the set of lethal genes. The Mouse Genome Informatic (MGI) database includes 78 such genes with 1-1 human orthologs. 81 clusters included or spanned at least one gene whose mouse 1-1 ortholog is haplo-insufficiently completely prenatally lethal (here after simply referred to as HISPNL genes).

HISPNL genes are not expected to be seen disrupted in patients or controls since their disruption would prevent the individual being born. I examined whether HISPNL genes tend to be located in genome-wide functional clusters not seen disrupted in any or the CNV datasets. For a genome-wide functional cluster to be "seen" in the CNV datasets at least two genes from the cluster must be affected by a single CNV from DECIPHER, GENCODYS or benign CNV datasets. To ensure my findings were not simply due to seen genome-wide functional clusters being larger and spanning more genes (some of which may be HISPNL genes), I randomly permuted the genes among genome-wide clusters. The observed number of lethal genes or clusters which span lethal genes which were seen in any of the CNV datasets was no different than the expectation based on 10,000 permutations ($p > 0.1$). This may be due to the small number of HISPNL genes that have been observed (only 78 unique genes) or due to errors when inferring this property from mice to humans, human development may be more robust than that of mice thus what is completely prenatally lethal in mice may be merely disease causing in humans.

Large CNVs can hit many functional clusters and many of these hits are likely to be incidental and have no relation to the pathogenicity of the CNV. The functional cluster with the most genes affected by the CNV is most likely to be responsible for the CNV's pathogenicity since it is the most disrupted; and I have shown previously that the largest functional cluster is most significantly unusual within *de novo* CNVs (Figure 4.4 in Chapter 4). Therefore I defined a cluster 'hit' by each CNV if it was the largest cluster in the CNV (and include at least two genes in the cluster).

Considering only the largest cluster affected by each CNV, 242, 175, and 95 different clusters of functionally-related genes were hit by DECIPHER, NIJMEGEN and control CNVs respectively. Within these sets of affected clusters of functionally-related genes, there is significant overlap between clusters of functionally-related genes hit by DECIPHER *de novo* CNVs and NIJMEGEN *de novo* CNVs; 76 clusters were affected by CNVs in both DECIPHER and NIJMEGEN but not control CNVs compared to 40

expected under a binomial model ($p = 2.7 \times 10^{-6}$, Figure 5.10), suggesting these clusters are particularly important in developmental disorders. There are 26 clusters of functionally-related genes hit by CNVs in all three CNV datasets which is significantly more than the five expected by chance ($p = 4.13 \times 10^{-11}$). When I considered all overlaps by benign CNVs affecting at least one gene in a functional cluster (Figure 5.11) then they affect 169 of the 315 clusters 'hit' by the *de novo* CNVs, more than the 143 expected under a binomial model ($p = 0.002$). This overlap between functional clusters affected by *de novo* and benign CNVs may be explained by particular functional clusters located in regions prone to the formation of CNVs due to the presence of segmental duplications as I discussed above.

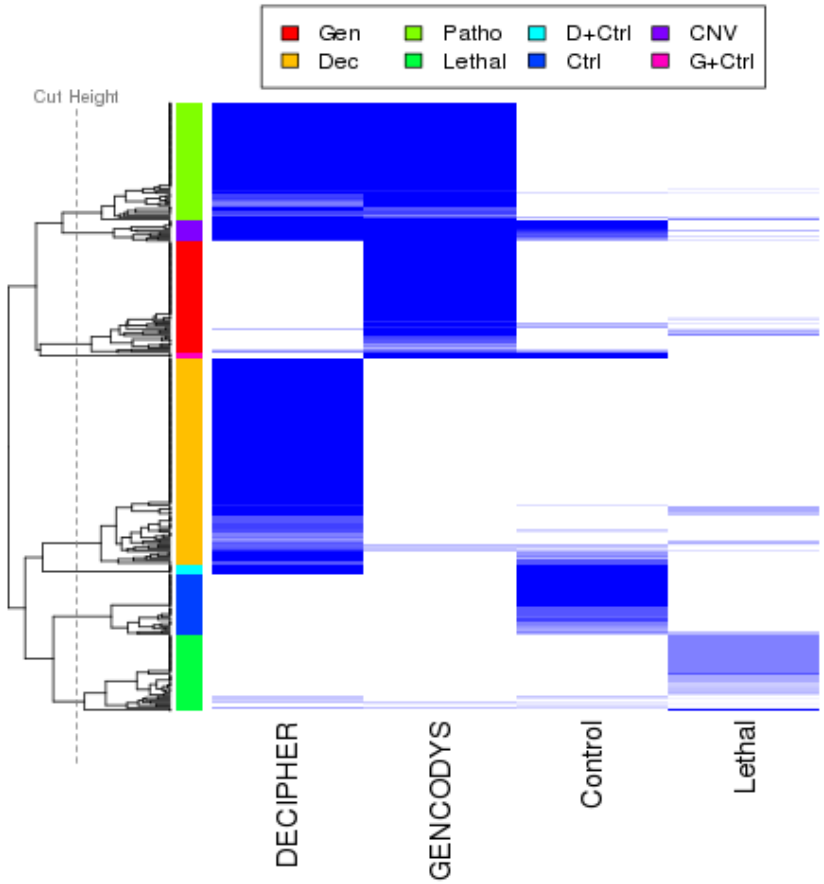


Figure 5.10: Grouping Functional Clusters by Largest Hit per CNV. Each of the 398 genome-wide functional clusters affected by any of these datasets are represented as a row. Intensity of blue indicates the proportion of cluster gene 'hit' by each dataset (bright blue = 100% of genes affected). A CNV 'hit' was defined as the largest number of genes (minimum 2) belonging to a cluster affected by a single CNV in each dataset. Coloured bar on the left shows the groupings identified using average-linkage hierarchical clustering when cut to produce 8 clusters (dashed grey line).

However, *de novo* CNVs affect more genes in the pathogenic clusters than the apparently benign CNVs; for the 54 clusters which were the largest cluster hit by a *de novo* CNV and the largest cluster hit by a benign CNV, the *de novo* CNVs affected on average 2.8 more genes within the cluster than the benign CNV ($p = 0.0006$, t-test); and for the 169 clusters hit by *de novo* CNVs and overlapped at all by a benign CNV the *de novo* CNV affected on average 2.2 more genes within the cluster than the benign CNV ($p = 1.1 \times 10^{-11}$, t-test). Furthermore, of the 315 cluster hit by a *de novo* CNV in 260 (83%) all genes in the cluster were affected by one of the *de novo* CNVs. In contrast, only 37 of the 95 (40%) of clusters hit by benign CNVs were contained in by at least one benign CNV. Considering all CNVs in each dataset 67% of DECIPHER *de novo* CNVs completely contain their largest cluster, similarly 72% of NIJMEGEN *de novo* CNVs affect all genes in their largest cluster. Control CNVs were significantly less likely to completely contain their largest cluster of functionally-related genes with only 29% of CNVs doing so ($p < 10 \times 10^{-15}$). This is consistent with the model where functional clusters contribute to CNV pathogenicity through compounding deleterious effects of each gene, thus it is only when the majority of the cluster has been affected by the CNV that the respective patient exhibits a disease phenotype.

To further examine this model I considered the extent to which different functional clusters were affected by CNVs obtained from a case-control study of CNVs in patients with developmental disorders (4). Here I considered all overlaps between CNVs and functional clusters (at least 1 gene affected) and considered the average proportion of cluster genes affected by each of these overlaps (Figure 5.11). Again it is evident that CNVs found in patients with developmental disorders (DEC, GEN, EIC-P) affect a greater proportion of genes in each functional cluster than CNVs found in controls (CTRL, EIC-C). However, since the case-control CNV dataset (EIC) did not distinguish the inheritance of CNVs found in patients there is much more noise in that dataset than among the *de novo* CNVs. It is important to note that CNVs found in patients larger than those found in controls (4) thus this could be simply a result of the size of the CNVs, however in Chapter 4 I showed that functional clusters predict pathogenicity

of CNVs beyond the effect of CNV size and are a particularly strong predictor in very large CNVs (Table 4.2, 4.1).

Pathogenicity of clusters was determined by grouping clusters together based on the extent they were hit by each of these CNV datasets and the extent to which they span HISPNL genes. Average-linkage hierarchical clustering was used to form the groups since it is a simple, standard clustering method. To ensure all were measured on a common scale from 0 to 1 each functional cluster was scored by the proportion of all cluster genes which were 'hit' by CNVs from each dataset and the ratio of the number of HISPNL genes spanned by the cluster divided by the total number of cluster genes. Cutting the resulting dendrogram at the longest branch resulted in eight different groups (Figure 5.10): clusters containing HISPNL genes but not 'hit' by CNVs (Lethal), clusters affected by just CNVs from DECIPHER or GENCODYS (Dec, Gen respectively) or by both set of *de novo* CNVs but not controls (Patho), those affected by all three CNV sets (CNV) or a combination of one set of *de novo* CNVs and controls (D+Ctrl, G+Ctrl) or by just controls (Ctrl). The first three of the groups (Lethal, Dec, Gen, Patho) are likely to be pathogenic functional clusters whereas the Ctrl group is least likely to be pathogenic. I tested the robustness of this clustering by repeating it using Ward's minimum variance hierarchical clustering, cutting the tree to produce eight clusters to match the results from average-linkage clustering resulted in nearly identical groups, only 24 of 398 clusters were placed in different groups using this other method. Notably, the majority of genome-wide functional clusters, 544 of the 942, were not place into any of these categories since they were not 'hit' by any of the CNVs nor contained HISPNL genes.

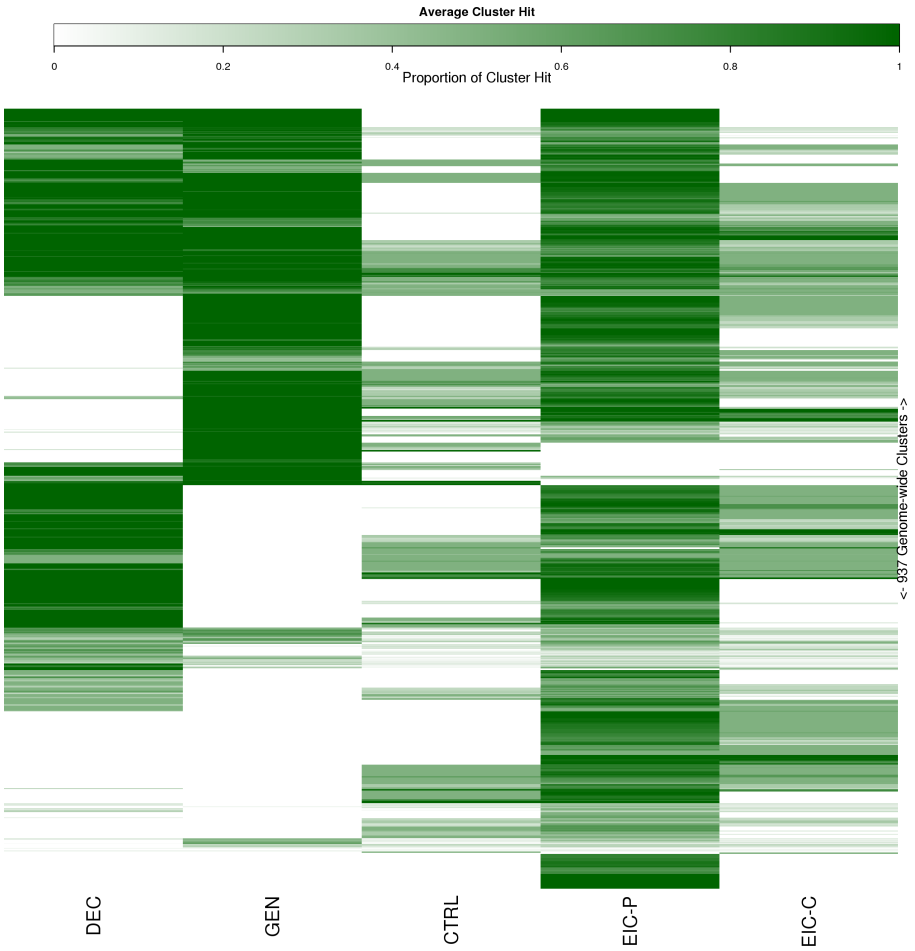


Figure 5.11: Average proportion of each cluster affected by CNVs from various datasets. Any amount of overlap of the cluster by a CNV was considered. Each row represents a single genome-wide functional cluster. DEC = DECIPHER *de novo* CNVs, GEN = GENCODYS *de novo* CNVs, CTRL = CNVs from healthy individuals (103), EIC-P & EIC-C = CNVs identified in patients and in controls, respectively, from (4).

Defining a Pathogenicity Score

In order to predict which of the remaining 544 genome-wide functional clusters, which were not the largest cluster hit by a single CNV nor spanned HISPNL genes, were likely to be pathogenic I identified functional annotations which were unevenly distributed between clusters assigned to the different groups defined by average-linkage hierarchical clustering (Figure 5.10). I considered both Gene Ontology (GO, (34)) and mouse phenotypes (MGI, (36; 37)) annotations, however to reduce the harshness of the multiple-testing correction applied, only 398 genome-wide functional cluster containing 2,473 unique genes participated in the groups, only terms present in GOSlims (34) and the 29 overarching mouse phenotypes (36; 37) were tested. All genes from all

Table 5.2: Functional annotations significantly unevenly distributed across genome-wide functional cluster categories identified in Figure 5.10.

Term	Description	P-Value
GO:0005622	intracellular	4.450×10^{-7}
GO:0003677	DNA binding	2.642×10^{-6}
GO:0043226	organelle	7.711×10^{-9}
GO:0005623	cell	1.418×10^{-6}
GO:0005575	cellular component	0.0003092
GO:0048856	anatomical structure development	4.759×10^{-7}
GO:0001071	nucleic acid binding transcription factor activity	0.000211
GO:0005634	nucleus	2.541×10^{-9}
MP:0005389	reproductive system phenotype	0.0001208
MP:0005384	cellular phenotype	0.001230
MP:0005387	immune system phenotype	5.014×10^{-5}
MP:0005397	hematopoietic system phenotype	0.0006269
MP:0005378	growth/size/body phenotype	0.0002116
MP:0005382	craniofacial phenotype	0.0005107
MP:0010768	mortality/aging	0.0002706
MP:0005380	embryogenesis phenotype	0.001514

genome-wide functional clusters assigned to the same group were combined to create unique gene lists for each group. One group (G+CTRL) was excluded due to its small size (only 3 functional clusters with a total of 9 unique genes) which is insufficient to detect enrichments. First I considered how terms were distributed between the seven remaining groups using a Fisher's Test (Table 5.2). Eight GO terms and eight MGI terms were significantly unevenly distributed between the seven groups after a Bonferroni correction for multiple tests. All of these terms were most common in clusters which span HISPNL genes and least common in clusters belonging to any of the groups hit by benign CNVs (Figure 5.12). Thus these significant terms are associated with functional cluster pathogenicity.

Since the three groups containing clusters hit predominantly by patient CNVs (called: Patho, Gen, Dec in Figure 5.10) have a similar proportion of genes with the relevant annotations, as do the three groups hit predominantly by benign CNVs (called: CNV, Ctrl, D+Ctrl in Figure 5.10), these groups were combined together into 'pathogenic' and 'benign' genome-wide functional clusters (clusters belonging to the Lethal group

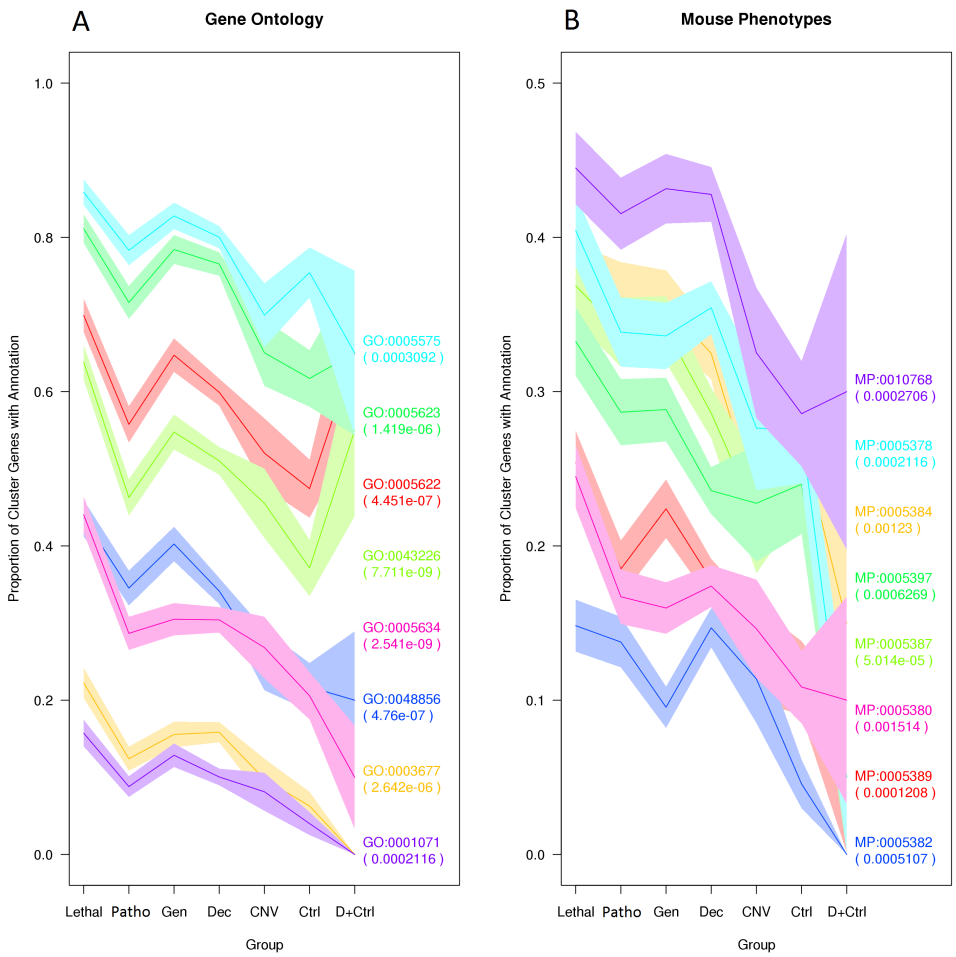


Figure 5.12: Distribution of functional annotations by genome-wide functional cluster groups. The proportion of genes in genome-wide functional clusters in each group from (Figure 5.10) with each of the significantly (after Bonferroni correction) unevenly distributed GO (A) or mouse phenotype (B) annotations. Almost all terms show a consistently decreasing trend with Lethal clusters the most enriched and clusters affected in controls least enriched. Shaded areas indicate the standard errors of the proportion given the total unique genes in the category. Values in brackets are the uncorrected p-values for each term.

were removed). Using Fisher's exact test, I identified which functional annotations were significantly enriched in 'pathogenic' functional clusters after a Bonferroni multiple testing correction. I then defined a 'pathogenicity score' equal to the average proportion of cluster genes annotated with each of the 3 GO and 6 MGI functional annotations found significant between both 'pathogenic' and 'benign' functional clusters as well as between the 7 groups identified using average-linkage hierarchical clustering (Table 5.3).

As an additional control, I also used all genes spanned by a cluster (including the cluster genes themselves) rather than just genes belonging to the cluster, which I expect to

give a noisier, less useful pathogenicity score. As above, enrichments of GOSlims and over-arching mouse phenotypes were determines using Fisher's test with a Bonferroni multiple testing correction. The 16 GO and 8 MGI functional annotations found significant between both 'pathogenic' and 'benign' functional clusters as well as between the 8 groups identified using average-linkage hierarchical clustering were used in the score (Table 5.3). A larger number of terms were significant when all spanned genes were considered compared to when just genome-wide cluster genes were considered since the larger number of spanned genes increased the power of the test. To validate this functional-annotation-based measure of pathogenicity I employed an additional independent set of CNVs (EIC-CNVs) from patients with developmental disorders and healthy controls from (4). The inheritance of the CNVs were not investigated for the EIC-CNVs, thus they contain both inherited and *de novo* CNVs unlike the DECIPHER and GENCODYS CNVs used to define the pathogenicity score. In all 13,798 of 58,012 (24%) patient CNVs 'hit' a genome-wide functional cluster compared to 380 of 886,767 (0.4%) control CNVs (Figure 5.13). This difference reinforces the association between functional clusters and pathogenic CNVs. In addition, the genome-wide functional clusters affected by patient EIC-CNVs had a significantly higher pathogenicity score (based on just genome-wide functional cluster genes) than those hit by control EIC-CNVs (Figure 5.13 A) ($p = 5.5 \times 10^{-31}$). When the pathogenicity score based on all spanned genes was used there was a still significant difference between clusters hit by patient EIC-CNVs and those hit by control CNVs but the difference was much smaller and less significant ($p = 0.0096$) than when only cluster genes were used to define the pathogenicity score (Figure 5.13 B). The all-spanned genes score was less strongly associated with case-CNVs despite combining information from more terms demonstrating that the pathogenicity score is not simply an artefact of study bias. Thus, I have validated the functional enrichments identified among putatively pathogenic functional clusters in an independent CNV dataset.

Table 5.3: Functional annotations defining the pathogenicity score. Multi-Group is a Fisher's test on all 8 categories, Case-Control is a Fisher's test after grouping DEC, GEN, Patho together to be 'Cases' and Ctrl, D+Ctrl, CNV together to be 'Controls'

Score	Term	Description	P-Values	
			Case-Control	Multi-Group
Only	GO:0003677	DNA binding	0.0001453	2.642×10^{-6}
Cluster	GO:0005623	cell	5.332×10^{-6}	1.419×10^{-6}
Genes	GO:0048856	anatomical structure development	3.968×10^{-6}	4.760×10^{-7}
	MP:0005378	growth/size phenotype	0.001303	0.0002116
	MP:0005382	craniofacial phenotype	0.001065	0.0005107
	MP:0005384	cellular phenotype	0.0001825	0.001230
	MP:0005387	immune system phenotype	0.0007733	5.014×10^{-5}
	MP:0005389	reproductive system	0.0003938	0.0001208
	MP:0010768	mortality/aging	6.730×10^{-6}	0.0002706
All	GO:0003677	DNA binding	0.0002118	1.512×10^{-8}
Spanned	GO:0005575	cellular component	4.912×10^{-19}	2.935×10^{-23}
Genes	GO:0005622	intracellular	7.894×10^{-12}	4.619×10^{-14}
	GO:0005623	cell	9.24×10^{-20}	5.296×10^{-28}
	GO:0005737	cytoplasm	1.011×10^{-8}	3.353×10^{-10}
	GO:0005886	plasma membrane	1.244×10^{-5}	2.311×10^{-7}
	GO:0006950	response to stress	3.496×10^{-6}	8.107×10^{-5}
	GO:0007165	signal transduction	1.499×10^{-5}	3.246×10^{-5}
	GO:0007267	cell-cell signaling	0.0003146804	3.440×10^{-5}
	GO:0008150	biological process	3.205×10^{-12}	2.577×10^{-14}
	GO:0030154	cell differentiation	3.256914e-5	2.889×10^{-5}
	GO:0034641	nitrogen compound metabolic process	0.0002968	3.000×10^{-6}
	GO:0043226	organelle	1.479×10^{-7}	3.656×10^{-10}
	GO:0043234	protein complex	1.838×10^{-5}	0.0001026
	GO:0048856	anatomical structure development	8.010×10^{-9}	2.554×10^{-7}
	GO:0050877	neurological system	0.0003422	1.438×10^{-6}
	MP:0003631	nervous system phenotype	1.331×10^{-7}	1.036×10^{-9}
	MP:0005378	growth/size phenotype	1.130×10^{-5}	8.237×10^{-6}
	MP:0005382	craniofacial phenotype	0.0008094	0.0003704
	MP:0005384	cellular phenotype	6.331×10^{-5}	0.0007735
	MP:0005386	behavior/neurological	4.633×10^{-7}	2.323×10^{-6}
	MP:0005387	immune system phenotype	2.967×10^{-6}	0.0006303
	MP:0005388	respiratory system	4.766×10^{-6}	0.0005436
	MP:0010768	mortality/aging	8.212×10^{-11}	3.477×10^{-11}

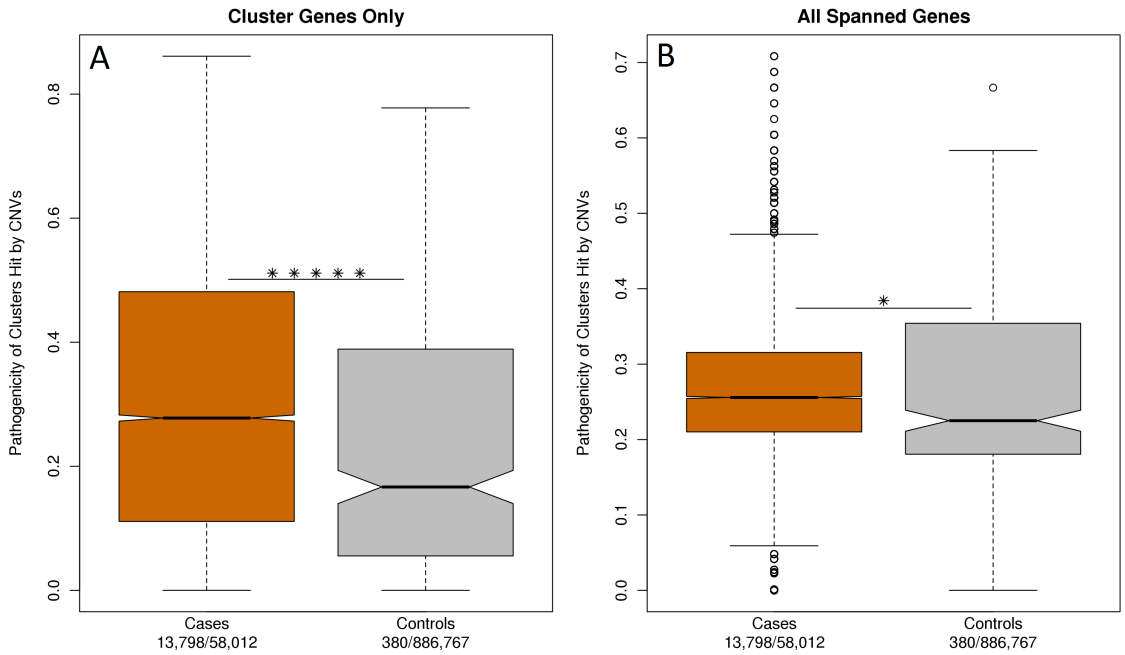


Figure 5.13: Validation of the pathogenicity score with an independent CNV dataset. The pathogenicity score as calculated using cluster genes only (A) or all spanned genes (B) of genome-wide functional clusters completely contained within CNVs from patients or controls, weighted by the number of CNVs which completely overlap the cluster. Stars indicate CNVs from cases (orange) contain clusters with a significantly higher pathogenicity score than controls (two-sided Wilcox-rank-sum test), one star = $p < 0.05$, two stars = $p < 0.005$, three stars = $p < 0.0005$ etc. . . up to a maximum of 5 stars. The number of CNVs of each type which 'hit' any cluster (affect at least two genes in the cluster) is noted in the figure margin.

Predictive ability of Pathogenicity Score

Next I attempted to use the pathogenicity score (based on only cluster genes) to predict pathogenic CNVs. Three different predictors were used in three different pathogenicity tests (Figure 5.14). First I predicted the 625 *de novo* CNVs from DECIPHER from 2,464 inherited CNVs from the same database, this reduces biases due to different study designs and calling technology/procedure between different datasets (17). Similarly I predicted the 426 *de novo* CNVs from the 639 inherited CNVs in GENCODYS (18), since this dataset was called using the same array platform and quality control procedure it avoids biases due to technical differences. Since inherited CNVs have been subjected to selection and most developmental disorders occur in families with no history of the disorder, *de novo* CNVs should be more likely to be pathogenic and have a larger/more deleterious effect than inherited CNVs (12; 14). Finally I predicted the 58,012 CNVs identified in patient from the 886,767 CNVs from controls from the EIC-

CNV dataset (4).

Predictions were made by ranking all CNVs by their score and predicting the top ranked CNVs as *de novo* or from patients respectively. Predictions were made at each rank and the number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) were calculated. Predictions were evaluated using the area under the precision (TP/(TP+FP)) versus recall (TP/(TP+FN)) (Figure 5.14). For comparison, I also predicted pathogenicity using the total length of CNVs as well as permuting CNV ranks.

Both functional clustering (MaxCluster: size of largest cluster) and pathogenicity (MaxC-Score:pathogenicity score of the largest cluster) are much better than random ranking at predicting CNV pathogenicity. However, neither is as good as total CNV length in bp.Combining both pathogenicity and clustering, by multiplying the pathogenicity of the largest cluster by the number of genes from that cluster affected by a CNV (MaxCSizeScore), performed much better than either measure alone and performed only slightly worse than CNV length (Figure 5.14). Two disadvantages of using functional clustering or my pathogenicity score as opposed to CNV length is that i) only those CNVs which affect a functional cluster can be scored, this is highlighted by the EIC-CNVs where less than 20% of the CNVs affected at least two genes in a functional cluster leaving the remaining 80% tied for the lowest rank and ii) only information about protein-coding genes could be used, which ignores many genomic elements which might contribute to developmental disorders such as non-coding RNAs, microRNAs, enhancers and other regulatory sequences. This second limitation might be alleviated as more functional information becomes available for these other genomic elements.

The poor performance of functional clustering alone in predicting CNV pathogenicity agrees with my earlier results (Table 4.2, 4.1 in Chapter 4) which showed that the presence of a functional cluster was a weak but significant predictor of pathogenicity

when the total number of CNV genes was included in the model and only became a strong predictor of pathogenicity when restricting to relatively large CNVs affecting at least 10 genes.

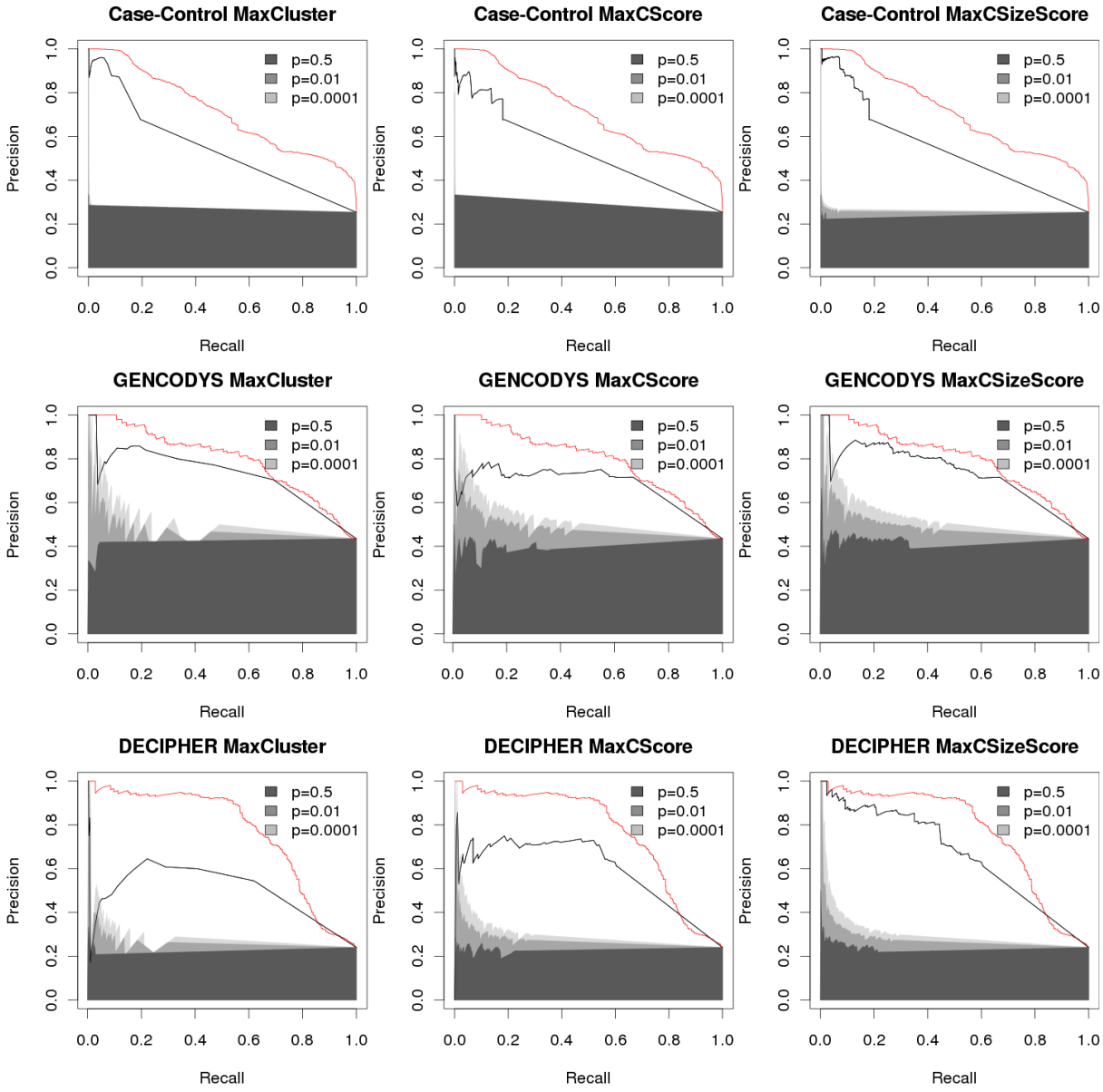


Figure 5.14: Functional clustering predicts pathogenicity of CNVs. Maximum cluster size (left column) the pathogenicity score of the largest cluster (centre column) and the product of these two values (right column) were used to predict pathogenic CNVs from more benign CNVs. (top row) CC-CNVs from patients were considered pathogenic and CC-CNVs from controls were considered benign. (middle row) *de novo* GENCODYS CNVs were considered pathogenic and inherited GENCODYS CNVs were considered benign. (bottom row) *de novo* DECIPHER CNVs were considered pathogenic and inherited DECIPHER CNVs were considered benign. Shaded areas give the probability of observing in randomly permuted rankings. Black line is the result of the respective ranking. Red line is the result of ranking by total CNV length in bp.

5.3.5 Functional clusters and phenotype

Finally, I looked at the contribution the genome-wide functional clusters to the specific phenotypes patients exhibited. Genome-wide functional clusters could be used to bring together patients with similar but not necessarily overlapping CNVs which are affecting the same biological function and thus resulting in similar phenotypes. Thus I examined the similarity of phenotypes exhibited by patients with CNVs affecting the same functional cluster in both DECIPHER and GENCODYS *de novo* CNV sets. To increase power both DECIPHER and GENCODYS *de novo* CNVs were combined into a single dataset. DECIPHER patient phenotypes were recorded using LDDB terms, whereas GENCODYS patient phenotypes were recorded using HPO terms; to make the phenotypes comparable I used an existing mapping file from the HPO website (`http://compbio.charite.de/svn/hpo/trunk/src/mappings/`) to map the LDDB terms used by DECIPHER to HPO, any phenotypes (from either ontology) for which a mapping did not exist were excluded from analysis. This may also reduce the missing data problem in DECIPHER, since less well known phenotypes (which are less likely to have been examined by all clinicians) are less likely to be covered by both ontologies.

I have shown previously that the largest cluster within a CNV is the most significantly large within *de novo* CNVs (Figure 4.4C in Chapter 4). In addition, the functional cluster with the largest number of genes affected by the CNV is most likely to be responsible for the CNV's pathogenicity since it is the most disrupted. Therefore, I defined a cluster 'hit' by a CNV if it was functional cluster with the largest number of genes affected by the CNV and at least two genes in the cluster were affected.

First I considered whether CNVs which hit the same genome-wide functional cluster are found patients with common phenotypes. To control for the distribution of phenotype frequencies and the number of CNVs hitting each cluster, I permuted the phenotypes assigned to each patient while preserving the number of unique phenotypes per

patient. On average over all functional clusters and all phenotypes, each phenotype is common to 30% of CNVs hitting the same functional cluster, which was significantly higher than the 29% for the 1,000 phenotype-permutations ($p = 0.002$).

However, CNVs which hit the same cluster may or may not affect the same genes and affecting the same genes or the same known disease genes (from the Online Mendelian Inheritance in Man(112) - OMIM) may be the cause of the shared patient phenotypes. To examine this question, *de novo* CNVs were considered on a pairwise basis rather than considering all CNVs affecting the same functional cluster at once. CNV pairs were grouped into six categories (Figure 5.15): i) those which hit the same cluster and at least one of the same genes (Cluster-and-Genes), ii) those which hit the same cluster and at least one of the same OMIM disease genes (Cluster-and-OMIM), iii) those which hit the same cluster but affect none of the same genes (Cluster-only), iv) those which affect at least one of the same genes but don't hit the same cluster (Genes-only), v) those which affect at least one of the same OMIM morbid genes but don't hit the same cluster (OMIM-only), vi) those which affect none of the same genes nor hit the same cluster (Shared-Nothing, red line in figures).

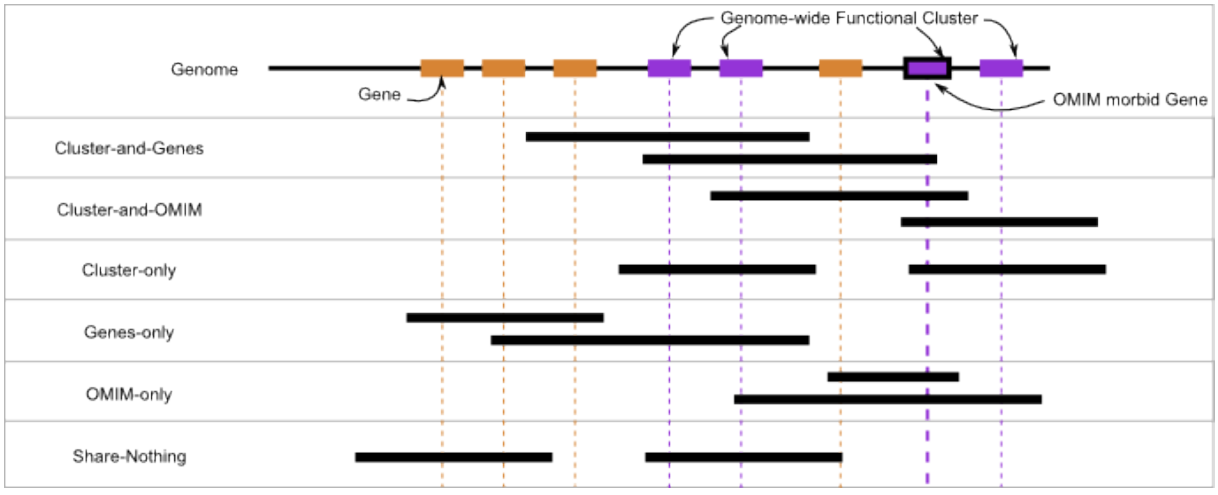


Figure 5.15: Examples of pairwise patient comparison categories. Beaded string represents the genes along the genome with purple indicating a functional cluster. The black outlined gene is an OMIM morbid gene. Black segments represent *de novo* CNV for simplicity one CNV per patient is depicted.

Phenotypic similarity between pairs of patients was calculated using the Goodall3 met-

ric (110) which weighs the similarity of a shared presence or shared absence of a phenotype according to its overall frequency in the population, thus two patients which both have a rare phenotype (eg. Hypertelorism) or which don't have a common phenotype (eg. Intellectual Disability) would receive a high similarity score. The resulting distributions of similarities among each category described above were compared using a Wilcox rank-sum test.

To ensure results were not due to a bias in the total number of patient phenotypes where *de novo* CNVs which hit the same genome-wide functional cluster are associated with a larger number of patient phenotypes than CNVs which affect the same genes because they affect a larger number of genes; all CNVs which affected less than 2 genes were excluded from the analysis. This was sufficient to ensure CNVs hitting the a genome-wide functional cluster did not have significantly more patient phenotypes than those which do not (Figure 5.16).

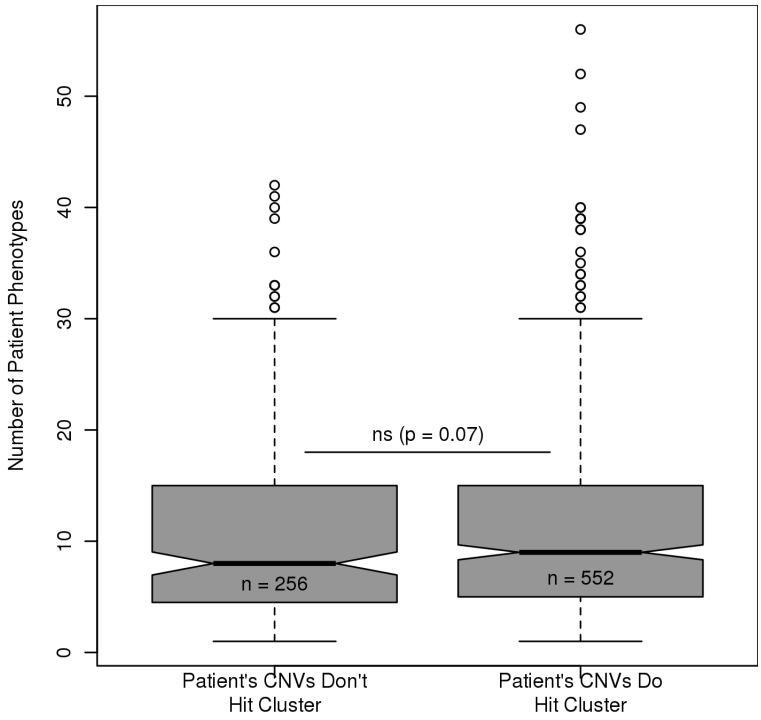


Figure 5.16: CNVs which hit a cluster are not found in patients with significantly more phenotypes. Significance evaluated using a two-sided Wilcox-rank-sum test.

Patients whose CNVs affect the same functional cluster were significantly more phe-

notypically similar than those whose CNVs affect the same genes (Figure 5.17 A). This remained significant even if the patients' CNVs did not affect any of the same genes (Cluster-only, $p = 0.048$). Interestingly affecting an OMIM disease gene (Cluster-and-OMIM or OMIM-only) was not significantly better than sharing any gene (Cluster-and-Genes or Genes-only, $p = 0.843$ and $p = 0.741$ respectively). Patients whose CNVs affected either the same genes or same functional cluster were more phenotypically similar to each other than patients who shared nothing in common; this was significant for those affecting the same functional cluster ($p = 3 \times 10^{-8}$) but not for those affecting only the same genes ($p = 0.19$) . However, using just the *de novo* CNVs there were only 145 cases where patients' CNVs affected the same functional cluster but not the same genes to increase this sample I repeated the analysis adding in inherited CNVs and CNVs with unknown inheritance from each dataset (Figure 5.17 B). While the differences were smaller when all patients with CNVs were included the larger number of patients replicated the finding that patients with CNVs affecting the same functional cluster were the most phenotypically similar. Patients whose CNVs affect only the same functional cluster, not any of the same genes, still were significantly more phenotypically similar than patients whose CNVs affect the same genes but not the same cluster ($p = 0.006$). Again patients whose CNVs affect the same functional cluster were significantly more phenotypically similar than patients whose share nothing ($p = 2.3 \times 10^{-11}$) but not those where only genes were shared ($p > 0.9$).

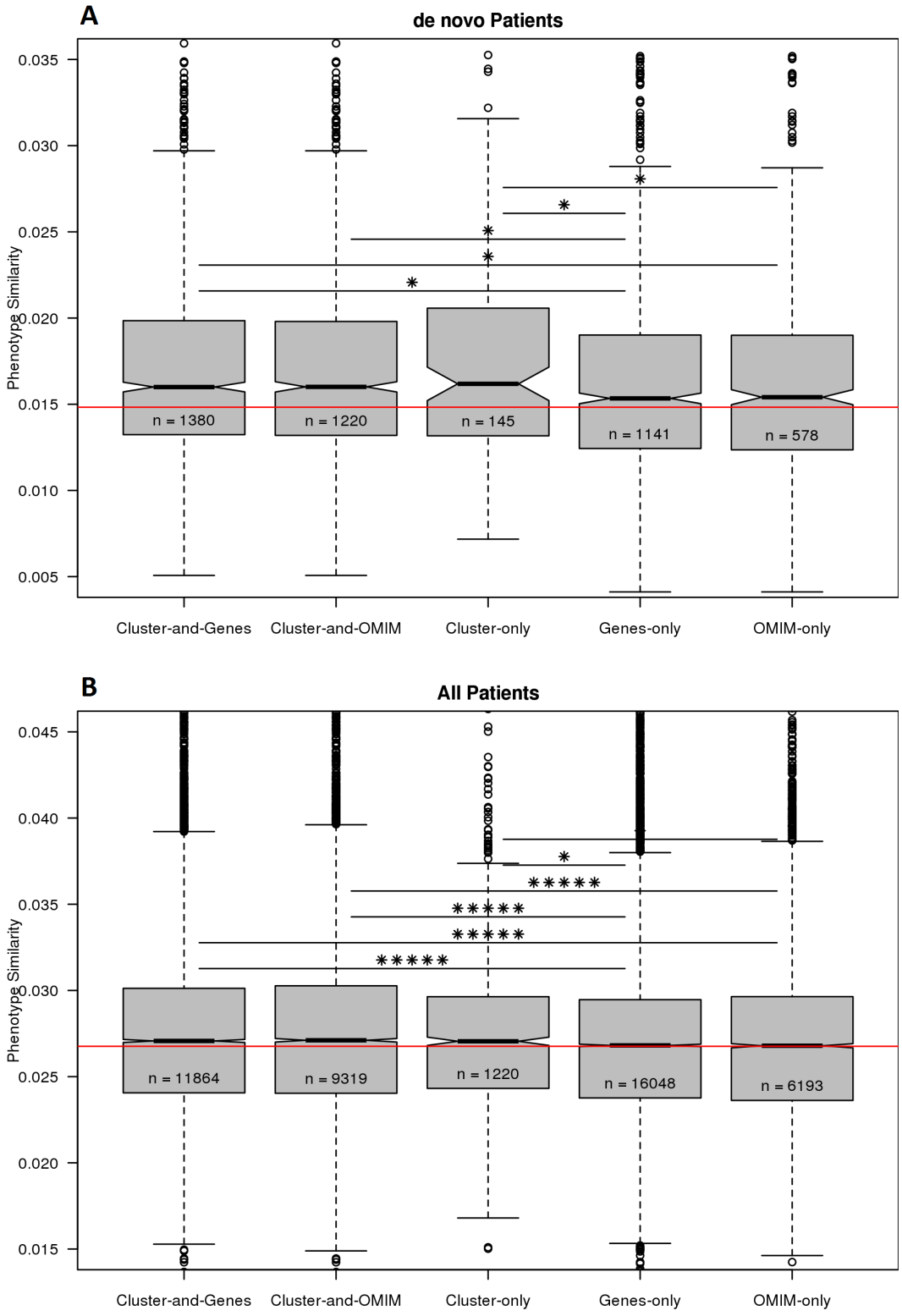


Figure 5.17: Patients whose CNVs hit the same functional cluster are more phenotypically similar than patients whose CNVs affect the same genes. Stars indicate significant differences using a two-sided Wilcox-rank-sum test. One star = $p < 0.05$, two stars = $p < 0.005$, three stars = $p < 0.0005$ etc.. up to a maximum of 5 stars. Red line is the median phenotypic similarity between patients which Share-Nothing. CNVs containing less than two genes were excluded. A *de novo* CNVs pairs. B patient pairs including all types of CNV.

Robustness of Phenotypic Similarity

To ensure these findings are not artefacts of the network or parameters used to define the functional clusters, I replicated the results using two different networks, two different distance thresholds, and two different functional similarity thresholds. (Figure 5.18) shows replications of *de novo* CNV only comparisons an (Figure 5.19) shows replications of all CNV patient comparisons. If paralogs within genome-wide functional clusters are not collapsed to a single copy or if only the original edges of the PLN are used rather than shortest-path similarities, the results stay the same with only slight changes in significance due to slight changes in sample sizes (top row) in both *de novo* and all patient comparisons.

Next I considered the effects of changing the two clustering parameters, the maximum distance between cluster members (D) and the minimum similarity between cluster members (T). Using either the 95th percentile distance threshold (D =1.3 Mb) or the 100th percentile distance threshold (D = 5.0 Mb) or the more permissive similarity threshold (T =top 5%) or more restrictive similarity threshold (T =top 0.1%) does not qualitatively change the results when all patients were considered (Figure 5.19, middle rows) but results in a loss of significance (but the trend remains) in the difference between Cluster-only and Genes-only patient pairs frequently due to a loss of power (small number of patients falling in the Cluster-only category) or due to a small decrease in the difference between the median phenotypic similarity for Cluster-only patient pairs (Figure 5.18, middle rows). However, Cluster-and-Genes patient pairs remain significantly more phenotypically similar than Genes-only patients.

Lastly I replicated the analysis using two other networks: HumanNet(53), an integrated functional network, and COXPRESdb(46), a human co-expression network. Both HumanNet and COXPRESdb replicated the trend of Cluster-and-Genes being significantly more phenotypically similar than patients in the Genes-only category both when all patients were considered and when only patients with *de novo* CNVs were

considered (Figure 5.19, 5.18 bottom row). In addition when only patients with *de novo* CNVs are considered both HumanNet and COXPRESdb replicate the trend (but lose significance) of Cluster-only patients being more phenotypically similar than Genes-only patients. However, when all patients are considered this trend remains and is significant only in HumanNet; whereas COXPRESdb exhibits a significant reverse pattern with Cluster-only patient pairs less phenotypically similar than Genes-only patients.

Thus overall patients with CNVs affecting the same functional cluster exhibited similar phenotypic outcomes more than patients whose CNVs affected only at least one of the same genes. This was robust when the PLN or HumanNet was used to identify the functional clusters but not for functional clusters identified using COXPRESdb.

5.4 Conclusion

In this chapter, I have shown that the human genome is functionally clustered using the PLN and other networks. That those clusters identified with the PLN are not bias towards housekeeping genes and have a high degree of chromatin interactions and are located close to the nuclear periphery. Furthermore I identified functional annotations associated with functional clusters affected by *de novo* but not benign CNVs and validated them in an independent set of case-control CNVs. Finally, I demonstrated that patients whose CNVs affect the same functional cluster are significantly more phenotypically similar that patients whose CNVs affect the same genes.

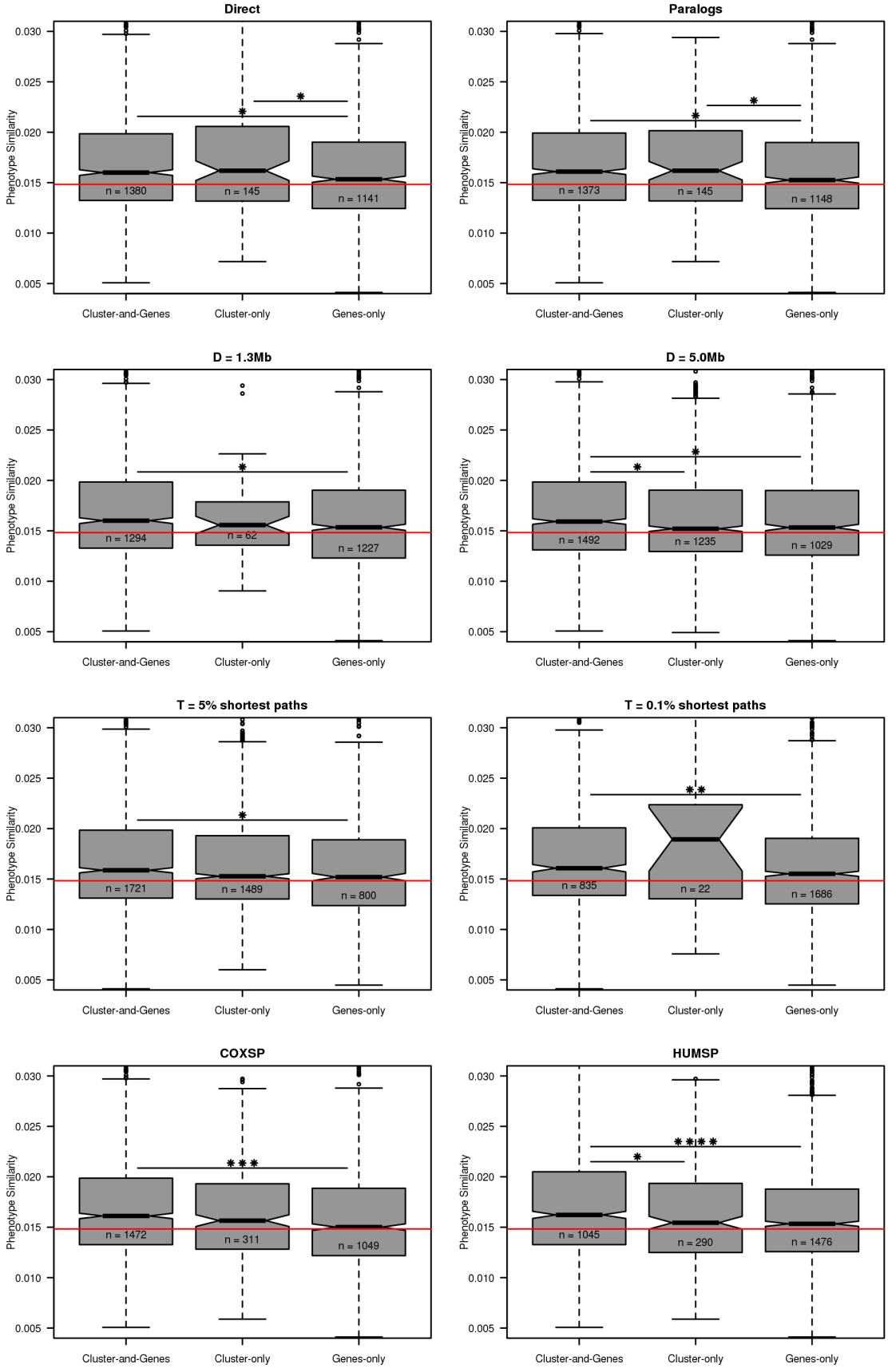


Figure 5.18: Robustness of pairwise patient comparisons considering only *de novo* CNVs. Stars indicate significant differences using a two-sided Wilcox-rank-sum test. One star = $p < 0.05$, two stars = $p < 0.005$, three stars = $p < 0.0005$ etc.. up to a maximum of 5 stars. *de novo* CNVs containing less than two genes were excluded.

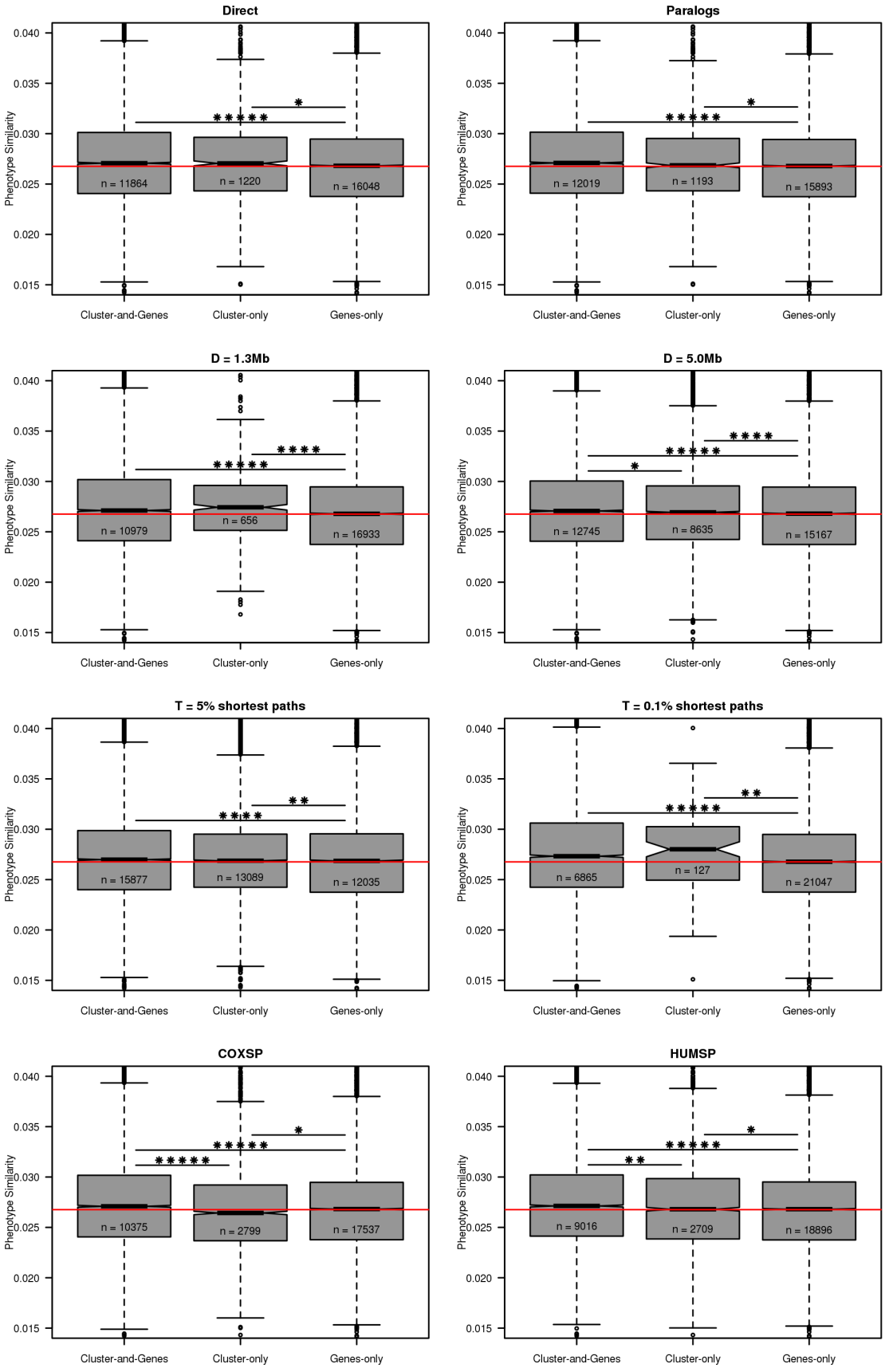


Figure 5.19: Robustness of pairwise patient comparisons considering all recorded CNVs. Stars indicate significant differences using a two-sided Wilcox-rank-sum test. One star = $p < 0.05$, two stars = $p < 0.005$, three stars = $p < 0.0005$ etc.. up to a maximum of 5 stars. *De novo*, inherited or CNVs of unknown inheritance containing less than two genes were excluded.

Chapter 6

Spanning the Phenome

6.1 Introduction

Mouse phenotype information is currently being used on a genome-wide scale. The Mouse Genome Informatics (MGI) database of mouse phenotypes has been used in gene set analysis of microarray data (183) and copy number variant data (20; 143), validating an integrated functional gene network (113), and prioritizing human disease genes (184). Ascertainment bias in this database may inflate the significance of results reported in such studies contributing to the number of false positive results in the biomedical literature (185).

There are two main sources of ascertainment bias: the selection of genes to knock-out in mice (selection-bias) and the choice of phenotypes to examine in the mice (examination-bias). Several mouse phenotyping projects are attempting to eliminate examination-bias by performing a battery of standardized phenotype assays on all the knockout-mice they generate, including the International Mouse Phenotyping Consortium (IMPC, (186)) and Europhenome (187). However, the selection of gene knockouts chosen for these projects were taken from suggestions from the research community, thus is susceptible to selection-bias. Only one project makes any attempt to reduce selection-bias, namely the Knockout Mouse Project (KOMP) which deliberately chooses un-

studied genes in addition to taking suggestions from the academic community (188). Selection-bias results in persistent gaps in our knowledge of gene function and potentially wastes effort gathering information on genes which have already been studied by other groups or which could have been inferred from the results of similar genes (189; 190).

The structural genomics initiative (191) and the encyclopaedia of bacteria & archaea project (192), have demonstrated the value obtained by prioritizing poorly studied distinct organisms or proteins. The information gained by filling in the gaps resulting from ascertainment bias in those fields has been used to guide further experimental work and improve predictive modelling. By deciphering the 3D structure of 752 proteins from diverse protein families the structural genomics initiatives enabled reliable modelling of over 9,000 distinct gene sequences (191). Similarly sequencing 56 phylogenetically diverse bacterial and archaeal genomes revealed 1,768 completely novel gene families (192). Taking a similar approach to characterizing mouse knockouts could be used to improve gene functional predictions, and identify correlated phenotypic classes to aid future phenotyping efforts.

In this chapter I examine the existing mouse phenotype projects for evidence of selection bias and propose a network-based method which identifies poorly studied genes unrelated to previously examined mouse-knockouts to aid prioritization of knockouts for these projects.

6.2 Specific Methods

6.2.1 Mouse Phenotyping Projects

Gene lists for all six mouse phenotyping projects were downloaded on 8 Jan 2014 (Table 6.1). Mouse genes were mapped from HUGO gene symbols to Ensembl IDs as necessary then converted to human genes using 1-1 orthologs from Ensembl70. Enrichments

Table 6.1: Knockout mouse phenotype projects/databases obtained on 8 Jan 2014.

Dataset	No. Genes	Status	Ref	Started
German Mouse Clinic (GMC)	171	Phenotyped	(193)	2001
Knockout Mouse Project (KOMP)	246	Live Mice Produced	(188)	2009 (phenotyping)
Europhenome (Euro)	538	Phenotyping Begun	(187)	2008 (first paper)
Wellcome Trust Sanger Institute Mouse Portal (WTSI)	752	Phenotyped	(194)	Europhenome member
International Mouse Phenotyping Consortium (IMPC-P)	819	Phenotyping Started or Attempt Registered	(186)	2011
International Mouse Phenotyping Consortium (IMPC-M)	3128	Mice Produced	(186)	2011
Mouse Genome Informatics (MGI)	7932	Phenotype(s) Recorded	(36; 37)	2002

were tested against the background of all Ensembl mouse genes with a HUGO gene symbol and a human 1-1 ortholog.

6.2.2 Functional Networks

Five different functional networks were used to detect selection bias within mouse phenotyping projects: a co-citation network (Pubmed) created from the frequency of genes being associated with the same abstract as recorded in `ftp://ftp.ncbi.nih.gov/gene/DATA/gene2pubmed.gz` (126), STRING which combines mathematical predictions and literature-based protein-protein interactions (51), HumanNet a publicly available integrated functional network (53), the human COXPRESdb co-expression network after filtering out all pearson correlations < 0.5(46), and iRefIndex which combines protein-protein interactions from multiple publicly available databases (52) (Table 6.2). All gene names were mapped to Ensembl IDs. HumanNet, COXPRESdb and iRefIndex were used to identify unstudied genes because they had the least contribution of direct literature co-citation.

Table 6.2: Functional Networks used to evaluate selection-bias in mouse-phenotyping projects.

Network	Data Type	No. Genes	Edges	Ref
Pubmed co-citation	Number of curated papers with both genes	25,920	238,760,859	(126)
STRING	Literature-based interactions	18,488	2,037,338	(51)
HumanNet	Integrated functional network	15,931	460,537	(53)
COXPRESdb	Human co-exprssion (Pearson correlation >0.5)	13,129	1,771,841	(46)
iRefIndex	Experimental protein-protein interactions	11,543	931,704	(52)

6.2.3 Selection Bias

For each of the five functional networks (Table 6.2) I calculated the average edge weight between genes included in the same phenotyping project (missing edges were not included), significance was determined using Z-scores with population standard deviation calculated from all the genes with 1-1 orthologs (all sample sizes were sufficiently large for the Central Limit Theorem to hold). I also considered the density of edges in each network between the genes included in each phenotyping project. Edge density is the number of edges present between a set of genes divided by the maximum number of edges possible between those genes (calculated as $\binom{n}{2}$, where n is the number of genes), significance was calculated using a two-sided binomial test.

6.2.4 Experimental Bias

I considered four potential sources of experimental bias which may contribute to the observed selection bias in mouse knockouts: coding sequence (CDS) length, linkage disequilibrium (LD), array expression broadness and RNASeq gene expression level (Table 6.3). CDS length was obtained from Ensembl(105). For linkage disequilibrium I used recombination hotspots from the deCODE project which are defined as regions with at least 10-fold higher recombination rate (sex-averaged) than the genome average (169). The distance to the closest recombination hotspot upstream and downstream (without crossing a centromere or telomere) was calculated for each gene; genes with recombination hotspots both upstream & downstream were assigned the average dis-

Table 6.3: Datasets for examining experimental biases that might explain the selection-bias in mouse phenotyping projects.

Dataset	Description	Source
Coding sequence length (CDS)	Length of longest cDNA for each gene	Ensembl(105)
Recombination hotspots (Recomb)	Regions with at least 10-fold higher recombination rate than the genome-wide average, male and female combined	deCODE(169)
Expression broadness	Proportion of 106 normal human tissues/cell-types with detectable expression	Gene Expression Barcode (128)
Expression RNASeq	Average FPKM across 16 normal human tissues	Illumina Body Map 2.0 downloaded from Gene Expression Atlas (127; 129)

tance. Array expression broadness was calculated as the proportion of the 106 normal human tissues with detectable expression in the Gene Expression Barcode v2 (128) for the Affy HGU133A plus 2 array. Finally RNASeq expression was quantified as the average FPKM across all 16 human tissues in the Illumina Body Map (127), which was downloaded from the Gene Expression Atlas (129; 130). The median value across the genes in each phenotyping project was compared to the distribution of all genes with human 1-1 orthologs and with all human genes using a Wilcox-rank-sum (aka Mann-Whitney-U) test.

6.2.5 Identifying Biological Modules

Biological modules were identified using the Infomap community detection algorithm (69) in each of COXPRESdb, HumanNet and iRefIndex. Infomap is based on efficiently compressing random-walks in the network and is one of the best performing algorithms on large complex networks (132). There is stochasticity in the optimization algorithm used by Infomap to identify the most efficient groupings of nodes in the network, thus the algorithm must be run multiple times to identify the globally optimal partition of the network. Thus, I considered the simple clustering of the network as the best partition as defined by the algorithm over 10,000 runs of the algorithm on the original network. In addition, I considered the consensus clustering obtained using

the Cluster-based Similarity Partitioning Algorithm (135) where the original clustering algorithm (best partition over 10,000 runs of Infomap) is applied to the association matrix, which is obtained from the pairwise frequencies of genes being grouped in the same network over 100 partitions obtained from the optimal partition over 100 attempts of the algorithm. The robustness of the clustering in each network to algorithmic stochasticity was measured using the variation of information between these two partitions (195).

The biological modules were validated by attempting to predict the 8,288 new mouse phenotype annotations (to 5,544 genes) added to the MGI database between February 2012 and January 2014 using just the annotations present in the February 2012 phenotype annotations. Mouse phenotype annotations from both releases were mapped to the 1-1 human orthologs and any terms not present in both releases, due to changes to the Mammalian Phenotype Ontology, were excluded. Predictions were made using a simple majority-rule: if >50% of genes with phenotype annotations within the same module were annotated with a term then that term was predicted to be annotated to all other genes in the module. Prediction quality was summarized using the F-measure (196) which is the harmonic mean of precision and recall (see equations below). The significance was determined by comparing to 1,000 permutations of cluster labels.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{6.1}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{6.2}$$

$$\text{F-measure} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \tag{6.3}$$

where TP = number of true positive predictions (phenotype is predicted and present), FP = number of false positives (phenotype is predicted but not present), FN = number of false negatives (phenotype is present but not predicted).

6.2.6 Transferability of Phenotype Information

I further considered the transferability of phenotype information within clusters of different size. Clusters were binned by the number of genes they contain and the average transferability over all clusters in the bin was calculated and compared to 1,000 permutations of cluster labels. Transferability was evaluated using: i) the F-measure of majority-rule (see above) leave-one-out cross validations using all of the MGI database phenotype annotations and ii) the average semantic similarity of mouse phenotype annotations between genes in the cluster (109). Semantic similarity (see: Methods 2.2.1) was calculated as the average information content (IC) of the highest IC disjoint common ancestors (GraSM) between terms which was combined using the average of the average-best-match (BMA) and the single-best-match (MAX) as defined in (109). These two methods are complementary since the cross-validations used and arbitrary threshold (majority-rule) where as semantic similarity does not, but cross-validations more closely reflect various prediction and imputation strategies that could be used to extend phenotype information since specific phenotypic terms are being predicted rather than the just a general measure of phenotypic similarity. Future work could determine which of the many methods of network-based gene-function prediction (see: (197) for a review) would be most useful for mouse phenotypes.

6.2.7 Saturation

If sll mouse phenotypes in a biological modules have been discovered then I expect to see the number of unique mouse phenotypes per gene to decrease as the number of genes examined increases. If there is no saturation I expect a linear relationship between the number of genes examined and the number of unique MPO terms. Thus I counted the number of unique mouse phenotype terms annotated to genes in each module as well as the number of genes with any phenotype annotations in the module. I tested for saturation using the lack-of-fit sum of squares test on the linear regression constrained to pass through the origin across all biological modules identified in each network.

6.3 Results

6.3.1 Biases in mouse phenotyping projects

I considered seven different mouse phenotype resources: the German Mouse Clinic (GMC, (193)), the Knockout Mouse Project (KOMP, (188)), Europhenome (EuP, (187)), the Wellcome Trust Sanger Mouse Portal (WTSI, (194)), the International mouse phenotyping consortium (IMPC-P & IMPC-M, genes undergoing phenotyping and those with live mice respectively, (186)), and the Mouse Genome Informatics database (MGI, (36; 37)). I collected from the respective websites all genes for which the phenotyping was completed or in progress when this information was available, otherwise all genes for which live mice have been produced (Table 6.1). In each dataset roughly 90% of the genes had 1-1 orthologs in humans as defined in Ensembl 70 (105), a significant deviation compared to all mouse genes of which only 44% have 1-1 orthologs ($p < 2.8 \times 10^{-9}$). The main reason for using mice is to improve our understanding of human biology and the relevance of information about a one-to-one ortholog is much easier to infer across species; thus we will only consider those genes with human 1-1 orthologs for the rest of this chapter.

First, I examined the overlap between these project to examine the redundancy between datasets. The IMPC attempts to combine effort from the earlier projects and contains records for 89% of the genes examined by GMC, KOMP, EuP, and WTSI (Table 6.4). However, there is considerable redundancy between the projects with 28% of genes studied by the GMC also being considered by the WTSI and 76% overlap with EuroPhenome; and 34% overlap between genes considered by the WTSI also being examined by EuroPhenome. In contrast, KOMP was almost completely independent from the other projects with only 7 of the 243 genes considered by this project also being examined by GMC, EuP or WTSI. However, the inclusion of the IMPC result into the MGI database has been slow with only 53% (440/834) of IMPC knockouts with phenotyping in progress having annotations recorded in the MGI database.

Table 6.4: Overlap between six mouse phenotyping projects

	GMC	KOMP	EuP	WTSI	IMPC-P	IMPC-M	MGI
GMC	107						
KOMP	0	243					
EuP	81	1	462				
WTSI	30	6	160	753			
IMPC-P	50	18	68	353	834		
IMPC-M	84	234	361	715	834	3050	
MGI	80	136	406	540	440	1560	7932

Since mouse knockouts are frequently developed to be used as models of human disease, I examined the proportion of genes in each phenotyping project which were orthologs of known human disease genes (112). Orthologs of known disease genes listed in the Online Mendian Inheritance in Man database (OMIM) comprised 12-18% of genes studied in the mouse phenotyping projects which was a significant enrichment over the 9% of all mouse genes with 1-1 orthologs ($p = 0.04 \times 10^{-10}$, Table 6.5). Indeed an OMIM disease gene was twice as likely to have been studied in mouse as another human gene with a 1-1 mouse ortholog. Recent phenotyping projects have a lower proportion of OMIM genes (12-16%) than MGI (18%) since they are not solely focused on developing disease models.

All the phenotyping projects accept suggestions from the academic community, and genes for which there are publications linking the gene to a biological process or disease may be more likely to be suggested. Thus I examined the number of papers which are associated with the human ortholog of each mouse gene in Pubmed (using `ftp://ftp.ncbi.nih.gov/gene/DATA/gene2pubmed.gz`). Genes selected for study by the mouse-phenotyping projects were associated with more papers than expected (Figure 6.1A), indicating a selection-bias towards well known genes. The IMPC did not quite reach significance but has a strong trend (IMPC-P $p = 0.067$, IMPC-M $p = 0.055$); this could be due to a delay in the curation of Pubmed abstract-gene associations or a result of a weaker selection bias in this project. KOMP also does not exhibit a significant bias towards well studied genes but this could be a result of lower power since it only contains 243 different genes.

I also considered biases within each project, since each one likely draws on their own pool of collaborators to solicit suggestions from. Thus I expected the subset of genes selected by each study to be related to each other in addition to being well-studied overall. With the exception of KOMP, genes selected by the same mouse-phenotyping project were associated with the same abstract significantly more often than expected (Figure 6.1B). KOMP is exceptional because it is the only project which, along with accepting suggestions from the community, prioritizes poorly studied genes (188).

I also considered co-expression (COXPRESdb, (46)), protein-protein interactions (STRING (51) & iRefIndex (52)), and an integrated functional network (HumanNet, (53)) as other sources of prior biological knowledge that may have been used to select genes (Table 6.2). Again I considered the strength of edges between the human 1-1 orthologs of genes selected in the same phenotyping project compared to the average over all genes with human-mouse 1-1 orthologs (Figure 6.1C). None of the datasets were significantly biased with respect to HumanNet, which makes sense because it was published in 2011 well after most of these studies had begun thus contains more recent data that may not have been available at the time the phenotyping projects were selecting genes (Table 6.1). There was little bias with respect to co-expression as well, although MGI and GMC did contain genes significantly more tightly co-expressed than expected. However, four of the phenotyping projects (GMC, Europhenome, WTSI, and MGI) showed significant or borderline significant bias with respect to STRING, a weighted protein-protein interaction network.

Unlike the patterns of co-citation some of these functional networks were quite sparse with only a few hundred thousand edges compared to the more than 140 million possible edges. Thus I also considered the density of edges between the human orthologs of genes examined in the same phenotyping project compared to all genes with human-mouse 1-1 orthologs (Figure 6.1D). As expected all datasets with the exception of KOMP and GMC were consistently significantly biased towards a higher density of

Table 6.5: Mouse phenotyping projects were significantly enriched in one-to-one orthologs of human genes and known disease genes

Phenotyping Project	No. Genes	No. 1-1 Orthologs	%	No. with OMIM	%	p
German Mouse Clinic	107	97	90.7%	15	15.5%	0.035
KOMP	243	225	92.6%	36	16.0%	0.00094
Europhenome	462	430	93.1%	56	13.0%	0.0066
WTSI Mouse Portal	753	672	89.2%	87	12.9%	0.0010
IMPC-P	834	743	89.1%	96	12.9%	0.00061
IMPC-M	3050	2768	90.8%	332	12.0%	1.8×10^{-7}
MGI	7932	7265	91.6%	1332	18.3%	6.2×10^{-284}
All mappable	38293	16732	43.7%	1561	9.3%	

edges between their selected genes in all networks.

KOMP was the only phenotyping project to consistently show little if any bias with respect to other biological data due to the deliberate prioritization of unstudied genes. In contrast, MGI, which makes no attempt to relieve either source of ascertainment bias, was consistently the most significantly biased across all datasets. This was particularly concerning since MGI is the most commonly used source of mouse phenotype information for genome-wide applications (20; 53; 183; 184). However, since much of the phenotype information in MGI has been available to researchers for follow-up longer than other projects I was unable to determine if this bias was a result of the selection procedure for the mouse phenotyping or due the phenotype information being used to inform molecular studies. However, I noted that MGI was significantly biased with respect to co-expression which was determined using genome-wide arrays and thus would not be subject to ascertainment-bias. Interestingly, the IMPC-P consistently shows a lower extent of selection bias than WTSI despite containing a similar number of phenotyped genes, 834 for IMPC-P and 753 for WTSI, suggesting their selection procedure may suffer from less bias.

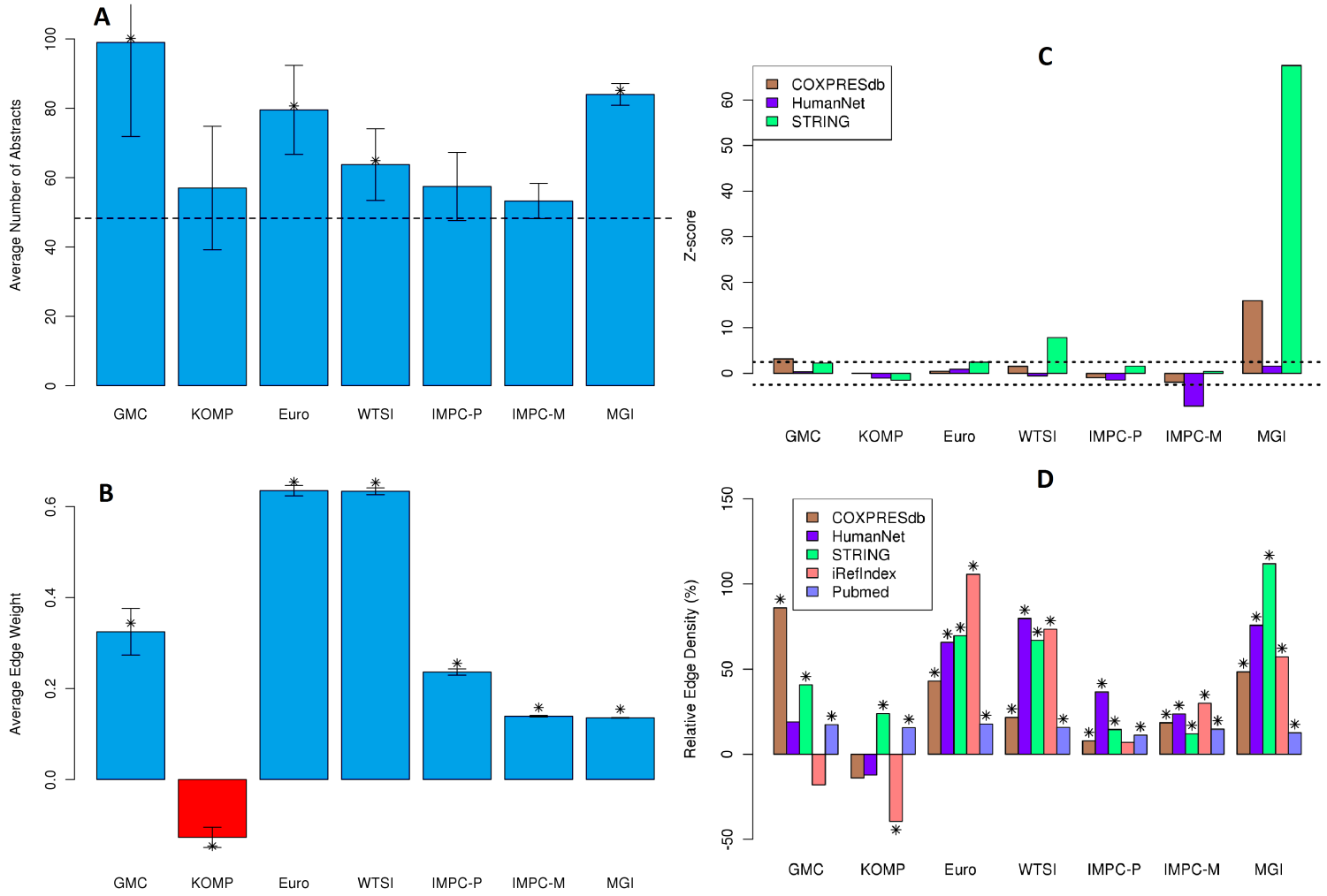


Figure 6.1: Phenotyped genes are biased towards well studied genes that are known to be associated with each other * = $p < 0.05$. Error bars indicate 95% CIs using the population standard error.(A) The average number of papers associated with the human 1-1 ortholog of each phenotyped mouse gene in the curated NCBI Gene2Pubmed resource. Dashed line is the average for all genes with mouse-human 1-1 orthologs. (B) The difference between the average link weight, based on co-occurrence of associated papers, between gene included in each phenotyping study and the average across all genes human genes with 1-1 mouse orthologs. (C) Average weight of edges between genes selected for each phenotyping project in 4 different networks, The Z-score was calculated for the observed mean for each project compared to all the average of genes with 1-1 orthologs. Dotted lines indicate significant at $p < 0.05/3$. (D) Edge density (number of observerd edges/number of possible edges) between genes investigated by each project relative to the density of edges between all genes with 1-1 mouse orthologs (% difference)

6.3.2 Experimental Biases

In addition to published information, experimental results may be used directly to prioritize genes for mouse phenotyping. Experiments such as differential expression, exome, GWAS, and linkage studies are commonly used to identify genes important for human disease. Exome and differential expression studies can be biased towards long, highly expressed genes (65; 67). Furthermore linkage and GWA studies may be biased with respect to linkage disequilibrium (LD). I examined the length, expression, and LD of the human orthologs of genes from each mouse phenotyping-project compared to all human genes with 1-1 mouse orthologs (Figure 6.2).

Long genes have a greater opportunity to incur *de novo* mutations making them more likely to be identified in exome studies (65); they also have more reads mapping to them thus there is greater power to detect differential expression using RNASeq (66; 67). Human genes with 1-1 mouse orthologs are 13% longer than other human genes ($p < 1 \times 10^{-10}$, Wilcox-rank-rum test, Figure 6.2A), this is likely due to the difficulty for a long gene to be duplicated in its entirety resulting in long genes having fewer paralogs than short genes (198). However, genes examined in all the mouse phenotyping projects (with the exception of KOMP) are significantly longer than other genes with 1-1 orthologs ($p < 0.0001$).

Genes expressed in many tissues or at a high level are more likely to be identified as differentially expressed (66; 67). I calculated breath of expression as the proportion of 106 normal tissues/cell-types with microarray expression data with detectable expression in the Gene Expression Barcodes database (128). Level of expression was calculated as the average FPKM across the 16 tissues in the Illumina Body Map 2.0 (127). All the larger mouse phenotyping projects contained significantly more broadly expressed genes (Figure 6.2C). In addition four contained significantly more highly expressed genes (Figure 6.2D). In both cases genes all genes with 1-1 orthologs were significantly more broadly and more highly expressed than all human genes but the

mouse phenotype projects were significantly more biased than them. However, the magnitude of these differences tended to be small (5%-71%, most < 30%).

Recombination rates and LD determine the size of LD blocks tagged by a single nucleotide variant (SNP) in GWAS studies, large LD blocks can harbour multiple causal variants contributing to a single significant signal (199; 200). In contrast, regions of high LD make it difficult to pinpoint a single causal gene in linkage studies (201; 202). I used the distance to deCODE recombination hotspots as a measure of LD block size (Figure 6.2D, (169)). Only MGI was significantly biased towards small LD blocks (small distance to recombination hotspot), and this is likely a reflection of the large number of disease genes it includes many of which were identified using linkage studies.

Mirroring my results for selection-bias above, KOMP was the only mouse phenotype project to not show any bias with respect to the gene length, expression or recombination. In contrast the genes annotated in the MGI database were significantly longer, more highly and more broadly expression and located closer to recombination hotspots than other genes with 1-1 orthologs. Other phenotyping projects were in between these two extremes showing weaker biases in the same direction as MGI. These biases in phenotype information with respect to gene length and expression level could result in biases in functional analyses of exome, and expression studies which include mouse phenotype information since experimentally identified genes could appear to participate in a single functional module simply because more mouse phenotype information is available for them.

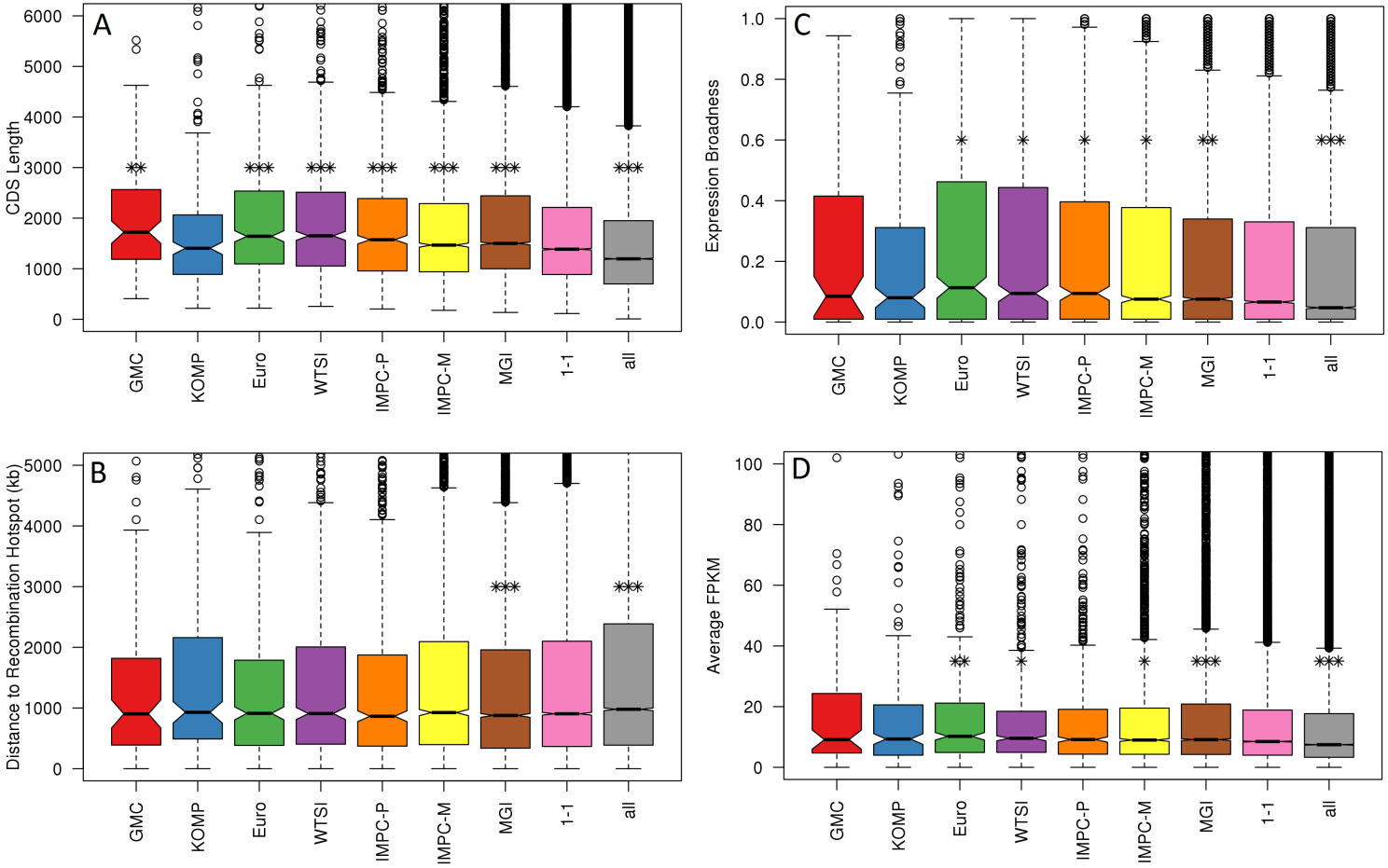


Figure 6.2: Mouse phenotyping prioritization may be influenced by experimental biases. (A) CDS Length for the longest transcript/gene was obtained from Ensembl; long genes are more likely to contain rare/*de novo* deleterious point mutations detected by exome-studies, (B) Distance to the closest recombination hotspot (averaged over upstream & downstream directions) as determined by an increase of at least 10-fold over the genome average from deCODE(169); recombination hotspots affect the linkage around the gene which affects the ability to identify the gene using GWAS and/or linkage studies,(C) Expression broadness was calculated as the proportion of 106 normal tissues with detectable expression in the Gene Expression Barcode database (array-based); breadth of expression affects the likelihood of detecting differential expression of the gene using arrays, (D) Average FPKM across 16 tissues of the Illumina BodyMap (RNASeq) (127); FPKM affects the likelihood of detecting differential expression of the gene using RNASeq, Stars indicate significant difference relative to all genes with 1-1 orthologs, * = $p < 0.05$, ** = $p < 0.005$, *** = $p < 0.0005$

6.3.3 Identifying biological modules

To correct for these biases, phenotyping projects should prioritize genes without functional information and unrelated to those that have already been studied. KOMP is the only project to incorporate such a goal and is also the only dataset without significant selection-bias, demonstrating the efficacy of such an approach (Figure 6.1, 6.2). Here I will explore the possibility of using existing functional information directly to identify understudied genes by exploiting the modular nature of gene function (28; 63; 73; 74; 75).

This can be achieved by identifying regions (modules) within gene networks containing few if any genes with mouse phenotype information available. I employed three publicly available networks which represent the different types of networks most commonly used to examine gene function: COXPRESdb, a co-expression networks (46); iRefIndex, a protein-protein interaction (PPI) network (52); and HumanNet, an integrated functional network which combines gene expression, protein-protein interactions, and co-citation from several model organisms as well as humans (53). Modules were identified in each of these networks using Infomap (69), which has been shown to be one of the top performing community-detection algorithms (132). To check the robustness of the resulting modules to the stochasticity inherent in the algorithm, I compared the result from the best clustering over 10,000 attempts (best-clustering), according to the metric used by the algorithm, on the original network to the consensus clustering (135), where the algorithm is applied to the association matrix representing the frequency two genes are located in the same module across 100 separate clusterings of the network (Figure 6.3).

The resulting modules were validated by using the known mouse phenotype annotations from MGI in February 2012 to predict new annotation in the January 2014 edition using a majority-rule algorithm within each module (Figure 6.3). Both the best-clustering and the consensus clustering on all three networks were able to predict

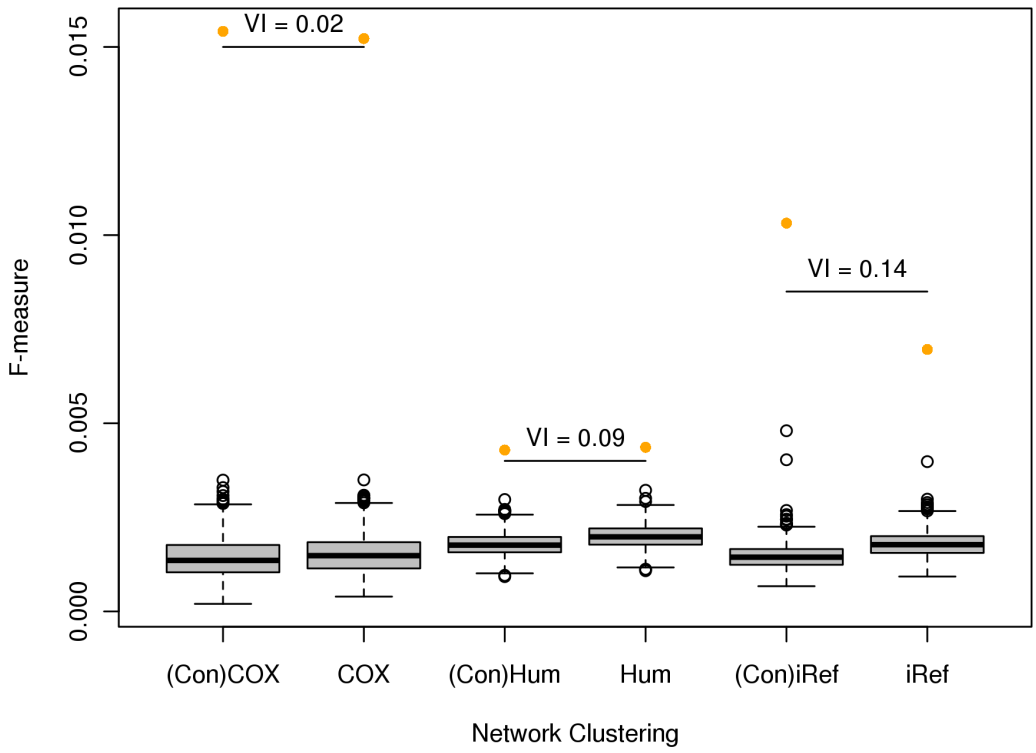


Figure 6.3: Validating network modules. Recently added mouse phenotype annotations (since 2012) were predicted from previous mouse phenotype information (prior to 2012) using a simple majority-rule prediction algorithm amongst the modules identified in each network using the Infomap community-detection algorithm(69). F-measure (aka F1 score) is the harmonic mean of the precision and recall of these predictions. (Con) = consensus clustering. COX = COXPRESdb, Hum = HumanNet. iRef = iRefIndex. Boxplots are the results from 1,000 permutations of module membership. VI = variation of information (195), a measure of the difference between the best individual clustering and the consensus clustering.

new phenotype annotations significantly better than 1,000 gene-label permutations. In addition, COXPRESdb and HumanNet showed very high robustness to noise in the clustering algorithm with almost no difference between the best-clustering and the consensus clustering. In contrast, the consensus clustering for iRefIndex was noticeably better at predicting phenotypes than best-clustering of the original network. Of the three different networks, COXPRESdb was the best at predicting the novel mouse phenotypes. COXPRESdb is also the least-biased network since only array expression data was used in its construction which far less biased than protein-protein interaction data (45; 203; 204). In contrast, HumanNet which is expected to be the most biased due to the inclusion of co-citation data as well as PPIs in its construction performs worst at predicting the novel phenotypes (53).

6.3.4 Identifying understudied genes

Next, I considered how reliably mouse phenotype information could be transferred between genes in modules of different sizes. Modules were binned according to the number of genes they contained, on a logarithmic scale; then I considered the quality of predictions using leave-one-out cross validations as well as the average semantic similarity (108; 109) between the phenotype annotations of genes within each module averaged over all modules in the same bin. As an example I have included the results for COXPRESdb consensus clustering (Figure 6.4) since this was the best performing network according to the validations performed above, however the trends remain consistent across all of the networks and clustering methods (see: Appendix A section A.2 for the respective figures). Precision of the predictions increased with increasing module size but the number of terms that were predicted dropped dramatically, resulting in a decreasing F-measure as module size increased. Semantic similarity of mouse phenotypes between genes in each module gradually declined toward the global average with increasing module size.

I determined whether the observed transferability was significantly better than expected controlling for the distribution of modules sizes by permuting module assignments for the genes 1,000 times. However, for both COXPRESdb and HumanNet the identified functional modules consistently performed better than expectation with a few exceptions of bins with relatively few modules in them resulting in higher variability among randomizations. In contrast, iRefIndex modules only performed significantly better than expected for those with less than 100 genes (Figure 6.5). Interestingly, the distribution of module sizes was different between networks. COXPRESdb contained a large number of small modules (2-3 genes) and the most very large modules (>500 genes). However, HumanNet contained mostly modules of 20-50 genes; iRefIndex was somewhere in between. This mirrored the networks' ability to predict new mouse phenotypes discussed earlier where COXPRESdb performed the best and HumanNet performed the worst. Thus the observed differences between the networks during validation is partly a result of differences in transferability between modules of

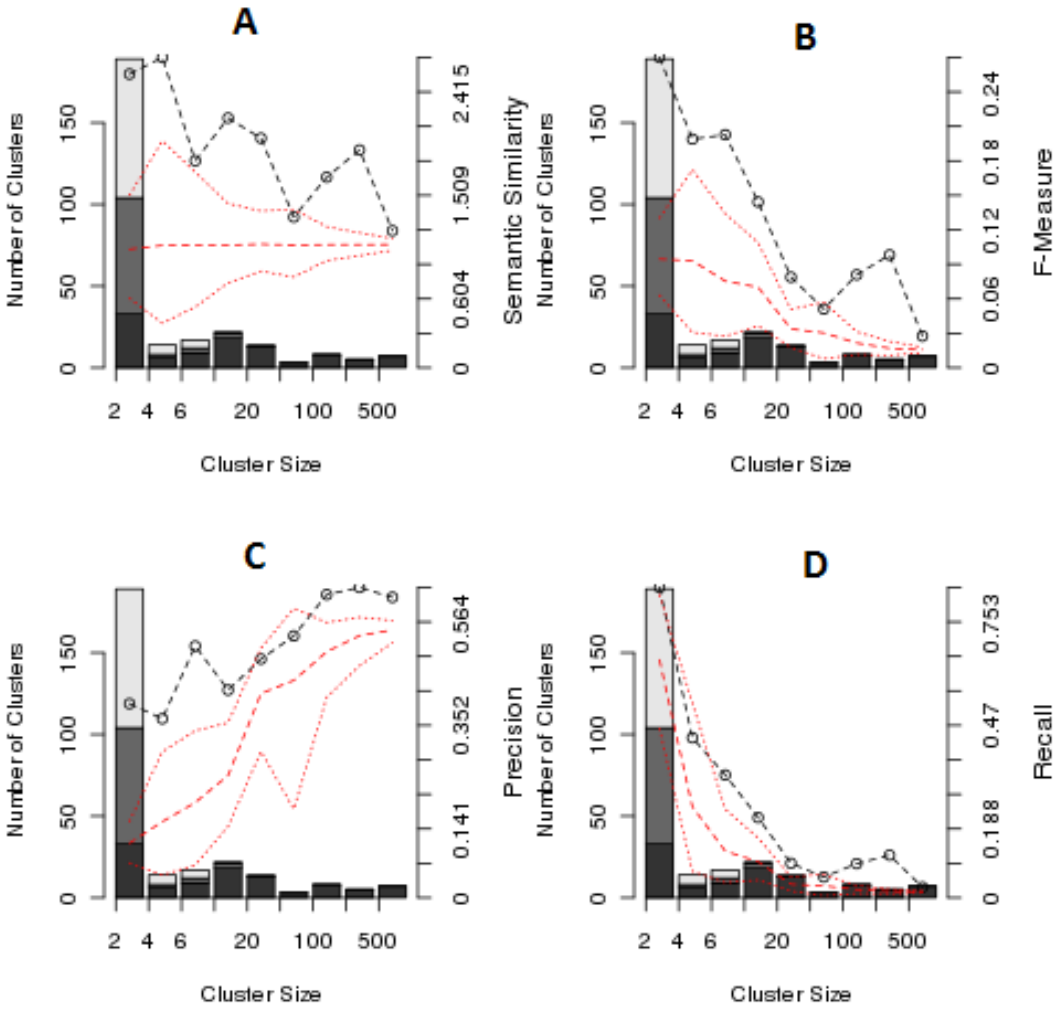


Figure 6.4: Transferability of phenotypic information using the COXPRESdb network consensus clustering modules across a broad range of module sizes. (A,C,D)Leave-one-out cross validations within each module using a majority-rule prediction algorithm. (B) Average semantic similarity. Black line and open circles indicates the observed results. Red dashed line indicates the median over 1,000 module-permutations. Dotted lines contain 95% of the permutations. Dark bars are the number of modules in that size class with at least 2 genes with mouse phenotype annotations thus contribute to the transferability estimate; medium grey bars are the number of modules with exactly one gene with mouse phenotype information available; light bars are the number of modules with no genes with mouse phenotype information available.

different sizes and the different distributions of modules sizes identified in each network.

Many modules in each network contained exclusively genes with no known mouse phenotype information; genes in these clusters with mouse orthologs should be prioritized in phenotyping efforts to fill in this missing information. However, due to the differences in network structure between the networks there is little overlap in the lists of genes to prioritize (Figure 6.5D); since COXPRESdb performed the best when predicting new mouse phenotypes and is the least susceptible to research bias it would be the preferred network to use. I examined the known function (if any) of these understudied genes using Gene Ontology (GO) and the Kyoto Encyclopaedia of Genes and Genomes (KEGG) pathways using a hypergeometric test and Bonferroni correction (Table B.11). There were 643 genes identified as understudied across all the networks. Of these, 412 (64%) had any GO biological process annotations which is significantly fewer than expected (74% of all genes, $p = 2.1 \times 10^{-8}$), and were enriched in genes involved in the mitochondrial inner membrane and glycosylphosphatidylinositol (GPI) biosynthesis. Molecular function and cellular component GO ontologies similarly show a depletion of annotations amongst the identified understudied genes as well as enrichment for related functions (Table B.11). Likewise only 107 (17%) of the understudied genes have annotations in KEGG which is a significant decrease with respect to the whole genome (25%, $p = 3.6 \times 10^{-7}$), in addition to the KEGG terms describing the GPI pathway, I found KEGG terms related to olfactory transduction enriched among under-studied genes. It makes sense that olfactory receptors and mitochondrial inner membrane proteins would be understudied since in the first case it is unlikely any interesting phenotypes would result and in the latter case it may be impossible to knock out the gene without killing the organism. However these two functional enrichments account for only 30 of the 643 understudied genes. Furthermore, GPI membrane anchors have been implicated in the formation of lipid rafts and signal transduction thus represent a potentially interesting but understudied pathway (205). When I focus on just the 286 genes which are identified as understudied using

the COXPRESdb network, I find a significant depletion of genes with GO biological process annotations ($p = 0.006$) or KEGG annotations (4.9×10^{-9}) and enrichments for genes involved in gene expression (Table B.12). Twenty-six genes were identified as understudied using more than one network; of which 20 have unique mouse orthologs and thus able to be investigated in mouse (Table B.13 in Appendix B). These 20 genes have an average size of 432 amino acids and fourteen were expressed in at least ten of the 16 tissues in the Illumina BodyMap 2.0 (127). Finally of the 20 identified genes with mouse 1-1 orthologs three have no GO annotations at all and four more have only computationally predicted annotations.

I have shown identifying distinct regions in functional networks using established community detection algorithms is an efficient method to find understudied but potentially interesting genes. One gene was identified as unknown in using all three networks: CCDC155, a broadly expressed coiled-coil domain containing protein with a 1-1 ortholog in mouse (105).

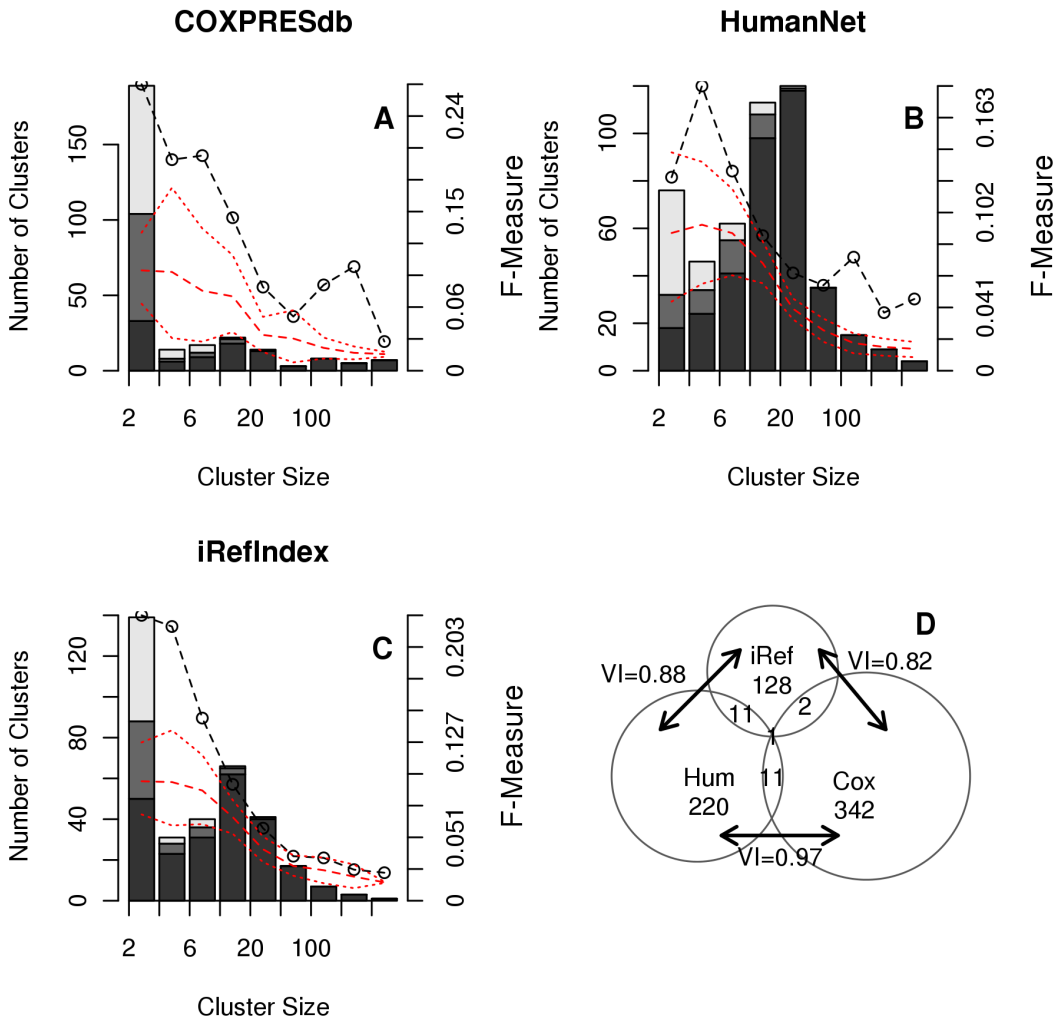


Figure 6.5: (A-C) Phenotypic information could be predicted using the COXPRESdb, HumanNet, and iRefIndex network consensus clustering modules across a broad range of module sizes. Black line and points indicate the average F-measure (harmonic mean of precision and recall) for leave-one-out cross validations within each module using a majority-rule prediction algorithm. Red dashed line indicates the median over 1,000 module-permutations. Dotted lines contain 95% of the permutations. Dark bars are the number of modules in that size class with at least 2 genes with mouse phenotype annotations thus contribute to the transferability estimate; medium grey bars are the number of modules with exactly one gene with mouse phenotype information available; light bars are the number of modules with no genes with mouse phenotype information available. (D) Genes present in completely unphenotyped modules differed by network. iRef = iRefIndex, Cox = COXPRESdb, Hum = HumanNet. This is a reflection of the very different module structure identified in each network as indicated by the variation of information (VI) between the clusterings.

6.3.5 Completeness of existing phenotype information

Finally, I sought to determine the value in continuing to examine the phenotypes of genes in large well studied clusters by examining the number of unique phenotypes associated with genes in each cluster for signs of saturation (Figure 6.6). Saturation was tested using a lack-of-fit sum of squares test on a linear regression restricted to pass through the origin. Despite the largest cluster having over 1000 unique phenotypes, there was only evidence of saturation in the clustered PPI. Thus there is likely to be many more undetected phenotypes for even well studied mouse-knockouts.

6.4 Conclusion

I have shown all mouse phenotyping project to date, with the exception of KOMP, show significant ascertainment bias (Figure 6.1). I further showed these datasets reflect known experimental biases toward long, highly and broadly expressed genes (Figure 6.2). I proposed using the modular structure of functional networks to identify under-studied genes (Figure 6.5). Finally I showed that there are many more phenotypes to discover for existing mouse knockouts (Figure 6.6).

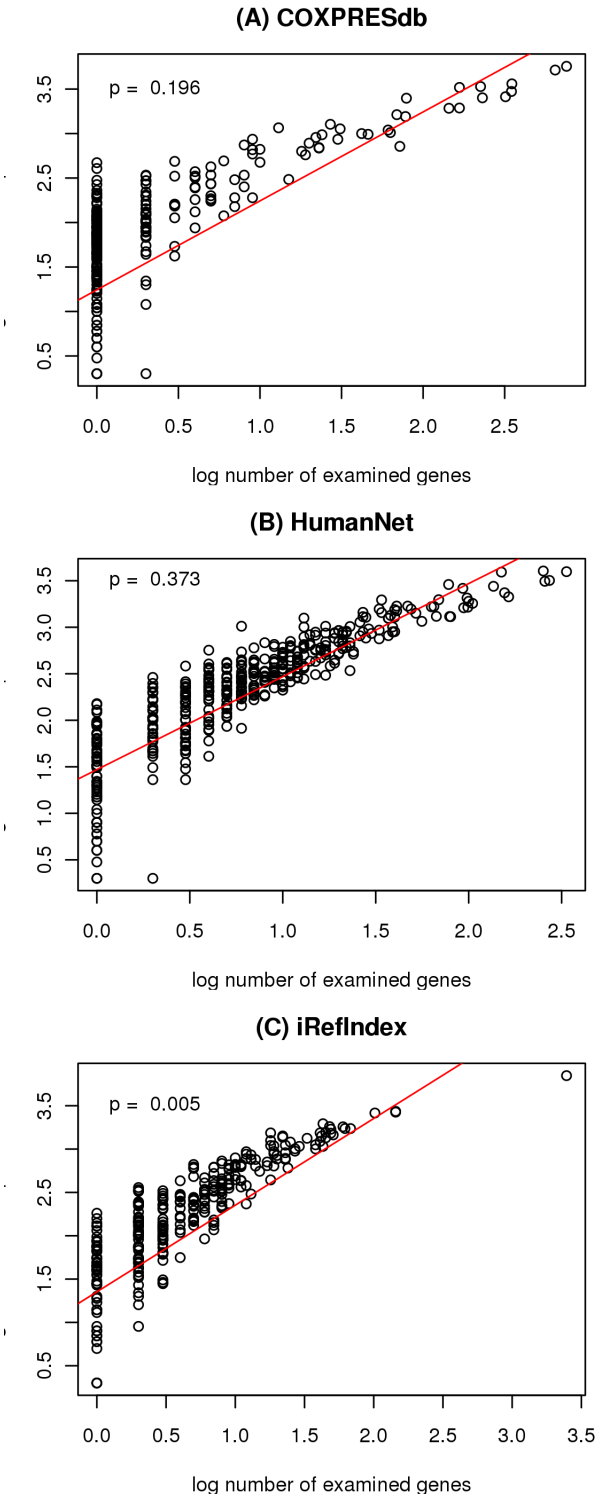


Figure 6.6: Saturation of unique mouse phenotype terms as the number of genes with annotations increases. Each point represents one cluster. If no saturation occurs we expect a linear relationship of slope = 1 in the log-log plot (red) the y-intercept is the slope of the linear relationship in the un-logged data as calculated using least-squares.

Chapter 7

Discussion and Conclusions

In this thesis, I have investigated the uses of gene networks in representing pathways underlying diverse phenotypic presentations. I identified pathways associated with specific phenotypes among patients with diverse developmental disorders, then showed that these pathways frequently also explained the patient's co-morbid phenotypes. Furthermore I showed that there are frequently multiple genes from the same pathway affected by a single copy-number variant in a single patient. By integrating multiple sources of functional information together and comparing to randomized variants I showed that patient-derived copy-number variants tend to affect a significantly large group of functionally related genes (on average 4-5 genes/CNV) which was enriched in known disease genes. Comparing to CNVs from healthy individuals, I showed the presence of a group of functionally-related genes was a significant predictor of pathogenicity of the CNV even when controlling for the total number of genes affected by a CNV and the presence of known disease genes. I then showed the functional clusters seen in patient CNVs were part of a larger set of functional clusters throughout the human genome the majority of which were expressed in a tissue specific manner. I identified various gene functional annotations enriched among those clusters affected only in patient *de novo* CNVs and validated their ability to predict pathogenic CNVs in an independent dataset. In addition, I showed that patients with CNVs affecting the same functional cluster had significantly similar phenotypes. Finally, I examined used the modular nature of gene networks to identify poorly studied

genes and pathways which should be prioritized to reduce the biases in available functional annotations.

7.1 Beyond the Monogenic Model of Developmental Disorders

The monogenic model assumes each patient has a variant in a single gene which causes their genetic disease, and all patients with that disease have a mutation in that gene (206). The monogenic model has dominated the investigation of rare developmental disorders for the past 65 years but has only succeeded explaining a third to one half of identified Mendelian diseases (206; 207). This was at least in part due to technical limitations which made large population scale studies infeasible, rather most studies focused on tracing linkage of a disease phenotype with genetic markers within one or a small number of families (206; 207). However, contiguous gene syndromes, disorders associated with large variants at a single locus but that encompass multiple genes, have been known of for almost the same length of time (208).

Recent work has challenged this primarily monogenic model of developmental disorders. Several disorders have been associated with variants in multiple distinct loci, for instance two loci (4q25 and 13q14) are linked to Rieger syndrome (209), mutations in two genes (*NKCC2*, *ROMK*) have been shown to cause Barter's syndrome (210) and Coffin-Siris syndrome has been linked to mutations in any of at least six different genes (14). In addition, multiple deleterious variants are frequently found within the same patient, for example patients with a 509kb deletion at 16p12.1 were significantly more likely to harbour a second CNV supporting a two-hit model (211), in a study of 20 ASD patients 8 had multiple *de novo* variants and 2 had variants in two different known autism-associated genes (212), and Golzio et al.(23) determined that a single CNV contained three genes which contributed epistatically to the associated phenotype.

My work further supports a polygenic cause for large number of patients with diverse developmental disorders. In Chaper 4, I showed that *de novo* CNVs found in patients with developmental disorders contain a larger number functionally related genes than expected given the total number of genes affected by the CNV (Figure 4.4). In contradiction with the monogenic model of developmental disorders these functional clusters were a better predictor of CNV pathogenicity than the presence of a single known disease gene. I further showed that the presence of a functional cluster was a significant (if small) predictor of CNV pathogenicity even after accounting for the total number genes affected by the CNV. Large *de novo* CNVs are found in 10-20% of patients (2; 8; 12; 15). I found that half of such variants contained a functional cluster, thus synergistic effects between multiple related genes may contribute to 5-10% of patients with developmental disorders. This is likely an underestimate since it does consider the role of functional clustering in rare inherited CNVs.

Advances in genetic technologies (eg. SNP-chips, exome sequencing, array CGH) and databases containing functional informations about large numbers of genes have enabled human genetics to move beyond the monogenic models of disease to examine biological pathways in large heterogeneous patient cohorts (27). This pathway approach assumes patients have variants in one or more genes in one (or more) common pathway(s) which are responsible for their phenotype, and other patients with variants in the same pathway(s) will have a similar phenotype (28). In Chapter 3, I tested this assumption of the pathway approach for developmental disorders using a rigorously phenotyped patient cohort. I found molecular pathways were significantly associated with 329/408 (81%) different phenotypes in patients with developmental disorders; and a third of the pathways showed broad phenotypic convergence where patients carrying CNVs affecting the pathway more phenotypically similar to each other than to those without such variants (Figure 3.6). In addition, a large number of patients had multiple genes contributing to the same enriched pathway, showing how the pathways can be combined with a polygenic model of disease.

7.1.1 Grouping Orphan Patients

Up to 80% of patients do not fit into a recognizable syndrome or are so rare traditional techniques, which depend on identifying multiple patients with the same syndrome, have failed to identify the underlying variants (3; 5). Several projects attempt to augment these techniques using automated comparison of detailed patient phenotypes such as 3D modelling of facial dysmorphisms to facilitate the identification of patients spread across multiple countries or continents (213; 214). Others advocate for the development of more detailed human phenotyping techniques as a solution to the difficulties presented by highly heterogeneous genetic diseases (215; 216). While I was able to identify pathways associated to over 80% of phenotypes annotated to at least 3 patients, fewer than half of all phenotype terms annotated to the patients could be tested as they were seen in fewer than 3 patients (see: Chapter 3). Further subdivision of patient cohorts through more detailed phenotyping must be balanced by even larger patient cohorts or additional phenotypic detail may be wasted by the absence of statistical power. In addition, I found that fewer than 100 patient phenotypes were consistently more common among patients with mutations in a particular biological pathway suggesting many currently examined phenotypes provide little information as to the genetic aetiology of the patient's disorder (Figure 3.13). Thus, efforts to improve patient phenotyping should ensure the refinements and deep phenotypes reflect biological systems which underlie development rather than chance events or environmental factors.

Another approach is to group patients on the basis of shared genetic variants which are inferred to be pathogenic regardless of whether the patients have similar phenotypes (208). However, even in large cohorts individual pathogenic variants are typically seen only once, eg. in a cohort of 1,133 patients 95/148 (64%) of single gene diagnoses were seen only once (217). Pathway approaches offer a solution to this problem as patients

with variants affecting the same pathway are expected to have similar phenotypic outcomes (28). Here, I showed the potential utility of pathways in this respect by demonstrating that patients which have variants affecting the same biological pathway or the same cluster of functionally-related genes have significantly similar phenotypes (Figures 3.6, 5.17). Since many variants affecting different genes may affect the same biological pathway, pathway approaches offer a solution to the extreme genetic heterogeneity seen among developmental disorders since they gain power from coverage of the pathway rather than relying on recurrent mutations of the same gene.

7.1.2 Exomes vs Copy Number Variants

Currently there are two major techniques for identifying variants associated with developmental disorders, namely array comparative genomic hybridization (aCGH) or genotype chips for identifying copy number variants (CNVs) and next generations sequencing targeting exons (exomes) for identifying rare/*de novo* single nucleotide variants (SNVs). CNV analysis is already commonly used in the clinic (2; 16) and exomes have been rapidly moving towards large-scale clinical use(208). Pathway approaches have been used with both CNVs (eg. (21)) and exome sequencing data (eg. (218)). However, the brunt of evidence for the multi-hit model of developmental disorders has come from CNVs (7; 14; 20; 21; 22; 23), where as many exome sequencing studies assume a monogenic cause of developmental disorders (146; 208; 219; 220).

Exome studies must assume a particular model of inheritance (dominant, recessive, *de novo*) for the disorder under investigation in order to identify novel causal loci due to the large number of variants identified in each patient. Exome studies typically find 20,000 SNVs per patient, roughly 100 of which will be rare or *de novo* protein-altering changes (11; 14; 206; 208; 217), which must be filtered down to a small number of likely candidates, thus an assumption must be made about the number of candidates per patient to determine the appropriate filtering. In contrast, there are roughly 150 CNVs

identified per individual of which a *de novo*/rare CNVs of size greater than 100kb (reliably detectable) occur in roughly 1 of every 50 individuals, thus minimal filtering is needed beyond quality control (14; 217). In addition, there is a clear burden of large rare/*de novo* CNVs among patients with developmental disorder (4; 162), where as several studies have failed to find a significant burden of *de novo* SNVs amongst patients when compared to controls or statistical models which account for variations in gene length and mutation rate (11; 14; 162; 218). Thus the brunt of the work presented here focused on large *de novo* CNVs rather than SNVs from exome sequencing studies due to the greater confidence in their individual pathogenicity.

Combining exome sequencing data with CNVs can be a powerful combination to identify novel disease genes by using the SNVs from the exome sequencing can to identify the particular genes within a CNV region while the CNV data would have a large contribution to the significant enrichment among patients vs controls(162; 217). I found that both the total number of genes and the presence of multiple functionally related genes were significant predictors of CNV pathogenicity beyond the presence of known disease or haploinsufficient genes (Table 4.1) which suggests combinatorial effects between multiple genes within the CNV region are contributing to the patient's phenotype. These combinatorial effects would be missed by most studies which include exome sequencing data since the large number of possible gene combinations demand extremely large cohorts to be statistically well-powered, thus SNV evidence is typically aggregated at the single gene level to test for association with disease (162; 206; 217).

7.2 Utility of Functional Genomics Datasources

All sources of functional information about genes are subject to errors and biases. High-throughput protein-protein interactions (PPI) contain many false positives (45), but also under-report PPIs involving membrane proteins (221). Whereas protein structures are biased towards genes known to be involved in disease (191). RNA sequencing

studies are biased towards long genes (67). Gene Ontology (GO) annotations likewise are biased towards 'interesting' thus better characterized genes (42). In addition, over 98% of GO annotations are inferred directly without curation using sequence homology and other features which is known to have a high error rate (222). Even curated databases such as KEGG contain many errors with 5-60% of proteins having incorrect annotations(43). Even experimental findings are often biased or false positives, with some estimating as many as half of all published findings could be false (185). Gene association studies often report false associations and those published in high impact journals are often then to be the most biased (223). Thus an important issue has been how best to use these datasets while avoiding or compensating for the errors and biases they contain.

7.2.1 Integrating vs Intersecting

Combining multiple sources of information is a common method to remove false-positives since as the number of independent replications of a finding there are the more likely it is to be correct since errors typically random occurrences, errors resulting from systematic biases in one particular resource/protocol can be eliminated by validating it with a different method/resource which does not have that bias, eg. it is common to validate PPIs using two or more different techniques to reduce false positives (191). Indeed multiple groups have shown that combining multiple sources of information together improves functional predictions over any single dataset alone (55; 56; 59). I used two different approaches to combine information: in Chapter 3 I took the intersection of genes identified using different functional annotations, in Chapters 4 and 5 I integrated various functional datasets together weighting each dataset based on its quality.

In Chapter 3, I identified pathways significantly enriched among patients with a particular phenotype using four different functional resources: GO annotations(34),

KEGG pathways (33), mouse knock-out phenotypes from the MGI(36), and co-expression in across various regions of the brain through time from Brainspan(123). Those genes signifcantly associated with a particular phenotype using various combinations of the datasets were identified and were examined for protein-protein interactions. The performance of each resource and the various combinations was evaluated by testing whether patients with CNVs affecting genes in the pathway were more phenotypically similar to each other than to those without CNVs affecting the pathway, which I call 'phenotypic convergence'. I found that genes identified using two different resources were not more likely to show phenotypic convergence that those identified using just one of the resources; and in some cases the intersection of two resources was less likely to show phenotypic convergence that either of the original resources (Figure 3.6). Likewise identifying those genes that were identified using at least two different resources and had significantly large numbers of PPIs between them did not significantly increase the proportion of pathways showing phenotypic convergence. This was surprising since each of these resources differ greatly in quantity and quality of functional information; Brainspan and GO cover most genes but with less precise information whereas KEGG and MGI cover only a few thousand genes but are curated from detailed experiments. However, I showed that taking the intersection of multiple different functional resources selects for the most well-studied genes frequently mentioned in the literature (Figure 3.11 A). This bias towards highly studied genes should be most pronounced in the functional resources with the least coverage of genes (ie. KEGG and MGI) and indeed these two resources consistently show the least phenotypic convergence (Figure 3.6). However another contributing factor could be the use of the same studies to inform each of the different functional annotation databases. For instance, it has been shown that GO annotations are often based on the same original literature as PPI database, 10-20% of literature sources were used by both affecting 66% of GO terms (42). This type of confound would greatly reduce the efficacy of intersecting different functional resources since it is likely they do not represent independent replications of an association but rather a single well-known study.

The integrated functional network used in Chapters 4 and‘ 5, and the construction of which is described in Chapter 2 and (65), down weighted each additional functional resource which included a given gene pair to reduce biases towards well studied genes. Instead it takes the majority of the final functional similarity value between a given pair of genes from the single most useful functional resource for that gene pair. The resulting network had a greater coverage of genes, whereas intersecting multiple resources results in lower coverage, and a stronger relationship with human phenotypic similarity than any of the integrated datasets did alone (Figure 2.3). Many other methods have been used to integrate various sources of functional information based on Bayesian statistics (53; 54; 224); which similarly use a training dataset of high confidence functional relationships to determine a measure of the probability a pair of genes are functionally related using Bayes Theorem and then sum the log of each probability over all input datasets. However, often these methods do not correct for the effect of some genes being represented in more of the source datasets (GO, PPIs, expression etc...) due to research bias thus may suffer more bias towards well-studied genes.

7.2.2 Gene expression vs protein-protein interactions

Protein-protein interactions (PPIs) are one of the most common types of gene networks used in the interpretation of genetic variants (eg. exomes (218), CNVs (21; 22), and GWAS-associated SNPs(225)) and in systems biology (eg. examining modular organization(226), lethality(227), and guilt-by-association predictions(228)). Some reasons for this include the ease of use and of interpretation of protein-protein interactions. Current PPI networks typically included 10-20 interactions per gene (Table 6.2) which are either present or absent between any given pair of genes/proteins; whereas co-expression networks contain hundreds of millions of weighted edges most of represent weak correlations. Since it is difficult to interpret the functional significance of weak gene expression correlations, they are often excluded using a relatively arbitrary threshold (119; 120; 121; 122). However PPIs seem intuitively to be more directly related to protein function since many important cellular functions are per-

formed by complexes of interacting proteins whereas genes that are co-expressed may not have direct functional relationship, eg. Glyceraldehyde 3-phosphate dehydrogenase (GAPDH) and ribosomal proteins/RNAs are both broadly expressed housekeeping genes but participate in very different biological processes, glycolysis and translation respectively (229).

However, in Chapter 3 and 6 I found that gene expression was a more useful representation of gene function (or functional similarity) than PPIs. The pathways identified using gene expression in the brain were most likely to have phenotypic convergence among patients with CNVs affecting genes in the pathway (Figure 3.6). In contrast after combining pathways and identifying those genes participating in significantly well connected PPIs, the resulting PPI-pathways were not more likely to have phenotypic convergence than the original pathways. Furthermore in the case of the extended pathways where both GO and gene expression identified many pathways showing phenotypic convergence, adding PPIs to the GO or gene expression pathways greatly reduced the number of pathways showing phenotypic convergence (Figure 3.6 C, GO-PPI & BS-PPI vs BS-GO). However I found that the specific network used is important as the brain-specific co-expression network showed much more phenotypic convergence than a more general co-expression network when looking among patients with neurodevelopmental disorders (Figure 3.2).

In Chapter 3, I used modules identified a co-expression network, a PPI network and an integrated network to predict the phenotypes exhibited by the knock-outs of the mouse-orthologs of the genes. The co-expression network out-performed both the PPI and integrated network in this regard (Figure 6.3). In addition, there were more differences between clustering methods in the PPI network than either of the others suggesting a less well-defined modular structure. This was surprising since many studies have previously identified robust modular structure of PPI networks eg. (226; 230; 231). However, despite the instability of the modules they did perform significantly better than expected by chance, and even out-performed HumanNet (53), the integrated net-

work in predicting phenotypes.

On the other hand, co-expression was found to be rather uninformative when constructing the phenotypic linkage network (PLN) and contributes very little to the final network (Figures 2.3, 5.6); whereas co-citation which is primarily derived from PPI information in the literature (51) and other protein-protein interaction databases are the third and fourth largest contributors to the PLN. In Chapter 5, I compared the PLN which to a co-expression only network for identifying functional clusters in the genome. I find that while the co-expression only network identifies a larger number and more significantly many genes participating in functional clusters (Figure 5.4, 5.5) it preferentially identifies clusters of housekeeping genes over clusters of tissue-specific genes whereas the PLN does not (Figure 5.7). This could represent leaky expression, where the chromatin modifications and other activating factors maintaining high expression from the house-keeping genes also raises the expression of neighbouring genes (92). This leaky expression is not necessarily useful to the cell but rather a by-product of particular mechanisms of transcriptional regulation. Another factor which could increase the co-expression between nearby genes in the genome are bidirectional promoters, where a single promoter region controls the transcription of two genes in opposing directions (164). Thus protein-protein interactions may be preferable to co-expression when looking at genes close together in the genome, where as co-expression is more useful when looking for functional relationships between genes at distinct loci.

7.2.3 Mouse Phenotypes

Another popular source of gene functional information are the phenotypes observed in model organisms which carry a mutant allele in the gene. Mice, being mammals and thus closely related to humans, should be most relevant to human disease(232). Indeed in support of this intuition, I found during the construction of the Phenotypic Linkage

Network (PLN), that mouse phenotype information was more strongly correlated with human disease phenotypes than the combined network after integrating all other data-sources together and contributed a substantial increase in the quality and quantity of links in the final PLN (Figure 2.3). However, when I used these data to identify pathways enriched among patients with particular developmental phenotypes in Chapter 3, I found mouse phenotype information performed relatively poorly. Mouse phenotypes identified the fewest number of significantly enriched pathways and those pathways showed almost no evidence of broader phenotypic similarity amongst patients with variants affecting the pathway (Figure 3.6).

One of the reasons for the poor performance of mouse phenotypes to identify pathways associated with disease may be the low coverage of genes with mouse phenotype information and a bias towards genes known to be associated with many human disease phenotypes since the Mouse Genome Informatics database only contains information for 8,000 different genes 18% of which are known human disease genes (Table 6.1). In addition to less than a quarter of genes having any mouse phenotype information, I also found little evidence of saturation in the number of unique phenotypes which suggests many phenotypes are missing from even those genes that have been examined (Figure 6.6). Consistent with this I was only able to test a small number of mouse-phenotype pathways for phenotypic convergence among the respective patients with variants affecting the pathway due to there being fewer than 10 patients in the cohort with variants affected genes known to cause the mouse phenotype (Figure 3.6).

An explanation for the lack of phenotypic convergence among those mouse-phenotype derived pathways, despite the reported agreements between mouse and human phenotype presentations (22), is the presence of ascertainment bias among mouse phenotype information. Ascertainment bias both in the genes selected for study and the choice of phenotypes to examine in the mouse would lead to an over-estimation of the agreement between mouse and human phenotype presentations for well known dis-

ease genes, such as those examined in (22), in comparison to genes more recently seen affected in human patient cohorts like the GENCODYS dataset examined here (18). I have showed that the Mouse Genome Informatics database is significantly biased towards known disease genes and the genes shown to be functionally related to them (Figure 6.1, Table 6.1). Furthermore the biases I observed towards long genes among mouse phenotype resources (Figure 6.2 A) could confound studies which integrate these data with variants obtained from exome sequencing (such as (217),(233)); as long genes are also more likely to harbour *de novo* SNVs (65). Thus, the results presented here suggest the current mouse phenotype resources should be used with caution and appropriate controls must be used to avoid false positives resulting from the ascertainment biases present.

Chapter 8

Bibliography

[1] P. A. Baird, T. W. Anderson, H. B. Newcombe, and R. B. Lowry, "Genetic disorders in children and young adults: a population study.," *American Journal of Human Genetics*, vol. 42, no. 5, pp. 677–693, 1988.

[2] D. T. Miller, M. P. Adam, S. Aradhya, L. G. Biesecker, A. R. Brothman, N. P. Carter, D. M. Church, J. A. Crolla, E. E. Eichler, C. J. Epstein, W. A. Faucett, L. Feuk, J. M. Friedman, A. Hamosh, L. Jackson, E. B. Kaminsky, K. Kok, I. D. Krantz, R. M. Kuhn, C. Lee, J. M. Ostell, C. Rosenberg, S. W. Scherer, N. B. Spinner, D. J. Stavropoulos, J. H. Tepperberg, E. C. Thorland, J. R. Vermeesch, D. J. Waggoner, M. S. Watson, C. L. Martin, and D. H. Ledbetter, "Consensus statement: Chromosomal microarray is a first-tier clinical diagnostic test for individuals with developmental disabilities or congenital anomalies," *The American Journal of Human Genetics*, vol. 86, pp. 749–764, 5/14 2010.

[3] T. Hart and P. Hart, "Genetic studies of craniofacial anomalies: clinical implications and applications," *Orthodontics & Craniofacial Research*, vol. 12, no. 3, pp. 212–220, 2009. 7000 genetic disorders, 30-40% craniofacial.

[4] G. M. Cooper, B. P. Coe, S. Girirajan, J. A. Rosenfeld, T. H. Vu, C. Baker, C. Williams, H. Stalker, R. Hamid, V. Hannig, H. Abdel-Hamid, P. Bader, E. McCracken, D. Niyazov, K. Leppig, H. Thiese, M. Hummel, N. Alexander, J. Gorski, J. Kussmann, V. Shashi, K. Johnson, C. Rehder, B. C. Ballif, L. G. Shaffer, and E. E.

Eichler, "A copy number variation morbidity map of developmental delay," *Nature genetics*, vol. 43, pp. 838–846, print 2011.

[5] H. Dolk, M. Loane, and E. Garne, "The prevalence of congenital anomalies in europe," *Adv Exp Med Biol*, vol. 686, pp. 349–64, 2010.

[6] B. P. Coe, S. Girirajan, and E. E. Eichler, "The genetic variability and commonality of neurodevelopmental disease," *American Journal of Medical Genetics Part C: Seminars in Medical Genetics*, vol. 160C, no. 2, pp. 118–129, 2012.

[7] S. Girirajan, J. A. Rosenfeld, B. P. Coe, S. Parikh, N. Friedman, A. Goldstein, R. A. Filipink, J. S. McConnell, B. Angle, W. S. Meschino, M. M. Nezarati, A. Asamoah, K. E. Jackson, G. C. Gowans, J. A. Martin, E. P. Carmany, D. W. Stockton, R. E. Schnur, L. S. Penney, D. M. Martin, S. Raskin, K. Leppig, H. Thiese, R. Smith, E. Aberg, D. M. Niyazov, L. F. Escobar, D. El-Khechen, K. D. Johnson, R. R. Lebel, K. Siefkas, S. Ball, N. Shur, M. McGuire, C. K. Brasington, J. E. Spence, L. S. Martin, C. Clericuzio, B. C. Ballif, L. G. Shaffer, and E. E. Eichler, "Phenotypic heterogeneity of genomic disorders and rare copy-number variants," *N Engl J Med*, vol. 367, no. 14, pp. 1321–1331, 2012.

[8] L. E. L. M. Vissers, J. A. Veltman, A. G. van Kessel, and H. G. Brunner, "Identification of disease genes by whole genome cgh arrays," *Human molecular genetics*, vol. 14, pp. R215–R223, 15 October 2005 2005.

[9] S. J. Sanders, A. G. Ercan-Sencicek, V. Hus, R. Luo, M. T. Murtha, D. Moreno-De-Luca, S. H. Chu, M. P. Moreau, A. R. Gupta, S. A. Thomson, C. E. Mason, K. Bilguvar, P. B. S. Celestino-Soper, M. Choi, E. L. Crawford, L. Davis, N. R. D. Wright, R. M. Dhodapkar, M. DiCola, N. M. DiLullo, T. V. Fernandez, V. Fielding-Singh, D. O. Fishman, S. Frahm, R. Garagaloyan, G. S. Goh, S. Kammela, L. Klei, J. K. Lowe, S. C. Lund, A. D. McGrew, K. A. Meyer, W. J. Moffat, J. D. Murdoch, B. J. O'Roak, G. T. Ober, R. S. Pottenger, M. J. Raubeson, Y. Song, Q. Wang, B. L. Yaspan, T. W. Yu, I. R. Yurkiewicz, A. L. Beaudet, R. M. Cantor, M. Curland, D. E. Grice, M. GÃ¼nel, R. P. Lifton, S. M. Mane, D. M. Martin, C. A. Shaw, M. Sheldon, J. A. Tischfield, C. A. Walsh, E. M. Morrow, D. H. Ledbetter, E. Fombonne,

C. Lord, C. L. Martin, A. I. Brooks, J. S. Sutcliffe, E. H. C. Jr., D. Geschwind, K. Roeder, B. Devlin, and M. W. State, "Multiple recurrent de novo cnvs, including duplications of the 7q11.23 williams syndrome region, are strongly associated with autism," *Neuron*, vol. 70, pp. 863–885, 6/9 2011.

[10] D. Levy, M. Ronemus, B. Yamrom, Y. ha Lee, A. Leotta, J. Kendall, S. Marks, B. Lakshmi, D. Pai, K. Ye, A. Buja, A. Krieger, S. Yoon, J. Troge, L. Rodgers, I. Iossifov, and M. Wigler, "Rare de novo and transmitted copy-number variation in autistic spectrum disorders," *Neuron*, vol. 70, pp. 886–897, 6/9 2011.

[11] J. de Ligt, M. H. Willemsen, B. W. van Bon, T. Kleefstra, H. G. Yntema, T. Kroes, A. T. Vulto-van Silfhout, D. A. Koolen, P. de Vries, C. Gilissen, M. del Rosario, A. Hoischen, H. Scheffer, B. B. de Vries, H. G. Brunner, J. A. Veltman, and L. E. Vissers, "Diagnostic exome sequencing in persons with severe intellectual disability," *New England Journal of Medicine*, vol. 367, no. 20, pp. 1921–1929, 2012. PMID: 23033978.

[12] D. Malhotra and J. Sebat, "Cnvs: Harbingers of a rare variant revolution in psychiatric genetics," *Cell*, vol. 148, pp. 1223–1241, 2014/07 2014. doi: 10.1016/j.cell.2012.02.039; 31.

[13] R. Hochstenbach, M. Poot, I. J. Nijman, I. Renkens, K. J. Duran, R. van'T Slot, E. van Binsbergen, der Zwaag van, M. J. Vogel, P. A. Terhal, van Amstel Ploos, W. P. Kloosterman, and E. Cuppen, "Discovery of variants unmasked by hemizygous deletions," *European journal of human genetics : EJHG*, vol. 20, pp. 748–753, print 2012.

[14] J. A. Veltman and H. G. Brunner, "De novo mutations in human genetic disease," *Nature reviews. Genetics*, vol. 13, pp. 565–575, print 2012. M3: 10.1038/nrg3241; 10.1038/nrg3241.

[15] J. Sebat, B. Lakshmi, D. Malhotra, J. Troge, C. Lese-Martin, T. Walsh, B. Yamrom, S. Yoon, A. Krasnitz, J. Kendall, A. Leotta, D. Pai, R. Zhang, Y.-H. Lee, J. Hicks, S. J. Spence, A. T. Lee, K. Puura, T. Lehtimäki, D. Ledbetter, P. K. Gregersen, J. Bregman, J. S. Sutcliffe, V. Jobanputra, W. Chung, D. Warburton, M.-C. King,

D. Skuse, D. H. Geschwind, T. C. Gilliam, K. Ye, and M. Wigler, "Strong association of de novo copy number mutations with autism," *Science*, vol. 316, pp. 445–449, April 20 2007.

[16] C. P. Schaaf, J. Wiszniewska, and A. L. Beaudet, "Copy number and snp arrays in clinical diagnostics," *Annu Rev Genomics Hum Genet*, vol. 12, pp. 25–51, 2011.

[17] E. Bragin, E. A. Chatzimichali, C. F. Wright, M. E. Hurles, H. V. Firth, A. P. Bevan, and G. J. Swaminathan, "Decipher: database for the interpretation of phenotype-linked plausibly pathogenic sequence and copy-number variation," *Nucleic Acids Res*, vol. 42, no. Database issue, pp. D993–D1000, 2014.

[18] A. T. Vulto-van Silfhout, J. Y. Hehir-Kwa, B. W. van Bon, J. H. Schuurs-Hoeijmakers, S. Meader, C. J. Hellebrekers, I. J. Thoonen, A. P. de Brouwer, H. G. Brunner, C. Webber, R. Pfundt, N. de Leeuw, and B. B. de Vries, "Clinical significance of de novo and inherited copy-number variation," *Hum Mutat*, vol. 34, no. 12, pp. 1679–87, 2013.

[19] P. RA, K. S, U. R, and et al, "Fecundity of patients with schizophrenia, autism, bipolar disorder, depression, anorexia nervosa, or substance abuse vs their unaffected siblings," *JAMA Psychiatry*, vol. 70, no. 1, pp. 22–30, 2013.

[20] H. Boulding and C. Webber, "Large-scale objective association of mouse phenotypes with human symptoms through structural variation identified in patients with developmental disorders," *Human mutation*, vol. 33, no. 5, pp. 874–883, 2012.

[21] H. J. Noh, C. P. Ponting, H. C. Boulding, S. Meader, C. Betancur, J. D. Buxbaum, D. Pinto, C. R. Marshall, A. C. Lionel, S. W. Scherer, and C. Webber, "Network topologies and convergent aetiologies arising from deletions and duplications observed in individuals with autism," *PLoS Genet*, vol. 9, no. 6, p. e1003523, 2013.

[22] S. C. Doelken, S. KÃ¶hler, C. J. Mungall, G. V. Gkoutos, B. J. Ruef, C. Smith, D. Smedley, S. Bauer, E. Klopocki, P. N. Schofield, M. Westerfield, P. N. Robinson, and S. E. Lewis, "Phenotypic overlap in the contribution of individual genes to

cnv pathogenicity revealed by cross-species computational analysis of single-gene mutations in humans, mice and zebrafish," *Disease Models & Mechanisms*, vol. 6, pp. 358–372, March 01 2013.

[23] C. Golzio, J. Willer, M. E. Talkowski, E. C. Oh, Y. Taniguchi, S. Jacquemont, A. Reymond, M. Sun, A. Sawa, J. F. Gusella, A. Kamiya, J. S. Beckmann, and N. Katsanis, "Kctd13 is a major driver of mirrored neuroanatomical phenotypes of the 16p11.2 copy number variant," *Nature*, vol. 485, pp. 363–367, 05/17 2012. M3: 10.1038/nature11091; 10.1038/nature11091.

[24] S. Girirajan, M. Y. Dennis, C. Baker, M. Malig, B. P. Coe, C. D. Campbell, K. Mark, T. H. Vu, C. Alkan, and Z. Cheng, "Refinement and discovery of new hotspots of copy-number variation associated with autism spectrum disorder," *The American Journal of Human Genetics*, vol. 92, pp. 221–237, 2013.

[25] A. Itsara, G. M. Cooper, C. Baker, S. Girirajan, J. Li, D. Absher, R. M. Krauss, R. M. Myers, P. M. Ridker, D. I. Chasman, H. Mefford, P. Ying, D. A. Nickerson, and E. E. Eichler, "Population analysis of large copy number variants and hotspots of human genetic disease," *American Journal of Human Genetics*, vol. 84, no. 2, pp. 148–161, 2009.

[26] M. Vidal, M. E. Cusick, and A. L. Barabasi, "Interactome networks and human disease," *Cell*, vol. 144, no. 6, pp. 986–98, 2011.

[27] V. K. Ramanan, L. Shen, J. H. Moore, and A. J. Saykin, "Pathway analysis of genomic data: concepts, methods, and prospects for future development," *Trends Genet*, vol. 28, no. 7, pp. 323–32, 2012.

[28] M. Oti and H. G. Brunner, "The modular nature of genetic diseases," *Clin Genet*, vol. 71, no. 1, pp. 1–11, 2007.

[29] D. Pinto, A. T. Pagnamenta, L. Klei, R. Anney, D. Merico, R. Regan, J. Conroy, T. R. Magalhaes, C. Correia, B. S. Abrahams, J. Almeida, E. Bacchelli, G. D. Bader, A. J. Bailey, G. Baird, A. Battaglia, T. Berney, N. Bolshakova, S. Bolte, P. F. Bolton, T. Bourgeron, S. Brennan, J. Brian, S. E. Bryson, A. R. Carson, G. Casallo, J. Casey,

B. H. Y. Chung, L. Cochrane, C. Corsello, E. L. Crawford, A. Crossett, C. Cytrynbaum, G. Dawson, M. de Jonge, R. Delorme, I. Drmic, E. Duketis, F. Duque, A. Estes, P. Farrar, B. A. Fernandez, S. E. Folstein, E. Fombonne, C. M. Freitag, J. Gilbert, C. Gillberg, J. T. Glessner, J. Goldberg, A. Green, J. Green, S. J. Guter, H. Hakonarson, E. A. Heron, M. Hill, R. Holt, J. L. Howe, G. Hughes, V. Hus, R. Igliozzi, C. Kim, S. M. Klauck, A. Kolevzon, O. Korvatska, V. Kustanovich, C. M. Lajonchere, J. A. Lamb, M. Laskawiec, M. Leboyer, A. L. Couteur, B. L. Leventhal, A. C. Lionel, X.-Q. Liu, C. Lord, L. Lotspeich, S. C. Lund, E. Maestrini, W. Mahoney, C. Mantoulan, C. R. Marshall, H. McConachie, C. J. McDougle, J. McGrath, W. M. McMahon, A. Merikangas, O. Migita, N. J. Minshew, G. K. Mirza, J. Munson, S. F. Nelson, C. Noakes, A. Noor, G. Nygren, G. Oliveira, K. Papanikolaou, J. R. Parr, B. Parrini, T. Paton, A. Pickles, M. Pilorge, J. Piven, C. P. Ponting, D. J. Posey, A. Poustka, F. Poustka, A. Prasad, J. Ragoussis, K. Renshaw, J. Rickaby, W. Roberts, K. Roeder, B. Roge, M. L. Rutter, L. J. Bierut, J. P. Rice, J. Salt, K. Sansom, D. Sato, R. Segurado, A. F. Sequeira, L. Senman, N. Shah, V. C. Sheffield, L. Soorya, I. Sousa, O. Stein, N. Sykes, V. Stoppioni, C. Strawbridge, R. Tancredi, K. Tansey, B. Thiruvahindrapduram, A. P. Thompson, S. Thomson, A. Tryfon, J. Tsiantis, H. V. Engeland, J. B. Vincent, F. Volkmar, S. Wallace, K. Wang, Z. Wang, T. H. Wassink, C. Webber, R. Weksberg, K. Wing, K. Wittemeyer, S. Wood, J. Wu, B. L. Yaspan, D. Zurawiecki, L. Zwaigenbaum, J. D. Buxbaum, R. M. Cantor, E. H. Cook, H. Coon, M. L. Cuccaro, B. Devlin, S. Ennis, L. Gallagher, D. H. Geschwind, M. Gill, J. L. Haines, J. Hallmayer, J. Miller, A. P. Monaco, J. I. N. Jr, A. D. Paterson, M. Pericak-Vance, G. D. Schellenberg, P. Szatmari, A. M. Vicente, V. J. Vieland, E. M. Wijsman, S. W. Scherer, J. S. Sutcliffe, and C. Betancur, "Functional impact of global rare copy number variation in autism spectrum disorders," *Nature*, vol. 466, pp. 368–372, 07/15 2010. 10.1038/nature09146.

[30] P. Stankiewicz and J. R. Lupski, "Structural variation in the human genome and its role in disease," *Annu Rev Med*, vol. 61, pp. 437–55, 2010.

[31] J. H. Schuurs-Hoeijmakers, E. C. Oh, L. E. Vissers, C. Swinkels, Mar-

iëlleÂ E.M.and Gillissen, M. A. Willemsen, M. Holvoet, M. Steehouwer, J. Veltman, B. deÂ Vries, H. vanÂ Bokhoven, A. deÂ Brouwer, N. Katsanis, K. Devriendt, and H. Brunner, "Recurrent de novo mutations in pacs1 cause defective cranial-neural-crest migration and define a recognizable intellectual-disability syndrome," *The American Journal of Human Genetics*, vol. 91, no. 6, pp. 1122–1127, 2012.

[32] "Prevalence of rare diseases: Bibliographic data," in *Orphannet Report Series, Rare diseases collection* (N. Marpillat, ed.), Springer New York, May 2014.

[33] M. Kanehisa and S. Goto, "Kegg: kyoto encyclopedia of genes and genomes," *Nucleic Acids Res*, vol. 28, no. 1, pp. 27–30, 2000.

[34] The Gene Ontology Consortium, M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock, "Gene ontology: tool for the unification of biology. the gene ontology consortium," *Nat Genet*, vol. 25, no. 1, pp. 25–9, 2000.

[35] P. N. Robinson, S. Köhler, S. Bauer, D. Seelow, D. Horn, and S. Mundlos, "The human phenotype ontology: A tool for annotating and analyzing human hereditary disease," *The American Journal of Human Genetics*, vol. 83, pp. 610–615, 11/17 2008.

[36] C. J. Bult, J. T. Eppig, J. A. Kadin, J. E. Richardson, J. A. Blake, and the Mouse Genome Database Group, "The mouse genome database (mgd): mouse biology and model systems," *Nucleic acids research*, vol. 36, pp. D724–D728, January 01 2008.

[37] J. T. Eppig, J. A. Blake, C. J. Bult, J. E. Richardson, J. A. Kadin, and M. Ringwald, "Mouse genome informatics (mgi) resources for pathology and toxicology," *Toxicol Pathol*, vol. 35, no. 3, pp. 456–7, 2007.

[38] C. Webber, "Functional enrichment analysis with structural variants: Pitfalls and

strategies," *Cytogenet Genome Res*, 2011.

[39] R. Kariminejad, A. Lind-Thomsen, Z. Tumer, F. Erdogan, H. H. Ropers, N. Tommerup, R. Ullmann, and R. S. Moller, "High frequency of rare copy number variants affecting functionally related genes in patients with structural brain malformations," *Hum Mutat*, vol. 32, no. 12, pp. 1427–35, 2011.

[40] X. Gai, H. M. Xie, J. C. Perin, N. Takahashi, K. Murphy, A. S. Wenocur, M. D'Arcy, R. J. O'Hara, E. Goldmuntz, D. E. Grice, T. H. Shaikh, H. Hakonarson, J. D. Buxbaum, J. Elia, and P. S. White, "Rare structural variation of synapse and neurotransmission genes in autism," *Mol Psychiatry*, vol. 17, no. 4, pp. 402–11, 2012.

[41] C. Webber, J. Y. Hehir-Kwa, D. Q. Nguyen, B. B. de Vries, J. A. Veltman, and C. P. Ponting, "Forging links between human mental retardation-associated cnvs and mouse gene knockout models," *PLoS Genet*, vol. 5, no. 6, p. e1000531, 2009.

[42] J. Gillis and P. Pavlidis, "Assessing identity, redundancy and confounds in gene ontology annotations over time," *Bioinformatics*, vol. 29, no. 4, pp. 476–82, 2013.

[43] A. M. Schnoes, S. D. Brown, I. Dodevski, and P. C. Babbitt, "Annotation error in public databases: misannotation of molecular function in enzyme superfamilies," *PLoS Comput Biol*, vol. 5, no. 12, p. e1000605, 2009.

[44] C. Andorf, D. Dobbs, and V. Honavar, "Exploring inconsistencies in genome-wide protein function annotations: a machine learning approach," *BMC Bioinformatics*, vol. 8, p. 284, 2007.

[45] E. Sprinzak, S. Sattath, and H. Margalit, "How reliable are experimental protein-protein interaction data?," *Journal of Molecular Biology*, vol. 327, no. 5, pp. 919–923, 2003.

[46] T. Obayashi, S. Hayashi, M. Shibaoka, M. Saeki, H. Ohta, and K. Kinoshita, "Coxpresdb: a database of coexpressed gene networks in mammals," *Nucleic Acids Res*, vol. 36, no. Database issue, pp. D77–82, 2008.

[47] R. Xulvi-Brunet and H. Li, "Co-expression networks: graph properties and topological comparisons," *Bioinformatics*, vol. 26, pp. 205–214, January 15 2010.

[48] J. M. Stuart, E. Segal, D. Koller, and S. K. Kim, "A gene-coexpression network for global discovery of conserved genetic modules," *Science*, vol. 302, pp. 249–255, October 10 2003.

[49] M. Bansal, V. Belcastro, A. Ambesi-Impiombato, and D. di Bernardo, "How to infer gene networks from expression profiles," *Mol Syst Biol*, vol. 3, p. 78, 02/13 2007. M3: 10.1038/msb4100120; 10.1038/msb4100120.

[50] B. Zhang and S. Horvath, "A general framework for weighted gene co-expression network analysis.," *Statistical Applications in Genetics and Molecular Biology*, vol. 4, no. 1, p. Article 17, 2005.

[51] D. Szklarczyk, A. Franceschini, M. Kuhn, M. Simonovic, A. Roth, P. Minguez, T. Doerks, M. Stark, J. Muller, P. Bork, L. J. Jensen, and C. von Mering, "The string database in 2011: functional interaction networks of proteins, globally integrated and scored," *Nucleic acids research*, vol. 39, pp. D561–D568, January 01 2011.

[52] S. Razick, G. Magklaras, and I. Donaldson, "irefindex: A consolidated protein interaction database with provenance," *BMC Bioinformatics*, vol. 9, no. 1, p. 405, 2008. M3: 10.1186/1471-2105-9-405.

[53] I. Lee, U. M. Blom, P. I. Wang, J. E. Shim, and E. M. Marcotte, "Prioritizing candidate disease genes by network-based boosting of genome-wide association data," *Genome research*, vol. 21, pp. 1109–1121, July 01 2011.

[54] A. Alexeyenko and E. L. L. Sonnhammer, "Global networks of functional coupling in eukaryotes from comprehensive data integration," *Genome research*, vol. 19, pp. 1107–1116, June 01 2009.

[55] I. Lee, S. V. Date, A. T. Adai, and E. M. Marcotte, "A probabilistic functional network of yeast genes," *Science*, vol. 306, pp. 1555–1558, November 26 2004.

[56] O. G. Troyanskaya, K. Dolinski, A. B. Owen, R. B. Altman, and D. Botstein, "A bayesian framework for combining heterogeneous data sources for gene function prediction (in saccharomyces cerevisiae)," *Proceedings of the National Academy of Sciences*, vol. 100, pp. 8348–8353, July 08 2003.

[57] B. Lehner and I. Lee, "Network-guided genetic screening: building, testing and using gene networks to predict gene function," *Briefings in Functional Genomics & Proteomics*, vol. 7, pp. 217–227, May 01 2008.

[58] J. C. Costello, M. M. Dalkilic, S. M. Beason, J. R. Gehlhausen, R. Patwardhan, S. Middha, B. D. Eads, and J. R. Andrews, "Gene networks in drosophila melanogaster: integrating experimental datas to predict gene function.," *Genome biology*, vol. 10, no. R97, 2009.

[59] M. Deng, T. Chen, and F. Sun, "An integrated porbabilistic model for functional prediction of proteins.," *Journal of Computational Biology*, vol. 11, no. 2-3, pp. 463–475, 2004.

[60] D. Warde-Farley, S. L. Donaldson, O. Comes, K. Zuberi, R. Badrawi, P. Chao, M. Franz, C. Grouios, F. Kazi, C. T. Lopes, A. Maitland, S. Mostafavi, J. Montojo, Q. Shao, G. Wright, G. D. Bader, and Q. Morris, "The genemania prediction server: biological network integration for gene prioritization and predicting gene function," *Nucleic acids research*, vol. 38, pp. W214–W220, July 01 2010.

[61] I. Lee and E. M. Marcotte, "Integrating functional genomics data," *Methods Mol Biol*, vol. 453, pp. 267–78, 2008.

[62] K. Wabnik, T. R. Hvidsten, A. Kedzierska, J. Van Leene, G. De Jaeger, G. T. Beemster, J. Komorowski, and M. T. Kuiper, "Gene expression trends and protein features effectively complement each other in gene function prediction," *Bioinformatics*, vol. 25, no. 3, pp. 322–30, 2009.

[63] A.-L. Barabási and Z. N. Oltvai, "Network biology: understanding the cell's functional organization," *Nature reviews. Genetics*, vol. 5, pp. 101–113, print 2004. M3: 10.1038/nrg1272; 10.1038/nrg1272.

[64] T. Opsahl, G. Agneessens, and J. Skvoretz, "Node centrality in weighted networks: Generalizing degree and shortest paths," *Social Networks*, vol. 32, no. 3, pp. 245–251, 2010.

[65] F. Honti, S. Meader, and C. Webber, "Unbiased functional clustering of gene vari-

ants with a phenotypic-linkage network," *PLoS Comput Biol*, vol. 10, p. e1003815, 08/28 2014.

[66] M. Young, M. Wakefield, G. Smyth, and A. Oshlack, "Gene ontology analysis for rna-seq: accounting for selection bias," *Genome biology*, vol. 11, no. 2, p. R14, 2010. M3: 10.1186/gb-2010-11-2-r14.

[67] A. Oshlack and M. Wakefield, "Transcript length bias in rna-seq data confounds systems biology," *Biology Direct*, vol. 4, no. 1, p. 14, 2009.

[68] E. W. Dijkstra, "A note on two problems in connexion with graphs," *Numerische Mathematik*, vol. 1, pp. 269–271, 12/01 1959. J2: Numer. Math.

[69] M. Rosvall and C. T. Bergstrom, "Multilevel compression of random walks on networks reveals hierarchical organization in large integrated systems," *PLoS ONE*, vol. 6, p. e18209, 04/08 2011. M3: doi:10.1371/journal.pone.0018209.

[70] U. Feige, G. Kortsarz, and D. Peleg, "The dense k-subgraph problem," *Algorithmica*, vol. 29, p. 2001, 1999.

[71] R. K. Darst, D. R. Reichman, P. Ronhovde, and Z. Nussinov, "An edge density definition of overlapping and weighted graph communities," *CoRR*, vol. abs/1301.3120, 2013.

[72] M. Girvan and M. E. J. Newman, "Community structure in social and biological networks," *Proceedings of the National Academy of Sciences*, vol. 99, no. 12, pp. 7821–7826, 2002.

[73] E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai, and A. L. Barabási, "Hierarchical organization of modularity in metabolic networks," *Science*, vol. 297, pp. 1551–1555, 08/30 2002. 10.1126/science.1073374.

[74] L. H. Hartwell, J. J. Hopfield, S. Leibler, and A. W. Murray, "From molecular to modular cell biology," *Nature*, vol. 402, pp. C7–C88, December 1999.

[75] A. W. Rives and T. Galitski, "Modular organization of cellular networks," *Pro-*

ceedings of the National Academy of Sciences, vol. 100, pp. 1128–1133, February 04 2003.

[76] J. F. Poyatos and L. D. Hurst, "Is optimal gene order impossible?," *Trends in Genetics*, vol. 22, no. 8, p. 420, 2006.

[77] C. Pal and L. D. Hurst, "Evidence for co-evolution of gene order and recombination rate," *Nature genetics*, vol. 33, pp. 392–395, print 2003. M3: 10.1038/ng1111; 10.1038/ng1111.

[78] B. A. Cohen, R. D. Mitra, J. D. Hughes, and G. M. Church, "A computational analysis of whole-genome expression data reveals chromosomal domains of gene expression," *Nature genetics*, vol. 26, pp. 183–186, print 2000. M3: 10.1038/79896; 10.1038/79896.

[79] Q. Li, B. Lee, and L. Zhang, "Genome-scale analysis of positional clustering of mouse testis-specific genes," *BMC Genomics*, vol. 6, no. 1, p. 7, 2005. M3: 10.1186/1471-2164-6-7.

[80] G. A. C. Singer, A. T. Lloyd, L. B. Huminiecki, and K. H. Wolfe, "Clusters of co-expressed genes in mammalian genomes are conserved by natural selection," *Molecular biology and evolution*, vol. 22, pp. 767–775, MAR 2005 2005. PT: J; TC: 60; UT: WOS:000227163100044.

[81] C. C. Weber and L. D. Hurst, "Support for multiple classes of local expression clusters in drosophila melanogaster but no evidence for gene order conservation," *Genome Biology*, vol. 12, no. 3, 2011.

[82] P. T. Spellman and G. M. Rubin, "Evidence for large domains of similarly expressed genes in the drosophila genome.," *Journal of biology*, vol. 1, p. 5, 2002 2002. PT: J; UT: MEDLINE:12144710.

[83] J. Mezey, S. Nuzhdin, F. Ye, and C. Jones, "Coordinated evolution of co-expressed gene clusters in the drosophila transcriptome," *BMC Evolutionary Biology*, vol. 8, no. 1, p. 2, 2008. M3: 10.1186/1471-2148-8-2.

[84] R. S. Kamath, A. G. Fraser, Y. Dong, G. Poulin, R. Durbin, M. Gotta,

A. Kanapin, N. L. Bot, S. Moreno, M. Sohrmann, D. P. Welchman, P. Zipperlen, and J. Ahringer, "Systematic functional analysis of the caenorhabidits elegans genome using rnai," *Nature*, vol. 421, no. 6920, pp. 231–237, 2003.

[85] Y. K. Ng, W. Wu, and L. Zhang, "Positive correlation between gene coexpression and positional clustering in the zebrafish genome," *Bmc Genomics*, vol. 10, p. 42, JAN 22 2009 2009. PT: J; TC: 5; UT: WOS:000264970300001.

[86] F. Al-Shahrour, P. Minguez, T. Marques-Bonet, E. Gazave, A. Navarro, and J. Dopazo, "Selection upon genome architecture: Conservation of functional neighborhoods with changing genes," *Plos Computational Biology*, vol. 6, p. e1000953, OCT 2010 2010. PT: J; TC: 2; UT: WOS:000283651900019.

[87] P. Michalak, "Coexpression, coregulation, and cofunctionality of neighboring genes in eukaryotic genomes," *Genomics*, vol. 91, pp. 243–248, 3 2008.

[88] Y. Fukuoka, H. Inaoka, and I. Kohane, "Inter-species differences of co-expression of neighboring genes in eukaryotic genomes," *BMC Genomics*, vol. 5, no. 1, p. 4, 2004.

[89] J. M. Lee and E. L. L. Sonnhammer, "Genomic gene clustering analysis of pathways in eukaryotes," *Genome research*, vol. 13, no. 5, pp. 875–882, 2003.

[90] H. Caron, B. van Schaik, M. van der Mee, F. Baas, G. Riggins, P. van Sluis, M.-C. Hermus, R. van Asperen, K. Boon, P. A. Voûte, S. Heisterkamp, A. van Kampen, and R. Versteeg, "The human transcriptome map: Clustering of highly expressed genes in chromosomal domains," *Science*, vol. 291, no. 5507, pp. 1289–1292, 2001.

[91] T. Makino and A. McLysaght, "Interacting gene clusters and the evolution of the vertebrate immune system," *Molecular biology and evolution*, vol. 25, pp. 1855–1862, SEP 2008 2008. PT: J; TC: 2; UT: WOS:000258473400007.

[92] T. Makino and A. McLysaght, *The Evolution of Functional Gene Clusters in Eukaryote Genomes*. 2009 2009. PT: B; CT: 12th Evolutionary Biology Meeting; CY: SEP 24-26, 2008; CL: Marseilles, FRANCE; TC: 0; UT: WOS:000270816600011.

[93] M. J. Lercher, A. O. Urrutia, and L. D. Hurst, "Clustering of housekeeping genes

provides a unified model of gene order in the human genome," *Nature genetics*, vol. 31, pp. 180–183, print 2002. M3: 10.1038/ng887; 10.1038/ng887.

[94] T. Yamashita, M. Honda, H. Takatori, R. Nishino, N. Hoshino, and S. Kaneko, "Genome-wide transcriptome mapping analysis identifies organ-specific gene expression patterns along human chromosomes," *Genomics*, vol. 84, 2004.

[95] N. N. Batada and L. D. Hurst, "Evolution of chromosome organization driven by selection for reduced gene expression noise," *Nature genetics*, vol. 39, pp. 945–949, AUG 2007 2007. PT: J; TC: 53; UT: WOS:000248446900006.

[96] T. Warnecke and L. D. Hurst, "Error prevention and mitigation as forces in the evolution of genes and genomes," *Nature reviews. Genetics*, vol. 12, pp. 875–881, print 2011. M3: 10.1038/nrg3092; 10.1038/nrg3092.

[97] N. N. Batada, A. O. Urrutia, and L. D. Hurst, "Chromatin remodelling is a major source of coexpression of linked genes in yeast," *Trends in Genetics*, vol. 23, pp. 480–484, 10 2007.

[98] P. M. Petkov, J. H. Graber, G. A. Churchill, K. DiPetrillo, B. L. King, and K. Paigen, "Evidence of a large-scale functional organization of mammalian chromosomes," *PLoS Genetics*, vol. 1, p. e33, 09 2005.

[99] S. R. Gilman, I. Iossifov, D. Levy, M. Ronemus, M. Wigler, and D. Vitkup, "Rare de novo variants associated with autism implicate a large functional network of genes involved in formation and function of synapses," *Neuron*, vol. 70, pp. 898–907, /6/9/ 2011.

[100] A. Itsara, H. Wu, J. D. Smith, D. A. Nickerson, I. Romieu, S. J. London, and E. E. Eichler, "De novo rates and selection of large copy number variation," *Genome research*, vol. 20, pp. 1469–1481, November 01 2010.

[101] T. Tucker, A. Montpetit, D. Chai, S. Chan, S. Chenier, B. Coe, A. Delaney, P. Eydoux, W. Lam, S. Langlois, E. Lemyre, M. Marra, H. Qian, G. Rouleau, D. Vincent, J. Michaud, and J. Friedman, "Comparison of genome-wide array genomic hybridization platforms for the detection of copy number variants in idiopathic

mental retardation," *BMC Medical Genomics*, vol. 4, no. 1, p. 25, 2011. M3: 10.1186/1755-8794-4-25.

[102] B. Rhead, D. Karolchik, R. M. Kuhn, A. S. Hinrichs, A. S. Zweig, P. A. Fujita, M. Diekhans, K. E. Smith, K. R. Rosenbloom, B. J. Raney, A. Pohl, M. Pheasant, L. R. Meyer, K. Learned, F. Hsu, J. Hillman-Jackson, R. A. Harte, B. Giardine, T. R. Dreszer, H. Clawson, G. P. Barber, D. Haussler, and W. J. Kent, "The ucsc genome browser database: update 2010," *Nucleic acids research*, vol. 38, pp. D613–D619, January 01 2010.

[103] T. H. Shaikh, X. Gai, J. C. Perin, J. T. Glessner, H. Xie, K. Murphy, R. O'Hara, T. Casalunovo, L. K. Conlin, M. D'Arcy, E. C. Frackelton, E. A. Geiger, C. Haldeman-Englert, M. Imielinski, C. E. Kim, L. Medne, K. Annaiah, J. P. Bradfield, E. Dabaghyan, A. Eckert, C. C. Onyiah, S. Ostapenko, F. G. Otieno, E. Santa, J. L. Shaner, R. Skraban, R. M. Smith, J. Elia, E. Goldmuntz, N. B. Spinner, E. H. Zackai, R. M. Chiavacci, R. Grundmeier, E. F. Rappaport, S. F. Grant, P. S. White, and H. Hakonarson, "High-resolution mapping and analysis of copy number variations in the human genome: a data resource for clinical and research applications," *Genome Res*, vol. 19, no. 9, pp. 1682–90, 2009.

[104] D. F. Conrad, D. Pinto, R. Redon, L. Feuk, O. Gokcumen, Y. Zhang, J. Aerts, T. D. Andrews, C. Barnes, P. Campbell, T. Fitzgerald, M. Hu, C. H. Ihm, K. Kristiansson, D. G. MacArthur, J. R. MacDonald, I. Onyiah, A. W. C. Pang, S. Robson, K. Stirrups, A. Valsesia, K. Walter, J. Wei, C. Tyler-Smith, N. P. Carter, C. Lee, S. W. Scherer, and M. E. Hurles, "Origins and functional impact of copy number variation in the human genome," *Nature*, vol. 464, pp. 704–712, 04/01 2010. M3: 10.1038/nature08516; 10.1038/nature08516.

[105] P. Flicek, B. L. Aken, B. Ballester, K. Beal, E. Bragin, S. Brent, Y. Chen, P. Clapham, G. Coates, S. Fairley, S. Fitzgerald, J. Fernandez-Banet, L. Gordon, S. Gräf, S. Haider, M. Hammond, K. Howe, A. Jenkinson, N. Johnson, A. Kähäri, D. Keefe, S. Keenan, R. Kinsella, F. Kokocinski, G. Koscielny, E. Kulesha, D. Lawson, I. Longden, T. Massingham, W. McLaren, K. Megy, B. Overduin,

B. Pritchard, D. Rios, M. Ruffier, M. Schuster, G. Slater, D. Smedley, G. Spudich, Y. A. Tang, S. Trevanion, A. Vilella, J. Vogel, S. White, S. P. Wilder, A. Zadissa, E. Birney, F. Cunningham, I. Dunham, R. Durbin, X. M. FernÃ¡ndez-Suarez, J. Herrero, T. J. P. Hubbard, A. Parker, G. Proctor, J. Smith, and S. M. J. Searle, "Ensembl's 10th year," *Nucleic acids research*, vol. 38, pp. D557–D562, January 01 2010.

[106] C. Webber, "Functional enrichment analysis with structural variants: Pitfalls and strategies," *Cytogenetic and Genome Research*, vol. 135, no. 3-4, pp. 277–285, 2011.

[107] C. L. Smith and J. T. Eppig, "The mammalian phenotype ontology: enabling robust annotation and comparative analysis," *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, vol. 1, no. 3, pp. 390–399, 2009.

[108] P. Resnik, "Using information content to evalutate semantic similarity in a taxonomy," *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, 1995.

[109] C. Pesquita, D. Faria, A. O. F. ao, P. Lord, and F. M. Couto, "Semantic similarity in biomedical ontologies," *PloS Computational Biology*, vol. 5, p. e1000443, 07/31 2009. M3: doi:10.1371/journal.pcbi.1000443.

[110] S. Boriah and V. Chandola, Varun ahd Kuman, "Similarity measures for categorical data: A commparative evaluation," *Proceedings of the eighth SIAM International Conference on Data Mining*, vol. 30, pp. 234–254, 2008.

[111] A. J. Ridley, "Rho {GTPases} and actin dynamics in membrane protrusions and vesicle trafficking," *Trends in Cell Biology*, vol. 16, no. 10, pp. 522 – 529, 2006. Membrane Dynamics.

[112] O. Online Mendelian Inheritance in Man, "Mckusick-nathans institute of genetic medicine, johns hopkins university (baltimore, md)," 2012.

[113] I. Lee, U. M. Blom, P. I. Wang, J. E. Shim, and E. M. Marcotte, "Prioritizing candidate disease genes by network-based boosting of genome-wide association data," *Genome research*, vol. 21, pp. 1109–1121, July 01 2011.

[114] B. Kampmann, C. Hemingway, A. Stephens, R. Davidson, A. Goodsall, S. Anderson, M. Nicol, E. SchÃ¶lvinck, D. Relman, S. Waddell, P. Langford, B. Sheehan, L. Semple, K. A. Wilkinson, R. J. Wilkinson, S. Ress, M. Hibberd, and M. Levin, "Acquired predisposition to mycobacterial disease due to autoantibodies to ifn-

γ

," *The Journal of Clinical Investigation*, vol. 115, pp. 2480–2488, 9 2005.

[115] R. Shyamsundar, Y. Kim, J. Higgins, K. Montgomery, M. Jorden, A. Sethuraman, M. van de Rijn, D. Botstein, P. Brown, and J. Pollack, "A dna microarray survey of gene expression in normal human tissues," *Genome Biology*, vol. 6, no. 3, p. R22, 2005.

[116] T. O. Nielsen, R. B. West, S. C. Linn, O. Alter, M. A. Knowling, J. O. X., S. Zhu, M. Fero, G. Sherlock, J. R. Pollack, P. O. Brown, D. Botstein, and de Rijn van 2002.

[117] A. A. Alizadeh, M. B. Eisen, R. E. Davis, C. Ma, I. S. Lossos, A. Rosenwald, J. C. Boldrick, H. Sabet, T. Tran, X. Yu, J. I. Powell, L. Yang, G. E. Marti, T. Moore, J. Hudson, L. Lu, D. B. Lewis, R. Tibshirani, G. Sherlock, W. C. Chan, T. C. Greiner, D. D. Weisenburger, J. O. Armitage, R. Warnke, R. Levy, W. Wilson, M. R. Grever, J. C. Byrd, D. Botstein, P. O. Brown, and L. M. Staudt, "Distinct types of diffuse large b-cell lymphoma identified by gene expression profiling," *Nature*, vol. 403, pp. 503–511, 02/03 2000. 10.1038/35000501.

[118] M. Schaner, B. Davidson, M. Skrede, R. Reich, V. Florenes, B. Risberg, A. Berner, I. Goldberg, V. Givant-Horwitz, C. Trope, G. Kristensen, J. Nesland, and A.-L. Borresen-Dale, "Variation in gene expression patterns in effusions and primary tumors from serous ovarian cancer patients," *Molecular Cancer*, vol. 4, no. 1, p. 26, 2005.

[119] R. R. Nayak, M. Kearns, R. S. Spielman, and V. G. Cheung, "Coexpression network based on natural variation in human gene expression reveals gene interactions and functions," *Genome Research*, vol. 19, no. 11, pp. 1953–1962, 2009.

[120] I. K. Jordan, L. Mariño-RamÃrez, Y. I. Wolf, and E. V. Koonin, "Conservation and coevolution in the scale-free human gene coexpression network," *Molecular Biology and Evolution*, vol. 21, no. 11, pp. 2058–2070, 2004.

[121] V. van Noort, B. Snel, and M. A. Huynen, "The yeast coexpression network has a small-world, scale-free architecture and can be explained by a simple model," *EMBO reports*, vol. 5, no. 3, pp. 280–284, 2004.

[122] A. Reverter, N. J. Hudson, Y. Wang, S.-H. Tan, W. Barris, K. A. Byrne, S. M. McWilliam, C. D. K. Bottema, A. Kister, P. L. Greenwood, G. S. Harper, S. A. Lehnert, and B. P. Dalrymple, "A gene coexpression network for bovine skeletal muscle inferred from microarray data," *Physiological Genomics*, vol. 28, no. 1, pp. 76–83, 2006.

[123] J. A. Miller, S.-L. Ding, S. M. Sunkin, K. A. Smith, L. Ng, A. Szafer, A. Ebbert, Z. L. Riley, J. J. Royall, K. Aiona, J. M. Arnold, C. Bennet, D. Bertagnolli, K. Brouner, S. Butler, S. Caldejon, A. Carey, C. Cuhaciyan, R. A. Dalley, N. Dee, T. A. Dolbeare, B. A. C. Facer, D. Feng, T. P. Fliss, G. Gee, J. Goldy, L. Gourley, B. W. Gregor, G. Gu, R. E. Howard, J. M. Jochim, C. L. Kuan, C. Lau, C.-K. Lee, F. Lee, T. A. Lemon, P. Lesnar, B. McMurray, N. Mastan, N. Mosqueda, T. Naluai-Cecchini, N.-K. Ngo, J. Nyhus, A. Oldre, E. Olson, J. Parente, P. D. Parker, S. E. Parry, A. Stevens, M. Pletikos, M. Reding, K. Roll, D. Sandman, M. Sarreal, S. Shapouri, N. V. Shapovalova, E. H. Shen, N. Sjoquist, C. R. Slaughterbeck, M. Smith, A. J. Sodt, D. Williams, L. Zollei, B. Fischl, M. B. Gerstein, D. H. Geschwind, I. A. Glass, M. J. Hawrylycz, R. F. Hevner, H. Huang, A. R. Jones, J. A. Knowles, P. Levitt, J. W. Phillips, N. Sestan, P. Wohnoutka, C. Dang, A. Bernard, J. G. Hohmann, and E. S. Lein, "Transcriptional landscape of the prenatal human brain," *Nature*, vol. 508, pp. 199–206, 04/10 2014.

[124] G. Consortium, "The genotype-tissue expression (gtex) project," *Nat Genet*, vol. 45, no. 6, pp. 580–5, 2013.

[125] E. J. Rossin, K. Lage, S. Raychaudhuri, R. J. Xavier, D. Tatar, Y. Benita, C. Cotsapas, and M. J. Daly, "Proteins encoded in genomic regions associated with

immune-mediated disease physically interact and suggest underlying biology," *PLoS Genet*, vol. 7, no. 1, p. e1001273, 2011.

[126] D. Maglott, J. Ostell, K. D. Pruitt, and T. Tatusova, "Entrez gene: gene-centered information at ncbi," *Nucleic acids research*, vol. 35, pp. D26–D31, January 01 2007.

[127] C. M. Farrell, N. A. O'Leary, R. A. Harte, J. E. Loveland, L. G. Wilming, C. Wallin, M. Diekhans, D. Barrell, S. M. J. Searle, B. Aken, S. M. Hiatt, A. Frankish, M.-M. Suner, B. Rajput, C. A. Steward, G. R. Brown, R. Bennett, M. Murphy, W. Wu, M. P. Kay, J. Hart, J. Rajan, J. Weber, C. Snow, L. D. Riddick, T. Hunt, D. Webb, M. Thomas, P. Tamez, S. H. Rangwala, K. M. McGarvey, S. Pujar, A. Shkeda, J. M. Mudge, J. M. Gonzalez, J. G. R. Gilbert, S. J. Trevanion, R. Baertsch, J. L. Harrow, T. Hubbard, J. M. Ostell, D. Haussler, and K. D. Pruitt, "Current status and new features of the consensus coding sequence database," *Nucleic acids research*, vol. 42, pp. D865–D872, January 01 2014.

[128] M. N. McCall, B. M. Bolstad, and R. A. Irizarry, "Frozen robust multiarray analysis (frma)," *Biostatistics*, vol. 11, pp. 242–253, April 01 2010.

[129] M. Kapushesky, T. Adamusiak, T. Burdett, A. Culhane, A. Farne, A. Filippov, E. Holloway, A. Klebanov, N. Kryvych, N. Kurbatova, P. Kurnosov, J. Malone, O. Melnichuk, R. Petryszak, N. Pultsin, G. Rustici, A. Tikhonov, R. S. Travillian, E. Williams, A. Zorin, H. Parkinson, and A. Brazma, "Gene expression atlas update: a value-added database of microarray and sequencing-based functional genomics experiments," January 01 2012.

[130] R. Petryszak, T. Burdett, B. Fiorelli, N. A. Fonseca, M. Gonzalez-Porta, E. Hastings, W. Huber, S. Jupp, M. Keays, N. Kryvych, J. McMurry, J. C. Marioni, J. Malone, K. Megy, G. Rustici, A. Y. Tang, J. Taubert, E. Williams, O. Mannion, H. E. Parkinson, and A. Brazma, "Expression atlas update: a database of gene and transcript expression from microarray- and sequencing-based functional genomics experiments," *Nucleic acids research*, vol. 42, pp. D926–D932, January 01 2014.

[131] J. Steinberg and C. Webber, "The roles of fmrp-regulated genes in autism spec-

trum disorder: single- and multiple-hit genetic etiologies," *Am J Hum Genet*, vol. 93, no. 5, pp. 825–39, 2013.

[132] A. Lancichinetti and S. Fortunato, "Community detection algorithms: A comparative analysis," *Physical Review E*, vol. 80, p. 056117, 2009.

[133] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2008, no. 10, p. P10008, 2008.

[134] S. Fortunato and M. BarthÃ©lemy, "Resolution limit in community detection," *Proceedings of the National Academy of Sciences*, vol. 104, pp. 36–41, January 02 2007.

[135] A. Strehl and J. Ghosh, "Cluster ensembles — a knowledge reuse framework for combining multiple partitions," *J.Mach.Learn.Res.*, vol. 3, pp. 583–617, mar 2003.

[136] J. B. Tomblin and P. R. Buckwalter, "Heritability of poor language achievement among twins," *Journal of Speech, Language, and Hearing Research*, vol. 41, no. 1, pp. 188–199, 1998.

[137] D. J. Fidler, J. N. Bailey, and S. L. Smalley, "Macrocephaly in autism and other pervasive developmental disorders," *Developmental Medicine & Child Neurology*, vol. null, pp. 737–740, 11 2000.

[138] A. Bailey, A. Le Couteur, I. Gottesman, P. Bolton, E. Simonoff, E. Yuzda, and M. Rutter, "Autism as a strongly genetic disorder: evidence from a british twin study," *Psychological Medicine*, vol. 25, pp. 63–77, 1 1995.

[139] C. Lee, A. J. Iafrate, and A. R. Brothman, "Copy number variations and clinical cytogenetic diagnosis of constitutional disorders," *Nature genetics*, 06/27 2007.

[140] I. Lee, Z. Li, and E. M. Marcotte, "An improved, bias-reduced probabilistic functional gene network of baker's yeast, *Saccharomyces cerevisiae*," *PLoS ONE*, vol. 2, p. e988, 10 2007. biases in GO, KEGG and other functional information.

[141] Y. Benjamini and Y. Hochbert, "Controlling the false discovery rate: a practical

and powerful approach to multiple testing," *J. Roy. Statist. Soc. Ser. B*, vol. 57, no. 1, pp. 289–300, 1995.

[142] C. E. Shannon, "A mathematical theory of communication," *The Bell System Technical Journal*, vol. 27, pp. 379–423,623–656, 1948.

[143] T. H. Shaikh, C. Haldeman-Englert, E. A. Geiger, C. P. Ponting, and C. Webber, "Genes and biological processes commonly disrupted in rare and heterogeneous developmental delay syndromes," *Hum Mol Genet*, vol. 20, no. 5, pp. 880–93, 2011.

[144] V. des Portes, N. Boddaert, S. Sacco, S. Briault, K. Maincent, N. Bahi, M. Gomot, N. Ronce, J. Bursztyn, C. Adamsbaum, M. Zilbovicius, J. Chelly, and C. Moraine, "Specific clinical and brain mri features in mentally retarded patients with mutations in the oligophrenin-1 gene," *Am J Med Genet A*, vol. 124A, no. 4, pp. 364–71, 2004.

[145] S. R. Lalani, A. M. Safiullah, S. D. Fernbach, K. G. Harutyunyan, C. Thaller, L. E. Peterson, J. D. McPherson, R. A. Gibbs, L. D. White, M. Hefner, S. L. Davenport, J. M. Graham, C. A. Bacino, N. L. Glass, J. A. Towbin, W. J. Craigen, S. R. Neish, A. E. Lin, and J. W. Belmont, "Spectrum of chd7 mutations in 110 individuals with charge syndrome and genotype-phenotype correlation," *Am J Hum Genet*, vol. 78, no. 2, pp. 303–14, 2006.

[146] S. B. Ng, A. W. Bigham, K. J. Buckingham, M. C. Hannibal, M. J. McMillin, H. I. Gildersleeve, A. E. Beck, H. K. Tabor, G. M. Cooper, H. C. Mefford, C. Lee, E. H. Turner, J. D. Smith, M. J. Rieder, K. Yoshiura, N. Matsumoto, T. Ohta, N. Niikawa, D. A. Nickerson, M. J. Bamshad, and J. Shendure, "Exome sequencing identifies mll2 mutations as a cause of kabuki syndrome," *Nat Genet*, vol. 42, no. 9, pp. 790–3, 2010.

[147] D. Germanaud, M. Rossi, G. Bussy, D. Gerard, L. Hertz-Pannier, P. Blanchet, H. Dollfus, F. Giuliano, V. Bennouna-Greene, P. Sarda, S. Sigaudy, A. Curie, M. C. Vincent, R. Touraine, and V. des Portes, "The renpenning syndrome spectrum:

new clinical insights supported by 13 new pqbp1-mutated males," *Clin Genet*, vol. 79, no. 3, pp. 225–35, 2011.

[148] I. Perrault, F. F. Hamdan, M. Rio, J. M. Capo-Chichi, N. Boddaert, J. C. Decarie, B. Maranda, R. Nabbout, M. Sylvain, A. Lortie, P. P. Roux, E. Rossignol, X. Gerard, G. Barcia, P. Berquin, A. Munnich, G. A. Rouleau, J. Kaplan, J. M. Rozet, and J. L. Michaud, "Mutations in dock7 in individuals with epileptic encephalopathy and cortical blindness," *Am J Hum Genet*, vol. 94, no. 6, pp. 891–7, 2014.

[149] Q. Ferry, J. Steinberg, C. Webber, D. R. FitzPatrick, C. P. Ponting, A. Zisserman, and C. Nellåker, "Diagnostically relevant facial gestalt information from ordinary photos," *eLife*, vol. 3, 2014.

[150] The Autism Genome Project Consortium, "Mapping autism risk loci using genetic linkage and chromosomal rearrangements," *Nature genetics*, vol. 39, pp. 319–328, print 2007.

[151] K. Lage, E. O. Karlberg, Z. M. Storling, P. I. Olason, A. G. Pedersen, O. Rigina, A. M. Hinsby, Z. Tumer, F. Pociot, N. Tommerup, Y. Moreau, and S. Brunak, "A human phenome-interactome network of protein complexes implicated in genetic disorders," *Nat Biotech*, vol. 25, pp. 309–316, print 2007.

[152] M. Sémon and L. Duret, "Evolutionary origin and maintenance of coexpressed gene clusters in mammals," *Molecular biology and evolution*, vol. 23, pp. 1715–1723, SEP 2006 2006. PT: J; TC: 48; UT: WOS:000239905100010.

[153] F. Zhang, W. Gu, M. E. Hurles, and J. R. Lupski, "Copy number variation in human health, disease, and evolution," *Annual Review of Genomics and Human Genetics*, vol. 10, pp. 451–481, 09/01; 2013/09 2009. doi: 10.1146/annurev.genom.9.081307.164217; M3: doi: 10.1146/annurev.genom.9.081307.164217; 21.

[154] J. Y. Hehir-Kwa, M. Egmont-Petersen, I. M. Janssen, D. Smeets, A. G. van Kessel, and J. A. Veltman, "Genome-wide copy number profiling on high-density bacterial artificial chromosomes, single-nucleotide polymorphisms, and oligonu-

cleotide microarrays: A platform comparison based on statistical power analysis," *DNA Research*, vol. 14, pp. 1–11, January 01 2007.

[155] J. R. Vermeesch, I. Balikova, C. Schrander-Stumpel, J.-P. Fryns, and K. Devriendt, "The causality of de novo copy number variants is overestimated," *European journal of human genetics : EJHG*, vol. 19, pp. 1112–1113, print 2011.

[156] A. J. Vilella, J. Severin, A. Ureta-Vidal, L. Heng, R. Durbin, and E. Birney, "Ensemblcompara genetrees: Complete, duplication-aware phylogenetic trees in vertebrates," *Genome research*, vol. 19, pp. 327–335, February 01 2009.

[157] A. Heger and C. P. Ponting, "Optic: orthologous and paralogous transcripts in clades," *Nucleic acids research*, vol. 36, pp. D267–D270, January 01 2008.

[158] V. T. Dang, K. S. Kassahn, A. E. Marcos, and M. A. Ragan, "Identification of human haploinsufficient genes and their genomic proximity to segmental duplications," *European journal of human genetics : EJHG*, vol. 16, pp. 1350–1357, 06/04 2008.

[159] R Development Core Team, *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2011. ISBN 3-900051-07-0.

[160] B. Linghu, E. Snitkin, Z. Hu, Y. Xia, and C. DeLisi, "Genome-wide prioritization of disease genes and identification of disease-disease associations from an integrated human functional linkage network," *Genome biology*, vol. 10, no. 9, p. R91, 2009. M3: 10.1186/gb-2009-10-9-r91.

[161] J. A. Buchanan and S. W. Scherer, "Contemplating effects of genomic structural variation," *Genetics in medicine : official journal of the American College of Medical Genetics*, vol. 10, pp. 639–647, print 2008.

[162] B. P. Coe, K. Witherspoon, J. A. Rosenfeld, B. W. M. van Bon, A. T. Vulto-van Silfhout, P. Bosco, K. L. Friend, C. Baker, S. Buono, L. E. L. M. Vissers, J. H. Schuurs-Hoeijmakers, A. Hoischen, R. Pfundt, N. Krumm, G. L. Carvill, D. Li, D. Amaral, N. Brown, P. J. Lockhart, I. E. Scheffer, A. Alberti, M. Shaw, R. Pettinato,

R. Tervo, N. de Leeuw, M. R. F. Reijnders, B. S. Torchia, H. Peeters, E. Thompson, B. J. O'Roak, M. Fichera, J. Y. Hehir-Kwa, J. Shendure, H. C. Mefford, E. Haan, J. Gecz, B. B. A. de Vries, C. Romano, and E. E. Eichler, "Refining analyses of copy number variation identifies specific genes associated with developmental delay," *Nature genetics*, vol. 46, pp. 1063–1071, 10 2014.

[163] E. Zotenko, J. Mestre, D. P. O'Leary, and T. M. Przytycka, "Why do hubs in the yeast protein interaction network tend to be essential: Reexamining the connection between the network topology and essentiality," *PLoS Comput Biol*, vol. 4, p. e1000140, 08/01 2008. M3: doi:10.1371/journal.pcbi.1000140.

[164] L. D. Hurst, C. Pal, and M. J. Lercher, "The evolutionary dynamics of eukaryotic gene order," *Nature Reviews Genetics*, vol. 5, no. 4, pp. 299–310, 2004.

[165] R. Horton, L. Wilming, V. Rand, R. C. Lovering, E. A. Bruford, V. K. Khodiyar, M. J. Lush, S. Povey, C. C. Talbot, M. W. Wright, H. M. Wain, J. Trowsdale, A. Ziegler, and S. Beck, "Gene map of the extended human mhc," *Nature reviews. Genetics*, vol. 5, pp. 889–899, print 2004. M3: 10.1038/nrg1489; 10.1038/nrg1489.

[166] L. Guelen, L. Pagie, E. Brasset, W. Meuleman, M. B. Faza, W. Talhout, B. H. Eussen, A. de Klein, L. Wessels, W. de Laat, and B. van Steensel, "Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions," *Nature*, vol. 453, pp. 948–951, 06/12 2008. M3: 10.1038/nature06947; 10.1038/nature06947.

[167] J. A. Bailey, A. M. Yavor, H. F. Massa, B. J. Trask, and E. E. Eichler, "Segmental duplications: Organization and impact within the current human genome project assembly," *Genome Research*, vol. 11, no. 6, pp. 1005–1017, 2001.

[168] J. A. Bailey, Z. Gu, R. A. Clark, K. Reinert, R. V. Samonte, S. Schwartz, M. D. Adams, E. W. Myers, P. W. Li, and E. E. Eichler, "Recent segmental duplications in the human genome," *Science*, vol. 297, no. 5583, pp. 1003–1007, 2002.

[169] A. Kong, G. Thorleifsson, D. F. Gudbjartsson, G. Masson, A. Sigurdsson, A. Jonasdottir, G. B. Walters, A. Jonasdottir, A. Gylfason, K. T. Kristinsson, S. A.

Gudjonsson, M. L. Frigge, A. Helgason, U. Thorsteinsdottir, and K. Stefansson, "Fine-scale recombination rate differences between sexes, populations and individuals," *Nature*, vol. 467, pp. 1099–1103, 10/28 2010.

[170] E. Lieberman-Aiden, N. L. van Berkum, L. Williams, M. Imakaev, T. Ragoczy, A. Telling, I. Amit, B. R. Lajoie, P. J. Sabo, M. O. Dorschner, R. Sandstrom, B. Bernstein, M. A. Bender, M. Groudine, A. Gnirke, J. Stamatoyannopoulos, L. A. Mirny, E. S. Lander, and J. Dekker, "Comprehensive mapping of long-range interactions reveals folding principles of the human genome," *Science*, vol. 326, pp. 289–293, October 09 2009.

[171] J. H. Ward, "Hierarchical grouping to optimize an objective function," *Journal of the American Statistical Association*, vol. 58, no. 301, pp. 236–244, 1963.

[172] S. V. Dongen, "Graph clustering via a discrete uncoupling process," *SIAM Journal on Matrix Analysis and Applications*, vol. 30, no. 1, p. 121, 2008. TY: JOURNAL ARTICLE; M3: 10.1137/040608635.

[173] W.-M. Boon, T. Beissbarth, L. Hyde, G. Smyth, J. Gunnersen, D. A. Denton, H. Scott, and S.-S. Tan, "A comparative analysis of transcribed genes in the mouse hypothalamus and neocortex reveals chromosomal clustering," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 41, pp. 14972–14977, 2004.

[174] C. S. Osborne, L. Chakalova, K. E. Brown, D. Carter, A. Horton, E. Debrand, B. Goyenechea, J. A. Mitchell, S. Lopes, W. Reik, and P. Fraser, "Active genes dynamically colocalize to shared sites of ongoing transcription," *Nature genetics*, vol. 36, pp. 1065–1071, print 2004. M3: 10.1038/ng1423; 10.1038/ng1423.

[175] L. Mirny, "The fractal globule as a model of chromatin architecture in the cell," *Chromosome Research*, vol. 19, no. 1, pp. 37–51, 2011.

[176] J. M. Zullo, I. A. Demarco, R. PiquÃ©-Regi, D. J. Gaffney, C. B. Epstein, C. J. Spooner, T. R. Luperchio, B. E. Bernstein, J. K. Pritchard, K. L. Reddy, and

H. Singh, "Dna sequence-dependent compartmentalization and silencing of chromatin at the nuclear lamina," *Cell*, vol. 149, pp. 1474–1487, 6/22 2012.

[177] P. Meister, B. D. Towbin, B. L. Pike, A. Ponti, and S. M. Gasser, "The spatial dynamics of tissue-specific promoters during c. elegans development," *Genes and Development*, vol. 24, no. 8, pp. 766–782, 2010. Cited By (since 1996):56.

[178] J. R. Dixon, S. Selvaraj, F. Yue, A. Kim, Y. Li, Y. Shen, M. Hu, J. S. Liu, and B. Ren, "Topological domains in mammalian genomes identified by analysis of chromatin interactions," *Nature*, vol. 485, pp. 376–380, 05/17 2012. 10.1038/nature11082.

[179] S. S. P. Rao, M. H. Huntley, N. C. Durand, E. K. Stamenova, I. D. Bochkov, J. T. Robinson, A. L. Sanborn, I. Machol, A. D. Omer, E. S. Lander, and E. L. Aiden, "A 3d map of the human genome at kilobase resolution reveals principles of chromatin looping," *Cell*, 2014/12. doi: 10.1016/j.cell.2014.11.021; 14.

[180] P. C. Phillips, "Epistasis - the essential role of gene interactions in the structure and evolution of genetic systems," *Nature reviews.Genetics*, vol. 9, pp. 855–867, 11 2008. J1: Nat Rev Genet.

[181] S. M. Fullerton, A. Bernardo Carvalho, and A. G. Clark, "Local rates of recombination are positively correlated with gc content in the human genome," *Molecular Biology and Evolution*, vol. 18, no. 6, pp. 1139–1142, 2001.

[182] A. Heger, C. Webber, M. Goodson, C. P. Ponting, and G. Lunter, "Gat: a simulation framework for testing the association of genomic intervals," *Bioinformatics*, vol. 29, no. 16, pp. 2046–2048, 2013.

[183] S. De, Y. Zhang, J. R. Garner, S. A. Wang, and K. G. Becker, "Disease and phenotype gene set analysis of disease-based gene expression in mouse and human," *Physiological Genomics*, vol. 42A, pp. 162–167, October 01 2010.

[184] J. Chen, H. Xu, B. Aronow, and A. Jegga, "Improved human disease candidate gene prioritization using mouse phenotype," *BMC Bioinformatics*, vol. 8, no. 1, p. 392, 2007. M3: 10.1186/1471-2105-8-392.

[185] J. P. A. Ioannidis, "Why most published research findings are false," *PLoS Med*, vol. 2, p. e124, 08/30 2005. M3: doi:10.1371/journal.pmed.0020124.

[186] S. D. M. Brown and M. W. Moore, "Towards an encyclopaedia of mammalian gene function: the international mouse phenotyping consortium," *Disease Models & Mechanisms*, vol. 5, pp. 289–292, May 01 2012.

[187] H. Morgan, T. Beck, A. Blake, H. Gates, N. Adams, G. Debouzy, S. Leblanc, C. Lengger, H. Maier, D. Melvin, H. Meziane, D. Richardson, S. Wells, J. White, J. Wood, T. E. Consortium, M. H. de Angelis, S. D. M. Brown, J. M. Hancock, and A.-M. Mallon, "Europhenome: a repository for high-throughput mouse phenotyping data," *Nucleic acids research*, vol. 38, pp. D577–D585, January 01 2010.

[188] T. C. K. M. P. Consortium, "The knockout mouse project," *Nature Genetics*, vol. 36, pp. 921–924, September 2004.

[189] Z.-P. Liu and L. Chen, "Proteome-wide prediction of protein-protein interactions from high-throughput data," *Protein & Cell*, vol. 3, pp. 508–520, 07/01 2012. J2: Protein Cell.

[190] R. Jansen, H. Yu, D. Greenbaum, Y. Kluger, N. J. Krogan, S. Chung, A. Emili, M. Snyder, J. F. Greenblatt, and M. Gerstein, "A bayesian networks approach for predicting protein-protein interactions from genomic data," *Science*, vol. 302, pp. 449–453, October 17 2003.

[191] A. E. Todd, R. L. Marsden, J. M. Thornton, and C. A. Orengo, "Progress of structural genomics initiatives: An analysis of solved target structures," *Journal of Molecular Biology*, vol. 348, pp. 1235–1260, 5/20 2005.

[192] D. Wu, P. Hugenholtz, K. Mavromatis, R. Pukall, E. Dalin, N. N. Ivanova, V. Kunin, L. Goodwin, M. Wu, B. J. Tindall, S. D. Hooper, A. Pati, A. Lykidis, S. Spring, I. J. Anderson, P. D'haeseleer, A. Zemla, M. Singer, A. Lapidus, M. Nolan, A. Copeland, C. Han, F. Chen, J.-F. Cheng, S. Lucas, C. Kerfeld, E. Lang, S. Gronow, P. Chain, D. Bruce, E. M. Rubin, N. C. Kyrpides, H.-P. Klenk, and J. A. Eisen, "A phylogeny-driven genomic encyclopaedia of bacteria and ar-

chaea," *Nature*, vol. 462, pp. 1056–1060, DEC 24 2009. PT: J; NR: 30; TC: 212; J9: NATURE; PG: 5; GA: 535UB; UT: WOS:000272996000047.

[193] V. Gailus-Durner, H. Fuchs, L. Becker, I. Bolle, M. Brielmeier, J. Calzada-Wack, R. Elvert, N. Ehrhardt, C. Dalke, T. J. Franz, E. Grundner-Culemann, S. Hammelbacher, S. M. Holter, G. Holzlwimmer, M. Horsch, A. Javaheri, S. vetoslav Kalaydjiev, M. Klempt, E. Kling, S. Kunder, C. Lengger, T. Lisse, T. Mijalski, B. Naton, V. Pedersen, C. Prehn, G. Przemeck, I. Racz, C. Reinhard, P. Reitmeir, I. Schneider, A. Schrewe, R. Steinkamp, C. Zybill, J. Adamski, J. Beckers, H. Behrendt, J. Favor, J. Graw, G. Heldmaier, H. Hofler, B. Ivandic, H. Katus, P. Kirchhof, M. Klingenspor, T. Klopstock, A. Lengeling, W. Muller, F. Ohl, M. Ollert, L. Quintanilla-Martinez, J. Schmidt, H. Schulz, E. Wolf, W. Wurst, A. Zimmer, D. H. Busch, and M. H. de Angelis, "Introducing the german mouse clinic: open access platform for standardized phenotyping," *Nat Meth*, vol. 2, pp. 403–404, print 2005. 10.1038/nmeth0605-403.

[194] A. Ayadi, M.-C. Birling, J. Bottomley, J. Bussell, H. Fuchs, M. Fray, V. Gailus-Durner, S. Greenaway, R. Houghton, N. Karp, S. Leblanc, C. Lengger, H. Maier, A.-M. Mallon, S. Marschall, D. Melvin, H. Morgan, G. Pavlovic, E. Ryder, W. C. Skarnes, M. Selloum, R. Ramirez-Solis, T. Sorg, L. Teboul, L. Vasseur, A. Walling, T. Weaver, S. Wells, J. K. White, A. Bradley, D. J. Adams, K. P. Steel, de Angelis Hrabě, S. Brown, and Y. Herault, "Mouse large-scale phenotyping initiatives: overview of the european mouse disease clinic (eumodic) and of the wellcome trust sanger institute mouse genetics project," *Mammalian Genome*, vol. 23, pp. 600–610, 10/01 2012. J2: Mamm Genome.

[195] M. Meilă, *Comparing Clusterings by the Variation of Information*, vol. 2777. Springer Berlin Heidelberg, 01/01 2003.

[196] C. J. van Rijsbergen, *Information Retrieval*. London: Butterworths, 2 ed., 1979.

[197] R. Sharan, I. Ulitsky, and R. Shamir, "Network-based prediction of protein function," *Molecular Systems Biology*, vol. 3, no. 1, pp. n/a–n/a, 2007.

[198] V. Grishkevich and I. Yanai, "Gene length and expression level shape genomic

novelties," *Genome research*, vol. 24, pp. 1497–1503, September 01 2014.

[199] S. P. Dickson, K. Wang, I. Krantz, H. Hakonarson, and D. B. Goldstein, "Rare variants create synthetic genome-wide associations," *PLoS Biol*, vol. 8, p. e1000294, 01/26 2010.

[200] K. Wang, S. P. Dickson, C. A. Stolle, I. D. Krantz, D. B. Goldstein, and H. Hakonarson, "Interpretation of association signals and identification of causal variants from genome-wide association studies," *The American Journal of Human Genetics*, vol. 86, pp. 730–742, 5/14 2010.

[201] M. Boehnke, "Estimating the power of a proposed linkage study: a practical computer simulation approach.," *American Journal of Human Genetics*, vol. 39, no. 4, pp. 513–527, 1986.

[202] E. R. Martin, S. A. Monks, L. L. Warren, and N. L. Kaplan, "A test for linkage and association in general pedigrees: The pedigree disequilibrium test," *The American Journal of Human Genetics*, vol. 67, pp. 146–154, 2014/09 2000. doi: 10.1086/302957; 24.

[203] G. Brito and D. Andrews, "Removing bias against membrane proteins in interaction networks," *BMC Systems Biology*, vol. 5, no. 1, p. 169, 2011.

[204] I. Lee, Z. Li, and E. M. Marcotte, "An improved, bias-reduced probabilistic functional gene network of baker's yeast, saccharomyces cerevisiae," *PLoS ONE*, vol. 2, p. e988, 10/03 2007.

[205] M. Ferguson, "The structure, biosynthesis and functions of glycosylphosphatidylinositol anchors, and the contributions of trypanosome research," *Journal of Cell Science*, vol. 112, no. 17, pp. 2799–2809, 1999.

[206] M. J. Bamshad, S. B. Ng, A. W. Bigham, H. K. Tabor, M. J. Emond, D. A. Nickerson, and J. Shendure, "Exome sequencing as a tool for mendelian disease gene discovery," *Nature Reviews Genetics*, vol. 12, no. 11, pp. 745–755, 2011.

[207] D. L. Rimoin and K. Hirschhorn, "A history of medical genetics in pediatrics," *Pediatr Res*, vol. 56, pp. 150–159, 07 2004.

[208] H. C. Mefford, M. L. Batshaw, and E. P. Hoffman, "Genomics, intellectual disability, and autism," *New England Journal of Medicine*, vol. 366, no. 8, pp. 733–743, 2012.

[209] J. C. Phillips, E. A. del Bono, A. M. Pralea, J. S. Cohen, L. J. Greff, and J. L. Wiggs, "A second locus for rieger syndrome maps to chromosome 13q14.," *American Journal of Human Genetics*, vol. 59, no. 3, pp. 613–619, 1996.

[210] D. B. Simon, F. E. Karet, J. Rodriguez-Soriano, J. H. Hamdan, A. DiPietro, H. Trachtman, S. A. Sanjad, and R. P. Lifton, "Genetic heterogeneity of barter's syndrome revealed by mutations in the k+ channel, romk," *Nature genetics*, vol. 14, pp. 152–156, print 1996. M3: 10.1038/ng1096-152; 10.1038/ng1096-152.

[211] S. Girirajan, J. A. Rosenfeld, G. M. Cooper, F. Antonacci, P. Siswara, A. Itsara, L. Vives, T. Walsh, S. E. McCarthy, C. Baker, H. C. Mefford, J. M. Kidd, S. R. Browning, B. L. Browning, D. E. Dickel, D. L. Levy, B. C. Ballif, K. Platky, D. M. Farber, G. C. Gowans, J. J. Wetherbee, A. Asamoah, D. D. Weaver, P. R. Mark, J. Dickerson, B. P. Garg, S. A. Ellingwood, R. Smith, V. C. Banks, W. Smith, M. T. McDonald, J. J. Hoo, B. N. French, C. Hudson, J. P. Johnson, J. R. Ozmore, J. B. Moeschler, U. Surti, L. F. Escobar, D. El-Khechen, J. L. Gorski, J. Kussmann, B. Salbert, Y. Lacassie, A. Biser, D. McDonald-McGinn, E. H. Zackai, M. A. Deardorff, T. H. Shaikh, E. Haan, K. L. Friend, M. Fichera, C. Romano, J. Gecz, L. E. DeLisi, J. Sebat, M.-C. King, L. G. Shaffer, and E. E. Eichler, "A recurrent 16p12.1 microdeletion supports a two-hit model for severe developmental delay," *Nature genetics*, vol. 42, pp. 203–209, print 2010. M3: 10.1038/ng.534; 10.1038/ng.534.

[212] B. O. J., P. Deriziotis, C. Lee, L. Vives, J. J. Schwartz, S. Girirajan, E. Karakoc, A. P. MacKenzie, S. B. Ng, C. Baker, M. J. Rieder, D. A. Nickerson, R. Bernier, S. E. Fisher, J. Shendure, and E. E. Eichler, "Exome sequencing in sporadic autism spectrum disorders identifies severe de novo mutations," *Nature genetics*, vol. 43, pp. 585–589, print 2011. M3: 10.1038/ng.835; 10.1038/ng.835.

[213] P. Hammond and M. Suttie, "Large-scale objective phenotyping of 3d facial morphology," *Human Mutation*, vol. 33, no. 5, pp. 817–825, 2012.

[214] G. Baynam, M. Walters, P. Claes, S. Kung, P. LeSouef, H. Dawkins, D. Gillett, and J. Goldblatt, "The facial evolution: Looking backward and moving forward," *Human Mutation*, vol. 34, no. 1, pp. 14–22, 2013.

[215] D. Houle, D. R. Govindaraju, and S. Omholt, "Phenomics: the next challenge," *Nat Rev Genet*, vol. 11, no. 12, pp. 855–66, 2010.

[216] A. B. Gjuvsland, J. O. Vik, D. A. Beard, P. J. Hunter, and S. W. Omholt, "Bridging the genotype-phenotype gap: what does it take?," *J Physiol*, vol. 591, no. Pt 8, pp. 2055–66, 2013.

[217] T. D. D. D. Study, "Large-scale discovery of novel genetic causes of developmental disorders," *Nature*, vol. advance online publication, 12 2014.

[218] B. J. O'Roak, L. Vives, S. Girirajan, E. Karakoc, N. Krumm, B. P. Coe, R. Levy, A. Ko, C. Lee, J. D. Smith, E. H. Turner, I. B. Stanaway, B. Vernot, M. Malig, C. Baker, B. Reilly, J. M. Akey, E. Borenstein, M. J. Rieder, D. A. Nickerson, R. Bernier, J. Shendure, and E. E. Eichler, "Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations," *Nature*, vol. 485, pp. 246–250, 05 2012.

[219] S. Topper, C. Ober, and D. S, "Exome sequencing and the genetics of intellectual disability," *Clinical Genetics*, vol. 80, pp. 117–126, 06 2011.

[220] S. B. Ng, K. J. Buckingham, C. Lee, A. W. Bigham, H. K. Tabor, K. M. Dent, C. D. Huff, P. T. Shannon, E. W. Jabs, D. A. Nickerson, J. Shendure, and M. J. Bamshad, "Exome sequencing identifies the cause of a mendelian disorder," *Nature Genetics*, vol. 42, pp. 30–35, 01 2010.

[221] G. Brito and D. Andrews, "Removing bias against membrane proteins in interaction networks," *BMC Systems Biology*, vol. 5, no. 1, p. 169, 2011.

[222] N. Škunca, A. Altenhoff, and C. Dessimoz, "Quality of computationally inferred gene ontology annotations," *PLoS Comput Biol*, vol. 8, p. e1002533, 05 2012.

[223] M. R. Munafò, G. Stothart, and J. Flint, "Bias in genetic association studies and impact factor," *Molecular psychiatry*, vol. 14, pp. 119–120, print 2009.

[224] L. Franke, H. van Bakel, L. Fokkens, E. D. de Jong, M. Egmont-Petersen, and C. Wijmenga, "Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes," *American Journal of Human Genetics*, vol. 78, pp. 1011–1025, 06/01 2006.

[225] P. Jia, S. Zheng, J. Long, W. Zheng, and Z. Zhao, "dmgwas: dense module searching for genome-wide association studies in protein-protein interaction networks," *Bioinformatics*, vol. 27, no. 1, pp. 95–102, 2011.

[226] J. Chen and B. Yuan, "Detecting functional modules in the yeast protein-protein interaction network," *Bioinformatics*, vol. 22, no. 18, pp. 2283–2290, 2006.

[227] H. Jeong, S. P. Mason, A.-L. Barabasi, and Z. N. Oltvai, "Lethality and centrality in protein networks," *Nature*, vol. 411, pp. 41–42, 03/05 2001.

[228] J.-F. Rual, K. Venkatesan, T. Hao, T. Hirozane-Kishikawa, A. Dricot, N. Li, G. F. Berriz, F. D. Gibbons, M. Dreze, N. Ayivi-Guedehoussou, N. Klitgord, C. Simon, M. Boxem, S. Milstein, J. Rosenberg, D. ebra S. Goldberg, L. V. Zhang, S. L. Wong, G. Franklin, S. Li, J. S. Albala, J. Lim, C. Fraughton, E. Llamosas, S. Cevik, C. Bex, P. Lamesch, R. S. Sikorski, J. Vandenhaute, H. Y. Zoghbi, A. Smolyar, S. Bosak, R. Sequerra, L. Doucette-Stamm, M. E. Cusick, D. E. Hill, F. P. Roth, and M. Vidal, "Towards a proteome-scale map of the human protein-protein interaction network," *Nature*, vol. 437, no. 7062, pp. 1173–1178, 2005.

[229] O. Thellin, W. Zorzi, B. Lakaye, B. D. Borman, B. Coumans, G. Hennen, T. Grisar, A. Igout, and E. Heinen, "Housekeeping genes as internal standards: use and limits," *Journal of Biotechnology*, vol. 75, no. 2-3, pp. 291–295, 1999.

[230] J.-D. J. Han, N. Bertin, T. Hao, D. S. Goldberg, G. F. Berriz, L. V. Zhang, D. Dupuy, A. J. M. Walhout, M. E. Cusick, F. P. Roth, and M. Vidal, "Evidence for dynamically organized modularity in the yeast protein-protein interaction network," *Nature*, vol. 430, pp. 88–93, 07/01 2004. M3: 10.1038/nature02555; 10.1038/nature02555.

[231] M. T. Dittrich, G. W. Klau, A. Rosenwald, T. Dandekar, and T. MÃ¼ller, "Identi-

fying functional modules in protein-protein interaction networks: an integrated exact approach," *Bioinformatics*, vol. 24, no. 13, pp. i223–i231, 2008.

[232] P. N. Robinson and C. Webber, "Phenotype ontologies and cross-species analysis for translational research," *PLoS Genet*, vol. 10, p. e1004268, 04/03 2014.

[233] P. N. Robinson, S. KÃ¶hler, A. Oellrich, S. M. G. Project, K. Wang, C. J. Mungall, S. E. Lewis, N. Washington, S. Bauer, D. Seelow, P. Krawitz, C. Gilissen, M. Haendel, and D. Smedley, "Improved exome prioritization of disease genes through cross-species phenotype comparison," *Genome Research*, vol. 24, no. 2, pp. 340–348, 2014.

Appendices

Appendix A

Additional Figures

A.1 Chapter 4

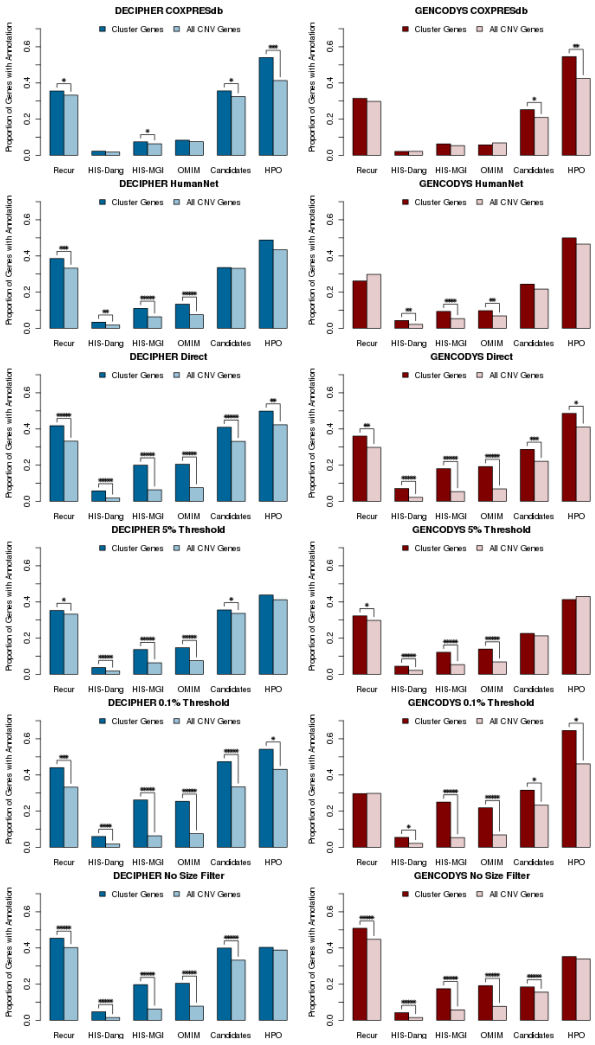


Figure A.1: Robustness of enrichments of various disease relevant annotations in functionally similar genes in DECIPHER and GENCODYS respectively compared to all genes in their CNVs. Recur indicates genes found in more than one *de novo* CNV in the same dataset, Dang-HIS are haplo-insufficient genes identified in (158), MGI-HIS are genes for which the mouse ortholog is annotated with at least one phenotype when heterozygous for the mutation (36; 37), OMIM are genes causally related to a disease in the Online Mendelian Inheritance in Man database (112), Candidates are those genes annotated with mouse phenotypes that were associated with the respective patient's symptoms in a previous study(20). HPO are those genes annotated with at least one of the respective patient's symptoms in the Human Phenotype Ontology database (35). One star indicates $p < 0.05$, two stars indicates $p < 0.0005$. (C) Functional clusters are found in highly recurrently affected regions of the genome. Direct is the direct edges in the PLN prior to calculation of shortest-paths, x% Threshold indicates the proportion of shortest-paths used to identify functional clusters. No Size Filter uses the PLN on all *de novo* CNVs in each dataset.

A.2 Chapter 6

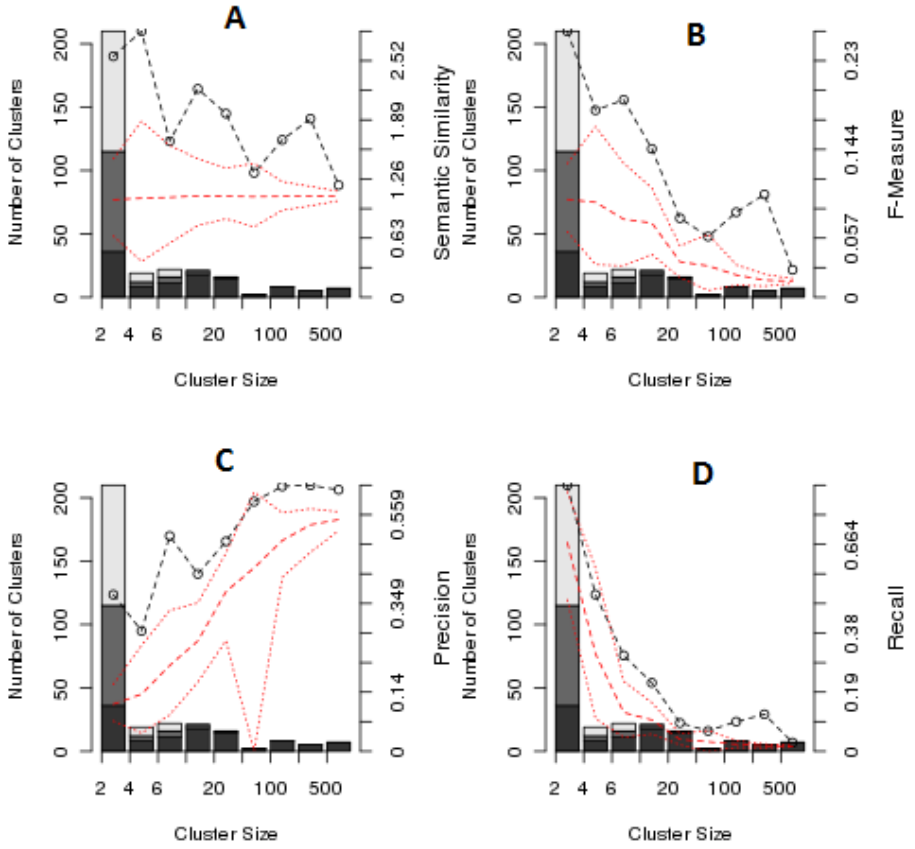


Figure A.2: Transferability of phenotypic information using the COXPRESdb network modules across a broad range of module sizes. (A,C,D)Leave-one-out cross validations within each module using a majority-rule prediction algorithm. (B) Average semantic similarity. Black line and open circles indicates the observed results. Red dashed line indicates the median over 1,000 module-permutations. Dotted lines contain 95% of the permutations. Dark bars are the number of modules in that size class with at least 2 genes with mouse phenotype annotations thus contribute to the transferability estimate; medium grey bars are the number of modules with exactly one gene with mouse phenotype information available; light bars are the number of modules with no genes with mouse phenotype information available.

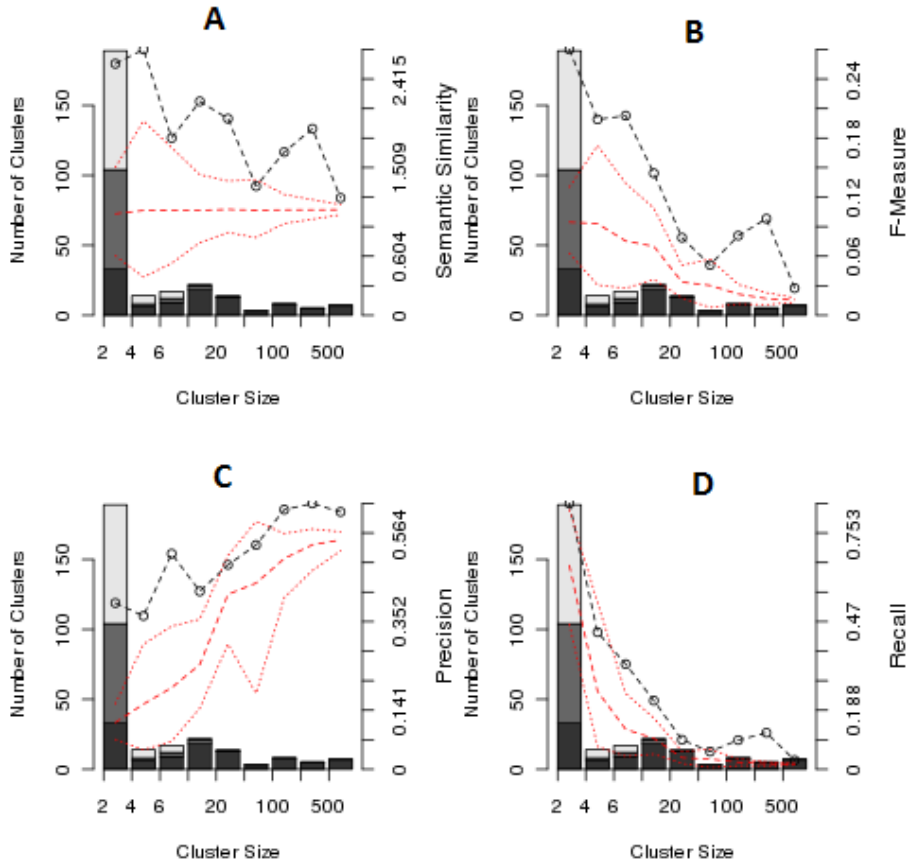


Figure A.3: Transferability of phenotypic information using the COXPRESdb network consensus clustering modules across a broad range of module sizes. (A,C,D)Leave-one-out cross validations within each module using a majority-rule prediction algorithm. (B) Average semantic similarity. Black line and open circles indicates the observed results. Red dashed line indicates the median over 1,000 module-permutations. Dotted lines contain 95% of the permutations. Dark bars are the number of modules in that size class with at least 2 genes with mouse phenotype annotations thus contribute to the transferability estimate; medium grey bars are the number of modules with exactly one gene with mouse phenotype information available; light bars are the number of modules with no genes with mouse phenotype information available.

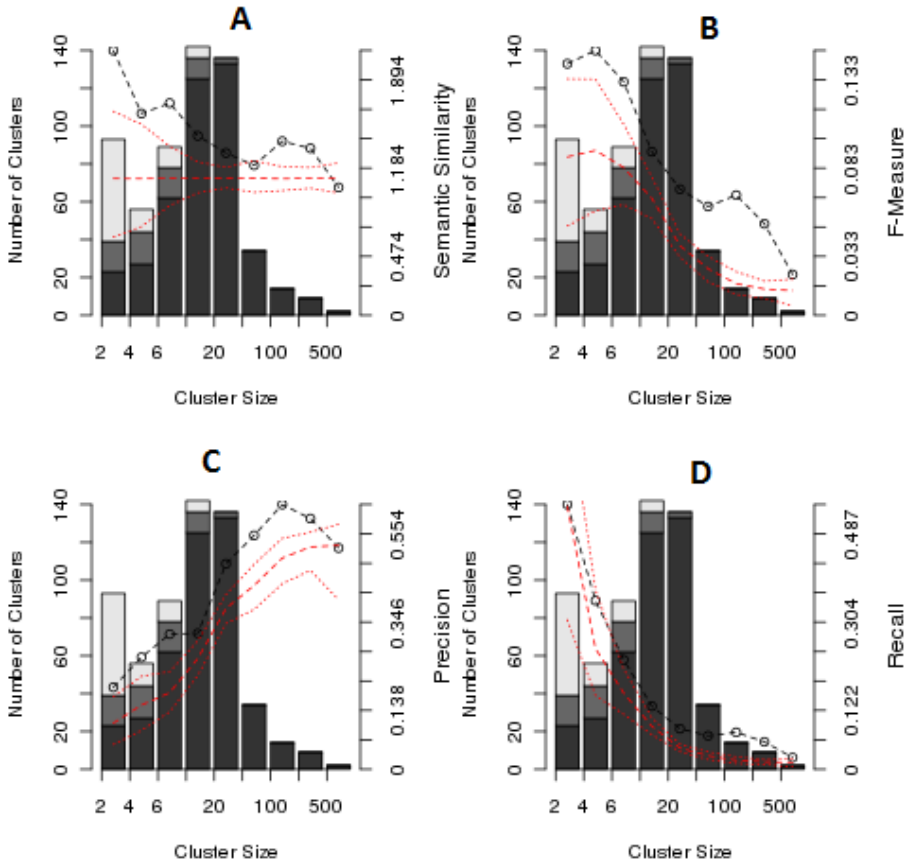


Figure A.4: Transferability of phenotypic information using the HumanNet network modules across a broad range of module sizes. (A,C,D)Leave-one-out cross validations within each module using a majority-rule prediction algorithm. (B) Average semantic similarity. Black line and open circles indicates the observed results. Red dashed line indicates the median over 1,000 module-permutations. Dotted lines contain 95% of the permutations. Dark bars are the number of modules in that size class with at least 2 genes with mouse phenotype annotations thus contribute to the transferability estimate; medium grey bars are the number of modules with exactly one gene with mouse phenotype information available; light bars are the number of modules with no genes with mouse phenotype information available.

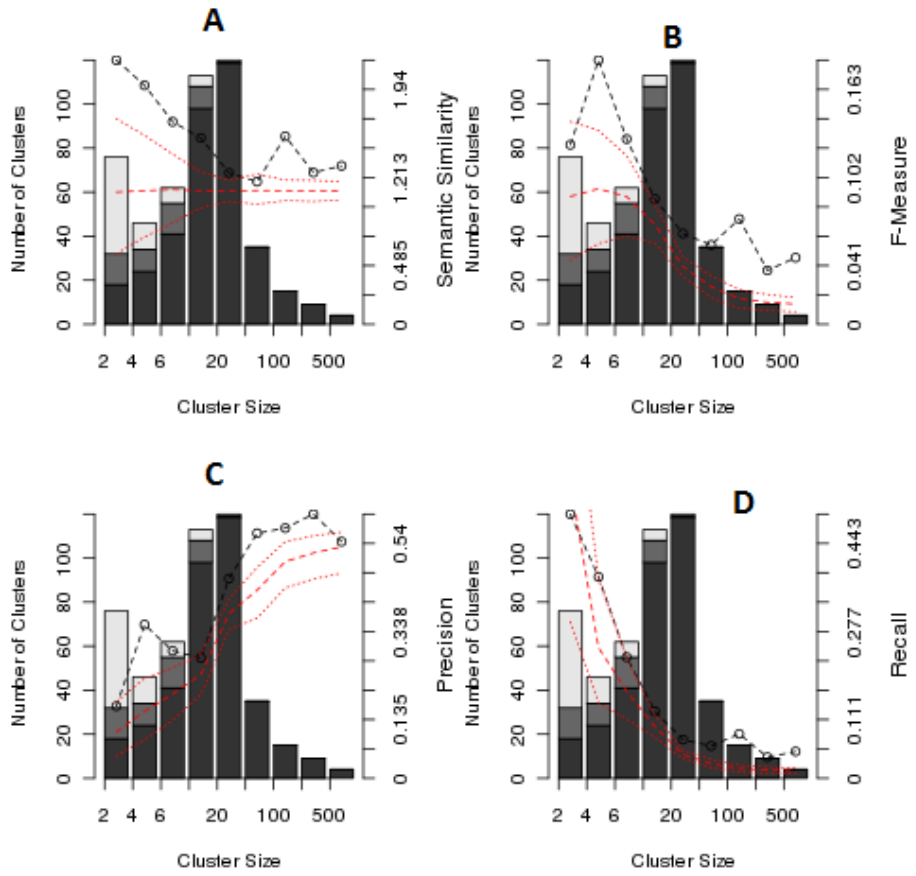


Figure A.5: Transferability of phenotypic information using the HumanNet network consensus clustering modules across a broad range of module sizes. (A,C,D)Leave-one-out cross validations within each module using a majority-rule prediction algorithm. (B) Average semantic similarity. Black line and open circles indicates the observed results. Red dashed line indicates the median over 1,000 module-permutations. Dotted lines contain 95% of the permutations. Dark bars are the number of modules in that size class with at least 2 genes with mouse phenotype annotations thus contribute to the transferability estimate; medium grey bars are the number of modules with exactly one gene with mouse phenotype information available; light bars are the number of modules with no genes with mouse phenotype information available.

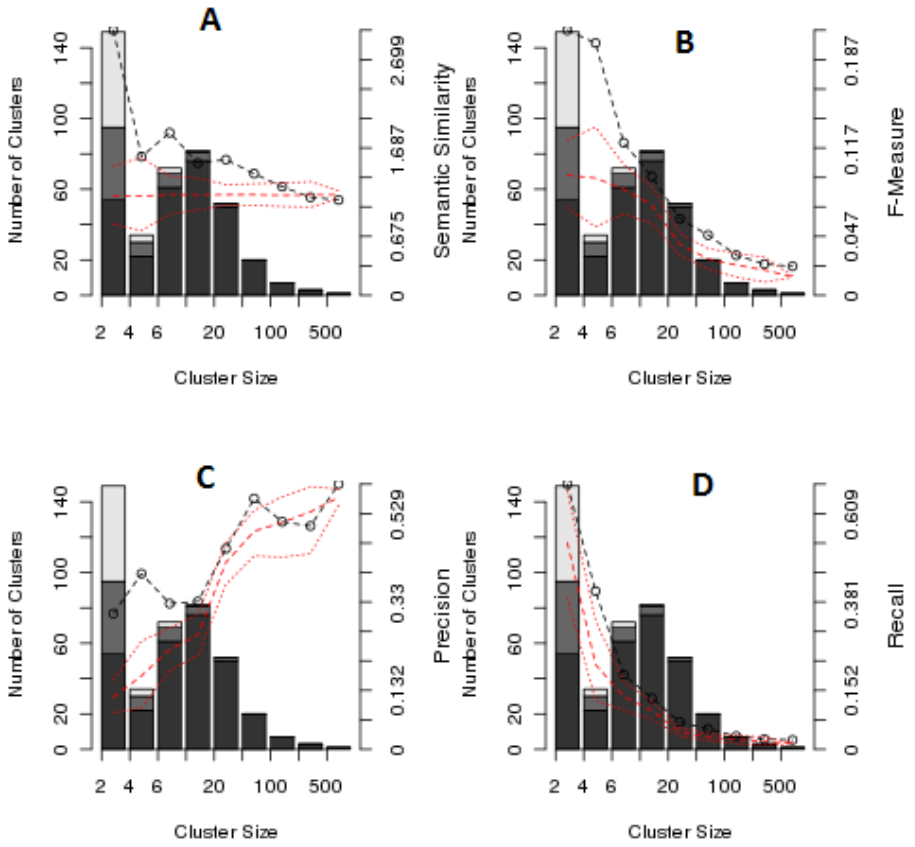


Figure A.6: Transferability of phenotypic information using the iRefIndex network modules across a broad range of module sizes. (A,C,D)Leave-one-out cross validations within each module using a majority-rule prediction algorithm. (B) Average semantic similarity. Black line and open circles indicates the observed results. Red dashed line indicates the median over 1,000 module-permutations. Dotted lines contain 95% of the permutations. Dark bars are the number of modules in that size class with at least 2 genes with mouse phenotype annotations thus contribute to the transferability estimate; medium grey bars are the number of modules with exactly one gene with mouse phenotype information available; light bars are the number of modules with no genes with mouse phenotype information available.

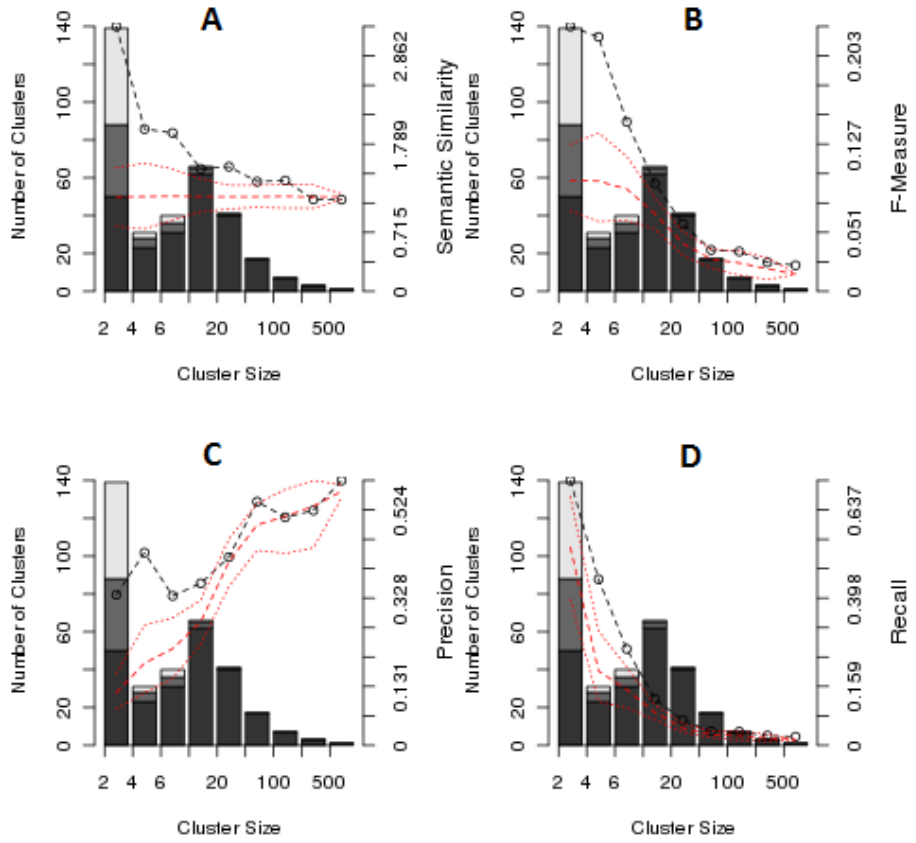


Figure A.7: Transferability of phenotypic information using the iRefIndex network consensus clustering modules across a broad range of module sizes. (A,C,D)Leave-one-out cross validations within each module using a majority-rule prediction algorithm. (B) Average semantic similarity. Black line and open circles indicates the observed results. Red dashed line indicates the median over 1,000 module-permutations. Dotted lines contain 95% of the permutations. Dark bars are the number of modules in that size class with at least 2 genes with mouse phenotype annotations thus contribute to the transferability estimate; medium grey bars are the number of modules with exactly one gene with mouse phenotype information available; light bars are the number of modules with no genes with mouse phenotype information available.

Appendix B

Additional Tables

B.1 Methods

Table B.1: Contiguous Gene Syndromes

Location	Name	No. Phenotypes
chr1:1-12762846	1p36 microdeletion syndrome	6
chr1:145000000-146350000	1q21.1 recurrent microdeletion	2
chr1:145000000-146350000	1q21.1 recurrent microduplication	2
chr1:144097863-144459424	1q21.1 susceptibility locus for TAR syndrome	0
chr2:59139200-61673319	2p15-16.1 microdeletion syndrome	10
chr2:44263955-44443088	2p21 Microdeletion Syndrome	0
chr2:196633366-204915184	2q33.1 deletion syndrome	8
chr2:239634800-239987580	2q37 monosomy	5
chr3:197211232-198829060	3q29 microdeletion syndrome	4
chr3:197211232-198829060	3q29 microduplication syndrome	0
chr4:1539257-2080034	Wolf-Hirschhorn Syndrome	5
chr5:112071100-112209835	Familial Adenomatous Polyposis	1
chr5:63001-12586304	Cri du Chat Syndrome	3
chr5:175657242-176984722	Sotos syndrome	5

Continued on next page

Location	Name	No. Phenotypes
chr5:126140213-126200611	Adult-onset autosomal dominant leukodystrophy	4
chr7:96156014-96177139	Split hand/foot malformation 1	11
chr7:72382391-73780608	Williams-Beuren Syndrome	7
chr7:72382391-73780608	7q11.23 duplication syndrome	3
chr8:8137465-11802038	8p23.1 deletion syndrome	7
chr8:8137465-11802038	8p23.1 duplication syndrome	2
chr8:77389019-77928794	8q21.11 Microdeletion Syndrome	0
chr9:139633264-139850399	9q subtelomeric deletion syndrome	9
chr11:31762915-32413663	WAGR 11p13 deletion syndrome	4
chr11:43951376-46009026	Potocki-Shaffer syndrome	5
chr12:1080000-1346471	12p13.33 Microdeletion Syndrome	0
chr12:63358186-66931792	12q14 microdeletion syndrome	3
chr15:28697598-30232699	15q13.3 microdeletion syndrome	3
chr15:72199696-73759966	15q24 recurrent microdeletion syndrome	7
chr15:97175493-100338915	15q26 overgrowth syndrome	3
chr15:20300718-26111861	Angelman syndrome (Type 1)	5
chr15:21171353-26111861	Angelman syndrome (Type 2)	5
chr15:20300718-26111861	Prader-Willi syndrome (Type 1)	5
chr15:21171353-26111861	Prader-Willi Syndrome (Type 2)	5
chr16:29514353-30107356	16p11.2 microduplication syndrome	0
chr16:21419563-3010735	16p11.2-p12.2 microdeletion syndrome	4
chr16:14894185-16394185	16p13.11 recurrent microdeletion	0
chr16:14894185-16394185	16p13.11 recurrent microduplication	0
chr16:1-774373	ATR-16 syndrome	2
chr16:21854025-22374785	Recurrent 16p12.1 microdeletion	3

Continued on next page

Location	Name	No. Phenotypes
chr16:3715056-3870122	Rubinstein-Taybi Syndrome	0
chr17:41060949-41650183	17q21.31 recurrent microdeletion syndrome	5
chr17:14038640-15411628	Hereditary Liability to Pressure Palsies	2
chr17:1-2535659	Miller-Dieker syndrome (MDS)	5
chr17:26131223-27287434	NF1-microdeletion syndrome	6
chr17:14038640-15411628	Charcot-Marie-Tooth syndrome type 1A	8
chr17:16713797-20162741	Potocki-Lupski syndrome	4
chr17:16713797-20162741	Smith-Magenis Syndrome	9
chr17:31889185-33290030	RCAD (renal cysts and diabetes)	3
chr21:26174731-26465317	Early-onset Alzheimer disease with cerebral amyloid angiopathy	3
chr22:17389792-19782445	Velocardiofacial / DiGeorge syndrome	5
chr22:17389792-19782445	22q11 duplication syndrome	3
chr22:20247117-22052445	22q11.2 distal deletion syndrome	4
chr22:49392382-49534710	Phelan-Mcdermid syndrome	5
chrX:671878-787875	Leri-Weill dyschondrostosis	4
chrX:380558-673877	Leri-Weill dyschondrostosis	4
chrX:102918094-102934203	Pelizaeus-Merzbacher disease	4
chrX:6465812-8093195	Steroid sulphatase deficiency	1
chrX:53417795-53700000	Xp11.22-linked intellectual disability	0
chrX:48219493-52134376	Xp11.22-p11.23 Microduplication	0
chrX:152940457-153016382	Xq28 (MECP2) duplication	6
chrX:153277757-153535047	Xq28 Microduplication	0

B.2 Chapter 3

Table B.2: Mapping HPO terms to MPO overarching categories

HPO	HPO Name	MPO	MPO Name
HP:0000141	Amenorrhea	MP:0005389	reproductive system phenotype
HP:0001028	Hemangiomas	MP:0002006	tumorigenesis
HP:0000076	Vesicoureteral reflux	MP:0005367	renal/urinary system phenotype
HP:0000262	Turricephaly	MP:0005382	craniofacial phenotype
HP:0008062	Aplasia/Hypoplasia affecting the anterior segment of the eye	MP:0005391	vision/eye phenotype
HP:0003241	Genital hypoplasia	MP:0005389	reproductive system phenotype
HP:0000481	Abnormality of the cornea	MP:0005391	vision/eye phenotype
HP:0002719	Recurrent infections	MP:0005387	immune system phenotype
HP:0000395	Prominent antihelix	MP:0005382	craniofacial phenotype
HP:0000239	Large fontanelles	MP:0005382	craniofacial phenotype
HP:0005751	Ridging of metopic suture	MP:0005382	craniofacial phenotype
HP:0007400	Irregular hyperpigmentation	MP:0001186	pigmentation phenotype
HP:0000248	Brachycephaly	MP:0005382	craniofacial phenotype
HP:z999018			
HP:0005927	Aplasia/Hypoplasia involving bones of the hand	MP:0005371	limbs/digits/tail phenotype
HP:0000598	Abnormality of the ear	MP:0005377	hearing/vestibular/ear phenotype

Continued on next page

HPO	HPO Name	MPO	MPO Name
HP:0002208	Coarse hair	MP:0010771	integument phenotype
HP:0001838	Vertical talus	MP:0005371	limbs/digits/tail phenotype
HP:0009122	Aplasia/Hypoplasia affecting bones of the axial skeleton	MP:0005390	skeleton phenotype
HP:0001597	Abnormality of the nail	MP:0010771	integument phenotype
HP:0007618	Subcutaneous calcification	MP:0010771	integument phenotype
HP:0005091	Asymmetric limb shortening	MP:0005371	limbs/digits/tail phenotype
HP:0002967	Cubitus valgus	MP:0005390	skeleton phenotype
HP:0000215	Thick upper lip vermilion	MP:0005382	craniofacial phenotype
HP:0002286	Light colored hair	MP:0010771	integument phenotype
HP:0010554	Cutaneous syndactyly of the fingers	MP:0005371	limbs/digits/tail phenotype
HP:0001627	Abnormality of the heart	MP:0005385	cardiovascular system phenotype
HP:0008070	Sparse hair	MP:0010771	integument phenotype
HP:0009553	Abnormality of the hairline	MP:0010771	integument phenotype
HP:0001572	Macrodontia	MP:0005382	craniofacial phenotype
HP:0000325	Triangular facies	MP:0005382	craniofacial phenotype
HP:0002926	Abnormality of thyroid physiology	MP:0005379	endocrine/exocrine gland phenotype
HP:0001018	Abnormal palmar dermatoglyphics	MP:0010771	integument phenotype
HP:0002857	Genu valgum	MP:0005371	limbs/digits/tail phenotype

Continued on next page

HPO	HPO Name	MPO	MPO Name
HP:0000189	Narrow palate	MP:0005382	craniofacial phenotype
HP:0001382	Joint hypermobility	MP:0005390	skeleton phenotype
HP:0005111	Dilatation of the ascending aorta	MP:0005385	cardiovascular system phenotype
HP:0004362	Abnormality of the enteric ganglia	MP:0005386	behavior/neurological phenotype
		MP:0003631	nervous system phenotype
HP:0001065	Striae distensae	MP:0010771	integument phenotype
HP:0001197	Abnormality of prenatal development or birth	MP:0000001	mammalian phenotype
HP:0006316	Irregularly spaced teeth	MP:0005382	craniofacial phenotype
HP:0000327	Hypoplasia of the maxilla	MP:0005382	craniofacial phenotype
HP:0008661	Urethral stenosis	MP:0005367	renal/urinary system phenotype
HP:0000311	Round face	MP:0005382	craniofacial phenotype
HP:0000009	Functional abnormality of the bladder	MP:0005367	renal/urinary system phenotype
HP:0001339	Lissencephaly	MP:0005384	cellular phenotype
HP:0001260	Dysarthria	MP:0005386	behavior/neurological phenotype
		MP:0003631	nervous system phenotype
HP:0000499	Abnormality of the eyelashes	MP:0005391	vision/eye phenotype
HP:0000284	Abnormality of the ocular region	MP:0005382	craniofacial phenotype
HP:0000750	Impaired language development	MP:0005386	behavior/neurological phenotype
		MP:0003631	nervous system phenotype

Continued on next page

HPO	HPO Name	MPO	MPO Name
HP:0000098	Tall stature	MP:0005378	growth/size phenotype
HP:0000294	Low anterior hairline	MP:0010771	integument phenotype
HP:0200045	Abnormality of pigmentation	MP:0010771	integument phenotype
HP:0000656	Ectropion	MP:0005391	vision/eye phenotype
HP:0003011	Abnormality of musculature	MP:0005369	muscle phenotype
HP:0000119	Abnormality of the genitourinary system	MP:0005389	reproductive system phenotype
HP:0001333	Abnormality of the sensory nervous system	MP:0005386	behavior/neurological phenotype
		MP:0003631	nervous system phenotype
HP:0010973	Abnormality of erythroid lineage cell	MP:0000001	mammalian phenotype
HP:0008736	Hypoplasia of penis	MP:0005389	reproductive system phenotype
HP:0000366	Abnormality of the nose	MP:0005382	craniofacial phenotype
HP:0000508	Ptosis	MP:0005391	vision/eye phenotype
HP:0000429	Abnormality of the nasal alae	MP:0005382	craniofacial phenotype
HP:0000388	Otitis media	MP:0005377	hearing/vestibular/ear phenotype
HP:0000589	Coloboma	MP:0005391	vision/eye phenotype
HP:0100259	Postaxial polydactyly	MP:0005371	limbs/digits/tail phenotype
HP:0002650	Scoliosis	MP:0005390	skeleton phenotype
HP:0000708	Behavioural/Psychiatric Abnormality	MP:0005386	behavior/neurological phenotype

Continued on next page

HPO	HPO Name	MPO	MPO Name
		MP:0003631	nervous system phenotype
HP:0100037	Abnormality of the scalp hair	MP:0005382	craniofacial phenotype
HP:0000446	Narrow nasal bridge	MP:0005382	craniofacial phenotype
HP:0001439	Abnormality of the thigh	MP:0005371	limbs/digits/tail phenotype
HP:0010647	Abnormality of skin texture	MP:0010771	integument phenotype
HP:0001518	Low birth weight	MP:0005378	growth/size phenotype
HP:0002779	Tracheomalacia	MP:0005388	respiratory system phenotype
HP:0000078	Abnormality of the genital system	MP:0005389	reproductive system phenotype
HP:0000527	Long eyelashes	MP:0005391	vision/eye phenotype
HP:0002251	Congenital megacolon	MP:0005381	digestive/alimentary phenotype
HP:0100238	Synostosis involving bones of the upper limbs	MP:0005390	skeleton phenotype
HP:0000240	Abnormality of skull size	MP:0005382	craniofacial phenotype
HP:0100656	Thoracoabdominal wall defects	MP:0005381	digestive/alimentary phenotype
HP:0000506	Telecanthus	MP:0005391	vision/eye phenotype
HP:0001291	Abnormality of the cranial nerves	MP:0005386	behavior/neurological phenotype
		MP:0003631	nervous system phenotype
HP:0001724	Aortic dilatation	MP:0005385	cardiovascular system phenotype

Continued on next page

HPO	HPO Name	MPO	MPO Name
HP:0008564	External auditory canal stenosis/atresia	MP:0005377	hearing/vestibular/ear phenotype
HP:0000456	Bifid nasal tip	MP:0005382	craniofacial phenotype
HP:0001052	Nevus flammeus	MP:0005385	cardiovascular system phenotype
HP:0100790	Herniae	MP:0005378	growth/size phenotype
HP:0002143	Abnormality of the spinal cord	MP:0005386	behavior/neurological phenotype
		MP:0003631	nervous system phenotype
HP:0001438	Abnormality of the abdomen	MP:0005381	digestive/alimentary phenotype
HP:0000416	Choanal atresia or stenosis	MP:0005382	craniofacial phenotype
HP:0005469	Flat occiput	MP:0005382	craniofacial phenotype
HP:0001654	Abnormality of the heart valves	MP:0005385	cardiovascular system phenotype
HP:0000396	Overfolded helix	MP:0005382	craniofacial phenotype
HP:0001324	Muscle weakness	MP:0005369	muscle phenotype
HP:0010781	Skin dimples	MP:0010771	integument phenotype
HP:0009485	Radial deviation of the hand or of fingers of the hand	MP:0005371	limbs/digits/tail phenotype
HP:0009118	Aplasia/Hypoplasia of the mandible	MP:0005382	craniofacial phenotype
HP:0000691	Microdontia	MP:0005382	craniofacial phenotype
HP:0002012	Abnormality of the abdominal organs	MP:0005381	digestive/alimentary phenotype
		MP:0005370	liver/biliary system phenotype

Continued on next page

HPO	HPO Name	MPO	MPO Name
HP:0000965	Cutis marmorata	MP:0010771	integument phenotype
HP:0001555	Asymmetry of the thorax	MP:0005390	skeleton phenotype
HP:0000288	Abnormality of the philtrum	MP:0005382	craniofacial phenotype
HP:0011069	Increased number of teeth	MP:0005382	craniofacial phenotype
HP:0002778	Abnormality of the trachea	MP:0005388	respiratory system phenotype
HP:z999011			
HP:0001360	Holoprosencephaly	MP:0005386	behavior/neurological phenotype
		MP:0003631	nervous system phenotype
HP:0001837	Broad toes	MP:0005371	limbs/digits/tail phenotype
HP:0000326	Abnormality of the maxilla	MP:0005382	craniofacial phenotype
HP:0100555	Asymmetric growth	MP:0005378	growth/size phenotype
HP:0001850	Abnormality of the tarsal bones	MP:0005371	limbs/digits/tail phenotype
HP:0000400	Large ears	MP:0005382	craniofacial phenotype
HP:0006499	Abnormality of femoral epiphyses	MP:0005371	limbs/digits/tail phenotype
HP:0000309	Abnormality of the midface	MP:0005382	craniofacial phenotype
HP:0000213	Thin lips	MP:0005382	craniofacial phenotype
HP:0009778	Hypoplastic/small thumb	MP:0005371	limbs/digits/tail phenotype
HP:0009997	Partial/complete duplication of phalanges of the hand	MP:0005371	limbs/digits/tail phenotype

Continued on next page

HPO	HPO Name	MPO	MPO Name
HP:0001085	Papilledema	MP:0005391	vision/eye phenotype
HP:0002536	Abnormal cortical gyration	MP:0005384	cellular phenotype
HP:0001956	Truncal obesity	MP:0005378	growth/size phenotype
HP:0001560	Abnormality of the amniotic fluid	MP:0000001	mammalian phenotype
HP:0002031	Abnormality of the esophagus	MP:0005381	digestive/alimentary phenotype
HP:0006496	Aplasia/Hypoplasia involving bones of the upper limbs	MP:0005371	limbs/digits/tail phenotype
HP:0000356	Abnormality of the outer ear	MP:0005377	hearing/vestibular/ear phenotype
HP:0001845	Overriding toes	MP:0005371	limbs/digits/tail phenotype
HP:0007874	Almond-shaped palpebral fissures	MP:0005391	vision/eye phenotype
HP:0002577	Abnormality of the stomach	MP:0005381	digestive/alimentary phenotype
HP:0004323	Abnormality of body weight	MP:0005378	growth/size phenotype
HP:0001252	Muscular hypotonia	MP:0005369	muscle phenotype
HP:0001642	Pulmonic stenosis	MP:0005385	cardiovascular system phenotype
HP:0000053	Macroorchidism	MP:0005389	reproductive system phenotype
HP:0010733	Naevus flammeus of the eyelid	MP:0005385	cardiovascular system phenotype

Continued on next page

HPO	HPO Name	MPO	MPO Name
HP:0005599	Hypopigmentation of hair	MP:0010771	integument phenotype
HP:0000540	Hypermetropia	MP:0005391	vision/eye phenotype
HP:0001852	Gap between first and second toes	MP:0005371	limbs/digits/tail phenotype
HP:0000606	Abnormality of the periorbital region	MP:0005391	vision/eye phenotype
HP:0004122	Midline defect of the nose	MP:0005382	craniofacial phenotype
HP:0010628	Facial nerve palsy	MP:0005369	muscle phenotype
HP:0000470	Short neck	MP:0005382	craniofacial phenotype
HP:0000490	Deeply set eye	MP:0005391	vision/eye phenotype
HP:0009179	Deviation of the 5th finger	MP:0005371	limbs/digits/tail phenotype
HP:0001892	Bleeding diathesis	MP:0005376	homeostasis/metabolism phenotype
HP:0000202	Cleft lip/palate	MP:0005382	craniofacial phenotype
HP:0000960	Sacral dimple	MP:0010771	integument phenotype
HP:0100490	Camptodactyly (hands)	MP:0005369	muscle phenotype
HP:0006705	Abnormality of the atrioventricular valves	MP:0005385	cardiovascular system phenotype
HP:0002818	Abnormality of the radius	MP:0005371	limbs/digits/tail phenotype
HP:0002214	Blond hair	MP:0010771	integument phenotype
HP:0000069	Abnormality of the ureter	MP:0005367	renal/urinary system phenotype
HP:0001608	Abnormality of the voice	MP:0005388	respiratory system phenotype
HP:0000301	Abnormality of facial musculature	MP:0005369	muscle phenotype

Continued on next page

HPO	HPO Name	MPO	MPO Name
HP:0002071	Extrapyramidal signs	MP:0005386	behavior/neurological phenotype
		MP:0003631	nervous system phenotype
HP:0000639	Nystagmus	MP:0005391	vision/eye phenotype
HP:0004377	Hematological neoplasia	MP:0002006	tumorigenesis
HP:0000478	Abnormality of the eye	MP:0005391	vision/eye phenotype
HP:0000581	Blepharophimosis	MP:0005391	vision/eye phenotype
HP:0000001	All	MP:0000001	mammalian phenotype
HP:0001965	Abnormality of the scalp	MP:0005382	craniofacial phenotype
HP:0001176	Large hands	MP:0005371	limbs/digits/tail phenotype
HP:0002664	Neoplasia	MP:0002006	tumorigenesis
HP:0006493	Aplasia/Hypoplasia involving bones of the lower limbs	MP:0005371	limbs/digits/tail phenotype
HP:0000077	Abnormality of the kidney	MP:0005367	renal/urinary system phenotype
HP:0001780	Abnormality of the toes	MP:0005371	limbs/digits/tail phenotype
HP:0002250	Abnormality of the large intestine	MP:0005381	digestive/alimentary phenotype
HP:0000218	High palate	MP:0005382	craniofacial phenotype
HP:0009907	Adherent earlobe	MP:0005382	craniofacial phenotype
HP:0001156	Brachydactyly	MP:0005371	limbs/digits/tail phenotype
HP:0000377	Abnormality of the pinna	MP:0005382	craniofacial phenotype
HP:0000179	Thick lower lip vermilion	MP:0005382	craniofacial phenotype
HP:0000463	Nares, anteverted	MP:0005382	craniofacial phenotype

Continued on next page

HPO	HPO Name	MPO	MPO Name
HP:0009767	Aplasia/Hypoplasia of the phalanges of the hand	MP:0005371	limbs/digits/tail phenotype
HP:0001187	Hyperextensibility of the finger joints	MP:0005371	limbs/digits/tail phenotype
HP:0010959	Congenital cystic adenomatoid malformation of the lung	MP:0005388	respiratory system phenotype
HP:0000348	High forehead	MP:0005382	craniofacial phenotype
HP:0000670	Carious teeth	MP:0005382	craniofacial phenotype
HP:0100585	Teleangiectasia of the skin	MP:0005369	muscle phenotype
HP:0001991	Aplasia/Hypoplasia of the toes	MP:0005371	limbs/digits/tail phenotype
HP:0010866	Abdominal wall defect	MP:0005381	digestive/alimentary phenotype
HP:0002019	Constipation	MP:0005381	digestive/alimentary phenotype
HP:0001643	Patent ductus arteriosus	MP:0005385	cardiovascular system phenotype
HP:0000455	Broad nasal tip	MP:0005382	craniofacial phenotype
HP:0000372	Abnormality of the auditory canal	MP:0005377	hearing/vestibular/ear phenotype
HP:0000765	Abnormality of the thorax	MP:0005390	skeleton phenotype
HP:0004328	Abnormality of the anterior segment of the eye	MP:0005391	vision/eye phenotype
HP:0001500	Broad fingers	MP:0005371	limbs/digits/tail phenotype
HP:0002230	Generalized hirsutism	MP:0010771	integument phenotype
HP:0002002	Deep philtrum	MP:0005382	craniofacial phenotype

Continued on next page

HPO	HPO Name	MPO	MPO Name
HP:0001167	Abnormality of the fingers	MP:0005371	limbs/digits/tail phenotype
HP:0009601	Aplasia/Hypoplasia of the thumb	MP:0005371	limbs/digits/tail phenotype
HP:0000958	Dry skin	MP:0010771	integument phenotype
HP:0002022	Feeding difficulties	MP:0005381	digestive/alimentary phenotype
		MP:0005370	liver/biliary system phenotype
HP:0000766	Abnormality of the sternum	MP:0005390	skeleton phenotype
HP:0006292	Abnormality of dental eruption	MP:0005382	craniofacial phenotype
HP:0000426	Prominent nasal bridge	MP:0005382	craniofacial phenotype
HP:0000504	Abnormality of vision	MP:0005391	vision/eye phenotype
HP:0004987	Mesomelia of the lower limbs	MP:0005371	limbs/digits/tail phenotype
HP:0001163	Abnormality of the metacarpal bones	MP:0005371	limbs/digits/tail phenotype
HP:0006261	Abnormality of phalangeal joints of the hand	MP:0005371	limbs/digits/tail phenotype
HP:0002213	Fine hair	MP:0010771	integument phenotype
HP:z999016			
HP:0004467	Preauricular pit	MP:0010771	integument phenotype
HP:0010549	Paralysis due to lesions of the principle motor tracts	MP:0005386	behavior/neurological phenotype
		MP:0003631	nervous system phenotype

Continued on next page

HPO	HPO Name	MPO	MPO Name
HP:0006500	Abnormality involving the epiphyses of the lower limbs	MP:0005371	limbs/digits/tail phenotype
HP:0000430	Hypoplastic nasal alae	MP:0005382	craniofacial phenotype
HP:0001631	Atrial septal defect	MP:0005385	cardiovascular system phenotype
HP:0200033	patches	MP:0001186	pigmentation phenotype
HP:0002373	Febrile seizures	MP:0005386	behavior/neurological phenotype
		MP:0003631	nervous system phenotype
HP:0001388	Joint laxity	MP:0005390	skeleton phenotype
HP:0000457	Flat nose	MP:0005382	craniofacial phenotype
HP:0000521	Abnormality of tear glands or tear production	MP:0005391	vision/eye phenotype
HP:0000517	Abnormality of the lens	MP:0005391	vision/eye phenotype
HP:0003508	Proportionate short stature	MP:0005378	growth/size phenotype
HP:0001928	Abnormality of coagulation	MP:0005376	homeostasis/metabolism phenotype
HP:0001273	Abnormality of the corpus callosum	MP:0005386	behavior/neurological phenotype
		MP:0003631	nervous system phenotype
HP:0010827	Abnormality of the seventh cranial nerve	MP:0005386	behavior/neurological phenotype
		MP:0003631	nervous system phenotype
HP:0000358	Posteriorly rotated ears	MP:0005382	craniofacial phenotype
HP:0007018	Attention deficit hyperactivity disorder	MP:0005386	behavior/neurological phenotype
		MP:0003631	nervous system phenotype

Continued on next page

HPO	HPO Name	MPO	MPO Name
HP:0100667	Brachydactyly (hand)	MP:0005371	limbs/digits/tail phenotype
HP:0000047	Hypospadias	MP:0005367	renal/urinary system phenotype
HP:0000415	Abnormality of the choanae	MP:0005382	craniofacial phenotype
HP:0008872	Feeding problems in infancy	MP:0005381	digestive/alimentary phenotype
		MP:0005370	liver/biliary system phenotype
HP:0004097	Deviated fingers	MP:0005371	limbs/digits/tail phenotype
HP:0001251	Ataxia	MP:0005386	behavior/neurological phenotype
		MP:0003631	nervous system phenotype
HP:0001556	Sloping shoulders	MP:0005390	skeleton phenotype
HP:0009811	Abnormality of the elbow	MP:0005390	skeleton phenotype
HP:0001035	Abnormality of keratinization	MP:0010771	integument phenotype
HP:0008544	Abnormally folded helix	MP:0005382	craniofacial phenotype
HP:0000821	Hypothyroidism	MP:0005379	endocrine/exocrine gland phenotype
HP:0000925	Abnormality of the vertebral column	MP:0005390	skeleton phenotype
HP:0000365	Hearing impairment	MP:0005377	hearing/vestibular/ear phenotype
HP:0002007	Frontal bossing	MP:0005382	craniofacial phenotype
HP:0000174	Abnormality of the palate	MP:0005382	craniofacial phenotype

Continued on next page

HPO	HPO Name	MPO	MPO Name
HP:0001863	Clinodactyly of feet	MP:0005371	limbs/digits/tail phenotype
HP:0006505	Abnormality involving the epiphyses of the limbs	MP:0005371	limbs/digits/tail phenotype
HP:0010767	Sacrococcygeal pilonidal abnormality	MP:0005390	skeleton phenotype
HP:0010936	Abnormality of the lower urinary tract	MP:0005367	renal/urinary system phenotype
HP:0010809	Broad uvula	MP:0005382	craniofacial phenotype
HP:0001212	Prominent fingertip pads	MP:0005371	limbs/digits/tail phenotype
HP:0000818	Abnormality of the endocrine system	MP:0005379	endocrine/exocrine gland phenotype
HP:0000944	Abnormality of the metaphyses	MP:0005390	skeleton phenotype
HP:0010479	Patent urachus	MP:0005367	renal/urinary system phenotype
HP:0009237	Hypoplastic/small 5th finger	MP:0005371	limbs/digits/tail phenotype
HP:0002939	Lordosis	MP:0005390	skeleton phenotype
HP:0010311	Aplasia/Hypoplasia of the breasts	MP:0005379	endocrine/exocrine gland phenotype
HP:0000954	Transverse palmar creases	MP:0010771	integument phenotype
HP:0003196	Nasal hypoplasia	MP:0005382	craniofacial phenotype
HP:0007477	Abnormal dermatoglyphics	MP:0010771	integument phenotype
HP:0007370	Aplasia/Hypoplasia of the corpus callosum	MP:0005386	behavior/neurological phenotype

Continued on next page

HPO	HPO Name	MPO	MPO Name
		MP:0003631	nervous system phenotype
HP:0011014	Abnormal glucose homeostasis	MP:0005376	homeostasis/metabolism phenotype
HP:0010674	Abnormality of the curvature of the vertebral column	MP:0005390	skeleton phenotype
HP:0100491	Abnormality of the joints of the lower limbs	MP:0005371	limbs/digits/tail phenotype
HP:0010576	Cystic malformations affecting the central nervous system	MP:0005386	behavior/neurological phenotype
		MP:0003631	nervous system phenotype
HP:0005288	Abnormality of the nares	MP:0005382	craniofacial phenotype
HP:0000383	Abnormality of periauricular region	MP:0005377	hearing/vestibular/ear phenotype
HP:0100840	Aplasia/Hypoplasia of the eyebrows	MP:0005391	vision/eye phenotype
HP:0002036	Hiatus hernia	MP:0005378	growth/size phenotype
HP:0001007	Hirsutism	MP:0010771	integument phenotype
HP:0000340	Sloping forehead	MP:0005382	craniofacial phenotype
HP:0010747	Medial flaring of the eyebrow	MP:0005391	vision/eye phenotype
HP:0200005	Abnormal shape of the palpebral fissures	MP:0005391	vision/eye phenotype
HP:0002011	Abnormality of the central nervous system	MP:0005386	behavior/neurological phenotype
		MP:0003631	nervous system phenotype
HP:0001537	Umbilical hernia	MP:0005380	embryogenesis phenotype

Continued on next page

HPO	HPO Name	MPO	MPO Name
HP:0002236	Frontal hair upsweep	MP:0010771	integument phenotype
HP:0000414	Bulbous nose	MP:0005382	craniofacial phenotype
HP:0001638	Cardiomyopathy	MP:0005369	muscle phenotype
HP:0000178	Abnormality of lower lip	MP:0005382	craniofacial phenotype
HP:0001574	Abnormality of the integument	MP:0010771	integument phenotype
HP:0000436	Abnormality of the nasal tip	MP:0005382	craniofacial phenotype
HP:0000040	Enlarged penis	MP:0005367	renal/urinary system phenotype
HP:0003712	Muscle hypertrophy	MP:0005369	muscle phenotype
HP:0001671	Abnormality of the cardiac septa	MP:0005385	cardiovascular system phenotype
HP:0000998	Hypertrichosis	MP:0010771	integument phenotype
HP:0000276	Long face	MP:0005382	craniofacial phenotype
HP:0000050	Hypoplastic genitalia	MP:0005389	reproductive system phenotype
HP:0001799	Short nails	MP:0010771	integument phenotype
HP:0000443	Bulbous nasal tip	MP:0005382	craniofacial phenotype
HP:0001844	Abnormality of the hallux	MP:0005371	limbs/digits/tail phenotype
HP:0000786	Primary amenorrhea	MP:0005389	reproductive system phenotype
HP:0000204	Cleft lip	MP:0005382	craniofacial phenotype
HP:0011004	Abnormality of the systemic arterial tree	MP:0005385	cardiovascular system phenotype
HP:0010282	Thin lower lip vermilion	MP:0005382	craniofacial phenotype
HP:0009890	High anterior hairline	MP:0010771	integument phenotype

Continued on next page

HPO	HPO Name	MPO	MPO Name
HP:0008386	Aplasia/Hypoplasia of the nails	MP:0010771	integument phenotype
HP:0003498	Short stature, disproportionate	MP:0005378	growth/size phenotype
HP:0002311	Incoordination	MP:0005386	behavior/neurological phenotype
		MP:0003631	nervous system phenotype
HP:0001769	Broad feet	MP:0005371	limbs/digits/tail phenotype
HP:0001641	Abnormality of the pulmonary valve	MP:0005385	cardiovascular system phenotype
HP:0009484	Deviation of the hand or of fingers of the hand	MP:0005371	limbs/digits/tail phenotype
HP:0007364	Aplasia/Hypoplasia of the cerebrum	MP:0005386	behavior/neurological phenotype
		MP:0003631	nervous system phenotype
HP:0002808	Kyphosis	MP:0005390	skeleton phenotype
HP:0000256	Macrocephaly	MP:0005382	craniofacial phenotype
HP:0000347	Micrognathia	MP:0005382	craniofacial phenotype
HP:0001763	Pes planus	MP:0005371	limbs/digits/tail phenotype
HP:0000286	Epicanthus	MP:0005391	vision/eye phenotype
HP:0009381	Hypoplastic/small fingers	MP:0005371	limbs/digits/tail phenotype
HP:0000962	Hyperkeratosis	MP:0010771	integument phenotype
HP:0000175	Cleft palate	MP:0005382	craniofacial phenotype
HP:0001385	Hip dysplasia	MP:0005390	skeleton phenotype
HP:z999017			

Continued on next page

HPO	HPO Name	MPO	MPO Name
HP:0010461	Abnormality of the male genitalia	MP:0005389	reproductive system phenotype
HP:0001169	Broad hands	MP:0005371	limbs/digits/tail phenotype
HP:0002119	Ventriculomegaly	MP:0005386	behavior/neurological phenotype
		MP:0003631	nervous system phenotype
HP:0006494	Aplasia/Hypoplasia involving bones of the feet	MP:0005371	limbs/digits/tail phenotype
HP:0100547	Abnormality of the forebrain	MP:0005386	behavior/neurological phenotype
		MP:0003631	nervous system phenotype
HP:0001679	Abnormality of the aorta	MP:0005385	cardiovascular system phenotype
HP:0002263	Exaggerated cupid's bow	MP:0005382	craniofacial phenotype
HP:0000752	Hyperactivity	MP:0005386	behavior/neurological phenotype
		MP:0003631	nervous system phenotype
HP:0010808	Protruding tongue	MP:0005382	craniofacial phenotype
HP:0100276	Skin pits	MP:0010771	integument phenotype
HP:0002269	Neuronal migration disorder	MP:0005384	cellular phenotype
HP:0000518	Cataract	MP:0005391	vision/eye phenotype
HP:0000767	Pectus excavatum	MP:0005390	skeleton phenotype
HP:0004299	Hernia of the abdominal wall	MP:0005378	growth/size phenotype
HP:0010993	Abnormality of the cerebral subcortex	MP:0005386	behavior/neurological phenotype

Continued on next page

HPO	HPO Name	MPO	MPO Name
		MP:0003631	nervous system phenotype
HP:0001172	Abnormality of the thumb	MP:0005371	limbs/digits/tail phenotype
HP:0007766	Hypoplastic optic disks	MP:0005391	vision/eye phenotype
HP:0003366	Abnormality of the femoral neck and head region	MP:0005390	skeleton phenotype
HP:0000464	Abnormality of the neck	MP:0005382	craniofacial phenotype
HP:0000271	Abnormality of the face	MP:0005382	craniofacial phenotype
HP:0100277	Periauricular skin pits	MP:0010771	integument phenotype
HP:0000736	Short attention span	MP:0005386	behavior/neurological phenotype
		MP:0003631	nervous system phenotype
HP:0004207	Abnormality of the 5th finger	MP:0005371	limbs/digits/tail phenotype
HP:0001513	Obesity	MP:0005378	growth/size phenotype
HP:0000283	Broad face	MP:0005382	craniofacial phenotype
HP:0001943	Hypoglycemia	MP:0005376	homeostasis/metabolism phenotype
HP:0000520	Proptosis	MP:0005391	vision/eye phenotype
HP:0100538	Abnormality of the supraorbital ridges	MP:0005391	vision/eye phenotype
HP:0009773	Symphalangism affecting the phalanges of the hand	MP:0005371	limbs/digits/tail phenotype
HP:0000664	Synophrys	MP:0010771	integument phenotype
HP:0001659	Aortic insufficiency	MP:0005385	cardiovascular system phenotype

Continued on next page

HPO	HPO Name	MPO	MPO Name
HP:0008050	Abnormality of the palpebral fissures	MP:0005391	vision/eye phenotype
HP:0000717	Autism	MP:0005386	behavior/neurological phenotype
		MP:0003631	nervous system phenotype
HP:0000002	Abnormality of body height	MP:0005378	growth/size phenotype
HP:0000769	Abnormality of the breast	MP:0005379	endocrine/exocrine gland phenotype
HP:0000431	Broad nasal bridge	MP:0005382	craniofacial phenotype
HP:0010490	Abnormality of the palmar creases	MP:0010771	integument phenotype
HP:0010766	Ectopic calcifications	MP:0005390	skeleton phenotype
HP:0001646	Abnormality of the aortic valve	MP:0005385	cardiovascular system phenotype
HP:0002817	Abnormality of the upper limb	MP:0005371	limbs/digits/tail phenotype
HP:0100543	Cognitive impairment	MP:0005386	behavior/neurological phenotype
		MP:0003631	nervous system phenotype
HP:0002086	Abnormality of the respiratory system	MP:0005388	respiratory system phenotype
HP:0001630	Abnormality of the atrial septum	MP:0005385	cardiovascular system phenotype
HP:0005287	Elevated nasal bridge	MP:0005382	craniofacial phenotype
HP:0000036	Abnormality of the penis	MP:0005389	reproductive system phenotype
HP:0000475	Broad neck	MP:0005382	craniofacial phenotype

Continued on next page

HPO	HPO Name	MPO	MPO Name
HP:0000407	Sensorineural hearing impairment	MP:0005377	hearing/vestibular/ear phenotype
HP:0100022	Abnormality of movement	MP:0005386	behavior/neurological phenotype
		MP:0003631	nervous system phenotype
HP:0000587	Abnormality of the optic nerve	MP:0005391	vision/eye phenotype
HP:0002219	Facial hypertrichosis	MP:0010771	integument phenotype
HP:0000153	Abnormality of the mouth	MP:0005382	craniofacial phenotype
HP:0001760	Abnormality of the feet	MP:0005371	limbs/digits/tail phenotype
HP:0009904	Prominent ear helix	MP:0005382	craniofacial phenotype
HP:0006610	Wide intermamillary distance	MP:0005379	endocrine/exocrine gland phenotype
HP:0002020	Gastroesophageal reflux	MP:0005369	muscle phenotype
HP:0001373	Joint dislocation	MP:0005390	skeleton phenotype
HP:0001520	Macrosomia	MP:0005378	growth/size phenotype
HP:0001600	Abnormality of the larynx	MP:0005388	respiratory system phenotype
HP:0000442	High nasal bridge	MP:0005382	craniofacial phenotype
HP:0011039	Abnormality of the helix	MP:0005382	craniofacial phenotype
HP:0004279	Hypoplastic hand	MP:0005371	limbs/digits/tail phenotype
HP:0000384	Preauricular skin tag	MP:0010771	integument phenotype
HP:0001270	Motor delay	MP:0005369	muscle phenotype
HP:0000687	Widely spaced teeth	MP:0005382	craniofacial phenotype
HP:0011121	Abnormality of skin morphology	MP:0010771	integument phenotype

Continued on next page

HPO	HPO Name	MPO	MPO Name
HP:0001000	Abnormality of skin pigmentation	MP:0001186	pigmentation phenotype
HP:0000168	Abnormality of the gingiva	MP:0005382	craniofacial phenotype
HP:0004325	Decreased body weight	MP:0005378	growth/size phenotype
HP:0000684	Delayed eruption of teeth	MP:0005382	craniofacial phenotype
HP:0000428	Low nasal bridge	MP:0005382	craniofacial phenotype
HP:0000277	Abnormality of the mandible	MP:0005382	craniofacial phenotype
HP:0000398	Dysplastic ears	MP:0005382	craniofacial phenotype
HP:0010651	Abnormality of the meninges	MP:0005386	behavior/neurological phenotype
		MP:0003631	nervous system phenotype
HP:0000370	Abnormality of the middle ear	MP:0005377	hearing/vestibular/ear phenotype
HP:0001263	Developmental delay	MP:0005386	behavior/neurological phenotype
		MP:0003631	nervous system phenotype
HP:0010864	Intellectual disability, severe	MP:0005386	behavior/neurological phenotype
		MP:0003631	nervous system phenotype
HP:0010218	Congenital vertical talus	MP:0005371	limbs/digits/tail phenotype
HP:0003319	Abnormality of the cervical spine	MP:0005390	skeleton phenotype
HP:0000280	Coarse facial features	MP:0005382	craniofacial phenotype
HP:0000152	Abnormality of head and neck	MP:0005382	craniofacial phenotype

Continued on next page

HPO	HPO Name	MPO	MPO Name
HP:0000929	Abnormality of the skull	MP:0005382	craniofacial phenotype
HP:0010978	Abnormality of immune system physiology	MP:0005387	immune system phenotype
HP:0000951	Abnormality of the skin	MP:0010771	integument phenotype
HP:0006824	Cranial nerve paralysis	MP:0005386	behavior/neurological phenotype
		MP:0003631	nervous system phenotype
HP:0003368	Abnormality of the femoral head	MP:0005390	skeleton phenotype
HP:0008551	Hypoplasia of the external ear	MP:0005382	craniofacial phenotype
HP:0009115	Aplasia/Hypoplasia involving the skeleton	MP:0005390	skeleton phenotype
HP:0002648	Abnormality of skull shape	MP:0005382	craniofacial phenotype
HP:0002435	Meningocele	MP:0005386	behavior/neurological phenotype
		MP:0003631	nervous system phenotype
HP:0000427	Upturned nose	MP:0005382	craniofacial phenotype
HP:0100561	Spinal cord lesions	MP:0005386	behavior/neurological phenotype
		MP:0003631	nervous system phenotype
HP:0000505	Impaired vision	MP:0005391	vision/eye phenotype
HP:0001637	Abnormality of the myocardium	MP:0005385	cardiovascular system phenotype
HP:0010725	Asymmetry in eye size	MP:0005391	vision/eye phenotype
HP:0008065	Aplasia/Hypoplasia of the skin	MP:0010771	integument phenotype

Continued on next page

HPO	HPO Name	MPO	MPO Name
HP:0000163	Abnormality of the oral cavity	MP:0005382	craniofacial phenotype
HP:0000138	Ovarian cysts	MP:0005389	reproductive system phenotype
HP:0009933	Naris, narrow	MP:0005382	craniofacial phenotype
HP:0002827	Dislocated hips	MP:0005390	skeleton phenotype
HP:0000614	Abnormality of the lacrimal duct	MP:0005391	vision/eye phenotype
HP:0009794	Branchial anomaly	MP:0005382	craniofacial phenotype
HP:0001792	Nail hypoplasia	MP:0010771	integument phenotype
HP:0000316	Hypertelorism	MP:0005391	vision/eye phenotype
HP:0000486	Strabismus	MP:0005391	vision/eye phenotype
HP:0001010	Hypopigmentation of the skin	MP:0001186	pigmentation phenotype
HP:0009929	Abnormality of the columella	MP:0005382	craniofacial phenotype
HP:0002410	Aqueductal stenosis	MP:0005386	behavior/neurological phenotype
		MP:0003631	nervous system phenotype
HP:0000412	Prominent ears	MP:0005382	craniofacial phenotype
HP:0000629	Periorbital fullness	MP:0005391	vision/eye phenotype
HP:0100688	Decreased corneal diameter	MP:0005391	vision/eye phenotype
HP:0009891	Hypoplasia of the supraorbital ridges	MP:0005391	vision/eye phenotype
HP:0100851	Abnormal emotion/affect behaviour	MP:0005386	behavior/neurological phenotype
		MP:0003631	nervous system phenotype

Continued on next page

HPO	HPO Name	MPO	MPO Name
HP:0000322	Short philtrum	MP:0005382	craniofacial phenotype
HP:0000293	Full cheeks	MP:0005382	craniofacial phenotype
HP:0004322	Short stature	MP:0005378	growth/size phenotype
HP:0002977	Aplasia/Hypoplasia involving the central nervous system	MP:0005386	behavior/neurological phenotype
		MP:0003631	nervous system phenotype
HP:0002564	Cardiac malformation	MP:0005385	cardiovascular system phenotype
HP:0010460	Abnormality of the female genitalia	MP:0005389	reproductive system phenotype
HP:0006482	Abnormality of dental morphology	MP:0005382	craniofacial phenotype
HP:0000269	Prominent occiput	MP:0005382	craniofacial phenotype
HP:0010300	Abnormally low-pitched voice	MP:0005388	respiratory system phenotype
HP:0004298	Abnormality of the abdominal wall	MP:0005381	digestive/alimentary phenotype
HP:0000160	Small mouth	MP:0005382	craniofacial phenotype
HP:0000154	Wide mouth	MP:0005382	craniofacial phenotype
HP:0011138	Abnormality of skin adnexa	MP:0010771	integument phenotype
HP:0001250	Seizures	MP:0005386	behavior/neurological phenotype
		MP:0003631	nervous system phenotype
HP:0009142	Duplication of bones involving the upper extremities	MP:0005371	limbs/digits/tail phenotype

Continued on next page

HPO	HPO Name	MPO	MPO Name
HP:0000768	Pectus carinatum	MP:0005390	skeleton phenotype
HP:0010938	Abnormality of the external nose	MP:0005382	craniofacial phenotype
HP:0010055	Broad hallux	MP:0005371	limbs/digits/tail phenotype
HP:z999021			
HP:0003468	Abnormality of the vertebrae	MP:0005390	skeleton phenotype
HP:0006709	Aplasia/Hypoplasia of the nipples	MP:0005379	endocrine/exocrine gland phenotype
HP:0100560	Upper limb asymmetry	MP:0005378	growth/size phenotype
HP:0004701	Hypoplasia of the toes	MP:0005371	limbs/digits/tail phenotype
HP:0000307	Pointed chin	MP:0005382	craniofacial phenotype
HP:0001939	Abnormality of metabolism/homeostasis	MP:0005376	homeostasis/metabolism phenotype
HP:0000357	Abnormal location of ears	MP:0005377	hearing/vestibular/ear phenotype
HP:0200031	macules	MP:0001186	pigmentation phenotype
HP:0003037	Enlarged joints	MP:0005390	skeleton phenotype
HP:0001713	Abnormality of the cardiac ventricle	MP:0005385	cardiovascular system phenotype
HP:0000812	Abnormal internal genitalia	MP:0005389	reproductive system phenotype
HP:0007281	Developmental arrest	MP:0005386	behavior/neurological phenotype
		MP:0003631	nervous system phenotype
HP:0000539	Abnormality of refraction	MP:0005391	vision/eye phenotype

Continued on next page

HPO	HPO Name	MPO	MPO Name
HP:0000820	Abnormality of the thyroid gland	MP:0005379	endocrine/exocrine gland phenotype
HP:0009826	Hypoplasia involving bones of the extremities	MP:0005371	limbs/digits/tail phenotype
HP:0005922	Abnormal hand morphology	MP:0005371	limbs/digits/tail phenotype
HP:0001211	Abnormality of the fingertips	MP:0005371	limbs/digits/tail phenotype
HP:0010718	Abnormality of habitus	MP:0005378	growth/size phenotype
HP:0009116	Aplasia/Hypoplasia involving bones of the skull	MP:0005382	craniofacial phenotype
HP:0000479	Abnormality of the retina	MP:0005391	vision/eye phenotype
HP:0000525	Abnormality of the iris	MP:0005391	vision/eye phenotype
HP:0008365	Abnormality of the talus	MP:0005371	limbs/digits/tail phenotype
HP:0005011	Mesomelia of the upper limbs	MP:0005371	limbs/digits/tail phenotype
HP:0100742	Vascular neoplasia	MP:0005385	cardiovascular system phenotype
HP:0000035	Abnormality of the testis	MP:0005389	reproductive system phenotype
HP:0000718	Aggressive behavior	MP:0005386	behavior/neurological phenotype
		MP:0003631	nervous system phenotype
HP:0000403	Recurrent otitis media	MP:0005377	hearing/vestibular/ear phenotype

Continued on next page

HPO	HPO Name	MPO	MPO Name
HP:0000404	Deafness	MP:0005377	hearing/vestibular/ear phenotype
HP:0000980	Pallor	MP:0010771	integument phenotype
HP:0002715	Abnormality of the immune system	MP:0005387	immune system phenotype
HP:0000350	Small forehead	MP:0005382	craniofacial phenotype
HP:0002558	Supernumerary nipples	MP:0005389	reproductive system phenotype
HP:0000496	Abnormality of eye movement	MP:0005391	vision/eye phenotype
HP:0003121	Limb contractures	MP:0005369	muscle phenotype
HP:0004329	Abnormality of the posterior segment of the eye	MP:0005391	vision/eye phenotype
HP:0000411	Protruding ears	MP:0005382	craniofacial phenotype
HP:0002118	Abnormality of the cerebral ventricles	MP:0005386	behavior/neurological phenotype
		MP:0003631	nervous system phenotype
HP:0000601	Hypotelorism	MP:0005391	vision/eye phenotype
HP:0003764	Abnormal or excess nevi	MP:0010771	integument phenotype
HP:0100714	Abnormality of the long tubular bones	MP:0005390	skeleton phenotype
HP:0009738	Abnormality of the antihelix	MP:0005382	craniofacial phenotype
HP:0100704	Cortical visual impairment	MP:0005391	vision/eye phenotype
HP:0001507	Growth abnormality	MP:0005378	growth/size phenotype
HP:0009651	Broad phalanges of the thumb	MP:0005371	limbs/digits/tail phenotype

Continued on next page

HPO	HPO Name	MPO	MPO Name
HP:0005618	Asymmetric leg shortening	MP:0005371	limbs/digits/tail phenotype
HP:0000232	Everted lower lip vermilion	MP:0005382	craniofacial phenotype
HP:0008058	Aplasia/Hypoplasia of the optic nerve	MP:0005391	vision/eye phenotype
HP:0000795	Abnormality of the urethra	MP:0005367	renal/urinary system phenotype
HP:0009901	Crumpled ear helices	MP:0005382	craniofacial phenotype
HP:0011061	Abnormality of dental structure	MP:0005382	craniofacial phenotype
HP:0010653	Abnormality of the falx cerebri	MP:0005386	behavior/neurological phenotype
		MP:0003631	nervous system phenotype
HP:0001371	Contractures	MP:0005369	muscle phenotype
HP:0008871	Height less than 3rd percentile	MP:0005378	growth/size phenotype
HP:0000977	Soft skin	MP:0010771	integument phenotype
HP:0005743	Avascular necrosis of the capital femoral epiphysis	MP:0005390	skeleton phenotype
HP:0000425	Flattened nasal bridge	MP:0005382	craniofacial phenotype
HP:0000268	Dolichocephaly	MP:0005382	craniofacial phenotype
HP:0000137	Abnormality of the ovary	MP:0005389	reproductive system phenotype
HP:0006483	Abnormal number of teeth	MP:0005382	craniofacial phenotype
HP:0001274	Agenesis of corpus callosum	MP:0005386	behavior/neurological phenotype
		MP:0003631	nervous system phenotype

Continued on next page

HPO	HPO Name	MPO	MPO Name
HP:0100807	Long fingers	MP:0005371	limbs/digits/tail phenotype
HP:0000073	Ureteral duplication	MP:0005367	renal/urinary system phenotype
HP:0010109	Hypoplastic/small hallux	MP:0005371	limbs/digits/tail phenotype
HP:0009602	Abnormality of the phalanges of the thumb	MP:0005371	limbs/digits/tail phenotype
HP:0010478	Abnormality of the urachus	MP:0005367	renal/urinary system phenotype
HP:0010450	Esophageal stenosis	MP:0005381	digestive/alimentary phenotype
HP:0000480	Retinal coloboma	MP:0005391	vision/eye phenotype
HP:0000953	Hyperpigmentation of the skin	MP:0001186	pigmentation phenotype
HP:0000599	Abnormality of the frontal hairline	MP:0010771	integument phenotype
HP:0000014	Abnormality of the bladder	MP:0005367	renal/urinary system phenotype
HP:0000209	Abnormality of the jaws	MP:0005382	craniofacial phenotype
HP:0000369	Low-set ears	MP:0005377	hearing/vestibular/ear phenotype
HP:0006262	Aplasia/Hypoplasia of the 5th finger	MP:0005371	limbs/digits/tail phenotype
HP:0001098	Abnormality of the fundus	MP:0005391	vision/eye phenotype
HP:0001159	Syndactyly	MP:0005371	limbs/digits/tail phenotype
HP:0000162	Glossoptosis	MP:0005382	craniofacial phenotype

Continued on next page

HPO	HPO Name	MPO	MPO Name
HP:0005107	Abnormality of the sacrum	MP:0005390	skeleton phenotype
HP:0000184	Prominent lips	MP:0005382	craniofacial phenotype
HP:0000278	Retrognathia	MP:0005382	craniofacial phenotype
HP:0000034	Hydrozele testis	MP:0005389	reproductive system phenotype
HP:0002167	Neurological speech impairment	MP:0005386	behavior/neurological phenotype
		MP:0003631	nervous system phenotype
HP:0002087	Abnormality of the upper respiratory tract	MP:0005388	respiratory system phenotype
HP:0001611	Nasal speech	MP:0005388	respiratory system phenotype
HP:0002644	Abnormality of the pelvis	MP:0005390	skeleton phenotype
HP:0008053	Aplasia/Hypoplasia of the iris	MP:0005391	vision/eye phenotype
HP:0003272	Abnormality of the hip	MP:0005390	skeleton phenotype
HP:0000692	Misalignment of teeth	MP:0005382	craniofacial phenotype
HP:0010611	Abnormal feet morphology	MP:0005371	limbs/digits/tail phenotype
HP:0009803	Hypoplastic/small phalanges of the hand	MP:0005371	limbs/digits/tail phenotype
HP:0010720	Abnormal hair growth pattern	MP:0010771	integument phenotype
HP:0002948	Vertebral fusion	MP:0005390	skeleton phenotype
HP:0009810	Abnormality of the joints of the upper limbs	MP:0005390	skeleton phenotype
HP:0000534	Abnormality of the eyebrow	MP:0005391	vision/eye phenotype

Continued on next page

HPO	HPO Name	MPO	MPO Name
HP:0001903	Anemia	MP:0005397	hematopoietic system phenotype
HP:0001256	Intellectual disability, mild	MP:0005386	behavior/neurological phenotype
		MP:0003631	nervous system phenotype
HP:0010719	Abnormality of hair texture	MP:0010771	integument phenotype
HP:0001798	Anonychia	MP:0010771	integument phenotype
HP:0000811	Abnormal external genitalia	MP:0005389	reproductive system phenotype
HP:0004426	Abnormality of the cheeks	MP:0005382	craniofacial phenotype
HP:0004334	Dermal atrophy	MP:0010771	integument phenotype
HP:0000419	Abnormality of the nasal septum	MP:0005382	craniofacial phenotype
HP:0000924	Abnormality of the musculoskeletal system	MP:0005390	skeleton phenotype
HP:0010935	Abnormality of the upper urinary tract	MP:0005367	renal/urinary system phenotype
HP:0000553	Abnormality of the uvea	MP:0005391	vision/eye phenotype
HP:0001009	Telangiectasia	MP:0005369	muscle phenotype
HP:0000172	Abnormality of the uvula	MP:0005382	craniofacial phenotype
HP:0009887	Abnormality of hair pigmentation	MP:0010771	integument phenotype
HP:0001249	Intellectual disability	MP:0005386	behavior/neurological phenotype
		MP:0003631	nervous system phenotype
HP:z999015			
HP:0000243	Trigonocephaly	MP:0005382	craniofacial phenotype

Continued on next page

HPO	HPO Name	MPO	MPO Name
HP:0008771	Aplasia/Hypoplasia of the ear	MP:0005377	hearing/vestibular/ear phenotype
HP:0000159	Lip abnormality	MP:0005382	craniofacial phenotype
HP:0011015	Abnormality of blood glucose concentration	MP:0005376	homeostasis/metabolism phenotype
HP:0100278	Periauricular skin tag	MP:0005377	hearing/vestibular/ear phenotype
HP:0001629	Ventricular septal defect	MP:0005385	cardiovascular system phenotype
HP:0100559	Lower limb asymmetry	MP:0005371	limbs/digits/tail phenotype
HP:0001877	Abnormality of erythrocytes	MP:0005397	hematopoietic system phenotype
HP:0009473	Joint contractures involving the joints of the hand	MP:0005369	muscle phenotype
HP:0001174	Short broad hands	MP:0005371	limbs/digits/tail phenotype
HP:0010438	Abnormality of the ventricular septum	MP:0005385	cardiovascular system phenotype
HP:0001238	Slender fingers	MP:0005371	limbs/digits/tail phenotype
HP:0000465	Webbed neck	MP:0005382	craniofacial phenotype
HP:0000489	Abnormality of globe location or size	MP:0005391	vision/eye phenotype
HP:0001337	Tremor	MP:0005386	behavior/neurological phenotype
		MP:0003631	nervous system phenotype

Continued on next page

HPO	HPO Name	MPO	MPO Name
HP:0008373	Puberty and gonadal disorders	MP:0005379	endocrine/exocrine gland phenotype
HP:0002088	Abnormality of the lung	MP:0005388	respiratory system phenotype
HP:0100360	Contractures of the joints of the upper limbs	MP:0005369	muscle phenotype
HP:0001999	Facial dysmorphism	MP:0005382	craniofacial phenotype
HP:0001547	Abnormality of the morphology or size of the rib cage	MP:0005390	skeleton phenotype
HP:0004691	2-3 toe syndactyly	MP:0005371	limbs/digits/tail phenotype
HP:0000526	Aniridia	MP:0005391	vision/eye phenotype
HP:0001162	Postaxial polydactyly (hands)	MP:0005371	limbs/digits/tail phenotype
HP:0002678	Skull asymmetry	MP:0005382	craniofacial phenotype
HP:0000405	Conductive hearing impairment	MP:0005377	hearing/vestibular/ear phenotype
HP:0000272	Malar hypoplasia	MP:0005382	craniofacial phenotype
HP:0009924	Aplasia/Hypoplasia involving the nose	MP:0005382	craniofacial phenotype
HP:0100240	Synostosis of joints	MP:0005390	skeleton phenotype
HP:0000382	Large, prominent ears	MP:0005382	craniofacial phenotype
HP:0000359	Abnormality of the inner ear	MP:0005377	hearing/vestibular/ear phenotype
HP:0001166	Arachnodactyly	MP:0005371	limbs/digits/tail phenotype

Continued on next page

HPO	HPO Name	MPO	MPO Name
HP:0005914	Aplasia/Hypoplasia involving the metacarpal bones	MP:0005371	limbs/digits/tail phenotype
HP:0000023	Inguinal hernia	MP:0005378	growth/size phenotype
HP:0000157	Abnormality of the tongue	MP:0005382	craniofacial phenotype
HP:0005306	Capillary hemangiomas	MP:0002006	tumorigenesis
HP:0002115	Sparse or absent hair	MP:0010771	integument phenotype
HP:0004324	Increased body weight	MP:0005378	growth/size phenotype
HP:0008562	Poorly formed pinnae	MP:0005382	craniofacial phenotype
HP:0001367	Abnormality of the joints	MP:0005390	skeleton phenotype
HP:0010652	Abnormality of the dura mater	MP:0005386	behavior/neurological phenotype
		MP:0003631	nervous system phenotype
HP:0001257	Spasticity	MP:0005369	muscle phenotype
HP:0200006	Slanting of the palpebral fissures	MP:0005391	vision/eye phenotype
HP:0008362	Aplasia/Hypoplasia of the hallux	MP:0005371	limbs/digits/tail phenotype
HP:0005930	Abnormality of the epiphyses	MP:0005390	skeleton phenotype
HP:0005557	Abnormality of the zygomatic arch	MP:0005382	craniofacial phenotype
HP:0001764	Small feet	MP:0005371	limbs/digits/tail phenotype
HP:0000363	Abnormality of ear lobes	MP:0005382	craniofacial phenotype
HP:0001871	Abnormality of the hematopoietic system	MP:0000001	mammalian phenotype
HP:0001510	Growth delay	MP:0005378	growth/size phenotype

Continued on next page

HPO	HPO Name	MPO	MPO Name
HP:0003307	Hyperlordosis	MP:0005390	skeleton phenotype
HP:0005607	Abnormality of the tracheobronchial system	MP:0005388	respiratory system phenotype
HP:0005562	Multiple renal cysts	MP:0005367	renal/urinary system phenotype
HP:0009824	Hypoplasia involving bones of the upper limbs	MP:0005371	limbs/digits/tail phenotype
HP:0005585	Spotty hyperpigmentation	MP:0001186	pigmentation phenotype
HP:0002554	Thin eyebrows	MP:0005391	vision/eye phenotype
HP:0002342	Intellectual disability, moderate	MP:0005386	behavior/neurological phenotype
		MP:0003631	nervous system phenotype
HP:0002438	Cerebellar malformation	MP:0005386	behavior/neurological phenotype
		MP:0003631	nervous system phenotype
HP:0008057	Aplasia/Hypoplasia affecting the fundus	MP:0005391	vision/eye phenotype
HP:0004411	Deviated nasal septum	MP:0005382	craniofacial phenotype
HP:0002683	Abnormality of the calvarium	MP:0005382	craniofacial phenotype
HP:0000032	Abnormality of male external genitalia	MP:0005389	reproductive system phenotype
HP:0001384	Abnormality of the hip joint	MP:0005390	skeleton phenotype
HP:0003396	Syringomyelia	MP:0005386	behavior/neurological phenotype
		MP:0003631	nervous system phenotype
HP:0010609	Skin tags	MP:0010771	integument phenotype

Continued on next page

HPO	HPO Name	MPO	MPO Name
HP:0000574	Thick eyebrows	MP:0005391	vision/eye phenotype
HP:0002973	Abnormality of the forearm	MP:0005371	limbs/digits/tail phenotype
HP:0010439	Atrioventricular septal defect	MP:0005385	cardiovascular system phenotype
HP:0002815	Abnormality of the knees	MP:0005371	limbs/digits/tail phenotype
HP:0005105	Abnormal nasal morphology	MP:0005382	craniofacial phenotype
HP:0000252	Microcephaly	MP:0005382	craniofacial phenotype
HP:0008772	Aplasia/Hypoplasia of the external ear	MP:0005377	hearing/vestibular/ear phenotype
HP:0009466	Radial deviation of fingers	MP:0005371	limbs/digits/tail phenotype
HP:0000028	Cryptorchidism	MP:0005389	reproductive system phenotype
HP:0001120	Abnormality of corneal size or shape	MP:0005391	vision/eye phenotype
HP:0000353	Facial muscle weakness, mild	MP:0005369	muscle phenotype
HP:0003549	Abnormality of connective tissue	MP:0005390	skeleton phenotype
HP:0009906	Aplasia/Hypoplasia of the earlobes	MP:0005382	craniofacial phenotype
HP:0000957	Cafe-au-lait spots	MP:0001186	pigmentation phenotype
HP:0004209	Clinodactyly of the 5th finger	MP:0005371	limbs/digits/tail phenotype
HP:0000319	Flat philtrum	MP:0005382	craniofacial phenotype

Continued on next page

HPO	HPO Name	MPO	MPO Name
HP:0002292	Frontal balding	MP:0010771	integument phenotype
HP:0001155	Abnormality of the hand	MP:0005371	limbs/digits/tail phenotype
HP:z999010			
HP:0000022	Abnormality of male internal genitalia	MP:0005389	reproductive system phenotype
HP:0009768	Broad phalanges of the hand	MP:0005371	limbs/digits/tail phenotype
HP:0002938	Lumbar hyperlordosis	MP:0005390	skeleton phenotype
HP:0009923	Lateral thinning of eyebrows	MP:0005391	vision/eye phenotype
HP:0000303	Mandibular prognathia	MP:0005382	craniofacial phenotype
HP:0002242	Abnormality of the intestine	MP:0005381	digestive/alimentary phenotype
HP:0000582	Upslanting palpebral fissures	MP:0005391	vision/eye phenotype
HP:0000964	Eczema	MP:0010771	integument phenotype
HP:0002308	Arnold-Chiari malformation	MP:0005386	behavior/neurological phenotype
		MP:0003631	nervous system phenotype
HP:0009815	Aplasia/Hypoplasia of the extremities	MP:0005371	limbs/digits/tail phenotype
HP:0002823	Abnormality of the femur	MP:0005371	limbs/digits/tail phenotype
HP:0000494	Downward slanting palpebral fissures	MP:0005391	vision/eye phenotype
HP:0001551	Abnormality of the umbilicus	MP:0005381	digestive/alimentary phenotype

Continued on next page

HPO	HPO Name	MPO	MPO Name
HP:0000343	Long philtrum	MP:0005382	craniofacial phenotype
HP:0003016	Metaphyseal widening	MP:0005390	skeleton phenotype
HP:0000364	Hearing abnormality	MP:0005377	hearing/vestibular/ear phenotype
HP:0001770	Toe syndactyly	MP:0005371	limbs/digits/tail phenotype
HP:0007319	Malformation of the central nervous system	MP:0005386	behavior/neurological phenotype
		MP:0003631	nervous system phenotype
HP:0000140	Abnormality of the menstrual cycle	MP:0005389	reproductive system phenotype
HP:0100323	Juvenile aseptic necrosis	MP:0005390	skeleton phenotype
HP:0000234	Abnormality of the head	MP:0005382	craniofacial phenotype
HP:0000492	Abnormality of the eyelid	MP:0005391	vision/eye phenotype
HP:0001533	Asthenic habitus	MP:0005378	growth/size phenotype
HP:0006101	Finger syndactyly	MP:0005371	limbs/digits/tail phenotype
HP:0000482	Microcornea	MP:0005391	vision/eye phenotype
HP:0009905	Thin ear helix	MP:0005382	craniofacial phenotype
HP:z999013			
HP:0011013	Abnormality of carbohydrate metabolism/homeostasis	MP:0005376	homeostasis/metabolism phenotype
HP:0010721	Abnormal hair whorl	MP:0010771	integument phenotype
HP:0001626	Abnormality of the cardiovascular system	MP:0005385	cardiovascular system phenotype
HP:0001633	Abnormality of the mitral valve	MP:0005385	cardiovascular system phenotype

Continued on next page

HPO	HPO Name	MPO	MPO Name
HP:0003043	Abnormality of the shoulder	MP:0005390	skeleton phenotype
HP:0004275	Duplication of hand bones	MP:0005371	limbs/digits/tail phenotype
HP:0000158	Macroglossia	MP:0005382	craniofacial phenotype
HP:0000219	Thin upper lip vermilion	MP:0005382	craniofacial phenotype
HP:0011024	Abnormality of the gastrointestinal tract	MP:0005381	digestive/alimentary phenotype
		MP:0005370	liver/biliary system phenotype
HP:0000315	Abnormality of the orbital region	MP:0005391	vision/eye phenotype
HP:0008069	Neoplasm of the skin	MP:0010771	integument phenotype
HP:z999020			
HP:0002997	Abnormality of the ulna	MP:0005371	limbs/digits/tail phenotype
HP:0005120	Abnormality of the cardiac atria	MP:0005385	cardiovascular system phenotype
HP:0002060	Abnormality of the cerebrum	MP:0005386	behavior/neurological phenotype
		MP:0003631	nervous system phenotype
HP:0009765	Columella, low hanging	MP:0005382	craniofacial phenotype
HP:0004456	Prominent ear lobes	MP:0005382	craniofacial phenotype
HP:0000324	Facial asymmetry	MP:0005382	craniofacial phenotype
HP:0005918	Abnormality of the phalanges of the hand	MP:0005371	limbs/digits/tail phenotype
HP:0009928	Thick nasal alae	MP:0005382	craniofacial phenotype

Continued on next page

HPO	HPO Name	MPO	MPO Name
HP:0008055	Aplasia/Hypoplasia affecting the uvea	MP:0005391	vision/eye phenotype
HP:0001161	Polydactyly (hands)	MP:0005371	limbs/digits/tail phenotype
HP:0000107	Renal cysts	MP:0005367	renal/urinary system phenotype
HP:0100679	Lack of skin elasticity	MP:0010771	integument phenotype
HP:0000164	Abnormality of the teeth	MP:0005382	craniofacial phenotype
HP:z999019			
HP:0010885	Aseptic necrosis	MP:0005390	skeleton phenotype
HP:0200007	Abnormal size of the palpebral fissures	MP:0005391	vision/eye phenotype
HP:0005556	Abnormality of the metopic suture	MP:0005382	craniofacial phenotype
HP:0009746	Thick nasal septum	MP:0005382	craniofacial phenotype
HP:0001562	Oligohydramnios	MP:0000001	mammalian phenotype
HP:0000445	Broad nose	MP:0005382	craniofacial phenotype
HP:0009804	Reduced number of teeth	MP:0005382	craniofacial phenotype
HP:0100026	Arteriovenous malformations	MP:0005385	cardiovascular system phenotype
HP:0010751	Chin dimple	MP:0005382	craniofacial phenotype
HP:0000008	Abnormality of female internal genitalia	MP:0005389	reproductive system phenotype
HP:0002714	Downturned corners of mouth	MP:0005382	craniofacial phenotype
HP:0008094	Widely spaced toes	MP:0005371	limbs/digits/tail phenotype

Continued on next page

HPO	HPO Name	MPO	MPO Name
HP:0009900	Unilateral deafness	MP:0005377	hearing/vestibular/ear phenotype
HP:0000637	Long palpebral fissures	MP:0005391	vision/eye phenotype
HP:0004404	Abnormality of the nipple	MP:0005379	endocrine/exocrine gland phenotype
HP:0000177	Abnormality of upper lip	MP:0005382	craniofacial phenotype
HP:0100713	Abnormality of the tubular bones	MP:0005390	skeleton phenotype
HP:0002538	Abnormality of the cerebral cortex	MP:0005386	behavior/neurological phenotype
		MP:0003631	nervous system phenotype
HP:0001595	Abnormality of the hair	MP:0010771	integument phenotype
HP:0002162	Low posterior hairline	MP:0005382	craniofacial phenotype
HP:0000337	Broad forehead	MP:0005382	craniofacial phenotype
HP:0002974	Radioulnar synostosis	MP:0005390	skeleton phenotype
HP:0000079	Abnormality of the urinary system	MP:0005367	renal/urinary system phenotype
HP:0001182	Tapered fingers	MP:0005371	limbs/digits/tail phenotype
HP:0001276	Hypertonia	MP:0005369	muscle phenotype
HP:0000402	Stenotic external auditory canal	MP:0005377	hearing/vestibular/ear phenotype
HP:0000545	Myopia	MP:0005391	vision/eye phenotype
HP:0000118	Phenotypic abnormality	MP:0000001	mammalian phenotype
HP:0008056	Aplasia/Hypoplasia affecting the eye	MP:0005391	vision/eye phenotype
HP:0002553	Arched eyebrows	MP:0005391	vision/eye phenotype

Continued on next page

HPO	HPO Name	MPO	MPO Name
HP:0002814	Abnormality of the lower limb	MP:0005371	limbs/digits/tail phenotype
HP:0006135	Decreased finger mobility	MP:0005371	limbs/digits/tail phenotype
HP:0010574	Abnormality of the epiphysis of the femoral head	MP:0005371	limbs/digits/tail phenotype
HP:0006265	Aplasia/Hypoplasia of fingers	MP:0005371	limbs/digits/tail phenotype
HP:0000422	Abnormality of the nasal bridge	MP:0005382	craniofacial phenotype
HP:0002164	Nail dysplasia	MP:0010771	integument phenotype
HP:0000461	Large nose	MP:0005382	craniofacial phenotype
HP:0000054	Micropenis	MP:0005367	renal/urinary system phenotype
HP:0001821	Broad nails	MP:0010771	integument phenotype
HP:0001519	Dolichostenomelia	MP:0005378	growth/size phenotype
HP:0009117	Aplasia/Hypoplasia of the maxilla	MP:0005382	craniofacial phenotype
HP:0100627	Displacement of the external urethral meatus	MP:0005367	renal/urinary system phenotype
HP:0001317	Abnormality of the cerebellum	MP:0005386	behavior/neurological phenotype
		MP:0003631	nervous system phenotype
HP:0006829	Severe muscular hypotonia	MP:0005369	muscle phenotype
HP:0010442	Polydactyly	MP:0005371	limbs/digits/tail phenotype

Continued on next page

HPO	HPO Name	MPO	MPO Name
HP:0002561	Absent nipples	MP:0005389	reproductive system phenotype
HP:0010511	Increased length of toes	MP:0005371	limbs/digits/tail phenotype
HP:0002565	Complex cardiac malformations	MP:0005385	cardiovascular system phenotype
HP:0001669	Transposition of the great arteries	MP:0005385	cardiovascular system phenotype
HP:0000306	Abnormality of the chin	MP:0005382	craniofacial phenotype
HP:0001653	Mitral regurgitation	MP:0005385	cardiovascular system phenotype
HP:0003100	Thin long bones	MP:0005371	limbs/digits/tail phenotype
HP:0002500	Abnormality of the cerebral white matter	MP:0005386	behavior/neurological phenotype
		MP:0003631	nervous system phenotype
HP:0003027	Mesomelia	MP:0005371	limbs/digits/tail phenotype
HP:0007925	Lacrimal duct aplasia or stenosis	MP:0005391	vision/eye phenotype
HP:0008873	Short stature, disproportionate short-limbed	MP:0005378	growth/size phenotype
HP:0002813	Abnormality of the extremities	MP:0005371	limbs/digits/tail phenotype
HP:0000290	Abnormality of the forehead	MP:0005382	craniofacial phenotype

Continued on next page

HPO	HPO Name	MPO	MPO Name
HP:0000235	Abnormality of the fontanelles and cranial sutures	MP:0005382	craniofacial phenotype
HP:0000707	Abnormality of the nervous system	MP:0005386	behavior/neurological phenotype
		MP:0003631	nervous system phenotype
HP:0009121	Abnormality of the axial skeleton	MP:0005390	skeleton phenotype
HP:0000341	Narrow forehead	MP:0005382	craniofacial phenotype
HP:0002597	Abnormality of the vasculature	MP:0005385	cardiovascular system phenotype

Table B.3: Significant GO enrichments after 5% FDR

HPO	HPO Name	GO term	P-value
HP0005927	Aplasia/Hypoplasia involving bones of the hand	transcription regulator activity	0.00012
HP0005927	Aplasia/Hypoplasia involving bones of the hand	nucleus	0.00012
HP0001518	Low birth weight	receptor activity	9.9E-05
HP0001518	Low birth weight	signal transduction	0.00023
HP0001518	Low birth weight	plasma membrane	0.00094
HP0001438	Abnormality of the abdomen	receptor activity	1.7E-06
HP0001438	Abnormality of the abdomen	signal transduction	4.7E-05
HP0009118	Aplasia/Hypoplasia of the mandible	nutrient reservoir activity	0.00028

Continued on next page

HPO	HPO Name	GO term	P-value
HP0000213	Thin lips	oxygen binding	1.10E-05
HP0006496	Aplasia/Hypoplasia involving bones of the upper limbs	transcription regulator activity	0.00012
HP0006496	Aplasia/Hypoplasia involving bones of the upper limbs	nucleus	0.00012
HP0000581	Blepharophimosis	oxygen binding	5.9E-05
HP0000581	Blepharophimosis	protein binding	0.00027
HP0000581	Blepharophimosis	peptidase activity	0.001
HP0001780	Abnormality of the toes	oxygen binding	0.00011
HP0009907	Adherent earlobe	signal transduction	7.3E-12
HP0009907	Adherent earlobe	plasma membrane	1.7E-11
HP0000377	Abnormality of the pinna	receptor activity	0.00015
HP0000179	Thick lower lip vermilion	oxygen binding	2.2E-07
HP0001167	Abnormality of the fingers	receptor activity	5.5E-07
HP0000504	Abnormality of vision	signal transduction	1.5E-11
HP0000504	Abnormality of vision	plasma membrane	6.3E-10
HP0004467	Preauricular pit	anatomical structure morphogenesis	2.5E-05
HP0000358	Posteriorly rotated ears	motor activity	0.00023
HP0000818	Abnormality of the endocrine system	plasma membrane	2.9E-11
HP0001769	Broad feet	anatomical structure morphogenesis	1.7E-05
HP0009484	Deviation of the hand or of fingers of the hand	oxygen binding	9.7E-06

Continued on next page

HPO	HPO Name	GO term	P-value
HP0007364	Aplasia/Hypoplasia of the cerebrum	receptor activity	3.6E-05
HP0007364	Aplasia/Hypoplasia of the cerebrum	signal transduction	0.00046
HP0002263	Exaggerated cupid's bow	oxygen binding	1.4E-07
HP0002263	Exaggerated cupid's bow	cellular homeostasis	0.00072
HP0002269	Neuronal migration disorder	anatomical structure morphogenesis	0.00019
HP0001513	Obesity	anatomical structure morphogenesis	0.00042
HP0000769	Abnormality of the breast	signal transduction	6.4E-12
HP0000769	Abnormality of the breast	plasma membrane	9.9E-10
HP0001520	Macrosomia	ion channel activity	2.0E-05
HP0001520	Macrosomia	ion transport	0.00067
HP0004279	Hypoplastic hand	transcription regulator activity	2.3E-05
HP0004279	Hypoplastic hand	nucleus	0.00026
HP0004279	Hypoplastic hand	transcription factor activity	0.00053
HP0004279	Hypoplastic hand	multicellular organismal development	0.0016
HP0004325	Decreased body weight	receptor activity	0.00022
HP0004325	Decreased body weight	signal transduction	0.00038
HP0004325	Decreased body weight	nutrient reservoir activity	0.00065
HP0000277	Abnormality of the mandible	receptor activity	1.8E-10
HP0000277	Abnormality of the mandible	signal transduction	3.9E-07

Continued on next page

HPO	HPO Name	GO term	P-value
HP0010864	Intellectual disability, severe	receptor activity	5.3E-05
HP0010864	Intellectual disability, severe	oxygen binding	0.00041
HP0009115	Aplasia/Hypoplasia involving the skeleton	signal transduction	1.9E-05
HP0009115	Aplasia/Hypoplasia involving the skeleton	nutrient reservoir activity	0.00073
HP0002648	Abnormality of skull shape	ion channel activity	0.00011
HP0002648	Abnormality of skull shape	nutrient reservoir activity	0.00036
HP0001792	Nail hypoplasia	extracellular space	0.00035
HP0000486	Strabismus	oxygen binding	0.00011
HP0001250	Seizures	receptor activity	4.8E-06
HP0001250	Seizures	plasma membrane	0.0002
HP0001250	Seizures	signal transduction	0.00042
HP0001250	Seizures	peptidase activity	0.0022
HP0000812	Abnormal internal genitalia	receptor activity	3.10E-06
HP0000812	Abnormal internal genitalia	oxygen binding	6.6E-06
HP0000812	Abnormal internal genitalia	signal transduction	5.7E-05
HP0000812	Abnormal internal genitalia	plasma membrane	0.00015
HP0000812	Abnormal internal genitalia	peptidase activity	0.0028

Continued on next page

HPO	HPO Name	GO term	P-value
HP0009826	Hypoplasia involving bones of the extremities	transcription regulator activity	1.1E-06
HP0009826	Hypoplasia involving bones of the extremities	transcription factor activity	5.2E-05
HP0009826	Hypoplasia involving bones of the extremities	multicellular organismal development	0.00024
HP0000404	Deafness	signal transduction	1.6E-10
HP0000404	Deafness	plasma membrane	5.6E-09
HP0000184	Prominent lips	oxygen binding	4.2E-07
HP0010719	Abnormality of hair texture	receptor activity	5.1E-05
HP0001249	Intellectual disability	nutrient reservoir activity	0.0012
HP0000272	Malar hypoplasia	oxygen binding	1.0E-05
HP0001367	Abnormality of the joints	receptor activity	3.2E-09
HP0001367	Abnormality of the joints	plasma membrane	1.2E-06
HP0001367	Abnormality of the joints	signal transduction	2.6E-05
HP0005557	Abnormality of the zygomatic arch	oxygen binding	1.0E-05
HP0000252	Microcephaly	receptor activity	1.10E-06
HP0000252	Microcephaly	signal transduction	6.8E-05
HP0005105	Abnormal nasal morphology	receptor activity	3.8E-06
HP0005105	Abnormal nasal morphology	signal transduction	0.00053
HP0009466	Radial deviation of fingers	oxygen binding	9.7E-06
HP0000022	Abnormality of male internal genitalia	oxygen binding	1.4E-06

Continued on next page

HPO	HPO Name	GO term	P-value
HP0000022	Abnormality of male internal genitalia	peptidase activity	0.00014
HP0000303	Mandibular prognathia	signal transduction	1.2E-10
HP0000303	Mandibular prognathia	plasma membrane	1.0E-09
HP0000364	Hearing abnormality	plasma membrane	1.0E-07
HP0000364	Hearing abnormality	signal transduction	1.2E-07
HP0000492	Abnormality of the eyelid	protein binding	0.00012
HP0000219	Thin upper lip vermilion	oxygen binding	1.10E-05
HP0000315	Abnormality of the orbital region	signal transduction	3.4E-06
HP0000315	Abnormality of the orbital region	receptor activity	4.4E-06
HP0200007	Abnormal size of the palpebral fissures	oxygen binding	6.1E-05
HP0200007	Abnormal size of the palpebral fissures	protein binding	0.00027
HP0200007	Abnormal size of the palpebral fissures	peptidase activity	0.0011
HP0002553	Arched eyebrows	signal transduction	7.3E-12
HP0002553	Arched eyebrows	plasma membrane	2.4E-11
HP0002814	Abnormality of the lower limb	receptor activity	7.4E-06
HP0002814	Abnormality of the lower limb	plasma membrane	0.00044
HP0000422	Abnormality of the nasal bridge	receptor activity	5.7E-06
HP0000422	Abnormality of the nasal bridge	signal transduction	0.00026

Continued on next page

HPO	HPO Name	GO term	P-value
HP0000422	Abnormality of the nasal bridge	oxygen binding	0.0013
HP0009121	Abnormality of the axial skeleton	nutrient reservoir activity	5.4E-05
HP0009121	Abnormality of the axial skeleton	receptor activity	0.00031
HP0009121	Abnormality of the axial skeleton	signal transduction	0.00053
HP0009121	Abnormality of the axial skeleton	ion channel activity	0.0015
HP0009122	Aplasia/Hypoplasia affecting bones of the axial skeleton	nutrient reservoir activity	0.00031
HP0000215	Thick upper lip vermilion	oxygen binding	2.3E-07
HP0000311	Round face	plasma membrane	2.9E-11
HP0000750	Impaired language development	receptor activity	2.1E-05
HP0000750	Impaired language development	signal transduction	0.00024
HP0000750	Impaired language development	nutrient reservoir activity	0.00072
HP0000119	Abnormality of the genitourinary system	receptor activity	2.2E-05
HP0000119	Abnormality of the genitourinary system	oxygen binding	4.5E-05
HP0000119	Abnormality of the genitourinary system	signal transduction	0.0002

Continued on next page

HPO	HPO Name	GO term	P-value
HP0000119	Abnormality of the genitourinary system	plasma membrane	0.00025
HP0000508	Ptosis	oxygen binding	4.7E-07
HP0000078	Abnormality of the genital system	receptor activity	1.3E-05
HP0000078	Abnormality of the genital system	oxygen binding	2.3E-05
HP0000078	Abnormality of the genital system	plasma membrane	0.0002
HP0000078	Abnormality of the genital system	signal transduction	0.00057
HP0000240	Abnormality of skull size	receptor activity	1.9E-08
HP0000240	Abnormality of skull size	signal transduction	6.3E-06
HP0009485	Radial deviation of the hand or of fingers of the hand	oxygen binding	9.7E-06
HP0002012	Abnormality of the abdominal organs	receptor activity	7.1E-07
HP0002012	Abnormality of the abdominal organs	signal transduction	7.9E-06
HP0000309	Abnormality of the midface	oxygen binding	1.6E-05
HP0000356	Abnormality of the outer ear	receptor activity	0.00043
HP0009179	Deviation of the 5th finger	oxygen binding	9.7E-06
HP0000001	All	peptidase activity	0.00052

Continued on next page

HPO	HPO Name	GO term	P-value
HP0006493	Aplasia/Hypoplasia involving bones of the lower limbs	signal transduction	4.8E-08
HP0006493	Aplasia/Hypoplasia involving bones of the lower limbs	plasma membrane	1.3E-06
HP0006493	Aplasia/Hypoplasia involving bones of the lower limbs	transcription regulator activity	0.00022
HP0002213	Fine hair	receptor activity	2.8E-05
HP0004097	Deviated fingers	oxygen binding	9.7E-06
HP0000365	Hearing impairment	plasma membrane	1.0E-07
HP0000365	Hearing impairment	signal transduction	1.2E-07
HP0002007	Frontal bossing	ion channel activity	0.00039
HP0001863	Clinodactyly of feet	oxygen binding	6.9E-07
HP0000178	Abnormality of lower lip	oxygen binding	6.6E-06
HP0000178	Abnormality of lower lip	anatomical structure morphogenesis	0.00069
HP0000347	Micrognathia	nutrient reservoir activity	0.00028
HP0010461	Abnormality of the male genitalia	oxygen binding	4.7E-06
HP0010461	Abnormality of the male genitalia	peptidase activity	0.00061
HP0006494	Aplasia/Hypoplasia involving bones of the feet	signal transduction	4.8E-08
HP0006494	Aplasia/Hypoplasia involving bones of the feet	plasma membrane	1.3E-06

Continued on next page

HPO	HPO Name	GO term	P-value
HP0006494	Aplasia/Hypoplasia involving bones of the feet	transcription regulator activity	0.00022
HP0100276	Skin pits	anatomical structure morphogenesis	2.5E-05
HP0100277	Periauricular skin pits	anatomical structure morphogenesis	2.5E-05
HP0004207	Abnormality of the 5th finger	oxygen binding	4.4E-05
HP0004207	Abnormality of the 5th finger	ion channel activity	0.00032
HP0008050	Abnormality of the palpebral fissures	protein binding	0.00035
HP0000431	Broad nasal bridge	receptor activity	2.1E-08
HP0000431	Broad nasal bridge	signal transduction	5.3E-06
HP0000431	Broad nasal bridge	oxygen binding	0.00015
HP0000431	Broad nasal bridge	plasma membrane	0.00074
HP0002817	Abnormality of the upper limb	receptor activity	7.2E-05
HP0001760	Abnormality of the feet	receptor activity	3.2E-06
HP0001760	Abnormality of the feet	plasma membrane	0.0003
HP0001263	Developmental delay	receptor activity	2.1E-05
HP0001263	Developmental delay	signal transduction	0.00024
HP0001263	Developmental delay	nutrient reservoir activity	0.00072
HP0000929	Abnormality of the skull	nutrient reservoir activity	3.6E-05
HP0000929	Abnormality of the skull	receptor activity	8.2E-05
HP0000929	Abnormality of the skull	signal transduction	0.00079
HP0000163	Abnormality of the oral cavity	receptor activity	0.00022

Continued on next page

HPO	HPO Name	GO term	P-value
HP0000163	Abnormality of the oral cavity	nutrient reservoir activity	0.00048
HP0009794	Branchial anomaly	anatomical structure morphogenesis	2.5E-05
HP0000316	Hypertelorism	signal transduction	1.4E-05
HP0000316	Hypertelorism	receptor activity	1.4E-05
HP0000316	Hypertelorism	oxygen binding	0.00076
HP0002977	Aplasia/Hypoplasia involving the central nervous system	receptor activity	3.6E-05
HP0002977	Aplasia/Hypoplasia involving the central nervous system	signal transduction	0.00046
HP0000154	Wide mouth	cytosol	4.1E-05
HP0000154	Wide mouth	cytoplasm	0.00029
HP0000154	Wide mouth	motor activity	0.00098
HP0009116	Aplasia/Hypoplasia involving bones of the skull	nutrient reservoir activity	0.00031
HP0000035	Abnormality of the testis	oxygen binding	1.4E-06
HP0000035	Abnormality of the testis	peptidase activity	0.00014
HP0004329	Abnormality of the posterior segment of the eye	transcription regulator activity	1.4E-08
HP0004329	Abnormality of the posterior segment of the eye	multicellular organismal development	2.3E-08
HP0004329	Abnormality of the posterior segment of the eye	transcription factor activity	1.0E-06

Continued on next page

HPO	HPO Name	GO term	P-value
HP0000232	Everted lower lip vermilion	anatomical structure morphogenesis	0.00023
HP0000209	Abnormality of the jaws	receptor activity	2.9E-10
HP0000209	Abnormality of the jaws	signal transduction	8.2E-07
HP0001098	Abnormality of the fundus	transcription regulator activity	1.4E-08
HP0001098	Abnormality of the fundus	multicellular organismal development	2.3E-08
HP0001098	Abnormality of the fundus	transcription factor activity	1.0E-06
HP0000811	Abnormal external genitalia	oxygen binding	6.5E-06
HP0000489	Abnormality of globe location or size	signal transduction	3.4E-06
HP0000489	Abnormality of globe location or size	receptor activity	4.4E-06
HP0001999	Facial dysmorphism	receptor activity	1.2E-06
HP0001999	Facial dysmorphism	plasma membrane	1.5E-06
HP0001999	Facial dysmorphism	oxygen binding	0.00043
HP0001999	Facial dysmorphism	signal transduction	0.00057
HP0001257	Spasticity	plasma membrane	1.3E-10
HP0200006	Slanting of the palpebral fissures	oxygen binding	0.00043
HP0001764	Small feet	signal transduction	4.10E-08
HP0001764	Small feet	plasma membrane	2.8E-06
HP0001764	Small feet	transcription regulator activity	0.00017
HP0000363	Abnormality of ear lobes	plasma membrane	2.5E-11

Continued on next page

HPO	HPO Name	GO term	P-value
HP0000032	Abnormality of male external genitalia	oxygen binding	4.7E-06
HP0000032	Abnormality of male external genitalia	peptidase activity	0.00061
HP0000028	Cryptorchidism	oxygen binding	1.3E-06
HP0000028	Cryptorchidism	peptidase activity	0.00013
HP0004209	Clinodactyly of the 5th finger	oxygen binding	9.7E-06
HP0001155	Abnormality of the hand	receptor activity	4.8E-05
HP0009815	Aplasia/Hypoplasia of the extremities	plasma membrane	3.3E-06
HP0009815	Aplasia/Hypoplasia of the extremities	signal transduction	1.2E-05
HP0009815	Aplasia/Hypoplasia of the extremities	transcription regulator activity	0.00071
HP0000494	Downward slanting palpebral fissures	oxygen binding	3.6E-06
HP0011024	Abnormality of the gastrointestinal tract	signal transduction	1.0E-10
HP0011024	Abnormality of the gastrointestinal tract	plasma membrane	1.0E-06
HP0004404	Abnormality of the nipple	signal transduction	6.4E-12
HP0004404	Abnormality of the nipple	plasma membrane	9.9E-10
HP0000337	Broad forehead	ion channel activity	0.00024
HP0001276	Hypertonia	plasma membrane	1.3E-10
HP0001182	Tapered fingers	signal transduction	2.10E-09
HP0001182	Tapered fingers	plasma membrane	6.10E-09
HP0001182	Tapered fingers	oxygen binding	9.5E-07

Continued on next page

HPO	HPO Name	GO term	P-value
HP0000118	Phenotypic abnormality	peptidase activity	0.00052
HP0000461	Large nose	plasma membrane	3.5E-12
HP0000306	Abnormality of the chin	plasma membrane	2.6E-08
HP0000306	Abnormality of the chin	signal transduction	4.8E-08

Table B.4: Significant KEGG enrichments after 5% FDR

HPO	HPO Name	KEGG term	% Enriched	P-value
HP:0000001	All	Drug metabolism - cytochrome P450	155.55	0.00001
HP:0000001	All	Glycolysis / Gluconeogenesis	132.25	0.00038
HP:0000001	All	Taste transduction	146.15	0.00039
HP:0000001	All	SUBCLASS Xenobiotics Biodegradation and Metabolism	92.59	0.00060
HP:0000002	Abnormality of body height	Olfactory transduction	89.65	0.00009
HP:0000002	Abnormality of body height	SUBCLASS Sensory System	82.51	0.00010
HP:0000002	Abnormality of body height	SUBCLASS Immune System Diseases	-73.87	0.00246
HP:0000023	Inguinal hernia	CLASS Genetic Information Processing	243.64	0.00425
HP:0000078	Abnormality of the genital system	Olfactory transduction	231	0.00000
HP:0000078	Abnormality of the genital system	CLASS Organismal Systems	30.44	0.00524
HP:0000078	Abnormality of the genital system	CLASS Metabolism	-30.81	0.01168

Continued on next page

HPO	HPO Name	KEGG term	% Enriched	P-value
HP:0000078	Abnormality of the genital system	SUBCLASS Sensory System	203.31	0.00000
HP:0000118	Phenotypic abnormality	Drug metabolism - cytochrome P450	155.55	0.00001
HP:0000118	Phenotypic abnormality	Glycolysis / Gluconeogenesis	132.25	0.00038
HP:0000118	Phenotypic abnormality	Taste transduction	146.15	0.00039
HP:0000118	Phenotypic abnormality	SUBCLASS Xenobiotics Biodegradation and Metabolism	92.59	0.00060
HP:0000119	Abnormality of the genitourinary system	Olfactory transduction	205.56	0.00000
HP:0000119	Abnormality of the genitourinary system	CLASS Organismal Systems	29.77	0.00396
HP:0000119	Abnormality of the genitourinary system	CLASS Metabolism	-33.36	0.00429
HP:0000119	Abnormality of the genitourinary system	SUBCLASS Sensory System	195.31	0.00000
HP:0000152	Abnormality of head and neck	Taste transduction	200.09	0.00006
HP:0000152	Abnormality of head and neck	SUBCLASS Translation	-75.8	0.00104
HP:0000153	Abnormality of the mouth	SUBCLASS Translation	-88.6	0.00107
HP:0000153	Abnormality of the mouth	SUBCLASS Sensory System	54.77	0.00153

Continued on next page

HPO	HPO Name	KEGG term	% Enriched	P-value
HP:0000159	Lip abnormality	Olfactory transduction	-90.66	0.00016
HP:0000159	Lip abnormality	SUBCLASS Sensory System	-83.38	0.00030
HP:0000163	Abnormality of the oral cavity	CLASS Organismal Systems	23.61	0.00100
HP:0000163	Abnormality of the oral cavity	SUBCLASS Sensory System	72.63	0.00015
HP:0000163	Abnormality of the oral cavity	SUBCLASS Translation	-100	0.00032
HP:0000164	Abnormality of the teeth	Olfactory transduction	-100	0.00013
HP:0000164	Abnormality of the teeth	Insulin signaling pathway	267.52	0.00017
HP:0000164	Abnormality of the teeth	CLASS Environmental Information Processing	41.41	0.00412
HP:0000164	Abnormality of the teeth	SUBCLASS Sensory System	-100	0.00004
HP:0000164	Abnormality of the teeth	SUBCLASS Signal Transduction	66.91	0.00135
HP:0000177	Abnormality of upper lip	Olfactory transduction	-90.4	0.00021
HP:0000177	Abnormality of upper lip	SUBCLASS Sensory System	-82.91	0.00041
HP:0000209	Abnormality of the jaws	Olfactory transduction	144.73	0.00000

Continued on next page

HPO	HPO Name	KEGG term	% Enriched	P-value
HP:0000209	Abnormality of the jaws	Taste transduction	438.46	0.00001
HP:0000209	Abnormality of the jaws	CLASS Organismal Systems	40.15	0.00003
HP:0000209	Abnormality of the jaws	SUBCLASS Sensory System	181.92	0.00000
HP:0000234	Abnormality of the head	Taste transduction	200.09	0.00006
HP:0000234	Abnormality of the head	SUBCLASS Translation	-75.8	0.00104
HP:0000240	Abnormality of skull size	Taste transduction	428.16	0.00000
HP:0000240	Abnormality of skull size	Olfactory transduction	111.8	0.00001
HP:0000240	Abnormality of skull size	CLASS Organismal Systems	35.49	0.00004
HP:0000240	Abnormality of skull size	CLASS Metabolism	-34.96	0.00029
HP:0000240	Abnormality of skull size	SUBCLASS Sensory System	151.38	0.00000
HP:0000240	Abnormality of skull size	SUBCLASS Glycan Biosynthesis and Metabolism	-89.5	0.00054
HP:0000240	Abnormality of skull size	SUBCLASS Translation	-100	0.00292
HP:0000252	Microcephaly	Olfactory transduction	181.59	0.00000

Continued on next page

HPO	HPO Name	KEGG term	% Enriched	P-value
HP:0000252	Microcephaly	CLASS Organismal Systems	35.92	0.00032
HP:0000252	Microcephaly	CLASS Metabolism	-32.06	0.00372
HP:0000252	Microcephaly	SUBCLASS Sensory System	164.58	0.00000
HP:0000256	Macrocephaly	Taste transduction	1481.73	0.00000
HP:0000271	Abnormality of the face	Taste transduction	203.08	0.00006
HP:0000271	Abnormality of the face	SUBCLASS Translation	-75.56	0.00115
HP:0000277	Abnormality of the mandible	Olfactory transduction	144.73	0.00000
HP:0000277	Abnormality of the mandible	Taste transduction	438.46	0.00001
HP:0000277	Abnormality of the mandible	CLASS Organismal Systems	40.15	0.00003
HP:0000277	Abnormality of the mandible	SUBCLASS Sensory System	181.92	0.00000
HP:0000278	Retrognathia	NOD-like receptor signaling pathway	588.4	0.00020
HP:0000280	Coarse facial features	CLASS Genetic Information Processing	272.27	0.00258
HP:0000284	Abnormality of the ocular region	Taste transduction	205.33	0.00020
HP:0000284	Abnormality of the ocular region	SUBCLASS Translation	-81.06	0.00116
HP:0000288	Abnormality of the philtrum	NOD-like receptor signaling pathway	575.36	0.00006

Continued on next page

HPO	HPO Name	KEGG term	% Enriched	P-value
HP:0000288	Abnormality of the philtrum	Chemokine signaling pathway	246.31	0.00028
HP:0000288	Abnormality of the philtrum	CLASS Environmental Information Processing	44.62	0.00692
HP:0000288	Abnormality of the philtrum	SUBCLASS Nervous System	207.47	0.00078
HP:0000290	Abnormality of the forehead	Taste transduction	767.69	0.00000
HP:0000290	Abnormality of the forehead	CLASS Metabolism	-36.4	0.00315
HP:0000303	Mandibular prognathia	Olfactory transduction	630.3	0.00000
HP:0000303	Mandibular prognathia	CLASS Organismal Systems	106.86	0.00000
HP:0000303	Mandibular prognathia	CLASS Metabolism	-49.12	0.00755
HP:0000303	Mandibular prognathia	SUBCLASS Sensory System	550.08	0.00000
HP:0000306	Abnormality of the chin	Olfactory transduction	463.83	0.00000
HP:0000306	Abnormality of the chin	CLASS Organismal Systems	82.13	0.00000
HP:0000306	Abnormality of the chin	CLASS Metabolism	-44.88	0.00484
HP:0000306	Abnormality of the chin	SUBCLASS Sensory System	401.9	-0.00000

Continued on next page

HPO	HPO Name	KEGG term	% Enriched	P-value
HP:0000311	Round face	Olfactory transduction	952.49	-0.00000
HP:0000311	Round face	CLASS Organismal Systems	132.94	0.00000
HP:0000311	Round face	CLASS Environmental Information Processing	-84.28	0.00005
HP:0000311	Round face	CLASS Human Diseases	-89.88	0.00018
HP:0000311	Round face	CLASS Metabolism	-68.34	0.00156
HP:0000311	Round face	CLASS Cellular Processes	-69.53	0.00585
HP:0000311	Round face	CLASS Genetic Information Processing	-83.45	0.01186
HP:0000311	Round face	SUBCLASS Sensory System	836.89	-0.00000
HP:0000311	Round face	SUBCLASS Signal Transduction	-100	0.00019
HP:0000315	Abnormality of the orbital region	Taste transduction	424.3	0.00000
HP:0000315	Abnormality of the orbital region	Olfactory transduction	99.47	0.00002
HP:0000315	Abnormality of the orbital region	CLASS Metabolism	-24.79	0.00623
HP:0000315	Abnormality of the orbital region	SUBCLASS Sensory System	139.95	0.00000
HP:0000316	Hypertelorism	Olfactory transduction	167.49	0.00000

Continued on next page

HPO	HPO Name	KEGG term	% Enriched	P-value
HP:0000316	Hypertelorism	SUBCLASS Sensory System	145.11	0.00000
HP:0000319	Flat philtrum	SUBCLASS Cell Communication	235.76	0.00095
HP:0000319	Flat philtrum	SUBCLASS Nervous System	334.08	0.00217
HP:0000322	Short philtrum	NOD-like receptor signaling pathway	1017.14	0.00007
HP:0000322	Short philtrum	Chemokine signaling pathway	410.36	0.00034
HP:0000341	Narrow forehead	CLASS Environmental Information Processing	203.02	0.00269
HP:0000341	Narrow forehead	SUBCLASS Signaling Molecules and Interaction	422.22	0.00078
HP:0000347	Micrognathia	Taste transduction	1214.68	0.00000
HP:0000356	Abnormality of the outer ear	Olfactory transduction	88.56	0.00019
HP:0000357	Abnormal location of ears	SUBCLASS Sensory System	-100	0.00120
HP:0000363	Abnormality of ear lobes	Olfactory transduction	850.12	0.00000
HP:0000363	Abnormality of ear lobes	CLASS Organismal Systems	129.06	0.00000
HP:0000363	Abnormality of ear lobes	CLASS Metabolism	-92.42	0.00000

Continued on next page

HPO	HPO Name	KEGG term	% Enriched	P-value
HP:0000363	Abnormality of ear lobes	CLASS Environmental Information Processing	-69.91	0.00086
HP:0000363	Abnormality of ear lobes	CLASS Cellular Processes	-70.83	0.00402
HP:0000363	Abnormality of ear lobes	CLASS Human Diseases	-51.59	0.03637
HP:0000363	Abnormality of ear lobes	SUBCLASS Sensory System	745.76	0.00000
HP:0000364	Hearing abnormality	Olfactory transduction	531.49	-0.00000
HP:0000364	Hearing abnormality	CLASS Organismal Systems	85.09	0.00000
HP:0000364	Hearing abnormality	CLASS Genetic Information Processing	-80.14	0.00150
HP:0000364	Hearing abnormality	CLASS Human Diseases	-51.46	0.00860
HP:0000364	Hearing abnormality	CLASS Metabolism	-43.01	0.01045
HP:0000364	Hearing abnormality	SUBCLASS Sensory System	462.13	-0.00000
HP:0000365	Hearing impairment	Olfactory transduction	531.49	-0.00000
HP:0000365	Hearing impairment	CLASS Organismal Systems	85.09	0.00000
HP:0000365	Hearing impairment	CLASS Genetic Information Processing	-80.14	0.00150
HP:0000365	Hearing impairment	CLASS Human Diseases	-51.46	0.00860

Continued on next page

HPO	HPO Name	KEGG term	% Enriched	P-value
HP:0000365	Hearing impairment	CLASS Metabolism	-43.01	0.01045
HP:0000365	Hearing impairment	SUBCLASS Sensory System	462.13	-0.00000
HP:0000377	Abnormality of the pinna	Olfactory transduction	104.48	0.00003
HP:0000377	Abnormality of the pinna	SUBCLASS Sensory System	82.02	0.00037
HP:0000400	Large ears	SUBCLASS Sensory System	-100	0.00131
HP:0000404	Deafness	Olfactory transduction	640.03	0.00000
HP:0000404	Deafness	CLASS Organismal Systems	99.2	0.00000
HP:0000404	Deafness	CLASS Human Diseases	-64.45	0.00224
HP:0000404	Deafness	CLASS Genetic Information Processing	-76.73	0.00561
HP:0000404	Deafness	CLASS Environmental Information Processing	-50.28	0.00564
HP:0000404	Deafness	CLASS Metabolism	-38.78	0.03080
HP:0000404	Deafness	SUBCLASS Sensory System	558.75	0.00000
HP:0000411	Protruding ears	NOD-like receptor signaling pathway	661.95	0.00003
HP:0000411	Protruding ears	SUBCLASS Sensory System	-100	0.00110

Continued on next page

HPO	HPO Name	KEGG term	% Enriched	P-value
HP:0000412	Prominent ears	NOD-like receptor signaling pathway	661.95	0.00003
HP:0000412	Prominent ears	SUBCLASS Sensory System	-100	0.00110
HP:0000422	Abnormality of the nasal bridge	Olfactory transduction	138	0.00000
HP:0000422	Abnormality of the nasal bridge	NOD-like receptor signaling pathway	283.98	0.00046
HP:0000422	Abnormality of the nasal bridge	CLASS Metabolism	-31.99	0.00195
HP:0000422	Abnormality of the nasal bridge	SUBCLASS Sensory System	130.01	0.00000
HP:0000426	Prominent nasal bridge	CLASS Environmental Information Processing	88.04	0.00026
HP:0000426	Prominent nasal bridge	SUBCLASS Signal Transduction	135.13	0.00013
HP:0000431	Broad nasal bridge	Olfactory transduction	245.93	0.00000
HP:0000431	Broad nasal bridge	NOD-like receptor signaling pathway	410.69	0.00014
HP:0000431	Broad nasal bridge	CLASS Metabolism	-38.4	0.00243
HP:0000431	Broad nasal bridge	CLASS Organismal Systems	27.81	0.01136
HP:0000431	Broad nasal bridge	SUBCLASS Sensory System	216.99	0.00000
HP:0000436	Abnormality of the nasal tip	CLASS Organismal Systems	-34.61	0.00792

Continued on next page

HPO	HPO Name	KEGG term	% Enriched	P-value
HP:0000436	Abnormality of the nasal tip	SUBCLASS Sensory System	-100	0.00034
HP:0000445	Broad nose	CLASS Environmental Information Processing	93.86	0.00164
HP:0000445	Broad nose	CLASS Metabolism	-65.53	0.01264
HP:0000445	Broad nose	CLASS Organismal Systems	-54.3	0.01611
HP:0000461	Large nose	Olfactory transduction	726.96	0.00000
HP:0000461	Large nose	CLASS Organismal Systems	115.1	0.00000
HP:0000461	Large nose	CLASS Environmental Information Processing	-60.71	0.00209
HP:0000461	Large nose	CLASS Metabolism	-53.83	0.00675
HP:0000461	Large nose	CLASS Human Diseases	-57.87	0.01218
HP:0000461	Large nose	CLASS Cellular Processes	-57.69	0.01269
HP:0000461	Large nose	CLASS Genetic Information Processing	-72.42	0.01775
HP:0000461	Large nose	SUBCLASS Sensory System	636.13	-0.00000
HP:0000478	Abnormality of the eye	SUBCLASS Translation	-82.48	0.00052
HP:0000486	Strabismus	CLASS Cellular Processes	55.94	0.00093

Continued on next page

HPO	HPO Name	KEGG term	% Enriched	P-value
HP:0000486	Strabismus	SUBCLASS Immune System	68.74	0.00123
HP:0000489	Abnormality of globe location or size	Taste transduction	424.3	0.00000
HP:0000489	Abnormality of globe location or size	Olfactory transduction	99.47	0.00002
HP:0000489	Abnormality of globe location or size	CLASS Metabolism	-24.79	0.00623
HP:0000489	Abnormality of globe location or size	SUBCLASS Sensory System	139.95	0.00000
HP:0000490	Deeply set eye	Taste transduction	1329.14	0.00000
HP:0000490	Deeply set eye	CLASS Metabolism	-53.9	0.00066
HP:0000492	Abnormality of the eyelid	Olfactory transduction	-90.82	0.00000
HP:0000492	Abnormality of the eyelid	CLASS Cellular Processes	30.11	0.00371
HP:0000492	Abnormality of the eyelid	SUBCLASS Sensory System	-79.58	0.00000
HP:0000504	Abnormality of vision	Olfactory transduction	630.3	0.00000
HP:0000504	Abnormality of vision	CLASS Metabolism	-83.04	0.00000
HP:0000504	Abnormality of vision	CLASS Organismal Systems	79.88	0.00000
HP:0000504	Abnormality of vision	CLASS Human Diseases	-71.11	0.00073
HP:0000504	Abnormality of vision	CLASS Cellular Processes	-49.22	0.02119

Continued on next page

HPO	HPO Name	KEGG term	% Enriched	P-value
HP:0000504	Abnormality of vision	SUBCLASS Sensory System	550.08	0.00000
HP:0000517	Abnormality of the lens	NOD-like receptor signaling pathway	2896.58	0.00000
HP:0000517	Abnormality of the lens	Chemokine signaling pathway	877.83	0.00001
HP:0000517	Abnormality of the lens	Cytokine-cytokine receptor interaction	595.83	0.00012
HP:0000518	Cataract	NOD-like receptor signaling pathway	2896.58	0.00000
HP:0000518	Cataract	Chemokine signaling pathway	877.83	0.00001
HP:0000518	Cataract	Cytokine-cytokine receptor interaction	595.83	0.00012
HP:0000534	Abnormality of the eyebrow	Olfactory transduction	736.53	0.00000
HP:0000534	Abnormality of the eyebrow	CLASS Organismal Systems	116.34	0.00000
HP:0000534	Abnormality of the eyebrow	CLASS Metabolism	-80.57	0.00002
HP:0000534	Abnormality of the eyebrow	CLASS Cellular Processes	-58.45	0.01079
HP:0000534	Abnormality of the eyebrow	CLASS Environmental Information Processing	-48.57	0.01297
HP:0000534	Abnormality of the eyebrow	SUBCLASS Sensory System	644.64	-0.00000

Continued on next page

HPO	HPO Name	KEGG term	% Enriched	P-value
HP:0000539	Abnormality of refraction	CLASS Environmental Information Processing	58.47	0.00504
HP:0000540	Hypermetropia	CLASS Environmental Information Processing	165.14	0.00104
HP:0000581	Blepharophimosis	Olfactory transduction	-91.54	0.00005
HP:0000692	Misalignment of teeth	CLASS Environmental Information Processing	107.95	0.00804
HP:0000707	Abnormality of the nervous system	Drug metabolism - cytochrome P450	155.3	0.00001
HP:0000707	Abnormality of the nervous system	Glycolysis / Gluconeogenesis	142.58	0.00022
HP:0000707	Abnormality of the nervous system	Taste transduction	157.09	0.00023
HP:0000707	Abnormality of the nervous system	SUBCLASS Xenobiotics Biodegradation and Metabolism	93.41	0.00074
HP:0000708	Behavioural/Psychiatric Abnormality	Drug metabolism - cytochrome P450	169.87	0.00000
HP:0000708	Behavioural/Psychiatric Abnormality	Glycolysis / Gluconeogenesis	156.41	0.00010
HP:0000708	Behavioural/Psychiatric Abnormality	Taste transduction	171.76	0.00012

Continued on next page

HPO	HPO Name	KEGG term	% Enriched	P-value
HP:0000708	Behavioural/Psychiatric Abnormality	SUBCLASS Xenobiotics Biodegradation and Metabolism	104.44	0.00031
HP:0000717	Autism	CLASS Environmental Information Processing	44.75	0.00129
HP:0000718	Aggressive behavior	CLASS Genetic Information Processing	-59.01	0.00598
HP:0000718	Aggressive behavior	CLASS Environmental Information Processing	36.22	0.01252
HP:0000736	Short attention span	Folate biosynthesis	3318.18	0.00007
HP:0000736	Short attention span	CLASS Genetic Information Processing	-100	0.00228
HP:0000750	Impaired language development	Taste transduction	306.17	0.00001
HP:0000750	Impaired language development	SUBCLASS Sensory System	89.6	0.00000
HP:0000752	Hyperactivity	Folate biosynthesis	3318.18	0.00007
HP:0000752	Hyperactivity	CLASS Genetic Information Processing	-100	0.00228
HP:0000765	Abnormality of the thorax	NOD-like receptor signaling pathway	1032.04	0.00001
HP:0000765	Abnormality of the thorax	Chemokine signaling pathway	454.1	0.00002
HP:0000765	Abnormality of the thorax	Neurotrophin signaling pathway	457.03	0.00063

Continued on next page

HPO	HPO Name	KEGG term	% Enriched	P-value
HP:0000765	Abnormality of the thorax	CLASS Environmental Information Processing	72.83	0.00260
HP:0000765	Abnormality of the thorax	CLASS Genetic Information Processing	115.09	0.00487
HP:0000767	Pectus excavatum	CLASS Genetic Information Processing	219.09	0.00661
HP:0000769	Abnormality of the breast	Olfactory transduction	586.7	0.00000
HP:0000769	Abnormality of the breast	CLASS Organismal Systems	94.51	0.00000
HP:0000769	Abnormality of the breast	CLASS Metabolism	-57.47	0.00125
HP:0000769	Abnormality of the breast	CLASS Cellular Processes	-65.89	0.00139
HP:0000769	Abnormality of the breast	CLASS Environmental Information Processing	-41.95	0.01727
HP:0000769	Abnormality of the breast	SUBCLASS Sensory System	511.27	0.00000
HP:0000811	Abnormal external genitalia	CLASS Environmental Information Processing	49.03	0.00168
HP:0000811	Abnormal external genitalia	SUBCLASS Sensory System	-88.19	0.00138
HP:0000812	Abnormal internal genitalia	Olfactory transduction	307.16	0.00000

Continued on next page

HPO	HPO Name	KEGG term	% Enriched	P-value
HP:0000812	Abnormal internal genitalia	CLASS Organismal Systems	42.91	0.00059
HP:0000812	Abnormal internal genitalia	CLASS Metabolism	-36.96	0.00688
HP:0000812	Abnormal internal genitalia	CLASS Human Diseases	-35.57	0.02374
HP:0000812	Abnormal internal genitalia	SUBCLASS Sensory System	273.09	0.00000
HP:0000818	Abnormality of the endocrine system	Olfactory transduction	1016.4	0.00000
HP:0000818	Abnormality of the endocrine system	CLASS Organismal Systems	147.9	0.00000
HP:0000818	Abnormality of the endocrine system	CLASS Environmental Information Processing	-91.16	0.00002
HP:0000818	Abnormality of the endocrine system	CLASS Cellular Processes	-100	0.00004
HP:0000818	Abnormality of the endocrine system	CLASS Human Diseases	-88.62	0.00057
HP:0000818	Abnormality of the endocrine system	CLASS Metabolism	-73.28	0.00131
HP:0000818	Abnormality of the endocrine system	CLASS Genetic Information Processing	-81.38	0.02215
HP:0000818	Abnormality of the endocrine system	SUBCLASS Sensory System	893.77	0.00000
HP:0000818	Abnormality of the endocrine system	SUBCLASS Signaling Molecules and Interaction	-100	0.00272

Continued on next page

HPO	HPO Name	KEGG term	% Enriched	P-value
HP:0000924	Abnormality of the musculoskeletal system	Taste transduction	216.96	0.00001
HP:0000929	Abnormality of the skull	Taste transduction	293.67	0.00000
HP:0000929	Abnormality of the skull	CLASS Organismal Systems	18.04	0.00554
HP:0000929	Abnormality of the skull	CLASS Metabolism	-17.88	0.01513
HP:0000929	Abnormality of the skull	SUBCLASS Sensory System	73.99	0.00003
HP:0000951	Abnormality of the skin	Taste transduction	419.13	0.00000
HP:0000951	Abnormality of the skin	Olfactory transduction	-88.43	0.00000
HP:0000980	Pallor	CLASS Environmental Information Processing	107.95	0.00804
HP:0001000	Abnormality of skin pigmentation	NOD-like receptor signaling pathway	1088.64	0.00000
HP:0001000	Abnormality of skin pigmentation	Chemokine signaling pathway	398.69	0.00005
HP:0001010	Hypopigmentation of the skin	CLASS Environmental Information Processing	107.95	0.00804
HP:0001155	Abnormality of the hand	Olfactory transduction	87.02	0.00023

Continued on next page

HPO	HPO Name	KEGG term	% Enriched	P-value
HP:0001155	Abnormality of the hand	SUBCLASS Sensory System	81.17	0.00022
HP:0001167	Abnormality of the fingers	Olfactory transduction	138.38	0.00000
HP:0001167	Abnormality of the fingers	CLASS Metabolism	-28.03	0.00694
HP:0001167	Abnormality of the fingers	SUBCLASS Sensory System	130.92	0.00000
HP:0001169	Broad hands	NOD-like receptor signaling pathway	2134.29	0.00000
HP:0001169	Broad hands	Chemokine signaling pathway	774.9	0.00003
HP:0001169	Broad hands	Cytokine-cytokine receptor interaction	522.59	0.00024
HP:0001182	Tapered fingers	Olfactory transduction	642.08	0.00000
HP:0001182	Tapered fingers	CLASS Organismal Systems	82.78	0.00000
HP:0001182	Tapered fingers	CLASS Human Diseases	-70.64	0.00089
HP:0001182	Tapered fingers	CLASS Cellular Processes	-70.51	0.00094
HP:0001182	Tapered fingers	CLASS Environmental Information Processing	-54.38	0.00322
HP:0001182	Tapered fingers	CLASS Metabolism	-42.55	0.02066
HP:0001182	Tapered fingers	SUBCLASS Sensory System	580	-0.00000

Continued on next page

HPO	HPO Name	KEGG term	% Enriched	P-value
HP:0001249	Intellectual disability	Drug metabolism - cytochrome P450	176.63	0.00000
HP:0001249	Intellectual disability	Taste transduction	191.83	0.00005
HP:0001249	Intellectual disability	Glycolysis / Gluconeogenesis	144.76	0.00050
HP:0001249	Intellectual disability	MAPK signaling pathway	62.19	0.00051
HP:0001249	Intellectual disability	Neurotrophin signaling pathway	88.18	0.00123
HP:0001249	Intellectual disability	SUBCLASS Xenobiotics Biodegradation and Metabolism	110.77	0.00027
HP:0001249	Intellectual disability	SUBCLASS Immune System Diseases	-54.44	0.00125
HP:0001250	Seizures	Olfactory transduction	172.11	0.00000
HP:0001250	Seizures	CLASS Metabolism	-36.12	0.00081
HP:0001250	Seizures	SUBCLASS Sensory System	155.31	0.00000
HP:0001252	Muscular hypotonia	Olfactory transduction	98.03	0.00004
HP:0001252	Muscular hypotonia	NOD-like receptor signaling pathway	279.64	0.00011
HP:0001252	Muscular hypotonia	SUBCLASS Sensory System	95.86	0.00002
HP:0001252	Muscular hypotonia	SUBCLASS Immune System Diseases	-81.31	0.00110

Continued on next page

HPO	HPO Name	KEGG term	% Enriched	P-value
HP:0001256	Intellectual disability, mild	SUBCLASS Sensory System	-100	0.00017
HP:0001257	Spasticity	Olfactory transduction	693.26	0.00000
HP:0001257	Spasticity	CLASS Organismal Systems	95.38	0.00000
HP:0001257	Spasticity	CLASS Environmental Information Processing	-57.33	0.00264
HP:0001257	Spasticity	CLASS Metabolism	-57.01	0.00293
HP:0001257	Spasticity	CLASS Human Diseases	-60.77	0.00629
HP:0001257	Spasticity	SUBCLASS Sensory System	606.13	0.00000
HP:0001263	Developmental delay	Taste transduction	306.17	0.00001
HP:0001263	Developmental delay	SUBCLASS Sensory System	89.6	0.00000
HP:0001273	Abnormality of the corpus callosum	CLASS Cellular Processes	105.62	0.00091
HP:0001274	Agenesis of corpus callosum	CLASS Cellular Processes	105.62	0.00091
HP:0001276	Hypertonia	Olfactory transduction	693.26	0.00000
HP:0001276	Hypertonia	CLASS Organismal Systems	95.38	0.00000
HP:0001276	Hypertonia	CLASS Environmental Information Processing	-57.33	0.00264

Continued on next page

HPO	HPO Name	KEGG term	% Enriched	P-value
HP:0001276	Hypertonia	CLASS Metabolism	-57.01	0.00293
HP:0001276	Hypertonia	CLASS Human Diseases	-60.77	0.00629
HP:0001276	Hypertonia	SUBCLASS Sensory System	606.13	0.00000
HP:0001317	Abnormality of the cerebellum	SUBCLASS Infectious Diseases	574.87	0.00071
HP:0001367	Abnormality of the joints	Olfactory transduction	214.29	0.00000
HP:0001367	Abnormality of the joints	CLASS Metabolism	-42.55	0.00030
HP:0001367	Abnormality of the joints	CLASS Organismal Systems	24.29	0.01547
HP:0001367	Abnormality of the joints	SUBCLASS Sensory System	179.77	0.00000
HP:0001367	Abnormality of the joints	SUBCLASS Immune System Diseases	-100	0.00091
HP:0001371	Contractures	CLASS Environmental Information Processing	120.95	0.00048
HP:0001371	Contractures	SUBCLASS Signaling Molecules and Interaction	296.01	0.00000
HP:0001382	Joint hypermobility	SUBCLASS Signaling Molecules and Interaction	131.51	0.00025

Continued on next page

HPO	HPO Name	KEGG term	% Enriched	P-value
HP:0001388	Joint laxity	SUBCLASS Signaling Molecules and Interaction	167.13	0.00002
HP:0001438	Abnormality of the abdomen	Olfactory transduction	172.15	0.00000
HP:0001438	Abnormality of the abdomen	CLASS Organismal Systems	31.36	0.00123
HP:0001438	Abnormality of the abdomen	SUBCLASS Sensory System	155.72	0.00000
HP:0001507	Growth abnormality	SUBCLASS Immune System Diseases	-73.72	0.00043
HP:0001507	Growth abnormality	SUBCLASS Translation	-88.34	0.00130
HP:0001510	Growth delay	SUBCLASS Immune System Diseases	-71.53	0.00114
HP:0001510	Growth delay	SUBCLASS Translation	-87.36	0.00246
HP:0001510	Growth delay	SUBCLASS Sensory System	52.9	0.00344
HP:0001518	Low birth weight	Olfactory transduction	131.03	0.00000
HP:0001518	Low birth weight	CLASS Organismal Systems	27.14	0.00243
HP:0001518	Low birth weight	SUBCLASS Sensory System	117.41	0.00000
HP:0001520	Macrosomia	CLASS Environmental Information Processing	149.54	0.00035

Continued on next page

HPO	HPO Name	KEGG term	% Enriched	P-value
HP:0001520	Macrosomia	SUBCLASS Signaling Molecules and Interaction	244.05	0.00091
HP:0001537	Umbilical hernia	NOD-like receptor signaling pathway	2581.15	0.00000
HP:0001537	Umbilical hernia	Chemokine signaling pathway	1066.53	0.00000
HP:0001537	Umbilical hernia	Cytokine-cytokine receptor interaction	522.59	0.00024
HP:0001537	Umbilical hernia	CLASS Metabolism	-100	0.00188
HP:0001537	Umbilical hernia	SUBCLASS Immune System	205.19	0.00106
HP:0001551	Abnormality of the umbilicus	NOD-like receptor signaling pathway	2581.15	0.00000
HP:0001551	Abnormality of the umbilicus	Chemokine signaling pathway	1066.53	0.00000
HP:0001551	Abnormality of the umbilicus	Cytokine-cytokine receptor interaction	522.59	0.00024
HP:0001551	Abnormality of the umbilicus	CLASS Metabolism	-100	0.00188
HP:0001551	Abnormality of the umbilicus	SUBCLASS Immune System	205.19	0.00106
HP:0001574	Abnormality of the integument	Taste transduction	419.13	0.00000
HP:0001574	Abnormality of the integument	Olfactory transduction	-88.43	0.00000
HP:0001595	Abnormality of the hair	Taste transduction	1077.1	0.00000

Continued on next page

HPO	HPO Name	KEGG term	% Enriched	P-value
HP:0001595	Abnormality of the hair	CLASS Metabolism	-50.3	0.00142
HP:0001595	Abnormality of the hair	CLASS Environmental Information Processing	39.76	0.01532
HP:0001597	Abnormality of the nail	NOD-like receptor signaling pathway	763.42	0.00005
HP:0001597	Abnormality of the nail	Cytokine-cytokine receptor interaction	267.57	0.00014
HP:0001629	Ventricular septal defect	Folate biosynthesis	3581.11	0.00005
HP:0001629	Ventricular septal defect	Chronic myeloid leukemia	799.82	0.00020
HP:0001629	Ventricular septal defect	mTOR signaling pathway	938.26	0.00053
HP:0001669	Transposition of the great arteries	Folate biosynthesis	4250.41	0.00003
HP:0001669	Transposition of the great arteries	mTOR signaling pathway	1127.03	0.00028
HP:0001671	Abnormality of the cardiac septa	Folate biosynthesis	2954.54	0.00010
HP:0001671	Abnormality of the cardiac septa	Chronic myeloid leukemia	646.66	0.00049
HP:0001713	Abnormality of the cardiac ventricle	Folate biosynthesis	3581.11	0.00005
HP:0001713	Abnormality of the cardiac ventricle	Chronic myeloid leukemia	799.82	0.00020

Continued on next page

HPO	HPO Name	KEGG term	% Enriched	P-value
HP:0001713	Abnormality of the cardiac ventricle	mTOR signaling pathway	938.26	0.00053
HP:0001760	Abnormality of the feet	Olfactory transduction	160.51	0.00000
HP:0001760	Abnormality of the feet	SUBCLASS Sensory System	138.33	0.00000
HP:0001763	Pes planus	Folate biosynthesis	2558.58	0.00015
HP:0001764	Small feet	Olfactory transduction	539.01	0.00000
HP:0001764	Small feet	CLASS Organismal Systems	61.33	0.00012
HP:0001764	Small feet	CLASS Metabolism	-60.42	0.00042
HP:0001764	Small feet	SUBCLASS Sensory System	468.82	0.00000
HP:0001780	Abnormality of the toes	CLASS Organismal Systems	-35.85	0.00317
HP:0001780	Abnormality of the toes	SUBCLASS Sensory System	-88.63	0.00100
HP:0001792	Nail hypoplasia	NOD-like receptor signaling pathway	1829.61	0.00000
HP:0001792	Nail hypoplasia	Chemokine signaling pathway	655.59	0.00009
HP:0001792	Nail hypoplasia	Cytokine-cytokine receptor interaction	437.69	0.00060
HP:0001852	Gap between first and second toes	CLASS Environmental Information Processing	104.67	0.00643

Continued on next page

HPO	HPO Name	KEGG term	% Enriched	P-value
HP:0001956	Truncal obesity	CLASS Environmental Information Processing	253.52	0.00638
HP:0001999	Facial dysmorphism	Olfactory transduction	216.89	0.00000
HP:0001999	Facial dysmorphism	CLASS Organismal Systems	32.69	0.00159
HP:0001999	Facial dysmorphism	SUBCLASS Sensory System	220.2	0.00000
HP:0002007	Frontal bossing	Taste transduction	1035.29	0.00000
HP:0002007	Frontal bossing	CLASS Metabolism	-50.07	0.00040
HP:0002011	Abnormality of the central nervous system	Drug metabolism - cytochrome P450	157.17	0.00002
HP:0002011	Abnormality of the central nervous system	Taste transduction	171.3	0.00012
HP:0002011	Abnormality of the central nervous system	SUBCLASS Xenobiotics Biodegradation and Metabolism	95.94	0.00080
HP:0002011	Abnormality of the central nervous system	SUBCLASS Translation	-65.82	0.00218
HP:0002012	Abnormality of the abdominal organs	Olfactory transduction	197.04	0.00000
HP:0002012	Abnormality of the abdominal organs	CLASS Organismal Systems	38.2	0.00021

Continued on next page

HPO	HPO Name	KEGG term	% Enriched	P-value
HP:0002012	Abnormality of the abdominal organs	CLASS Genetic Information Processing	-59.14	0.00069
HP:0002012	Abnormality of the abdominal organs	SUBCLASS Sensory System	179.1	0.00000
HP:0002019	Constipation	NOD-like receptor signaling pathway	1598.06	0.00000
HP:0002019	Constipation	Chemokine signaling pathway	564.92	0.00020
HP:0002060	Abnormality of the cerebrum	Olfactory transduction	114.6	0.00000
HP:0002060	Abnormality of the cerebrum	CLASS Organismal Systems	31.04	0.00032
HP:0002060	Abnormality of the cerebrum	SUBCLASS Sensory System	101.64	0.00001
HP:0002213	Fine hair	Taste transduction	2967.59	0.00000
HP:0002213	Fine hair	CLASS Metabolism	-78.41	0.00174
HP:0002213	Fine hair	CLASS Organismal Systems	71.7	0.00257
HP:0002213	Fine hair	SUBCLASS Sensory System	265.02	0.00022
HP:0002242	Abnormality of the intestine	NOD-like receptor signaling pathway	1543.28	0.00000
HP:0002242	Abnormality of the intestine	Chemokine signaling pathway	525.6	0.00008
HP:0002263	Exaggerated cupid's bow	NOD-like receptor signaling pathway	1269.4	0.00002
HP:0002263	Exaggerated cupid's bow	Chemokine signaling pathway	525.6	0.00008

Continued on next page

HPO	HPO Name	KEGG term	% Enriched	P-value
HP:0002342	Intellectual disability, moderate	Taste transduction	1063.57	0.00000
HP:0002500	Abnormality of the cerebral white matter	CLASS Cellular Processes	105.62	0.00091
HP:0002553	Arched eyebrows	Olfactory transduction	830.33	0.00000
HP:0002553	Arched eyebrows	CLASS Organismal Systems	130.19	0.00000
HP:0002553	Arched eyebrows	CLASS Metabolism	-77.74	0.00015
HP:0002553	Arched eyebrows	CLASS Environmental Information Processing	-70.53	0.00066
HP:0002553	Arched eyebrows	CLASS Cellular Processes	-80.96	0.00070
HP:0002553	Arched eyebrows	CLASS Human Diseases	-52.6	0.03128
HP:0002553	Arched eyebrows	SUBCLASS Sensory System	728.14	0.00000
HP:0002565	Complex cardiac malformations	Folate biosynthesis	4250.41	0.00003
HP:0002565	Complex cardiac malformations	mTOR signaling pathway	1127.03	0.00028
HP:0002648	Abnormality of skull shape	Taste transduction	645.91	0.00000
HP:0002715	Abnormality of the immune system	NOD-like receptor signaling pathway	961.29	0.00000
HP:0002719	Recurrent infections	NOD-like receptor signaling pathway	961.29	0.00000

Continued on next page

HPO	HPO Name	KEGG term	% Enriched	P-value
HP:0002814	Abnormality of the lower limb	Olfactory transduction	124.49	0.00000
HP:0002814	Abnormality of the lower limb	SUBCLASS Sensory System	105.38	0.00001
HP:0002817	Abnormality of the upper limb	SUBCLASS Sensory System	78.99	0.00029
HP:0002977	Aplasia/Hypoplasia involving the central nervous system	Olfactory transduction	149.82	0.00000
HP:0002977	Aplasia/Hypoplasia involving the central nervous system	CLASS Organismal Systems	33.66	0.00033
HP:0002977	Aplasia/Hypoplasia involving the central nervous system	CLASS Metabolism	-26.94	0.00887
HP:0002977	Aplasia/Hypoplasia involving the central nervous system	SUBCLASS Sensory System	134.73	0.00000
HP:0003011	Abnormality of musculature	SUBCLASS Immune System Diseases	-83.58	0.00028
HP:0003011	Abnormality of musculature	SUBCLASS Sensory System	72.08	0.00038
HP:0003121	Limb contractures	SUBCLASS Signaling Molecules and Interaction	244.05	0.00091
HP:0003549	Abnormality of connective tissue	Chemokine signaling pathway	432.79	0.00001

Continued on next page

HPO	HPO Name	KEGG term	% Enriched	P-value
HP:0003549	Abnormality of connective tissue	NOD-like receptor signaling pathway	879.65	0.00002
HP:0003549	Abnormality of connective tissue	Cytokine-cytokine receptor interaction	279.14	0.00022
HP:0003549	Abnormality of connective tissue	CLASS Environmental Information Processing	97.15	0.00002
HP:0003549	Abnormality of connective tissue	CLASS Metabolism	-72.6	0.00026
HP:0003549	Abnormality of connective tissue	CLASS Genetic Information Processing	100.45	0.00705
HP:0003549	Abnormality of connective tissue	SUBCLASS Signaling Molecules and Interaction	223.37	0.00000
HP:0004207	Abnormality of the 5th finger	CLASS Environmental Information Processing	43.74	0.00701
HP:0004298	Abnormality of the abdominal wall	NOD-like receptor signaling pathway	1719.35	0.00000
HP:0004298	Abnormality of the abdominal wall	Chemokine signaling pathway	691.57	0.00000
HP:0004298	Abnormality of the abdominal wall	CLASS Metabolism	-74.56	0.00672
HP:0004298	Abnormality of the abdominal wall	CLASS Genetic Information Processing	139.32	0.00863
HP:0004298	Abnormality of the abdominal wall	CLASS Environmental Information Processing	76.76	0.01205

Continued on next page

HPO	HPO Name	KEGG term	% Enriched	P-value
HP:0004299	Hernia of the abdominal wall	NOD-like receptor signaling pathway	1719.35	0.00000
HP:0004299	Hernia of the abdominal wall	Chemokine signaling pathway	691.57	0.00000
HP:0004299	Hernia of the abdominal wall	CLASS Metabolism	-74.56	0.00672
HP:0004299	Hernia of the abdominal wall	CLASS Genetic Information Processing	139.32	0.00863
HP:0004299	Hernia of the abdominal wall	CLASS Environmental Information Processing	76.76	0.01205
HP:0004322	Short stature	Olfactory transduction	95.58	0.00004
HP:0004322	Short stature	SUBCLASS Sensory System	88.21	0.00005
HP:0004323	Abnormality of body weight	SUBCLASS Immune System Diseases	-82.78	0.00048
HP:0004323	Abnormality of body weight	SUBCLASS Sensory System	71.43	0.00058
HP:0004323	Abnormality of body weight	SUBCLASS Infectious Diseases	-86.85	0.00335
HP:0004324	Increased body weight	CLASS Environmental Information Processing	71.4	0.00036
HP:0004325	Decreased body weight	Olfactory transduction	111.8	0.00001
HP:0004325	Decreased body weight	SUBCLASS Sensory System	99.01	0.00002

Continued on next page

HPO	HPO Name	KEGG term	% Enriched	P-value
HP:0004325	Decreased body weight	SUBCLASS Immune System Diseases	-80.01	0.00203
HP:0004328	Abnormality of the anterior segment of the eye	NOD-like receptor signaling pathway	2581.15	0.00000
HP:0004328	Abnormality of the anterior segment of the eye	Chemokine signaling pathway	774.9	0.00003
HP:0004328	Abnormality of the anterior segment of the eye	Cytokine-cytokine receptor interaction	522.59	0.00024
HP:0004404	Abnormality of the nipple	Olfactory transduction	586.7	0.00000
HP:0004404	Abnormality of the nipple	CLASS Organismal Systems	94.51	0.00000
HP:0004404	Abnormality of the nipple	CLASS Metabolism	-57.47	0.00125
HP:0004404	Abnormality of the nipple	CLASS Cellular Processes	-65.89	0.00139
HP:0004404	Abnormality of the nipple	CLASS Environmental Information Processing	-41.95	0.01727
HP:0004404	Abnormality of the nipple	SUBCLASS Sensory System	511.27	0.00000
HP:0005105	Abnormal nasal morphology	Olfactory transduction	156.01	0.00000
HP:0005105	Abnormal nasal morphology	SUBCLASS Sensory System	127.89	0.00000

Continued on next page

HPO	HPO Name	KEGG term	% Enriched	P-value
HP:0005918	Abnormality of the phalanges of the hand	CLASS Environmental Information Processing	67.77	0.00128
HP:0005918	Abnormality of the phalanges of the hand	SUBCLASS Signaling Molecules and Interaction	123.05	0.00060
HP:0005927	Aplasia/Hypoplasia involving bones of the hand	CLASS Organismal Systems	-53.87	0.00492
HP:0005927	Aplasia/Hypoplasia involving bones of the hand	CLASS Environmental Information Processing	64.43	0.00810
HP:0005927	Aplasia/Hypoplasia involving bones of the hand	CLASS Genetic Information Processing	90.46	0.02343
HP:0006261	Abnormality of phalangeal joints of the hand	SUBCLASS Signaling Molecules and Interaction	244.05	0.00091
HP:0006482	Abnormality of dental morphology	Folate biosynthesis	10943.35	0.00000
HP:0006493	Aplasia/Hypoplasia involving bones of the lower limbs	Olfactory transduction	530.26	0.00000
HP:0006493	Aplasia/Hypoplasia involving bones of the lower limbs	CLASS Organismal Systems	59.12	0.00019

Continued on next page

HPO	HPO Name	KEGG term	% Enriched	P-value
HP:0006493	Aplasia/Hypoplasia involving bones of the lower limbs	CLASS Metabolism	-60.96	0.00033
HP:0006493	Aplasia/Hypoplasia involving bones of the lower limbs	SUBCLASS Sensory System	461.03	0.00000
HP:0006494	Aplasia/Hypoplasia involving bones of the feet	Olfactory transduction	530.26	0.00000
HP:0006494	Aplasia/Hypoplasia involving bones of the feet	CLASS Organismal Systems	59.12	0.00019
HP:0006494	Aplasia/Hypoplasia involving bones of the feet	CLASS Metabolism	-60.96	0.00033
HP:0006494	Aplasia/Hypoplasia involving bones of the feet	SUBCLASS Sensory System	461.03	0.00000
HP:0006496	Aplasia/Hypoplasia involving bones of the upper limbs	CLASS Organismal Systems	-53.87	0.00492
HP:0006496	Aplasia/Hypoplasia involving bones of the upper limbs	CLASS Environmental Information Processing	64.43	0.00810
HP:0006496	Aplasia/Hypoplasia involving bones of the upper limbs	CLASS Genetic Information Processing	90.46	0.02343

Continued on next page

HPO	HPO Name	KEGG term	% Enriched	P-value
HP:0006829	Severe muscular hypotonia	SUBCLASS Transport and Catabolism	92.43	0.00100
HP:0007018	Attention deficit hyperactivity disorder	Folate biosynthesis	3318.18	0.00007
HP:0007018	Attention deficit hyperactivity disorder	CLASS Genetic Information Processing	-100	0.00228
HP:0007364	Aplasia/Hypoplasia of the cerebrum	Olfactory transduction	149.82	0.00000
HP:0007364	Aplasia/Hypoplasia of the cerebrum	CLASS Organismal Systems	33.66	0.00033
HP:0007364	Aplasia/Hypoplasia of the cerebrum	CLASS Metabolism	-26.94	0.00887
HP:0007364	Aplasia/Hypoplasia of the cerebrum	SUBCLASS Sensory System	134.73	0.00000
HP:0007370	Aplasia/Hypoplasia of the corpus callosum	CLASS Cellular Processes	105.62	0.00091
HP:0008050	Abnormality of the palpebral fissures	Olfactory transduction	-89.13	0.00000
HP:0008050	Abnormality of the palpebral fissures	CLASS Cellular Processes	30.29	0.00709
HP:0008050	Abnormality of the palpebral fissures	SUBCLASS Sensory System	-75.81	0.00001
HP:0008050	Abnormality of the palpebral fissures	SUBCLASS Cancers	70.22	0.00214
HP:0008050	Abnormality of the palpebral fissures	SUBCLASS Cell Motility	95.74	0.00281

Continued on next page

HPO	HPO Name	KEGG term	% Enriched	P-value
HP:0008050	Abnormality of the palpebral fissures	SUBCLASS Nervous System	87.69	0.00564
HP:0008386	Aplasia/Hypoplasia of the nails	NOD-like receptor signaling pathway	1269.4	0.00002
HP:0008871	Height less than 3rd percentile	Olfactory transduction	101.89	0.00002
HP:0008871	Height less than 3rd percentile	SUBCLASS Sensory System	94.28	0.00002
HP:0008871	Height less than 3rd percentile	SUBCLASS Immune System Diseases	-81.46	0.00101
HP:0009115	Aplasia/Hypoplasia involving the skeleton	Olfactory transduction	184	0.00000
HP:0009115	Aplasia/Hypoplasia involving the skeleton	Taste transduction	524.88	0.00000
HP:0009115	Aplasia/Hypoplasia involving the skeleton	CLASS Organismal Systems	34.66	0.00075
HP:0009115	Aplasia/Hypoplasia involving the skeleton	CLASS Cellular Processes	-32.3	0.01444
HP:0009115	Aplasia/Hypoplasia involving the skeleton	SUBCLASS Sensory System	227.16	0.00000
HP:0009116	Aplasia/Hypoplasia involving bones of the skull	Taste transduction	1214.68	0.00000

Continued on next page

HPO	HPO Name	KEGG term	% Enriched	P-value
HP:0009118	Aplasia/Hypoplasia of the mandible	Taste transduction	1214.68	0.00000
HP:0009121	Abnormality of the axial skeleton	Taste transduction	271	0.00001
HP:0009121	Abnormality of the axial skeleton	CLASS Organismal Systems	19.4	0.00236
HP:0009121	Abnormality of the axial skeleton	CLASS Metabolism	-20.75	0.00440
HP:0009121	Abnormality of the axial skeleton	SUBCLASS Sensory System	63.97	0.00018
HP:0009122	Aplasia/Hypoplasia affecting bones of the axial skeleton	Taste transduction	1214.68	0.00000
HP:0009473	Joint contractures involving the joints of the hand	SUBCLASS Signaling Molecules and Interaction	244.05	0.00091
HP:0009810	Abnormality of the joints of the upper limbs	CLASS Environmental Information Processing	112.11	0.00027
HP:0009810	Abnormality of the joints of the upper limbs	CLASS Metabolism	-76.25	0.00393
HP:0009810	Abnormality of the joints of the upper limbs	CLASS Genetic Information Processing	123.36	0.01397
HP:0009810	Abnormality of the joints of the upper limbs	SUBCLASS Signaling Molecules and Interaction	216.81	0.00006

Continued on next page

HPO	HPO Name	KEGG term	% Enriched	P-value
HP:0009815	Aplasia/Hypoplasia of the extremities	Olfactory transduction	405.59	0.00000
HP:0009815	Aplasia/Hypoplasia of the extremities	CLASS Metabolism	-56.94	0.00019
HP:0009815	Aplasia/Hypoplasia of the extremities	CLASS Organismal Systems	52.55	0.00019
HP:0009815	Aplasia/Hypoplasia of the extremities	SUBCLASS Sensory System	350.06	0.00000
HP:0009826	Hypoplasia involving bones of the extremities	CLASS Environmental Information Processing	152.51	0.00099
HP:0009907	Adherent earlobe	Olfactory transduction	850.12	0.00000
HP:0009907	Adherent earlobe	CLASS Organismal Systems	129.06	0.00000
HP:0009907	Adherent earlobe	CLASS Metabolism	-92.42	0.00000
HP:0009907	Adherent earlobe	CLASS Environmental Information Processing	-69.91	0.00086
HP:0009907	Adherent earlobe	CLASS Cellular Processes	-70.83	0.00402
HP:0009907	Adherent earlobe	CLASS Human Diseases	-51.59	0.03637
HP:0009907	Adherent earlobe	SUBCLASS Sensory System	745.76	0.00000
HP:0009929	Abnormality of the columella	Folate biosynthesis	15851.51	0.00000

Continued on next page

HPO	HPO Name	KEGG term	% Enriched	P-value
HP:0010438	Abnormality of the ventricular septum	Folate biosynthesis	3581.11	0.00005
HP:0010438	Abnormality of the ventricular septum	Chronic myeloid leukemia	799.82	0.00020
HP:0010438	Abnormality of the ventricular septum	mTOR signaling pathway	938.26	0.00053
HP:0010719	Abnormality of hair texture	Taste transduction	2877.37	0.00000
HP:0010719	Abnormality of hair texture	CLASS Metabolism	-79.04	0.00132
HP:0010719	Abnormality of hair texture	CLASS Organismal Systems	66.65	0.00419
HP:0010719	Abnormality of hair texture	SUBCLASS Sensory System	254.28	0.00029
HP:0010720	Abnormal hair growth pattern	CLASS Environmental Information Processing	84.12	0.00040
HP:0010720	Abnormal hair growth pattern	CLASS Genetic Information Processing	101.65	0.00879
HP:0010720	Abnormal hair growth pattern	CLASS Organismal Systems	-46.87	0.00950
HP:0010864	Intellectual disability, severe	Olfactory transduction	126.61	0.00000
HP:0010864	Intellectual disability, severe	SUBCLASS Sensory System	119.01	0.00000
HP:0010864	Intellectual disability, severe	SUBCLASS Immune System Diseases	-89	0.00082

Continued on next page

HPO	HPO Name	KEGG term	% Enriched	P-value
HP:0010866	Abdominal wall defect	NOD-like receptor signaling pathway	1719.35	0.00000
HP:0010866	Abdominal wall defect	Chemokine signaling pathway	691.57	0.00000
HP:0010866	Abdominal wall defect	CLASS Metabolism	-74.56	0.00672
HP:0010866	Abdominal wall defect	CLASS Genetic Information Processing	139.32	0.00863
HP:0010866	Abdominal wall defect	CLASS Environmental Information Processing	76.76	0.01205
HP:0010938	Abnormality of the external nose	Folate biosynthesis	1593.96	0.00005
HP:0010938	Abnormality of the external nose	Olfactory transduction	-100	0.00015
HP:0010938	Abnormality of the external nose	SUBCLASS Sensory System	-100	0.00004
HP:0010978	Abnormality of immune system physiology	NOD-like receptor signaling pathway	961.29	0.00000
HP:0010993	Abnormality of the cerebral subcortex	CLASS Cellular Processes	105.62	0.00091
HP:0011024	Abnormality of the gastrointestinal tract	Olfactory transduction	539.01	0.00000
HP:0011024	Abnormality of the gastrointestinal tract	NOD-like receptor signaling pathway	725.44	0.00001
HP:0011024	Abnormality of the gastrointestinal tract	CLASS Organismal Systems	81	0.00000

Continued on next page

HPO	HPO Name	KEGG term	% Enriched	P-value
HP:0011024	Abnormality of the gastrointestinal tract	CLASS Metabolism	-55.48	0.00126
HP:0011024	Abnormality of the gastrointestinal tract	CLASS Human Diseases	-55.76	0.00527
HP:0011024	Abnormality of the gastrointestinal tract	SUBCLASS Sensory System	468.82	0.00000
HP:0011121	Abnormality of skin morphology	NOD-like receptor signaling pathway	924.69	0.00000
HP:0011121	Abnormality of skin morphology	Chemokine signaling pathway	329.9	0.00019
HP:0011121	Abnormality of skin morphology	CLASS Environmental Information Processing	58.47	0.00504
HP:0011138	Abnormality of skin adnexa	Taste transduction	644.34	0.00000
HP:0011138	Abnormality of skin adnexa	Olfactory transduction	-100	0.00002
HP:0100022	Abnormality of movement	CLASS Genetic Information Processing	-88.36	0.00102
HP:0100360	Contractures of the joints of the upper limbs	SUBCLASS Signaling Molecules and Interaction	244.05	0.00091
HP:0100490	Camptodactyly (hands)	SUBCLASS Signaling Molecules and Interaction	244.05	0.00091
HP:0100491	Abnormality of the joints of the lower limbs	CLASS Genetic Information Processing	-100	0.00305

Continued on next page

HPO	HPO Name	KEGG term	% Enriched	P-value
HP:0100543	Cognitive impairment	Drug metabolism - cytochrome P450	167.47	0.00001
HP:0100543	Cognitive impairment	Taste transduction	182.17	0.00007
HP:0100543	Cognitive impairment	Glycolysis / Gluconeogenesis	136.66	0.00073
HP:0100543	Cognitive impairment	SUBCLASS Xenobiotics Biodegradation and Metabolism	103.79	0.00045
HP:0100547	Abnormality of the forebrain	Olfactory transduction	114.6	0.00000
HP:0100547	Abnormality of the forebrain	CLASS Organismal Systems	31.04	0.00032
HP:0100547	Abnormality of the forebrain	SUBCLASS Sensory System	101.64	0.00001
HP:0100656	Thoracoabdominal wall defects	NOD-like receptor signaling pathway	1719.35	0.00000
HP:0100656	Thoracoabdominal wall defects	Chemokine signaling pathway	691.57	0.00000
HP:0100656	Thoracoabdominal wall defects	CLASS Metabolism	-74.56	0.00672
HP:0100656	Thoracoabdominal wall defects	CLASS Genetic Information Processing	139.32	0.00863
HP:0100656	Thoracoabdominal wall defects	CLASS Environmental Information Processing	76.76	0.01205
HP:0100790	Herniae	NOD-like receptor signaling pathway	1719.35	0.00000

Continued on next page

HPO	HPO Name	KEGG term	% Enriched	P-value
HP:0100790	Herniae	Chemokine signaling pathway	691.57	0.00000
HP:0100790	Herniae	CLASS Metabolism	-74.56	0.00672
HP:0100790	Herniae	CLASS Genetic Information Processing	139.32	0.00863
HP:0100790	Herniae	CLASS Environmental Information Processing	76.76	0.01205
HP:0100851	Abnormal emotion/affect behaviour	CLASS Environmental Information Processing	43.56	0.00010
HP:0200007	Abnormal size of the palpebral fissures	Olfactory transduction	-91.54	0.00005
HP:0200045	Abnormality of pigmentation	NOD-like receptor signaling pathway	1065.33	0.00000
HP:0200045	Abnormality of pigmentation	Chemokine signaling pathway	388.91	0.00007
HP:0200045	Abnormality of pigmentation	CLASS Environmental Information Processing	59.43	0.00743
HP:z999010		Folate biosynthesis	3144.37	0.00000
HP:z999016		Taste transduction	361.75	0.00000
HP:z999016		Olfactory transduction	-81	0.00000
HP:z999017		Olfactory transduction	-90.63	0.00000
HP:z999017		Taste transduction	320.33	0.00001

Table B.5: Significant co-expression in BrainSpan

HPO	HPO Name	No. Genes in Network	P-value
HP0005927	Aplasia/Hypoplasia involving bones of the hand	126	0.002
HP0000598	Abnormality of the ear	697	0.001
HP0001597	Abnormality of the nail	93	0.047
HP0009553	Abnormality of the hairline	67	0.001
HP0001018	Abnormal palmar dermatoglyphics	153	0.001
HP0006316	Irregularly spaced teeth	18	0.004
HP0000284	Abnormality of the ocular region	1158	0.001
HP0000098	Tall stature	42	0.003
HP0003011	Abnormality of musculature	759	0.001
HP0000366	Abnormality of the nose	851	0.001
HP0001518	Low birth weight	477	0.001
HP0001438	Abnormality of the abdomen	389	0.001
HP0001654	Abnormality of the heart valves	39	0.032
HP0000396	Overfolded helix	22	0.002
HP0009118	Aplasia/Hypoplasia of the mandible	161	0.015
HP0000288	Abnormality of the philtrum	217	0.02
HP0000400	Large ears	200	0.001

Continued on next page

HPO	HPO Name	No. Genes in Network	P-value
HP0000213	Thin lips	252	0.001
HP0006496	Aplasia/Hypoplasia involving bones of the upper limbs	126	0.003
HP0004323	Abnormality of body weight	657	0.001
HP0001252	Muscular hypotonia	646	0.002
HP0000606	Abnormality of the periorbital region	882	0.001
HP0000202	Cleft lip/palate	68	0.001
HP0100490	Camptodactyly (hands)	57	0.001
HP0000478	Abnormality of the eye	1301	0.001
HP0000581	Blepharophimosis	413	0.001
HP0001780	Abnormality of the toes	322	0.001
HP0009907	Adherent earlobe	48	0.05
HP0000377	Abnormality of the pinna	555	0.001
HP0000179	Thick lower lip vermilion	112	0.001
HP0000670	Carious teeth	31	0.011
HP0001167	Abnormality of the fingers	510	0.001
HP0000504	Abnormality of vision	109	0.001
HP0004467	Preauricular pit	30	0.001
HP0006500	Abnormality involving the epiphyses of the lower limbs	73	0.019
HP0000430	Hypoplastic nasal alae	45	0.001
HP0001273	Abnormality of the corpus callosum	103	0.001
HP0000358	Posteriorly rotated ears	152	0.001

Continued on next page

HPO	HPO Name	No. Genes in Network	P-value
HP0007018	Attention deficit hyperactivity disorder	100	0.002
HP0008872	Feeding problems in infancy	266	0.002
HP0000954	Transverse palmar creases	95	0.002
HP0000383	Abnormality of periauricular region	109	0.001
HP0002011	Abnormality of the central nervous system	1668	0.001
HP0001574	Abnormality of the integument	623	0.001
HP0000436	Abnormality of the nasal tip	269	0.001
HP0001769	Broad feet	27	0.002
HP0009484	Deviation of the hand or of fingers of the hand	223	0.001
HP0007364	Aplasia/Hypoplasia of the cerebrum	454	0.001
HP0001763	Pes planus	151	0.001
HP0000175	Cleft palate	67	0.006
HPz999017		767	0.001
HP0002119	Ventriculomegaly	71	0.007
HP0002263	Exaggerated cupid's bow	83	0.003
HP0002269	Neuronal migration disorder	32	0.005
HP0010993	Abnormality of the cerebral subcortex	103	0.002

Continued on next page

HPO	HPO Name	No. Genes in Network	P-value
HP0003366	Abnormality of the femoral neck and head region	73	0.021
HP0000736	Short attention span	100	0.002
HP0001513	Obesity	116	0.001
HP0000002	Abnormality of body height	698	0.001
HP0001630	Abnormality of the atrial septum	30	0.005
HP0100022	Abnormality of movement	172	0.004
HP0001520	Macrosomia	22	0.013
HP0001600	Abnormality of the larynx	126	0.007
HP0004279	Hypoplastic hand	110	0.008
HP0004325	Decreased body weight	567	0.001
HP0000398	Dysplastic ears	31	0.004
HP0000277	Abnormality of the mandible	397	0.03
HP0010864	"Intellectual disability, severe"	565	0.001
HP0003368	Abnormality of the femoral head	73	0.019
HP0009115	Aplasia/Hypoplasia involving the skeleton	351	0.005
HP0002648	Abnormality of skull shape	490	0.001
HP0000486	Strabismus	316	0.006
HP0000412	Prominent ears	168	0.001
HP0000293	Full cheeks	13	0.004

Continued on next page

HPO	HPO Name	No. Genes in Network	P-value
HP0001250	Seizures	483	0.001
HP0000357	Abnormal location of ears	220	0.004
HP0000812	Abnormal internal genitalia	265	0.001
HP0100742	Vascular neoplasia	22	0.034
HP0000404	Deafness	79	0.042
HP0002118	Abnormality of the cerebral ventricles	71	0.013
HP0001507	Growth abnormality	921	0.001
HP0001371	Contractures	67	0.001
HP0008871	Height less than 3rd percentile	645	0.001
HP0005743	Avascular necrosis of the capital femoral epiphysis	73	0.016
HP0000425	Flattened nasal bridge	107	0.001
HP0001159	Syndactyly	70	0.002
HP0000184	Prominent lips	124	0.001
HP0002087	Abnormality of the upper respiratory tract	133	0.007
HP0000692	Misalignment of teeth	22	0.026
HP0009810	Abnormality of the joints of the upper limbs	80	0.001
HP0001249	Intellectual disability	1519	0.001
HP0000159	Lip abnormality	417	0.001
HP0009473	Joint contractures involving the joints of the hand	57	0.001
HP0000272	Malar hypoplasia	204	0.001
HP0005306	Capillary hemangiomas	22	0.034

Continued on next page

HPO	HPO Name	No. Genes in Network	P-value
HP0001367	Abnormality of the joints	371	0.001
HP0008562	Poorly formed pinnae	17	0.008
HP0005557	Abnormality of the zygomatic arch	204	0.001
HP0005930	Abnormality of the epiphyses	73	0.025
HP0010609	Skin tags	67	0.001
HP0000252	Microcephaly	389	0.001
HP0005105	Abnormal nasal morphology	466	0.001
HP0009466	Radial deviation of fingers	223	0.001
HP0003549	Abnormality of connective tissue	119	0.001
HP0000022	Abnormality of male internal genitalia	234	0.001
HP0000303	Mandibular prognathia	94	0.01
HP0000964	Eczema	79	0.011
HP0001770	Toe syndactyly	41	0.023
HP0000492	Abnormality of the eyelid	834	0.001
HP0000219	Thin upper lip vermilion	252	0.001
HP0008069	Neoplasm of the skin	22	0.024
HP0000315	Abnormality of the orbital region	617	0.001
HPz999020		29	0.048
HP0005120	Abnormality of the cardiac atria	30	0.015
HP0002060	Abnormality of the cerebrum	518	0.001

Continued on next page

HPO	HPO Name	No. Genes in Network	P-value
HP0005918	Abnormality of the phalanges of the hand	138	0.001
HP0000164	Abnormality of the teeth	298	0.001
HPz999019		382	0.001
HP0200007	Abnormal size of the palpebral fissures	416	0.001
HP0000177	Abnormality of upper lip	404	0.001
HP0000545	Myopia	125	0.001
HP0002814	Abnormality of the lower limb	581	0.001
HP0010574	Abnormality of the epiphysis of the femoral head	73	0.013
HP0000422	Abnormality of the nasal bridge	505	0.001
HP0001317	Abnormality of the cerebellum	71	0.007
HP0002500	Abnormality of the cerebral white matter	103	0.003
HP0000290	Abnormality of the forehead	359	0.001
HP0000235	Abnormality of the fontanelles and cranial sutures	85	0.03
HP0000707	Abnormality of the nervous system	1756	0.001
HP0009121	Abnormality of the axial skeleton	1005	0.001
HP0000341	Narrow forehead	11	0.005

Continued on next page

HPO	HPO Name	No. Genes in Network	P-value
HP0001028	Hemangiomas	22	0.038
HP0000239	Large fontanelles	74	0.014
HP0000248	Brachycephaly	38	0.021
HPz999018		12	0.022
HP0009122	Aplasia/Hypoplasia affecting bones of the axial skeleton	162	0.015
HP0000215	Thick upper lip vermilion	113	0.001
HP0000750	Impaired language development	843	0.001
HP0000119	Abnormality of the genitourinary system	364	0.001
HP0000429	Abnormality of the nasal alae	72	0.001
HP0000708	Behavioural/Psychiatric Abnormality	1652	0.001
HP0100037	Abnormality of the scalp hair	67	0.001
HP0001439	Abnormality of the thigh	73	0.016
HP0000078	Abnormality of the genital system	321	0.001
HP0000240	Abnormality of skull size	502	0.001
HP0009485	Radial deviation of the hand or of fingers of the hand	223	0.001
HP0002012	Abnormality of the abdominal organs	351	0.005

Continued on next page

HPO	HPO Name	No. Genes in Network	P-value
HP0006499	Abnormality of femoral epiphyses	73	0.018
HP0000309	Abnormality of the mid-face	219	0.001
HP0000356	Abnormality of the outer ear	641	0.001
HP0001852	Gap between first and second toes	29	0.001
HP0000490	Deeply set eye	189	0.001
HP0009179	Deviation of the 5th finger	223	0.001
HP0001608	Abnormality of the voice	126	0.003
HP0000001	All	1840	0.001
HP0001965	Abnormality of the scalp	67	0.001
HP0006493	Aplasia/Hypoplasia involving bones of the lower limbs	140	0.026
HP0002019	Constipation	45	0.03
HP0000455	Broad nasal tip	156	0.001
HP0001500	Broad fingers	20	0.048
HP0002022	Feeding difficulties	266	0.002
HP0006261	Abnormality of phalangeal joints of the hand	57	0.001
HPz999016		780	0.002
HP0001631	Atrial septal defect	30	0.01
HP0004097	Deviated fingers	223	0.001
HP0002007	Frontal bossing	258	0.002
HP0001863	Clinodactyly of feet	132	0.001

Continued on next page

HPO	HPO Name	No. Genes in Network	P-value
HP0006505	Abnormality involving the epiphyses of the limbs	73	0.021
HP0010936	Abnormality of the lower urinary tract	39	0.015
HP0007477	Abnormal dermatoglyphics	153	0.001
HP0007370	Aplasia/Hypoplasia of the corpus callosum	103	0.007
HP0000340	Sloping forehead	11	0.039
HP0000178	Abnormality of lower lip	202	0.001
HP0000276	Long face	18	0.009
HP0009890	High anterior hairline	37	0.001
HP0008386	Aplasia/Hypoplasia of the nails	49	0.024
HP0000347	Micrognathia	161	0.015
HP0001385	Hip dysplasia	73	0.016
HP0010461	Abnormality of the male genitalia	282	0.001
HP0006494	Aplasia/Hypoplasia involving bones of the feet	140	0.028
HP0100547	Abnormality of the forebrain	518	0.001
HP0000752	Hyperactivity	100	0.001
HP0100276	Skin pits	30	0.001
HP0000464	Abnormality of the neck	102	0.001
HP0000271	Abnormality of the face	1404	0.001
HP0100277	Periauricular skin pits	30	0.001

Continued on next page

HPO	HPO Name	No. Genes in Network	P-value
HP0004207	Abnormality of the 5th finger	274	0.001
HP0008050	Abnormality of the palpebral fissures	680	0.001
HP0000717	Autism	369	0.001
HP0000431	Broad nasal bridge	312	0.001
HP0010490	Abnormality of the palmar creases	153	0.001
HP0100543	Cognitive impairment	1600	0.001
HP0002817	Abnormality of the upper limb	665	0.001
HP0002086	Abnormality of the respiratory system	146	0.01
HP0000153	Abnormality of the mouth	926	0.001
HP0001760	Abnormality of the feet	500	0.001
HP0011039	Abnormality of the helix	124	0.001
HP0000384	Preauricular skin tag	67	0.003
HP0001263	Developmental delay	843	0.001
HP0000152	Abnormality of head and neck	1428	0.001
HP0000929	Abnormality of the skull	942	0.001
HP0000951	Abnormality of the skin	623	0.001
HP0000163	Abnormality of the oral cavity	772	0.001
HP0009794	Branchial anomaly	30	0.003
HP0000316	Hypertelorism	428	0.001
HP0100851	Abnormal emotion/affect behaviour	576	0.001

Continued on next page

HPO	HPO Name	No. Genes in Network	P-value
HP0004322	Short stature	667	0.001
HP0002977	Aplasia/Hypoplasia involving the central nervous system	454	0.001
HP0000160	Small mouth	51	0.002
HP0000154	Wide mouth	74	0.013
HP0011138	Abnormality of skin adnexa	313	0.001
HP0010938	Abnormality of the external nose	328	0.001
HPz999021		330	0.001
HP0000539	Abnormality of refraction	154	0.002
HP0009116	Aplasia/Hypoplasia involving bones of the skull	162	0.014
HP0000035	Abnormality of the testis	234	0.001
HP0000718	Aggressive behavior	292	0.001
HP0003121	Limb contractures	57	0.001
HP0000411	Protruding ears	168	0.001
HP0000601	Hypotelorism	30	0.026
HP0000232	Everted lower lip vermilion	29	0.039
HP0011061	Abnormality of dental structure	31	0.01
HP0001274	Agenesis of corpus callosum	103	0.001
HP0100807	Long fingers	38	0.02

Continued on next page

HPO	HPO Name	No. Genes in Network	P-value
HP0000599	Abnormality of the frontal hairline	56	0.001
HP0000209	Abnormality of the jaws	398	0.037
HP0000278	Retrognathia	182	0.002
HP0001611	Nasal speech	126	0.005
HP0010720	Abnormal hair growth pattern	159	0.001
HP0000811	Abnormal external genitalia	290	0.001
HP0004426	Abnormality of the cheeks	13	0.002
HP0000924	Abnormality of the musculoskeletal system	1428	0.001
HP0100278	Periauricular skin tag	67	0.001
HP0000489	Abnormality of globe location or size	617	0.001
HP0100360	Contractures of the joints of the upper limbs	57	0.001
HP0001999	Facial dysmorphism	361	0.001
HP0004691	2-3 toe syndactyly	41	0.04
HP0002678	Skull asymmetry	107	0.003
HP0009924	Aplasia/Hypoplasia involving the nose	112	0.021
HP0004324	Increased body weight	136	0.001
HP0200006	Slanting of the palpebral fissures	392	0.001
HP0001764	Small feet	140	0.014
HP0001510	Growth delay	853	0.001

Continued on next page

HPO	HPO Name	No. Genes in Network	P-value
HP0002683	Abnormality of the calvarium	85	0.035
HP0000032	Abnormality of male external genitalia	282	0.001
HP0000028	Cryptorchidism	232	0.001
HP0004209	Clinodactyly of the 5th finger	223	0.001
HP0001155	Abnormality of the hand	659	0.001
HP0000582	Upslanting palpebral fissures	247	0.001
HP0009815	Aplasia/Hypoplasia of the extremities	175	0.011
HP0002823	Abnormality of the femur	73	0.015
HP0000494	Downward slanting palpebral fissures	156	0.029
HP0000343	Long philtrum	55	0.011
HP0100323	Juvenile aseptic necrosis	73	0.022
HP0000234	Abnormality of the head	1428	0.001
HP0010885	Aseptic necrosis	73	0.018
HP0000445	Broad nose	96	0.001
HP0001595	Abnormality of the hair	221	0.001
HP0001182	Tapered fingers	114	0.002
HP0000118	Phenotypic abnormality	1840	0.001
HP0006829	Severe muscular hypotonia	483	0.001
HP0002813	Abnormality of the extremities	826	0.001

Table B.6: Significant PPIs among genes identified using at least two methods

HPO	HPO Name	No. Genes	No. PPIs
HP:0000492	Abnormality of the eyelid	125	175
HP:0002007	Frontal bossing	3	2
HP:0000240	Abnormality of skull size	16	15
HP:0009121	Abnormality of the axial skeleton	54	74
HP:0000811	Abnormal external genitalia	10	6
HP:0001520	Macrosomia	2	1
HP:0000277	Abnormality of the mandible	13	20
HP:0000163	Abnormality of the oral cavity	20	29
HP:0000422	Abnormality of the nasal bridge	13	13
HP:0000929	Abnormality of the skull	52	69
HP:0001438	Abnormality of the abdomen	17	19
HP:0008050	Abnormality of the palpebral fissures	95	125
HP:0000001	All	13	10
HP:0000252	Microcephaly	12	8
HP:0000119	Abnormality of the genitourinary system	15	9
HP:0001182	Tapered fingers	4	2
HP:0002263	Exaggerated cupid's bow	2	1
HP:0000118	Phenotypic abnormality	13	10

Continued on next page

HPO	HPO Name	No. Genes	No. PPIs
HP:0002648	Abnormality of skull shape	5	3
HP:0000209	Abnormality of the jaws	13	20
HP:0001999	Facial dysmorphism	20	16
HP:0001249	Intellectual disability	33	50
HP:0007364	Aplasia/Hypoplasia of the cerebrum	16	13
HP:0002977	Aplasia/Hypoplasia involving the central nervous system	16	13
HP:0002012	Abnormality of the abdominal organs	14	16
HP:0008871	Height less than 3rd percentile	2	1
HP:0009466	Radial deviation of fingers	2	1
HP:0200006	Slanting of the palpebral fissures	2	1
HP:0000924	Abnormality of the musculoskeletal system	3	3
HP:0009179	Deviation of the 5th finger	2	1
HP:0000184	Prominent lips	2	1
HP:0000022	Abnormality of male internal genitalia	2	1
HP:0000219	Thin upper lip vermilion	2	1
HP:0000213	Thin lips	2	1
HP:0009484	Deviation of the hand or of fingers of the hand	2	1

Continued on next page

HPO	HPO Name	No. Genes	No. PPIs
HP:0000494	Downward slanting palpebral fissures	2	1
HP:0004209	Clinodactyly of the 5th finger	2	1
HP:0002011	Abnormality of the central nervous system	3	3
HP:0000234	Abnormality of the head	3	3
HP:0100543	Cognitive impairment	6	5
HP:0001863	Clinodactyly of feet	2	1
HP:0000271	Abnormality of the face	3	3
HP:0000284	Abnormality of the ocular region	2	1
HP:0000215	Thick upper lip vermilion	2	1
HP:0000002	Abnormality of body height	2	1
HP:0000272	Malar hypoplasia	2	1
HP:0001510	Growth delay	2	1
HP:0005557	Abnormality of the zygomatic arch	2	1
HP:0000708	Behavioural/Psychiatric Abnormality	6	5
HP:0004322	Short stature	2	1
HP:0000707	Abnormality of the nervous system	6	5
HP:0009485	Radial deviation of the hand or of fingers of the hand	2	1
HP:0001780	Abnormality of the toes	2	1

Continued on next page

HPO	HPO Name	No. Genes	No. PPIs
HP:0003011	Abnormality of musculature	2	1
HP:0010461	Abnormality of the male genitalia	2	1
HP:0006829	Severe muscular hypotonia	4	2
HP:0000032	Abnormality of male external genitalia	2	1
HP:0000309	Abnormality of the midface	2	1
HP:0000164	Abnormality of the teeth	5	5
HP:0000179	Thick lower lip vermilion	2	1
HP:0004097	Deviated fingers	2	1
HP:0000035	Abnormality of the testis	2	1
HP:0000288	Abnormality of the philtrum	9	9
HP:0000152	Abnormality of head and neck	3	3
HP:0000028	Cryptorchidism	2	1

Table B.7: Significant MGI enrichments after 5% FDR

HPO	HPO Name	MPO term	P-value
HP0000262	Turricephaly	abnormal head shape	0.00017
HP0000286	Epicanthus	nuclear cataracts	3.5e-05
HP0002119	Ventriculomegaly	abnormal prepulse inhibition	8.9e-05
HP0002119	Ventriculomegaly	decreased prepulse inhibition	0.00025

Continued on next page

HPO	HPO Name	MPO term	P-value
HP0001250	Seizures	abnormal prepulse inhibition	7.1e-05
HP0001250	Seizures	decreased prepulse inhibition	0.00011
HP0001250	Seizures	absence seizures	0.00033
HP0001250	Seizures	abnormal synaptic transmission	0.00045
HP0002118	Abnormality of the cerebral ventricles	abnormal prepulse inhibition	8.9e-05
HP0002118	Abnormality of the cerebral ventricles	decreased prepulse inhibition	0.00025
HP0001256	Intellectual disability, mild	abnormal associative learning	3.6e-05
HP0001256	Intellectual disability, mild	abnormal learning/ memory	0.00023
HP0001256	Intellectual disability, mild	abnormal learning/memory/conditioning	0.00023
HP0001256	Intellectual disability, mild	impaired coordination	0.00025
HP0000256	Macrocephaly	abnormal head shape	8.6e-05
HP0000717	Autism	abnormal prepulse inhibition	8.7e-05
HP0100851	Abnormal emotion/affect behaviour	abnormal prepulse inhibition	9.8e-05
HP0000278	Retrognathia	abnormal foramen magnum morphology	0.00021
HP0002342	Intellectual disability, moderate	abnormal prepulse inhibition	0.00013

Continued on next page

HPO	HPO Name	MPO term	P-value
HP0010864	Intellectual disability, severe	abnormal prepulse inhibition	2e-05
HP0010864	Intellectual disability, severe	abnormal CNS synaptic transmission	2.3e-05
HP0010864	Intellectual disability, severe	abnormal synaptic transmission	2.9e-05
HP0010864	Intellectual disability, severe	decreased prepulse inhibition	3.8e-05
HP0010864	Intellectual disability, severe	abnormal nervous system physiology	0.00044

Table B.8: Number of child phenotypes (subterms) exhibited by patients in the cohort. In any = number of subterms exhibited by at least one patient in the entire cohort. In de novo = number of subterms exhibited by at least one patient with a small *de novo* CNV.

HPO	HPO Name	No. Subterms	In Any	In de novo
HP:0001018	Abnormal palmar dermatoglyphics	18	6	2
HP:0000284	Abnormality of the ocular region	124	45	29
HP:0003011	Abnormality of musculature	507	67	20
HP:0000366	Abnormality of the nose	146	58	35
HP:0001518	Low birth weight	2	0	0
HP:0001438	Abnormality of the abdomen	403	85	21
HP:0000288	Abnormality of the philtrum	12	6	4
HP:0004323	Abnormality of body weight	13	7	6

Continued on next page

HPO	HPO Name	No. Subterms	In Any	In de novo
HP:0001252	Muscular hypotonia	9	4	1
HP:0000606	Abnormality of the periorbital region	109	41	24
HP:0000478	Abnormality of the eye	795	166	68
HP:0001780	Abnormality of the toes	711	43	15
HP:0000377	Abnormality of the pinna	65	43	19
HP:0001167	Abnormality of the fingers	869	70	32
HP:0000504	Abnormality of vision	69	12	2
HP:0008872	Feeding problems in infancy	0	0	0
HP:0000818	Abnormality of the endocrine system	214	29	7
HP:0002011	Abnormality of the central nervous system	843	174	53
HP:0000436	Abnormality of the nasal tip	15	5	4
HP:0001574	Abnormality of the integument	743	172	78
HP:0007364	Aplasia/Hypoplasia of the cerebrum	20	6	3
HP:0009484	Deviation of the hand or of fingers of the hand	32	14	5
HP:z999017		0	0	0
HP:0001513	Obesity	3	1	1
HP:0000769	Abnormality of the breast	23	13	6
HP:0001600	Abnormality of the larynx	52	9	3
HP:0004325	Decreased body weight	6	2	1

Continued on next page

HPO	HPO Name	No. Subterms	In Any	In de novo
HP:0000277	Abnormality of the mandible	24	7	4
HP:0010864	Intellectual disability, severe	0	0	0
HP:0002648	Abnormality of skull shape	53	15	10
HP:0001250	Seizures	57	11	1
HP:0000812	Abnormal internal genitalia	117	25	11
HP:0001507	Growth abnormality	122	35	21
HP:0008871	Height less than 3rd percentile	0	0	0
HP:0001159	Syndactyly	34	8	4
HP:0001249	Intellectual disability	7	3	3
HP:0001367	Abnormality of the joints	542	77	30
HP:0005105	Abnormal nasal morphology	41	19	12
HP:0009466	Radial deviation of fingers	8	2	1
HP:0003549	Abnormality of connective tissue	187	29	9
HP:0000492	Abnormality of the eyelid	80	25	14
HP:0000315	Abnormality of the orbital region	26	11	8
HP:0002060	Abnormality of the cerebrum	149	28	11
HP:0000164	Abnormality of the teeth	155	36	13
HP:0002814	Abnormality of the lower limb	1079	122	43

Continued on next page

HPO	HPO Name	No. Subterms	In Any	In de novo
HP:0000422	Abnormality of the nasal bridge	16	10	7
HP:0000707	Abnormality of the nervous system	1488	273	71
HP:0009121	Abnormality of the axial skeleton	661	125	53
HP:0000750	Impaired language development	13	4	0
HP:0000119	Abnormality of the genitourinary system	479	131	40
HP:0000708	Behavioural/Psychiatric Abnormality	409	96	21
HP:0000078	Abnormality of the genital system	221	68	26
HP:0000240	Abnormality of skull size	13	5	2
HP:0009485	Radial deviation of the hand or of fingers of the hand	12	3	2
HP:0002012	Abnormality of the abdominal organs	357	69	13
HP:0000356	Abnormality of the outer ear	110	59	33
HP:0009179	Deviation of the 5th finger	5	2	1
HP:0001608	Abnormality of the voice	20	7	2
HP:0000455	Broad nasal tip	3	0	0
HP:0002022	Feeding difficulties	7	2	1
HP:z999016		0	0	0
HP:0004097	Deviated fingers	24	9	3

Continued on next page

HPO	HPO Name	No. Subterms	In Any	In de novo
HP:0007477	Abnormal dermatoglyphics	22	8	3
HP:0010461	Abnormality of the male genitalia	97	28	15
HP:0100547	Abnormality of the forebrain	157	29	12
HP:0000271	Abnormality of the face	782	286	154
HP:0004207	Abnormality of the 5th finger	134	12	5
HP:0008050	Abnormality of the palpebral fissures	13	9	8
HP:0000431	Broad nasal bridge	4	1	0
HP:0010490	Abnormality of the palmar creases	8	5	1
HP:0002817	Abnormality of the upper limb	1502	156	64
HP:0100543	Cognitive impairment	68	20	8
HP:0000153	Abnormality of the mouth	362	120	56
HP:0001760	Abnormality of the feet	843	79	25
HP:0011039	Abnormality of the helix	21	16	7
HP:0001263	Developmental delay	23	10	3
HP:0000152	Abnormality of head and neck	1014	342	185
HP:0000929	Abnormality of the skull	210	45	24
HP:0000951	Abnormality of the skin	729	165	76
HP:0000316	Hypertelorism	0	0	0

Continued on next page

HPO	HPO Name	No. Subterms	In Any	In de novo
HP:0002977	Aplasia/Hypoplasia involving the central nervous system	56	15	4
HP:0000154	Wide mouth	0	0	0
HP:0010938	Abnormality of the external nose	41	16	11
HP:0000209	Abnormality of the jaws	43	15	8
HP:0000811	Abnormal external genitalia	83	28	12
HP:0000924	Abnormality of the musculoskeletal system	4046	523	210
HP:0000489	Abnormality of globe location or size	18	6	5
HP:0001999	Facial dysmorphism	36	13	6
HP:0004324	Increased body weight	5	3	3
HP:0200006	Slanting of the palpebral fissures	4	3	2
HP:0000363	Abnormality of ear lobes	11	8	3
HP:0001510	Growth delay	91	19	7
HP:0000032	Abnormality of male external genitalia	61	21	10
HP:0004209	Clinodactyly of the 5th finger	3	1	0
HP:0001155	Abnormality of the hand	1131	111	49
HP:0000582	Upslanting palpebral fissures	0	0	0
HP:0000234	Abnormality of the head	981	328	177

Continued on next page

HPO	HPO Name	No. Subterms	In Any	In de novo
HP:0011024	Abnormality of the gastrointestinal tract	190	48	10
HP:0000445	Broad nose	1	0	0
HP:0004404	Abnormality of the nipple	10	6	4
HP:0001595	Abnormality of the hair	160	42	26
HP:0006829	Severe muscular hypotonia	0	0	0
HP:0002813	Abnormality of the extremities	2696	310	125

Table B.9: The 87 phenotypes consistently showing significant enrichments in patients with CNVs affecting particular biological pathways.

Phenotype	Extended	Pathway	De novo
"Abnormality of the orbital region"	0.0007925	0.003804	0
"Thin upper lip vermilion"	0.0004901	0	0
"Flat philtrum"	0.007774	0	0
"Feeding problems in infancy"	0	0	0
"Abnormality of the joints"	0.009104	0.1317	0
"Abnormality of the larynx"	0	0	0
"Abnormality of the ventricular septum"	0	0.003717	0.001311
"Abnormality of the jaws"	0	0	0
"Low birth weight"	0	0.007491	0.005376
"Abnormality of head and neck"	0	0	0
"Abnormality of the nipple"	0	0.0005319	0.0006964
"Abnormality of the head"	0	0	0
"Abnormality of the skull"	0.04807	0.01304	0

Continued on next page

Phenotype	Extended	Pathway	De novo
"Abnormality of the ocular region"	0	0	0
"Abnormality of the extremities"	0	0	0
"Abnormality of the axial skeleton"	0	0.01848	0
"Abnormality of the hand"	0.0006080	0.02674	0
"Upslanting palpebral fissures"	1	1	0.002762
"Abnormality of the periorbital region"	0	0	0
"Abnormality of musculature"	0	0.03858	0
"Flattened nasal bridge"	0	0	1
"Radial deviation of fingers"	0.001153	0.002185	0
"Clinodactyly of the 5th finger"	0.0004273	0.002111	0
"Abnormality of the lower limb"	0	0.3474	0
"Severe muscular hypotonia"	0.2073	0.01217	0
"Abnormality of male internal genitalia"	0.05013	0	0.6373
"Abnormality of the philtrum"	0	0	0
"Posteriorly rotated ears"	1	0	0
"Abnormality of the nose"	0	0	0
"Abnormality of the breast"	0	0.09197	0.2241
"Abnormality of the feet"	0	0.3460	0
"Aplasia/Hypoplasia of the nails"	1	0.01785	0.01
"Abnormality of skin adnexa"	1	0.2397	0
"Deviation of the hand or of fingers of the hand"	0.002833	0.1077	0
"Abnormality of the 5th finger"	0	0	0

Continued on next page

Phenotype	Extended	Pathway	De novo
"Epicanthus"	0.04009	0.9455	0.002868
"Abnormality of the oral cavity"	0	0	0
"Abnormality of the eye"	0	0	0
"Long philtrum"	0	0.001566	0
"Deviated fingers"	0.002833	0.02661	0
"Abnormality of the toes"	0.03627	0.5542	0.04960
"Abnormality of the outer ear"	0	0	0
"Impaired language development"	0	0	0
"Abnormality of the eyelid"	0	0	0
"Abnormality of the ear"	0	0	0
"Broad nasal bridge"	0	0	0
"Abnormality of upper lip"	0	0	0
"Cryptorchidism"	0	0	0.6373
"Abnormality of the forehead"	0.0002197	0	0
"Lip abnormality"	0	0	0
"Nasal speech"	1	0	0
"Abnormality of the testis"	0.05013	0	0.6373
"Abnormality of the mandible"	0	0	0
"Radial deviation of the hand or of fingers of the hand"	0.001153	0.002184	0
"Deviation of the 5th finger"	0.0004273	0.002111	0
"Ventricular septal defect"	0	0.003717	0.001310
"HP:z999021"	0	1	0.5777
"Cardiac malformation"	0.1741	0.008510	1
"Abnormality of the integument"	0	0	0
"Abnormal internal genitalia"	0.02370	0	0.6373

Continued on next page

Phenotype	Extended	Pathway	De novo
"Abnormality of the mouth"	0	0	0
"Growth abnormality"	0.001491	0	0
"Decreased body weight"	0	0.001208	0.0008019
"Abnormality of skull shape"	0	0.0006365	0
"Developmental delay"	0	0	0
"Abnormality of the pinna"	0.1204	0.06237	0
"Low-set ears"	0	0.00208877284	15953
"Abnormality of the face"	0	0	0
"Abnormality of the fingers"	0.0001936	0.02558	0
"Abnormality of the upper limb"	0.06814	0.02674	0
"Abnormality of the musculoskeletal system"	0	0	0
"Abnormality of the nail"	0.05222	0.1565	0
"Joint hypermobility"	1	1	0.04
"Thin lips"	0.0004900	0	0
"Abnormality of body weight"	0.1114	0.04637	0.0005813
"Facial dysmorphism"	1	0.1251	0
"Muscular hypotonia"	0	0.1027	0
"Abnormality of the skin"	0	0	0
"Abnormality of the nasal bridge"	0	0	0
"Abnormality of the cardiac septa"	0	0.001128	0
"Feeding difficulties"	0	0	0.002710
"Growth delay"	0	0	0.0007993
"Abnormality of the cardiovascular system"	0	0	0
"Abnormal location of ears"	0	0	0.1479

Continued on next page

Phenotype	Extended	Pathway	De novo
"Abnormal nasal morphology"	0.002840	0	0
"Abnormality of the voice"	0 0	0	
"Abnormality of the cardiac ventricle"	0 0.003717	0.001310	

B.3 Chapter 4

Table B.10: Significant Gene Ontology enrichments among genes in functional clusters present in *de novo* CNVs with Hypotonia in DECIPHER and GENCODYS after a Bonferroni Correction (q-value). BP = biological process, MF = molecular functions, CC = cellular component

Term	Type	Fold-enriched	p-value	q-value
negative regulation of biological process	BP	2.09	9.50E-17	1.38E-12
negative regulation of cellular process	BP	2.1	2.73E-15	3.98E-11
multicellular organismal process	BP	1.53	4.58E-14	6.68E-10
regulation of biological quality	BP	1.97	2.78E-12	4.05E-08
response to chemical stimulus	BP	1.88	1.28E-11	1.86E-07
anatomical structure development	BP	1.69	2.03E-11	2.96E-07
developmental process	BP	1.61	2.97E-11	4.33E-07
cell communication	BP	2.02	4.18E-11	6.09E-07
system development	BP	1.72	6.42E-11	9.36E-07
cell periphery	CC	1.61	9.27E-11	1.35E-06
regulation of cellular process	BP	1.33	1.06E-10	1.54E-06

Continued on next page

Term	Type	Fold-enriched	p-value	q-value
multicellular organismal development	BP	1.63	1.43E-10	2.08E-06
nervous system development	BP	2.09	3.73E-10	5.43E-06
synaptic transmission	BP	3.14	4.03E-10	5.87E-06
signaling	BP	1.48	5.64E-10	8.22E-06
regulation of biological process	BP	1.3	5.79E-10	8.43E-06
cell surface receptor linked signaling pathway	BP	1.76	6.72E-10	9.79E-06
organ development	BP	1.86	7.66E-10	1.12E-05
regulation of signaling	BP	2.04	7.73E-10	1.13E-05
plasma membrane	CC	1.58	7.76E-10	1.13E-05
molecular transducer activity	MF	2	1.38E-09	2.01E-05
signal transducer activity	MF	2	1.38E-09	2.01E-05
behavior	BP	3.32	1.60E-09	2.33E-05
neurogenesis	BP	2.39	1.87E-09	2.73E-05
cell differentiation	BP	1.8	1.90E-09	2.76E-05
generation of neurons	BP	2.44	1.90E-09	2.77E-05
cellular process	BP	1.11	2.44E-09	3.56E-05
cellular developmental process	BP	1.77	3.00E-09	4.37E-05
multicellular organismal signaling	BP	2.87	3.29E-09	4.79E-05
transmission of nerve impulse	BP	2.87	3.29E-09	4.79E-05
biological regulation	BP	1.27	3.40E-09	4.95E-05

Continued on next page

Term	Type	Fold-enriched	p-value	q-value
anatomical structure morphogenesis	BP	1.95	3.41E-09	4.97E-05
regulation of nervous system development	BP	3.76	3.71E-09	5.41E-05
regulation of cell communication	BP	2.25	3.92E-09	5.71E-05
regulation of multicellular organismal process	BP	2.09	4.15E-09	6.04E-05
response to organic substance	BP	2.06	4.27E-09	6.23E-05
cell-cell signaling	BP	2.36	6.28E-09	9.15E-05
cellular response to stimulus	BP	1.42	7.05E-09	1.03E-04
signal transduction	BP	1.47	1.41E-08	2.05E-04
response to stimulus	BP	1.32	1.87E-08	2.72E-04
cytosol	CC	1.84	2.73E-08	3.97E-04
plasma membrane part	CC	1.84	2.90E-08	4.23E-04
developmental growth	BP	4.2	3.09E-08	4.50E-04
cell development	BP	2.12	3.37E-08	4.91E-04
cell fraction	CC	2.12	6.74E-08	9.82E-04
brain development	BP	2.94	1.42E-07	2.07E-03
neuron differentiation	BP	2.27	2.34E-07	3.42E-03
neuron development	BP	2.44	2.46E-07	3.59E-03
system process	BP	1.78	2.84E-07	4.13E-03
synapse part	CC	3.42	3.21E-07	4.67E-03
MAPKKK cascade	BP	3.03	3.26E-07	4.75E-03
organ morphogenesis	BP	2.41	3.55E-07	5.17E-03
regulation of neurogenesis	BP	3.46	4.48E-07	6.53E-03

Continued on next page

Term	Type	Fold-enriched	p-value	q-value
positive regulation of cellular process	BP	1.63	4.66E-07	6.79E-03
integral to plasma membrane	CC	2.02	5.29E-07	7.71E-03
positive regulation of biological process	BP	1.58	5.38E-07	7.84E-03
regulation of developmental process	BP	2.09	5.54E-07	8.07E-03
regulation of primary metabolic process	BP	1.5	5.58E-07	8.13E-03
regulation of metabolic process	BP	1.45	8.17E-07	1.19E-02
regulation of cellular metabolic process	BP	1.49	8.35E-07	1.22E-02
protein binding	MF	1.19	8.35E-07	1.22E-02
regulation of cellular component organization	BP	2.18	9.25E-07	1.35E-02
intrinsic to plasma membrane	CC	1.99	9.44E-07	1.38E-02
regulation of cell differentiation	BP	2.31	9.76E-07	1.42E-02
forebrain development	BP	3.55	1.04E-06	1.51E-02
regulation of cellular localization	BP	2.72	1.10E-06	1.61E-02
regulation of molecular function	BP	1.87	1.20E-06	1.74E-02
negative regulation of gene expression	BP	2.65	1.25E-06	1.82E-02

Continued on next page

Term	Type	Fold-enriched	p-value	q-value
negative regulation of metabolic process	BP	2.15	1.41E-06	2.06E-02
cell projection part	CC	2.68	1.51E-06	2.20E-02
enzyme linked receptor protein signaling pathway	BP	2.23	1.66E-06	2.41E-02
Binding	MF	1.1	1.66E-06	2.42E-02
regulation of catalytic activity	BP	1.98	1.69E-06	2.46E-02
cell proliferation	BP	1.88	1.98E-06	2.88E-02
Synapse	CC	2.82	2.12E-06	3.09E-02
central nervous system development	BP	2.4	2.19E-06	3.19E-02
neurological system process	BP	1.81	2.66E-06	3.87E-02
response to external stimulus	BP	1.95	3.24E-06	4.72E-02
neuron projection	CC	2.58	3.33E-06	4.85E-02

B.4 Chapter 6

Table B.11: Functional enrichments amongst 643 genes identified as understudied using any of the three networks.

Term	Description	Fold-Enrichment	Pval
GO:0007007	inner mitochondrial membrane organization	25.4	2.5×10^{-8}
GO:0045039	protein import into mitochondrial inner membrane	38.1	3.2×10^{-10}
GO:0006505	GPI ancho metabolic process	12.7	1.9×10^{-11}
GO:0016254	preassembly of GPI anchor in ER membrane	22.4	2.3×10^{-12}
GO:0006501	C-terminal protein lipidation	15.5	3.2×10^{-11}
GO:0006497	protein lipidation	9.3	7.6×10^{-10}
GO:0042158	lipoprotein biosynthetic process	8.5	2.5×10^{-9}
GO:0006506	GPI anchor biosynthetic process	12.7	8.5×10^{-11}
GO:0006661	phosphatidylinositol biosynthetic process	9.3	3.5×10^{-9}
GO:0018410	C-terminal protein amino acid modification	12.7	4.1×10^{-10}
GO:0046474	glycerophospholipid biosynthetic process	5.4	1.9×10^{-6}
GO:0017176	phosphatidylinositol N-acetylglucosaminyltransferase activity	34.7	2.0×10^{-8}
GO:0042719	mitochondrial intermembrane space protein transporter complex	34.6	5.6×10^{-10}
KEGGSubclass	Sensory System	2.6	1.5×10^{-5}
KEGGSubclass	Glycan Biosynthesis and Metabolism	4.3	5.2×10^{-8}
path:hsa00563	Glycosylphosphatidylinositol (GPI)-anchor biosynthesis	21.6	5.1×10^{-13}
KEGGClass	Metabolism	1.6	6.0×10^{-6}
path:hsa04740	Olfactory transduction	2.9	2.2×10^{-6}

Table B.12: Functional enrichments amongst 286 genes identified as understudied using any of the COXPRESdb network.

Term	Description	Fold-Enrichment	Pval
GO:0010468	regulation of gene expression	1.8	1.6×10^{-6}
GO:001556	regulation of macromolecule biosynthetic process	1.9	1.1×10^{-6}
GO:0010467	gene expression	1.7	1.7×10^{-6}
GO:0034645	cellular macromolecule biosynthetic process	1.7	1.3×10^{-6}
GO:2000112	regulation of cellular macromolecule biosynthetic process	1.9	4.0×10^{-7}
GO:0031326	regulation of cellular biosynthetic process	1.8	3.0×10^{-6}
GO:0051252	regulation of RNA metabolic process	2.0	1.0×10^{-7}
GO:0006355	regulation of transcription, DNA-dependent	2.0	2.9×10^{-9}
GO:0032774	RNA biosynthetic process	2.0	5.7×10^{-8}
GO:0016339	calcium-dependent cell-cell adhesion	18.9	5.8×10^{-7}
GO:0019219	regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	1.8	2.2×10^{-6}
GO:0016070	RNA metabolic process	1.7	2.0×10^{-6}
GO:0006351	transcription, DNA-dependent	2.1	9.6×10^{-9}
GO:0009059	macromolecule biosynthtic process	1.7	3.5×10^{-6}
GO:0046914	transition metal ion binding	2.0	6.9×10^{-7}
GO:0003676	nucleic acid binding	1.9	2.7×10^{-9}
GO:0008270	zinc ion binding	2.1	1.0×10^{-7}
GO:0003677	DNA binding	2.1	6.4×10^{-8}
path:hsa04612	Antigen processing and presentation	11.2	2.1×10^{-6}

Table B.13: Twenty genes with 1-1 orthologs identified as understudied using multiple networks. AAs is number of amino acids in the longest protein produced from the gene, Orthologs is the oldest organism in Ensembl for which an ortholog was defined. No. Tissues is the number of tissues in the Human Body Map (127) with >1 FPKM. Evidence codes for experimental GO terms include IDA = inferred from direct assay, IMP = inferred from mutant phenotype, IEP = inferred from expression profile, TAS = traceable author statement. IPI = inferred from physical interaction.

Gene	Name	AAs	Orthologs	GO terms (experimental)	No. Tissues
CCDC155	coiled-coil domain containing 155	524	vertebrates	Computational only	1
MFI2	antigen p97 (melanoma associated) identified by monoclonal antibodies 133.2 and 96.5	738	fruitfly	iron ion import (IMP), negative regulation of substrate adhesion-dependent cell spreading (IDA), positive regulation of extracellular matrix disassembly (IDA), positive regulation of plasminogen activation (IDA), anchored component of plasma membrane (IDA)	7
WDR59	WD repeat domain 59	974	yeast	None	15
TM6SF1	transmembrane 6 superfamily member 1	370	vertebrates	None	15
KCTD21	potassium channel tetramerization domain containing 21	260	vertebrates	Computational only	15

Continued on next page

Gene	Name	AAs	Orthologs	GO terms (experimental)	No. Tissues
TCEAL1	transcription elongation factor A (SII)-like 1	159	placental mammals	negative regulation of transcription from RNA polymerase II promoter (TAS)	16
NUB1	negative regulator of ubiquitin-like proteins 1	639	fruitfly	protein ubiquitination(IMP), response to interferon-gamma (IEP), response to tumor necrosis factor (IEP), cytoplasm (IDA), nucleolus (IDA), nucleus (IDA)	16
TSEN2	TSEN2 tRNA splicing endonuclease subunit	465	yeast	tRNA splicing via endonucleolytic cleavage and ligation (IDA), centrosome (IDA), cytoplasm (IDA), nucleus (IDA)	16
GFOD1	glucose-fructose oxidoreductase domain containing 1	390	fruitfly	None	16
TIMM10	translocase of inner mitochondrial membrane 10 homolog (yeast)	90	yeast	Many involving mitochondion	16

Continued on next page

Gene	Name	AAs	Orthologs	GO terms (experimental)	No. Tissues
TIMM9	translocase of inner mitochondrial membrane 9 homolog (yeast)	9	yeast	Many involving mitochondion	12
TIMM10B	translocase of inner mitochondrial membrane 10 homolog B (yeast)	103	fruitfly	cell-matrix adhesion (TAS), cellular protein metabolic process (TAS), protein targeting to mitochondrion (TAS), mitochondial inner membrane (IDA), mitochondrial intermembrane space protein transporter complex (IDA)	13
TIMM22	translocase of inner mitochondrial membrane 22 homolog (yeast)	194	yeast	protein import into mitochondrial inner membrane (TAS)	16
MYBPC2	myosin binding protein C, fast type	1141	fruitfly	structural constituent of muscle (TAS), muscle filament sliding (TAS), cytosol (TAS)	5
MYBPH	myosin binding protein H	477	fruitfly	structural constituent of muscle (TAS), regulation of striated muscle contraction (TAS)	3
AMIGO1	adhesion molecule with Ig-like domain 1	493	vetebrates	Computational only	16

Continued on next page

Gene	Name	AAs	Orthologs	GO terms (experimental)	No. Tissues
AMIGO2	adhesion molecule with Ig-like domain 2	522	vetebrates	protein binding (IPI)	14
CEP19	centrosomal protein 19kDa	167	vertebrates	centriole (IDA), ciliary basal body (IDA), spindle pole (IDA)	13
CCDC87	coiled-coil domain containing 87	849	vertebrates	Computational only	1