Chapter 4

Protein complexes, dosage sensitivity and copy-number variations

4.1 Introduction

In the previous chapter, I described the bias towards dominant mutations amongst mutations in protein interaction interfaces. As I mentioned there, dominance can be explained by haploinsufficiency or dominant negative effects. In either case, a 0.5 fold change in gene dosage of the functional (or mis-functional) protein causes a visible phenotype. It has been estimated that at least 20% of the entries in the OMIM database cause a phenotype as a *heterozygous* mutation (Kondrashov and Koonin, 2004). In contrast, the popular hypothesis explaining gene dominance formulated by Wright (1934) states that dominance is caused by "bottlenecks" in metabolic pathways and should generally be rare (Orr, 1991). Apparently, there are far more proteins that are *dosage sensitive* than can be explained by perturbations of biochemical pathways alone.

Papp *et al.* (2003) attempted to explain a similar observation made by Steinmetz *et al.* (2002) in *S. cerevisiae*. The latter had systematically created heterozygous deletion mutants for a range of genes orthologous to human disease-related genes. Papp *et al.* found that many haploinsufficient genes were members of protein complexes. They postulated that multi-protein complexes need to maintain the stoichiometry of their subunits to perform their biological function (the *balance hypothesis*). If this balance is disturbed, the function of the entire complex is disrupted. This conveniently explains the enrichment of haploinsufficiency amongst members of protein complexes. A range of other experiments also lend support to the balance hypothesis. It has been noted that expression levels of interacting proteins are highly co-ordinated (Jansen *et al.*, 2002), hinting that proportionality of subunit abundances is important. It has also been argued that tolerance towards polyploidization, compared to the sometimes severe effects of smaller duplications can be explained by conservation of stoichiometry (Aury *et al.*, 2006). The proposition in this case is that single gene duplications or deletions will cause a stronger negative fitness effect than copying all components of the complex, maintaining stoichiometric balance. Finally, it has been noted that highly-interacting proteins in higher organisms belong to small gene families (Yang *et al.*, 2003), which could be conveniently explained by a bias against duplication acting on multi-protein complexes.

There have been, however, several conflicting reports. Deutschbauer *et al.* (2005) performed a heterozygous deletion screen in *S. cerevisiae* that incorporated all open reading frames (ORFs) available for cloning at the time. They reported only 3% of genes to be haploinsufficient. While these genes were enriched for members of protein complexes, their subsequent overexpression did not cause a similar phenotype as their deletion. Unfortunately, it is not clear from the publication how the well described whole genome duplication that is characteristic for the *S. cerevisiae* lineage (Kellis *et al.*, 2004) affects these results. Subsequently, Sopko *et al.* (2006) systematically

induced gene overexpression for all ORFs in *S. cerevisiae*. The genes found to be toxic when overexpressed did not overlap with the haploinsufficient genes described by Deutschbauer *et al.*, and were not significantly enriched for protein complexes. This is in conflict with the dosage hypothesis in so far as it shows that deletion and duplication of the same gene do not usually lead to loss-of-function of the entire complex, as was initially suggested by Papp *et al.*. One important issue that has to be noted about the study by Sopko *et al.* is related to their experimental set-up. To assure that overexpression of the gene is controllable, they used an inducible promoter. They found that duplication sensitive genes were highly enriched for cell cycle proteins. A likely explanation for this bias is that the untimely expression of the proteins due to the non-physiological promoter is responsible for the negative fitness effect, rather than the actual dosage. The second important fact to consider is that single-cellular eukaryotes such as *S. cerevisiae* which are able to sustain both a haploid and diploid life-cycle, are likely to have different regulatory and dosage-compensatory mechanisms than multi-cellular organisms. One hint towards this difference is the increasing constraint on the number of paralogs of highly-interacting proteins in higher organisms, as described by Yang *et al.* (2003).

In light of the above points, Birchler *et al.* (2007) argued for a more elaborate concept to explain dosage sensitivity that they refer to as *regulatory balance*. Experiments in plants and later in *D. melanogaster* showed that duplications or deletions of some chromosomal regions cause no change in gene expression (Birchler, 1981; Devlin *et al.*, 1982), while variations of other genes causes up- or downregulation of various distal genes (Birchler *et al.*, 2001). One example referred to by Birchler *et al.* is *D. melanogaster* white eye colour controlled by the single gene *white*. Over the years, duplications of some and deletions of other genes (47 in total so far) have all been found to affect the expression of *white*. The majority of modulators of *white* act as negative regulators, *i.e.* a duplication of the regulator leads to lower expression of *white*. Birchler

et al. suggest that these regulators form a complex regulatory network where information transfer happens mostly through protein interactions, see for example Figure 4.1.

Considering these findings, it appears that there are multiple possible causes of dosage sensitivity, whereby deletion and duplication of the same gene do not necessarily lead to the same outcome:

- A limited number of enzymes are sensitive to low dosage because they are the rate limiting factor in a biochemical reaction.
- A range of proteins are likely to cause non-physiological binding or even agglomeration as a result of overexpression, as exemplified by susceptibility to early-onset Alzheimer's disease as a result of duplication of the APP locus (Lee and Lupski, 2006).
- Haploinsufficiency as well as duplication sensitivity are likely to affect the regulators controlling the balanced expression of a range of other proteins. As I described above, these proteins are in fact often complexes.

Dosage sensitivity and the concept of regulatory balance have important implications for gene duplicability and thus for the understanding of gene evolution. The widely accepted paradigm states that gene duplications can either create a non-functional pseudogene (*nonfunctionalization*) or relax selection constraints on one of the paralogous sequences, allowing it to diverge into related (*subfunctionalization*) or, in rare cases, new functions (*neofunctionalization*) (Prince and Pickett, 2002). Historically, it was assumed in this context that most genes can be duplicated without substantial negative fitness effects. It has been shown, however, that there are distinct differences between genes as to their duplicability (Veitia, 2005; Yang *et al.*, 2003) and that duplicated genes are in many cases still under negative selection (Kondrashov *et al.*, 2002; Lynch and Conery, 2000). How exactly these pressures on gene evolution are linked to dosage sensitivity and thereby to protein complexes is the focus of this chapter.

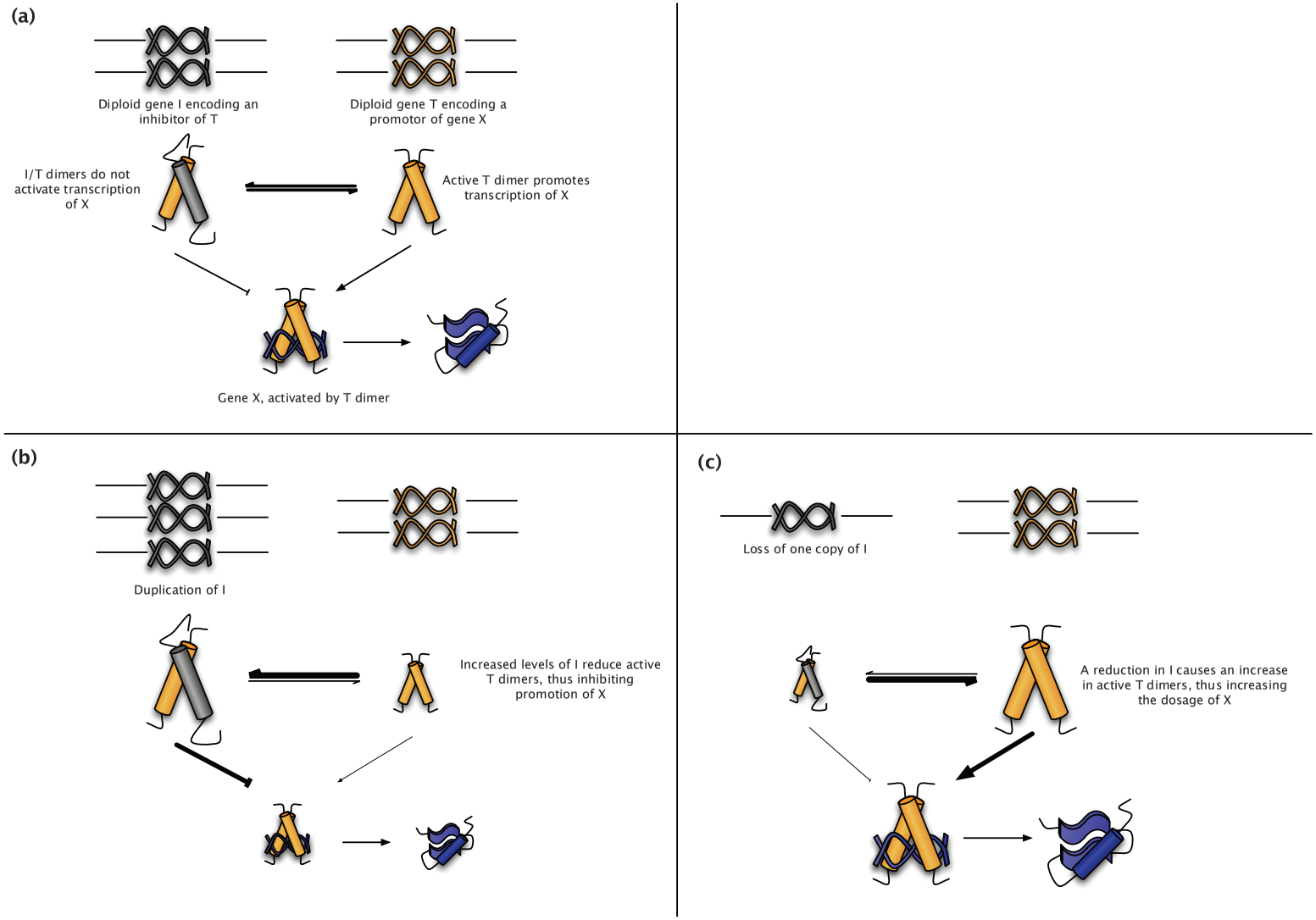

Figure 4.1: Schematic representation of a regulatory network controlling expression of gene X. Gene I encodes an inhibitor of transcription factor T. **(a)** In the normal, diploid state, I and T are in balance. **(b)** Duplication of I will cause lower than normal levels of X, denoted by smaller symbols and arrows. **(c)** Heterozygous deletion of I will lead to overactivation of T and thus to excess amounts of X, represented as large symbols. Figure adapted from Birchler *et al.* (2005).

It has been estimated that at least 2% of the human genome is affected by structural variations (Cooper *et al.*, 2007), such as inversions, small insertions/deletions or large copy-number variants (CNVs) (Conrad and Hurles, 2007). These sometimes large rearrangements may be seen as an important driving force of genome evolution. As a consequence, theories on gene evolution have to be re-evaluated in the context of such rapid and widespread large scale variation. Previous studies have already shown that the locations of CNVs and the function of genes inside CNV regions are biased (Cooper *et al.*, 2007; Nguyen *et al.*, 2006). CNVs are found more often in pericentromeric and subtelomeric regions, they overlap significantly with regions of segmental duplications and are more gene dense than the average for the genome. Genes within CNV regions are frequently involved in sensory perception and immune system activity, to a lesser extent in cell adhesion and in a number of cases signal transduction (Cooper *et al.*, 2007). Two theories have been postulated to explain this non-random distribution of CNVs. The *mutational hypothesis* states that most CNVs are in effect phenotypically neutral, but are carried by flanking genomic elements like ALU repeats which cause the bias in CNV distribution. The opposing theory could be called the *selection hypothesis*, stating that negative and positive selection shape the distribution of CNVs through the functional elements they encompass.

In this work, I use gene expression and copy-number variation data to study the relationship between protein complexes, dosage sensitivity and recent gene evolution in the human population. Firstly, I show that changes in gene copy number have a weak but measurable effect on gene expression. Next, I describe how genes involved in protein complexes are enriched for known dosage sensitive genes and exhibit substantially lower expressional noise than other genes. Consequentially, I observe that dosage sensitive genes tend to be underrepresented in CNV regions. Given these functional and positional biases on genes in CNV regions, I hypothesise that the regulatory balance of dosage sensitive genes exerts negative selective pressure on chromosomal structural

variations.

4.2 Methods

A wide range of diverse sources of data were combined in order to perform the analyses in this chapter. In the following paragraphs, I describe the provenance and composition of these different datasets. When no web URL is given, the data was extracted from supplementary materials files provided with the referenced publication.

4.2.1 Gene identifiers

A common problem when combining several independent data sets is inconsistencies in naming conventions. To assure that all gene identifiers were consistent, all data sets were mapped to the most recent HUGO Gene Nomenclature Committee (HGNC) identifiers in March 2008 (Bruford *et al.*, 2008). In case a gene name did not correspond to a primary gene symbol in HGNC, the HGNC *previous symbols* column was searched for an exact match, followed by a search in *aliases*. If no exact match could be found, the gene was removed from the set and not included in any further analysis.

4.2.2 Mammalian protein complexes

The CORUM database (Ruepp *et al.*, 2008) is a manually annotated resource, containing, at the time of writing, 1679 protein complexes from 10 mammalian species, with a strong focus on human. Entries are based on individual publications, not including high-throughput experiments. Table 4.1 lists Gene Ontology annotations for which CORUM deviates significantly from the rest of the genome. CORUM is enriched for nuclear proteins and contains a large number of transcriptional regulators. Conversely, extracellular and membrane proteins are underrepresented in the dataset. Figure 4.2 visually conveys an idea of the size distribution of this network of human complexes, as well

as reflecting its highly interconnected nature. Relationships for 2080 proteins in 1109 human complexes were downloaded from `http://mips.gsf.de/genre/proj/corum` on the 29th January 2008. 2028 proteins could be mapped to 1975 HGNC identifiers. Genomic coordinates for these gene identifiers were retrieved from Ensembl (v49) (`http://www.ensembl.org`) *via* BioMart.

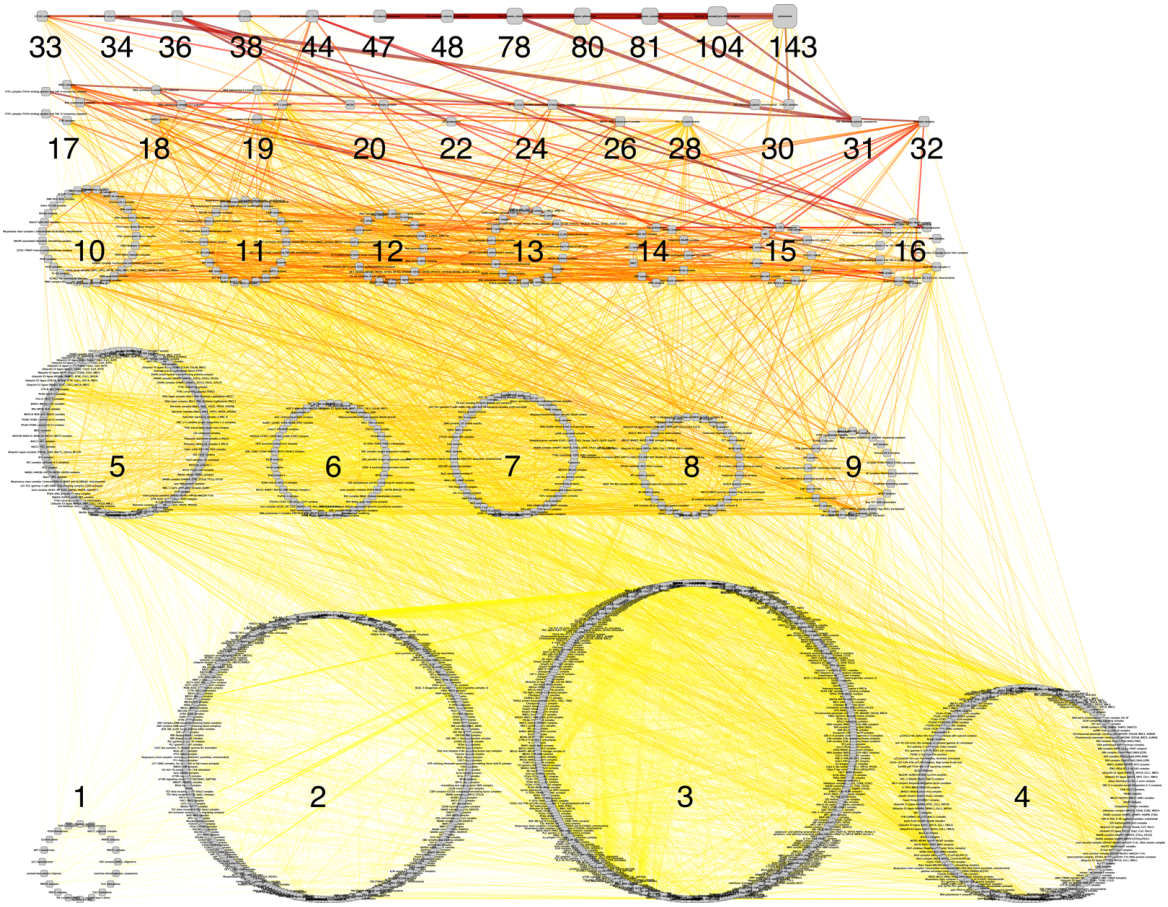

Figure 4.2: A network representation of the CORUM database. Nodes represent complexes and are ordered by number of unique components (shown as number next to groups). Edges denote shared components between complexes. The number of shared components is reflected in the colour (from yellow (few) to red (many) shared components) as well as in the line width. The large, highly overlapping complexes in the first row are mainly modules of the ribosome (6 out of 12) and spliceosome (3 out of 12). Other large complexes include RNA polymerase, respiratory chain complex and the proteasome. The group of complexes with only one member are homo-multimers.

Table 4.1: Composition of the CORUM database. Underrepresented terms are set in **bold font**. P-Values were calculated using Fisher's Exact Test, see Methods.

GO-Slim Term	Number of CORUM genes	P-Value
protein binding	1348	$1.78 \cdot 10^{-210}$
nucleus	1058	$3.73 \cdot 10^{-207}$
macromolecule metabolic process	1321	$1.59 \cdot 10^{-205}$
nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	852	$4.52 \cdot 10^{-148}$
nucleic acid binding	708	$5.73 \cdot 10^{-86}$
cytoplasm	933	$2.72 \cdot 10^{-62}$
regulation of biological process	722	$1.24 \cdot 10^{-51}$
chromosome	168	$7.95 \cdot 10^{-46}$
structural molecule activity	227	$5.51 \cdot 10^{-38}$
transcription regulator activity	301	$1.63 \cdot 10^{-30}$
biosynthetic process	279	$5.37 \cdot 10^{-26}$
helicase activity	53	$1.14 \cdot 10^{-15}$
cell death	146	$1.12 \cdot 10^{-12}$
protein transporter activity	45	$3.32 \cdot 10^{-11}$
response to stimulus	378	$3.42 \cdot 10^{-08}$
translation regulator activity	34	$2.29 \cdot 10^{-06}$
cell differentiation	232	$1.54 \cdot 10^{-05}$
extracellular region	77	$\mathbf{1.94 \cdot 10^{-06}}$
membrane	532	$\mathbf{3.35 \cdot 10^{-15}}$

4.2.3 Interaction and complex data

As an alternative to the manually compiled set of complexes in CORUM, an independent set of putative complexes was computationally derived from high-throughput protein interaction experiments by identifying highly connected clusters of proteins in an extended network of human protein interactions (Krogan *et al.*, 2006). Interaction data for three recent high-throughput studies (Ewing *et al.*, 2007; Rual *et al.*, 2005; Stelzl *et al.*, 2005) were retrieved from IntAct (Kerrien *et al.*, 2007) and subsequently

merged into a single network. As for CORUM, UniProt identifiers were mapped to HGNC identifiers to ensure consistency. This was achieved by extracting the HGNC annotations in the "cross-references" section of the UniProt flat-files. Clustering analysis was performed using the Markov clustering tool `mcl` (van Dongen, 2000) (parameter $I = 3.0$). The "alternative complex set" was defined as containing all clusters with more than three components (2325 unique genes).

4.2.4 Set of dosage sensitive genes

Dosage sensitive genes were extracted from the annotations of the Baylor College of Medicine Medical Genetics Laboratory 105k diagnostic Chromosomal Microarray (version 7), available at `http://www.bcm.edu/geneticlabs/cma/`. This post-natal screening tool comprises a manually compiled set of 146 genes (after mapping to HGNC) known to be sensitive to chromosomal imbalances (Cheung *et al.*, 2005). A complete list of the genes and the associated diseases can be found in Table H.1.

A separate set of genes overexpressed in cancer tissue was also used (Axelsen *et al.*, 2007). The dataset contains 2362 genes which are at least 4-fold overexpressed in brain (astrocytoma and glioblastoma), breast, colon, endometrium, kidney, liver, lung, ovary, prostate, skin, and thyroid cancers compared to healthy tissue of the same type.

4.2.5 Expression profiles

Gene expression can be measured on a large scale using expression arrays. Stranger *et al.* (2007) performed gene expression analysis on Eppstein-Barr virus transformed lymphoblast cell lines from each of the HapMap individuals. Gene expression was quantified using high-throughput human whole-genome expression arrays designed by Illumina (Kuhn *et al.*, 2004). These arrays consist of $\approx$ 48000 bead types, where each bead consists of several hundred thousand copies of a gene specific oligonucleotide probe. After RNA was extracted from the cell lines, it was carefully amplified and

labelled with Biotin-16-UTP. After hybridisation to the array, Cy3-streptavidin was applied to the array which binds to Biotin and subsequently allows the measurement of luminescence intensities for each bead type in a specially designed scanner. Kuhn *et al.* showed in a benchmark experiment that luminescence intensities are directly proportional to the expression strength within a defined dynamic range (Limit of Detection: ≈ 0.13pM, dynamic range: ≈ 3.2-fold). Each bead type is also replicated several times on the array, thus providing robustness and redundancy for quality control. Subsequent to data readout, the raw intensities for each redundant bead type were summarised by proprietary software provided by Illumina. Stranger *et al.* performed 4 replicate hybridisations per cell line, the results of which were summarised on a log scale using a quantile normalisation method across replicates of a single individual, followed by a median normalisation method across all 270 individuals. The resulting data, consisting of a matrix of gene expression values of 47293 probes over 270 individuals, were downloaded from `http://www.sanger.ac.uk/humgen/genevar/`.

Due to the sensitivity and dynamic range limitations of the Illumina WG6 expression arrays used by Stranger *et al.*, there is a correlation between detectable expression variation and total expression strength for genes with low overall expression, or no expression at all. Notably, there is a cluster of genes with both low detected expression and markedly lower coefficients of variation (CV, defined as the standard deviation of expression between individuals per gene, normalised to the mean absolute expression level) than the majority of genes, plotted in grey in Figure 4.3. These genes may be distinguished from the remaining genes by their lower absolute variation, that is the standard deviation between individuals before normalisation to the expression mean. In total, 6440 genes with an absolute population standard deviation ≤ 7 were removed from the dataset, as they are likely to be expressed below the confident detection threshold or not to be expressed at all.

A second set of expression data for 44760 probes applied to samples from 79 different

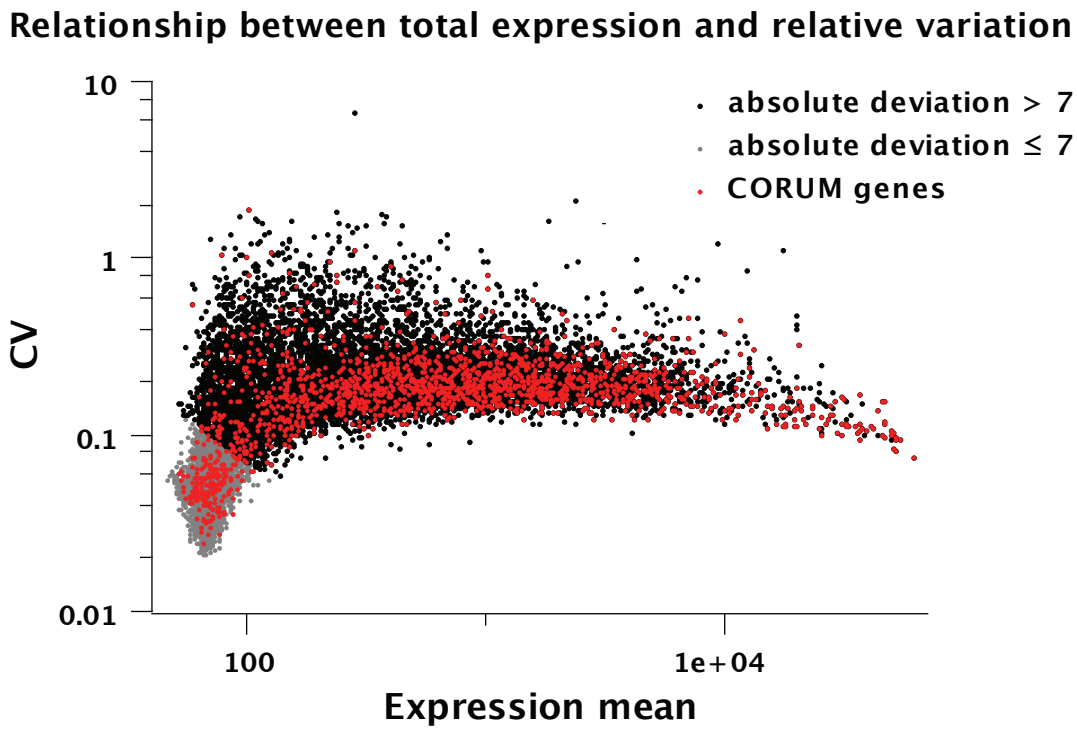

Figure 4.3: Coefficients of gene expression variation (CV) relative to absolute expression level. The measurable variation in gene expression is limited by the sensitivity of the employed array technology. Genes which are expressed at extremely low levels, or not expressed at all, cluster in the low expression/low CV region. Shown in grey are genes which were excluded from further calculations (standard deviation ≤ 7).

tissue types were provided by GNF SymAtlas (Su *et al.*, 2004) (`http://symatlas.gnf.org`). For the latter, different Affymetrix expression arrays were employed, raw results of which were normalised using global median scaling.

Probe identifiers for both data sets were mapped to HGNC gene names through Ensembl BioMart. Probes which could not be mapped to a gene name were exluded from further analysis. The resulting matrices contained expression data for 17122 genes (HapMap set) and 15012 genes (tissue set), respectively.

4.2.6 Correlation computation

As a measure of correlation between expression levels of two genes in different tissues/individuals, the Pearson product-moment correlation coefficient was employed. For two vectors x and y representing genes with n expression levels, the correlation r_{xy} is given by

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{(n-1)s_x s_y} \tag{4.1}$$

where $\overline{x}$ and $\overline{y}$ are the means and s_x and s_y are the standard deviations of x and y, respectively. For complexes with more than 2 components, correlations for all $n(n-1)/2$ combinations of gene pairs were averaged.

4.2.7 Copy-number variations

Chromosomal locations of variations relative to the NCBI36 human genome assembly were downloaded from the Database of Genomic Variants (DGV) (Iafrate *et al.*, 2004): `http://projects.tcag.ca/variation/`. This data also contains information on number of individuals and gain/loss annotation per CNV. CNV locations and whole genome tiling-path (WGTP) array hybridisation values for each HapMap individual were downloaded from `http://www.sanger.ac.uk/humgen/cnv/data`. The distribution of CNVs on selected human chromosomes is shown in Figure 4.4.

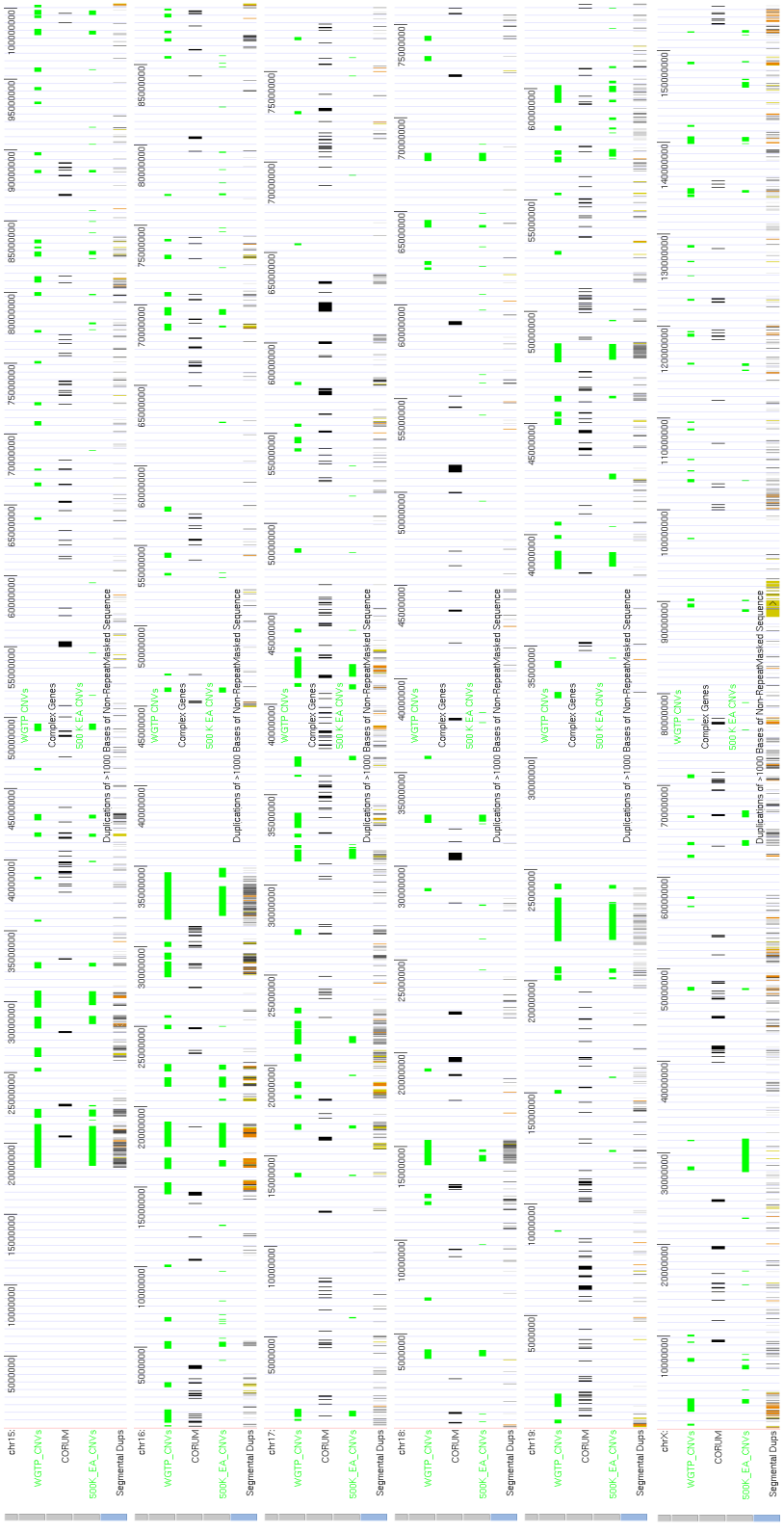

Figure 4.4: Position of CORUM genes (black), copy-number variants (green) and segmental duplications (shades of orange) on 5 autosomes and the X chromosome from human. CNVs from Redon *et al.* were derived by two different methods: WGTP and 500k array, which are shown separately. Graphics generated with the UCSC Genome Browser (Kent *et al.*, 2002)

4.2.8 Segmental duplications

Human segmental duplications of $\geq 90\%$ sequence identity and ≥ 1 kilobase length were provided by the segmental duplication database (She *et al.*, 2004) (`http://humanparalogy.gs.washington.edu`).

4.2.9 Gene Ontology analysis

181651 Gene Ontology (GO) annotations for 34591 human UniProt entries were provided by the GOA project (Camon *et al.*, 2004), available at `http://www.ebi.ac.uk/GOA/`. UniProt enries were mapped to HGNC identifiers through BioMart, resulting in 16213 annotated HGNC gene identifiers. There were 6775 unique GO terms in the full GOA dataset. The complexity of this hierarchical data structure was reduced by mapping GO terms to 64 GO-slim categories as defined by the GOA project themselves (`ftp://ftp.ebi.ac.uk/pub/databases/GO/goa/goslim/`).

4.2.10 Identification of paralogs

In-species paralogs for 10755 HGNC gene identifiers were downloaded from Ensembl Compara via BioMart. The paralog prediction uses automatically generated phylogenetic trees of all species in the Ensembl database. According to the Ensembl compara help website (`http://www.ensembl.org/info/about/docs/compara/homology_method.html`), the algorithm to identify orthologs comprises the following steps:

1. Align all pairs of full-length protein sequences of the longest transcript of two genes from two species using WUBlastp and subsequent Smith-Waterman.

2. Cluster genes by single-linkage clustering according to Best Reciprocal Hits and Best Score Ratio.

3. Create a multiple sequence alignment (MSA) for each cluster using MUSCLE (Edgar, 2004).

4. For each MSA, calculate a phylogenetic tree using TreeBeST (`http://treesoft.sourceforge.net/treebest.shtml`) and infer orthology and paralogy. TreeBeST in this case combines 5 tree building methods (maximum likelihood on protein and codon sequences *via* `phyml` (Guindon and Gascuel, 2003) and neighbour-joining on p-distance as well as dN and dS distances) and calculates a consensus tree.

4.2.11 Analysis of selection pressure

dN/dS values for human genes relative to mouse orthologs were acquired from Ensembl via BioMart. The calculation of dN/dS values is part of the automatic gene tree generation described above: dN/dS values are generated by `codeml` (model=0, NSsites=0) from the PAML package (Yang, 1997) for all genes from closely related species after the initial tree generation. In this analysis, only genes with a single unique ortholog in mouse were used in the analyses.

4.2.12 P-Values

Statistical significance of overlaps between gene sets was computed with Fisher's exact test (FET). The Mann-Whitney-U test (MWU) was employed to determine significance of differences between two distributions. In cases of multiple testing, Bonferroni correction was applied. All calculations were performed in R (R Development Core Team, 2006). Significance of differences in dN/dS ratios was calculated by random resampling: For the null hypothesis, 1000 sets of genes with identical size as the test set were each created by randomly drawing without replacement from the complete gene set. P-Values were calculated as the probability of observing a result at least as extreme, given the normally distributed null model derived from the resampling.

4.3 Results

In order to investigate the relationship between copy-number state, protein complexes and dosage, I need to assert several preconditions. Firstly, I investigate the impact of copy-number change on gene expression. Secondly, I analyse the relationship between protein interactions and dosage sensitivity. Finally, I combine these points to describe the effects of dosage sensitivity of protein complexes on the evolution of chromosomal structural variations.

4.3.1 Effects of CNVs on gene expression

Association studies (Stranger *et al.*, 2007) have shown both *cis* and *trans* effects of copy-number variations (CNVs) on genes. Stranger *et al.* also measured the relative contribution of single-nucleotide polymorphisms (SNPs) and CNVs on the observed variation in gene expression. They report that 83.6% of variation can be attributed to SNPs, whereas 17.7% of variation is associated with CNVs. However, the study was designed to identify associations between all genes and CNVs within a 2 million base-pair (MB) window simultaneously and thus had to use stringent multiple-testing correction. While Stranger *et al.* report 238 genes to be associated with a CNV within a 2MB window, it is not immediately clear what immediate effects CNVs have on contained genes, and whether there is a distinguishable effect between deletion and duplication polymorphisms.

I therefore focused my attention on the relationship between copy-number variations and gene dosage. I combined gene expression data derived from lymphoblast cell lines of 270 HapMap individuals (Stranger *et al.*, 2007) with the CNV dataset of Redon *et al.* (2006) on the same individuals.

I find that duplications and deletions have distinguishable profiles of expression ratios, see Figure 4.5. The expression ratio is defined as the average expression of a gene

in individuals with a CNV phenotype, divided by the average expression in unaffected individuals. Assuming a simple linear relationship between copy-number and expression level, one would expect a distribution with peaks at 0.5, 1 and 1.5, corresponding to a heterozygous deletion, balanced expression and heterozygous duplication, respectively. The observed distribution shown in Figure 4.5 reflects a more complex relationship.

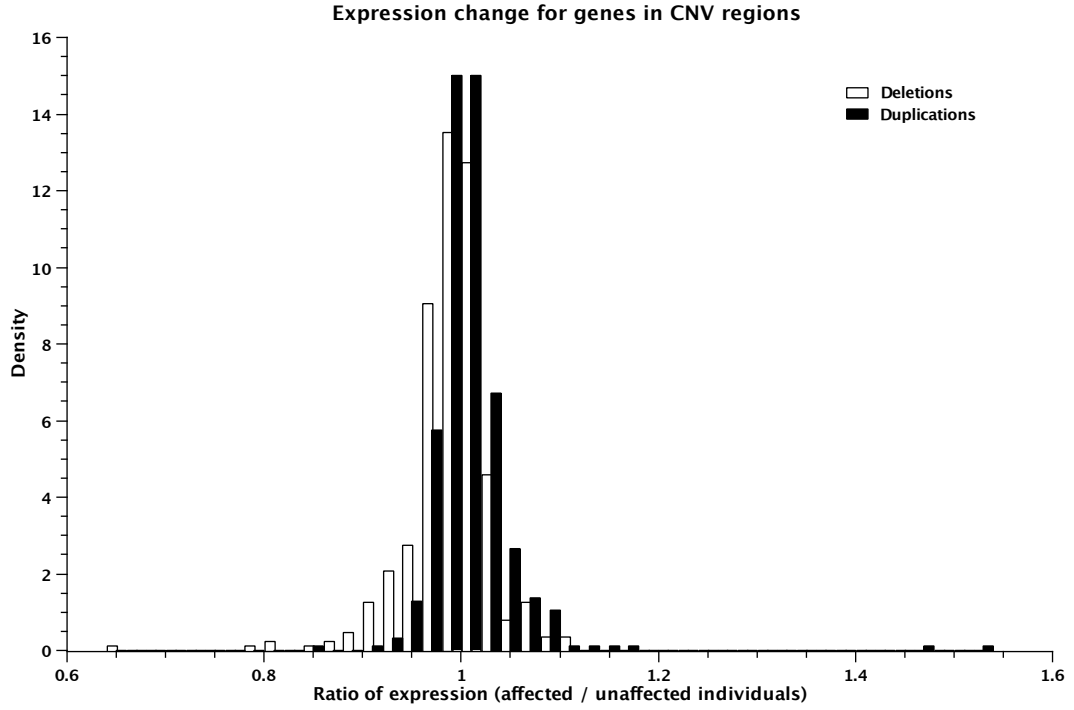

Figure 4.5: Difference between deletion (white) and duplication (black) variations in HapMap individuals. The histograms show the ratio of average expression levels between affected and unaffected individuals for all genes inside a copy number varied region. The shift between the two distributions is significantly larger than would be expected by chance (MWU: $P = 1.22 \cdot 10^{-11}$).

The magnitude of the expression difference between CNV and wild type individuals is smaller and more continuous than expected. However, the location shift between the two distributions is highly significant (MWU: $P = 1.22 \cdot 10^{-11}$). This indicates that deletions reduce gene expression, while duplications tend to increase expression.

As mentioned in the Methods, sensitivity and dynamic range of the expression arrays could partly account for the observed noise, but I did not find a correlation between absolute gene expression level and ratio of expression difference for genes overlapping CNV regions (Figure 4.6).

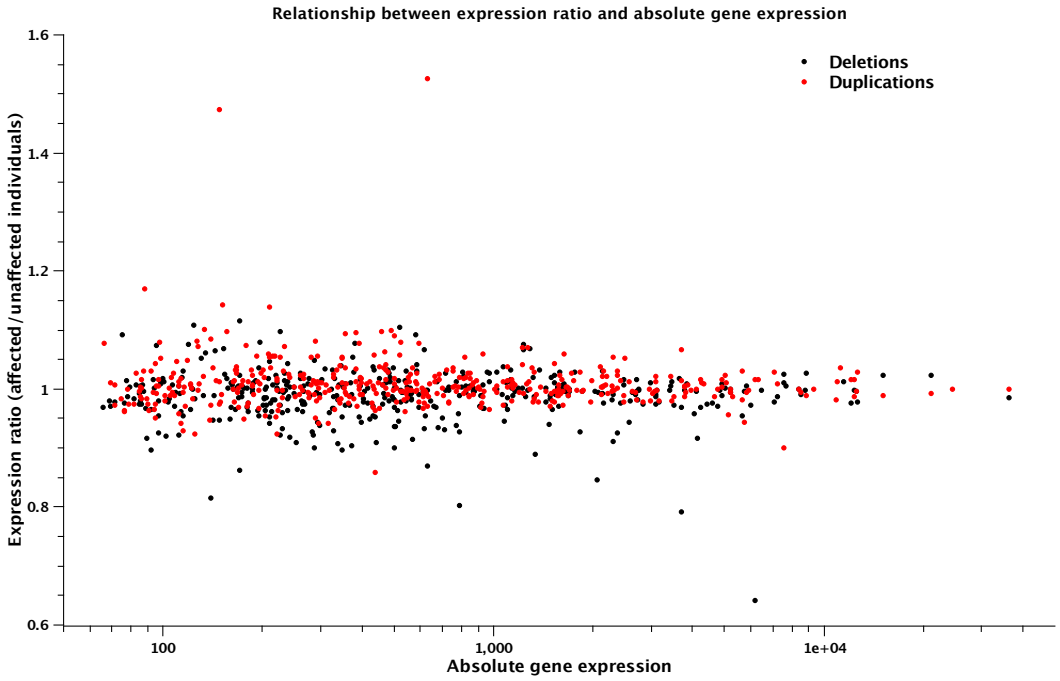

Figure 4.6: Relationship between effect of CNV on gene expression and absolute expression levels. The horizontal distribution suggests that there is no discernible correlation between absolute gene expression and expression ratio. A positive or negative correlation between absolute detection level and the fold expression change between affected and unaffected individuals could indicate a measurement-sensitivity induced bias, but within the analysed data no such relationship is detected.

The expression ratio distribution reflects a summary over a wide range of individuals. To elucidate the effects of CNVs on gene expression on a per-individual basis, I plotted the logarithm of hybridisation strength on the genomic hybridization arrays relative to the reference individual ($\log_2^H$) against the logarithm of expression, relative to the reference individual ($\log_2^E$). As a positive control, I compared two X-chromosomal

genes, one being inactivated (L1CAM, Figure 4.7a), the other being known to escape X-inactivation (UTX, Figure 4.7b). The latter exhibits a marked increase in expression in female individuals relative to the (male) reference individual. In contrast, L1CAM maintains equivalent expression in males and females levels due to inactivation of one gene copy in females.

I found 94 gene duplications and 98 gene deletions where the average log_2^H and log_2^E are at least one standard deviation below (deletions) or above (duplications) the mean of the unaffected individuals. Figures 4.7c and 4.7d show two examples of genes inside frequent CNVs exhibiting induced dosage effects. Deletions and duplications have clearly distinguishable expression levels. Notably, though, the expression ratios of the deletion/duplication individuals overlap with the expression ratios of unaffected individuals. In other words, CNVs only partly account for the differences in expression between individuals, while a large portion of the variance must stem from other sources. Figures 4.7e and 4.7f show two examples of rarer CNVs which also show a clear deviation of log_2^H and log_2^E relative to the majority of unaffected individuals.

Notably, several individuals were not called as CNVs, despite similar log_2^H and log_2^E ratios in the analysed region as the identified CNV individuals. These putative false negatives will reduce the magnitude of expression ratios between CNV and unaffected individuals. Summarising these individual effects leads to the conclusion that duplications and deletions have a measurable effect on gene expression, even though they are just one source of expression variation amongst others.

4.3.2 Limited expressional noise of protein-complex genes

It has previously been reported that expression levels of proteins within a complex are significantly more correlated across tissue types than would be expected by chance (Hahn *et al.*, 2005; Jansen *et al.*, 2002). Using both the expression from HapMap individuals mentioned above as well as a tissue-specific gene expression dataset, I verify

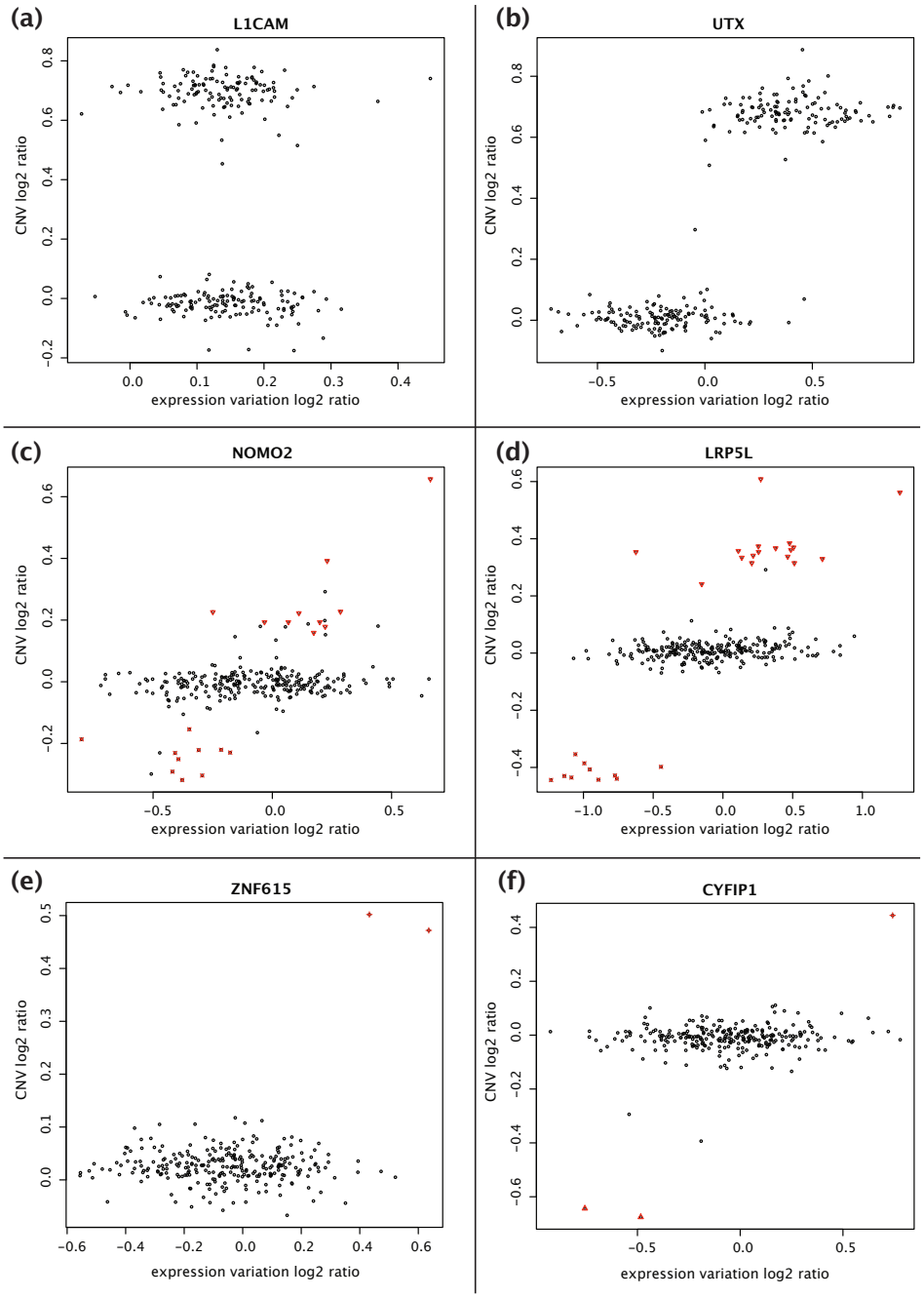

Figure 4.7: Ratio of WGTP array hybridisation intensity over relative expression level for four example genes. **(a)** L1CAM and **(b)** UTX. The increase in expression as a result of the copy-number increase in females is clearly visible for UTX which is known to escape X-inactivation. **(c)** and **(d)** Examples of autosomal genes with common CNV polymorphisms. Red crosses denote individuals in which a deletion phenotype has been called by Redon *et al.*, red triangles denote duplications. The plot highlights several potential false negatives with similar expression and hybridisation strength as the called deletions/duplications. Non-CNV related expression variation is substantial. **(e)** and **(f)** Examples of rare CNV genotypes with significant expression change.

that members of complexes from the CORUM database exhibit increased expression correlation (Figure 4.8).

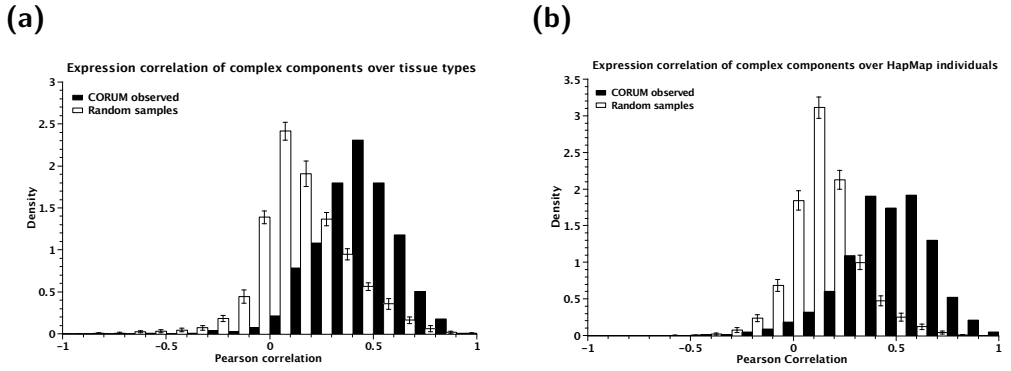

Figure 4.8: Distribution of average Pearson correlation coefficients between all members of known protein complexes as defined in CORUM (black), and randomly sampled proteins (white, N=100). **(a)** Expression intensities from 79 tissue types of different individuals. **(b)** Expression intensities from lymphoblast cell lines of 270 HapMap individuals.

In addition to that, the HapMap expression data allow me to perform a direct comparison of expression levels between individuals. I calculated coefficients of variation (CV), defined as the standard deviation of expression between individuals per gene, normalised to the mean absolute expression level. These values represent a dimensionless magnitude of variation for each gene. The CVs are significantly lower for CORUM genes than for the rest of the genome (MWU: $P = 2.67 \cdot 10^{-10}$), see Figure 4.9a/b. Interestingly, the average CV of genes within one complex decreases with the size of the complex, as shown in Figure 4.9c. This is independent of the mean absolute expression per gene, as shown in Figure 4.9d. I asserted that this effect is not a sampling artefact: When splitting all CORUM genes into sets with complexes of size ≥ 10 and size < 10 and comparing the distribution of CVs, it emerges that small complexes possess higher CVs (MWU: $P < 2.2 \cdot 10{-}16$). These results indicate that members of protein complexes are not just more likely to maintain relative expression levels between tissue types, but they are also more restricted as to their expression variation between

individuals within the same tissue.

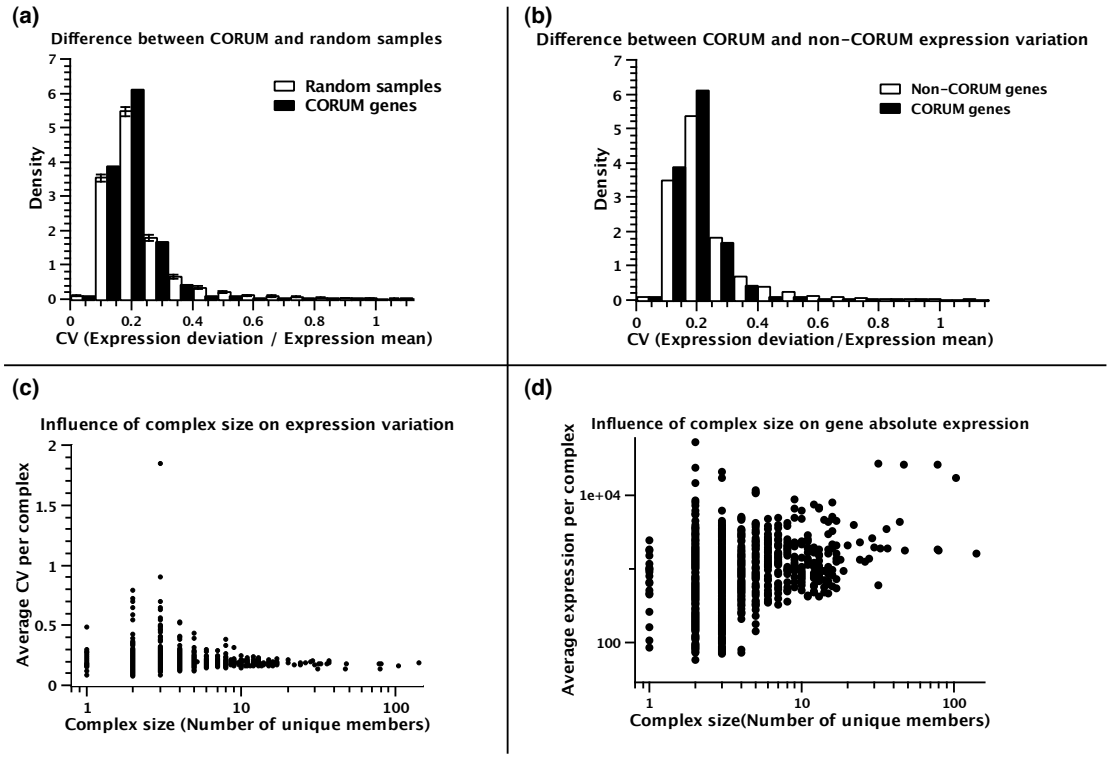

Figure 4.9: Coefficients of gene expression variation (CV) vary between CORUM and non-CORUM genes. **(a)** CORUM genes have significantly lower CVs than random sets of genes. **(b)** CORUM genes have significantly lower CVs than non-CORUM genes. Outliers beyond 1.4 are not shown. **(c)** Large CORUM complexes exhibit lower average CVs of their members. **(d)** Low absolute expression is not the reason for the lower noise in large complexes: mean absolute expression of large complexes is above average.

CORUM is a manually curated data source and thus prone to ascertainment bias. To ensure that these results are not biased by the composition of CORUM, I generated a separate dataset of putative protein complexes extracted from several high-throughput protein interaction detection experiments (see Section 4.2.3). The clusters represent an alternative set of "complexes" composed of 2325 proteins, 505 of which are also contained in CORUM. The CV distribution difference between these highly interacting proteins and the rest of the genome is also skewed towards lower CVs ($P = 7.0 \cdot 10^{-3}$).

This suggests that highly connected proteins in general avoid imbalances in protein expression.

Is there evidence that tight control of gene expression is actually relevant for human disease? Axelsen *et al.* (2007) compiled a list of 2362 genes which are overexpressed in various cancer tissues (see Section 4.2.4). I tested whether these cancer related genes are enriched for dosage sensitive genes, under the assumption that dosage sensitive genes are more likely to be causal in these diseases. In fact, I find that CORUM genes are overrepresented in these cancer related genes (356 genes, FET: $P = 6.56 \cdot 10^{-13}$). The fact that the tight regulation of expression of CORUM genes is disturbed in cancer tissue provides an interesting link between cancer, protein complexes and dosage sensitivity.

4.3.3 Dosage sensitive genes and CNVs

I have so far assembled evidence that protein complexes seem to be under constraint to maintain their relative expression levels and show limited expression variability between individuals. For the further analysis of dosage sensitivity, I also used an independently assembled set of 146 genes with known dosage-related disease phenotypes (see Section 4.2.4). There is a significant overlap between CORUM and this set of dosage sensitive genes (32 genes, FET: $P = 1.2 \cdot 10^{-5}$), further supporting the link between dosage sensitivity and protein complexes.

As previously stated, I found that CNVs can affect the expression levels of genes they contain. I therefore hypothesised that a CNV that encompasses a gene which is part of a protein complex will be more likely to have a negative effect on fitness. As the Redon *et al.* CNV data were derived from healthy individuals, I expect that genes encoding protein complexes will be underrepresented in CNV regions.

Out of 18534 protein coding genes for which both genomic locations and a unique gene name could be retrieved, 2311 genes are fully inside a CNV region. From 1975 proteins in the CORUM database, only 165 are found in a CNV region, significantly

fewer than one would expect by chance (FET: $P = 3.5 \cdot 10^{-10}$). The set of automatically clustered complexes were also underrepresented in CNV regions (256 out of 2325 genes, $P = 0.012$). Lastly, both the set of 146 dosage sensitive genes (8 genes inside CNV, $P = 4.7 \cdot 10^{-3}$) as well as the 2362 genes overexpressed in cancer (246 genes inside CNV, $P = 5.82 \cdot 10^{-4}$) are unlikely to be contained in CNV regions.

Nguyen *et al.* as well as Cooper *et al.* reported a highly significant depletion of genes with the Gene Ontology (GO) category "binding" within CNV regions, but they do not comment further on this fact. I verified independently that "binding" is the second most underrepresented GO category after "intracellular" amongst genes in CNV regions. This lends further support to the hypothesis that dosage sensitivity due to protein complex membership has an influence of the composition of CNV regions.

I speculated that a negative fitness effect due to a copy-number variation will increase the likelihood of subsequent removal of that CNV from the gene pool. The CNVs that contain CORUM genes occur in significantly fewer individuals (MWU: $P = 1.6 \cdot 10^{-4}$) than non-CORUM genes, indicating that purifying selection may have acted on some of the genes.

I also tested whether CORUM genes are underrepresented in gains compared to losses. Out of the 167 CORUM genes that overlap a CNV, 18.5% occur in a gain, compared to 29.8% of non-CORUM genes. This significant difference in ratios (FET: $P = 9.6 \cdot 10^{-4}$) suggests that amongst copy-number varied genes, there is indeed a bias against duplications for genes in protein complexes, supporting the notion that stoichiometric imbalance has a negative effect on protein complexes.

4.3.4 Compositional bias of copy-number varied genes

Various compositional biases on genes in CNV regions have been described (Cooper *et al.*, 2007; Nguyen *et al.*, 2006). Most notably, it has been reported that genes within CNV regions exhibit higher dN/dS than the rest of the genome. Is the observed low

frequency of CORUM and other dosage sensitive genes in CNV regions merely a result of a bias against slower evolving genes? I verified that dN/dS ratios of genes within CNV regions were elevated compared to their mouse orthologs (Median: 0.131, P-Value by resampling: $P = 3.2 \cdot 10^{-7}$). Conversely, CORUM genes exhibit lower than expected dN/dS (Median: 0.070, $P < 10^{-40}$). In contrast to non-complex genes, there is no significant difference in dN/dS between CORUM genes that overlap CNVs and those that do not. I therefore tested whether there is a causal relationship between complex membership, low dN/dS and CNV overlap.

Like CORUM genes, the automatically clustered complexes also exhibited low dN/dS (Median 0.08, $P = 1.9 \cdot 10^{-30}$). It has been argued that proteins with obligate interactions are under stronger selective pressure (Mintseris and Weng, 2005), which could explain the low dN/dS in both CORUM and the automatically clustered complexes. Interestingly, Cooper *et al.* showed that CNVs and segmental duplications (SDs) are of fundamentally similar nature and frequently overlap. I thus hypothesised that the reduction in negative selection within CNVs is related to the higher copy number of some genes which have been recently duplicated in a fixed SD. If I split the genes in CNV regions into those that overlap a SD and those that do not, it can be measured that dN/dS ratios are highly significantly elevated in the genes that overlap SDs (MWU: $P < 2.2 \cdot 10^{-16}$), but not in the group outside SDs ($P = 0.017$).

Subsequently, I analysed the distribution of numbers of paralogs for human genes. I found that genes in CNV regions have significantly more paralogs than would be expected by chance (MWU, $P = 1.45 \cdot 10^{-9}$),whereas genes from CORUM have significantly fewer ($P < 2.2 \cdot 10^{-16}$). As with the evolutionary rate, the increase in numbers of paralogs is largely driven by CNVs that overlap SDs. Removing all genes inside SDs reduced the number of paralogs substantially (P-value reduced from $1.45 \cdot 10^{-9}$ to 0.0033). Conversely, the genes that are in both CNVs and SDs have significantly more paralogs than genes only found in CNV regions ($P = 4.3 \cdot 10^{-11}$). I conclude that

the increase in dN/dS in CNV regions is driven by an increase in gene copy number and thus does not explain the underrepresentation of dosage sensitive genes in CNV regions.

If SDs are largely responsible for the increased dN/dS within CNVs and the increase in number of paralogs, can I still detect the underrepresentation of CORUM genes in CNVs that do not overlap a SD? After removing all genes that overlap a SD, CORUM genes were still significantly underrepresented ($P = 3.3 \cdot 10^{-4}$) in CNV regions, indicating that negative selective pressure not only affects regions of segmental duplication but also other types of CNVs.

4.4 Discussion

4.4.1 Protein complexes are sensitive to alterations in gene expression

Correlated gene expression of interacting proteins is a well known phenomenon, to the extent that correlation analysis is used to validate high-throughput protein interaction experiments (Hahn *et al.*, 2005). Usually, expression data is gathered under diverse physiological conditions, *e.g.* at different stages of the cell cycle. In this analysis, I have compared data from 79 different human tissue types. As expected, I observe strong correlation between the changes in gene expression for members of the same protein complex in different tissues. This observation hints at the importance of tightly regulated gene expression for the correct functioning of protein complexes.

However, it does not directly verify if the stoichiometry of complexes is under the same strong regulation. I therefore measured the variation in expression levels for interacting proteins in different HapMap individuals. Expressional noise of protein complexes has been analysed in *S. cerevisiae* and *D. melanogaster*(Lemos *et al.*, 2004), but the HapMap gene expression data allow the first systematic evaluation of protein complex expression in human. I find that genes in CORUM exhibit significantly

lower variation in expression than the rest of the genome. This is direct evidence that expression of complex genes is under tighter regulation than the rest of the genome. Furthermore, I find that genes in large complexes maintain particularly low expression variation. While I cannot rule out that this observation is due to functional constraints on the particular complexes, it does suggest that sensitivity to expressional noise is related to the number of subunits a complex maintains.

When I analysed the composition of genes in CNV regions, I made the curious observation that the small number of CORUM genes that overlap a CNV (165 genes in total) are biased towards deletions rather than duplications. If I assume that negative selection is acting on CNVs, the intuitive biological explanation for this phenomenon would be that CORUM genes are at least as sensitive to duplication as to deletion, which in turn supports the concept that members of protein complexes are sensitive not just to under- but also to overexpression.

I made another observation that supports this hypothesis. When comparing a manually curated set of dosage sensitive genes derived from the scientific literature, I found that a significantly larger than expected proportion of these genes were members of a protein complex as defined by the CORUM database. Taken together, these findings indicate that stoichiometric fluctuations negatively affect protein complexes.

4.4.2 CNVs affect expression levels of contained genes

A key proposition that underpins our understanding of dosage sensitivity is that duplication or deletion of the genomic region containing a gene will result in a significant up- or downregulation of expression of the gene. There have been previous reports of widespread expressional silencing of chromosomal amplifications (Platzer *et al.*, 2002). In contrast, I observed lower average gene expression in deletion CNVs compared to duplication CNVs (Figure 4.5). It has to be noted, though, that these differences in expression are small for the majority of genes within a CNV. Furthermore, there are

numerous cases where deletions seemingly result in increased expression and vice versa. Figures 4.7c and 4.7d exemplify how noisy the expression data for a gene can be, despite a visible expression difference between deletion and duplication genotypes. Sensitivity to detect expression differences at low concentration is not the main source of this variability in gene expression. Rather, I suspect there to be inherent fluctuations between the different cell lines used in the analysis (Blake *et al.*, 2003). Expressional noise alone does not explain that some CNVs seem not to affect gene expression at all. Rather, the inaccurate prediction of start and end coordinates of CNVs is likely to be largely responsible for the lack of correlation between CNVs and gene expression. Individuals with a CNV genotype falsely labelled as unaffected, or a gene erroneously placed inside a CNV, will skew the distribution of expression ratios.

I speculate, however, that there could also be a physiological explanation for the unexpectedly low change in gene expression upon copy-number variation. It is conceivable that the cell attempts to compensate changes in copy number on gene expression by *e.g.* increasing or decreasing transcription or modulating mRNA degradation. Such autosomal dosage compensation was first observed in *D. melanogaster* (Devlin *et al.*, 1982) and a general mechanism for dosage regulation has been proposed (Birchler *et al.*, 2005). According to this theory, dosage balance is achieved through a network of regulatory genes which themselves are therefore dosage sensitive. The enrichment of CORUM for regulatory and transcription related functions might thus explain its sensitivity to copy-number variation and the low effect of CNVs on gene expression at the same time. Interestingly, Kind *et al.* (2008) recently described the formation and binding properties of a dosage-regulatory complex in *D. melanogaster*. They note that the components of the complex are not only conserved in mammals, but there is also autosomal activity of the respective proteins which is not fully understood. With the arrival of new CNV datasets featuring improved breakpoint accuracy, it should become possible to better distinguish between false positive predictions and genes that are actually sub-

ject to dosage compensation. Subsequently, this will make it possible to determine the frequency of autosomal dosage compensation of copy-number varied genes.

4.4.3 CNVs as the source of recent duplications

It has been noted (Nguyen *et al.*, 2006) that genes within CNV regions exhibit higher than expected dN/dS ratios, suggesting a relaxation of selective pressure. On the contrary, complex genes, dosage sensitive genes and highly connected genes in general, show very low dN/dS ratios, irrespective of whether they overlap CNVs or not. Stronger selective constraints in highly connected proteins have previously been attributed to functional constraints on the protein surface in order to maintain multiple binding sites (Mintseris and Weng, 2005).

Interestingly, I also show that genes in CNV regions have significantly more paralogs than expected by chance, while genes in protein complexes possess, on average, fewer paralogs (Yang *et al.*, 2003). This suggests that CNV regions have been hot-spots of large scale variation for a prolonged period of time, as it has also been shown that gene-rich CNV regions correspond well with regions of segmental duplications (Cooper *et al.*, 2007). In fact, I found that those CNV regions that overlap segmental duplications are primarily (though not exclusively) responsible for the high number of paralogs.

Conversely, the reason for the increase in dN/dS in many genes within CNV regions could be attributed to their higher number of paralogous sequences: Even a partial relaxation of selection pressure due to an additional gene copy is likely to increase the observed dN/dS ratios. In fact, genes in CNVs overlapping segmental duplications are again primarily, but not exclusively, responsible for the elevated dN/dS ratios. These observations underline that CNV regions are a frequent source of gene duplicates which occasionally get fixed over the course of evolution and thus drive evolution of some gene families.

4.4.4 Dosage sensitivity and negative selection on CNVs

I observed that CNV regions are less likely to contain genes encoding protein complexes, as well as other dosage sensitive genes. Furthermore, CNVs which occur in multiple individuals and can thus be assumed to be older than unique CNVs are particularly depleted of CORUM genes. Hence, it appears that pressures on correct dosage limit the set of genes which can sustain variation in copy-number, even though the effect of CNVs on gene expression is not straightforward.

Dang *et al.* (2008) reported that haploinsufficient genes are seldom found between two regions of segmental duplication. These results shed new light on this finding: It seems that dosage sensitive genes in general are biased against regions in which they are prone to suffer from copy-number variation. Segmental duplications are the most common source of such rearrangements, however I show that other CNVs not related to segmental duplications are also depleted of dosage sensitive genes. This indicates that rearrangements due to CNVs are subject to negative selection.

These findings offer a partial but consistent explanation for the biased composition of CNV regions. In addition to that, the correlation between dosage sensitivity and protein complex membership provides a convenient way to predict which genes are likely to be important in diseases which involve genomic rearrangements. The enrichment of CORUM for genes upregulated in cancer clearly hints towards this possibility. Future investigations should focus on the involvement of CNVs of putative dosage sensitive genes in cancer and complex diseases.