PhD degree in Molecular Medicine

European School of Molecular Medicine (SEMM),

University of Milan and University of Naples "Federico II"

Faculty of Medicine

Settore disciplinare: BIO/10

Evolution Of Protein Interaction Networks Through Gene Duplication

Matteo D'Antonio

IFOM-IEO Campus, Milan

Matricola n. R07945

Supervisor: Dr. Francesca Ciccarelli

IFOM-IEO Campus, Milan

Anno accademico 2010-2011

Table of Contents

Table of Contents 1
Figure List 4
Table list 7
Acknowledgements 8
List of abbreviations 10
Abstract 12
Introduction 14
1. Aim of the Thesis 14
2. Biological networks 15
2.1. Protein-protein interactions 16
2.2. False positives in the detection of protein-protein interactions 20
2.3. Databases of protein interaction networks 22
2.4. Genetic interactions 22
2.5. Network classification and characteristics 25
3. Orthology and paralogy 29
3.1. Methods to detect orthology 31
3.2. Methods to identify paralogs 32
3.3. Evolution of paralogs 33
4. Gene duplication and genome evolution 35
4.1. Duplicability and essentiality 37
5. Duplicability and protein interaction networks 39
6. Cancer as a genetic disease 40
7. The hallmarks of cancer 44
7.1. Enabling characteristics of cancer 46
8. Known and candidate cancer genes 47
8.1. Cancer Gene Census 47
8.2. Novel candidate cancer genes 50
8.3. Methods to identify driver mutations 51
9. Systems-level properties of cancer genes 56
Methods 59
1. Identification of unique gene sets 59
2. Identification of cancer genes 62

3. Reconstruction of protein interaction networks 65
4. Orthology and paralogy assignment 68
5. Evolutionary origin, conservation and duplicability 71
6. Comparison of gene and network properties 74
7. Functional analysis 76
8. Identification of ohnologs 78
9. miRNA targets 79
10. Tissue-selectivity of human genes 80
11. Analysis of the mechanisms that control duplicated hubs 82
12. Identification of recent paralogs of cancer genes 82
13. Tools 83
Results 85
1. Gene and network properties undergo modifications during evolution 85
1.1. Origin distribution, conservation and duplicability change in evolution 85
1.2. Networks have different levels of completeness 91
1.3. Human duplicated genes encode highly connected and central proteins 95
2. An ancient network core is conserved in all species 98
2.1. Old proteins are highly connected and central 98
2.2. Conserved proteins are highly connected and central 103
2.3. A randomization test confirms the relationships between evolutionary and network properties 106
2.4. A core of singleton hubs is conserved in evolution 108
3. A novel group of duplicated hubs is acquired in the human network 109
4. Ancient and recent human hubs are involved in different functions 112
5. Gene dosage of human duplicated hubs is regulated through three different mechanisms 134
6. Dominant and recessive cancer genes are representative of ancient and recent human hubs 137
7. Cancer genes are neighbors in the human protein interaction network 140
8. Cancer genes are depleted in highly conserved paralogs 143
Discussion 148
Appendix 1 – Network of Cancer Genes: a web resource for integration and analysis of gene and network properties of cancer genes 159
1. Database description 159
2. Web interface 163
Appendix 2 – The human genetic interaction network 167

1. Introduction 167
2. Genetic interactions between cancer genes and their paralogs 167
2.1. Dataset of cancer cell lines 174
2.2. Detection of candidates for synthetic lethal screenings 175
3. The human genetic interaction network 178
3.1. Construction of a cancer gene-centered network of genetic interactions 178
3.2. Genetic interactors of cancer genes are ancient duplicated hubs 181
3.3. The integration of protein-protein interactions and genetic interactions allows the identification of putative cancer genes 184
References 187

Figure List

Figure 1: Detection of protein-protein interactions with in a yeast-two-hybrid system 18
Figure 2: Detection of protein-protein interactions with TAP 19
Figure 3: Detection of protein-protein interactions with FRET 20
Figure 4: Detection of genetic interactions 24
Figure 5: Network models 27
Figure 6: orthologs and paralogs 30
Figure 7: duplication-divergence model 34
Figure 8: mechanisms of duplication 36
Figure 9: Dosage balance of protein complexes 38
Figure 10: non-synonymous mutations 42
Figure 11: Hallmarks of cancer 44
Figure 12: translocations in the Cancer Gene Census 49
Figure 13: tumor types in the Cancer Gene Census 49
Figure 14: dominant and recessive genes from the Cancer Gene Census 49
Figure 15: Pipeline to detect candidate cancer genes in whole-exome mutational screenings 52
Figure 16: Pipeline to detect true mutations in whole genome sequencing studies 55
Figure 17: Heterogeneity of cancer genes identified in different cancer types 57
Figure 18: Cancer genes in Rb and p53 pathways 58
Figure 19: definition of gene 59
Figure 20: Pipeline to identify unique human genes 62
Figure 21: Tree of life 69
Figure 22: resolution of the clusters of orthologs 70
Figure 23: Origin and conservation assignment 72
Figure 24: Pipeline to assess enrichment in GO terms 77
Figure 25: Representative species used in the analysis 85
Figure 26: gene origin in evolution 86
Figure 27: gene conservation in evolution 88
Figure 28: gene duplicability in evolution 90
Figure 29: degree distribution of the *E. coli* network 93
Figure 30: degree distribution of the *S. cerevisiae* network 94
Figure 31: degree distribution of the *D. melanogaster* network 94

Figure 32: degree distribution of the *H. sapiens* network 94
Figure 33: network properties of singleton and duplicated genes 96
Figure 34: degree of proteins with different origins 99
Figure 35: relationships between origin and network properties 99
Figure 36: Relationships between conservation and network properties 103
Figure 37: Relationships between origin and network properties using a randomization test 107
Figure 38: Relationships between conservation and network properties using a randomization test 107
Figure 39: Relationships between conservation and network properties 110
Figure 40: Functional differences between the two classes of human hubs 113
Figure 41: dosage regulation of hubs 135
Figure 42: relationships between ohnologs and singleton hubs 136
Figure 43: miRNA regulation of singleton and duplicated hubs 136
Figure 44: tissue-selectivity of singleton and duplicated hubs 136
Figure 45: relationships between housekeeping genes and duplicability of hubs 137
Figure 46: origin of cancer genes 138
Figure 47: relationships between origin and duplicability of cancer genes 139
Figure 48: distance between cancer proteins 140
Figure 49: conservation of duplicates of cancer genes 144
Figure 50: conservation of genic duplicates of cancer genes 145
Figure 51: conservation of genomic duplicates of cancer genes 145
Figure 52: length of cancer genes 145
Figure 53: conservation of normalized duplicates of cancer genes 146
Figure 54: conservation of normalized genic duplicates of cancer genes 146
Figure 55: conservation of normalized genomic duplicates of cancer genes 147
Figure 56: date and party hubs 150
Figure 57: three models of network evolution 152
Figure 58: miRNA regulation of PTEN and PTENP1 155
Figure 59: cancer hubs 157
Figure 60: Loss of genes in the genome alignment pipeline 161
Figure 61: Orthology visualization in NCG 162
Figure 62: Orthology visualization in NCG 165
Figure 63: protein interaction network visualization in NCG 165
Figure 64: miRNA-gene interactions in NCG 166
Figure 65: KDM5C and KDM5D 170

Figure 66: CLTC and its translocations in cancer 171
Figure 67: domain composition of SMARCA4 and SMARCA2 172
Figure 68: VHL and VHLL pathway 173
Figure 69: essential genes and cancer cell lines 181
Figure 70: origin of essential genes 183
Figure 71: duplicability of essential genes and cancer cell lines 183

Table list

Table 1: Identification of unique genes 60
Table 2: experiments to identify cancer genes 64
Table 3: sources of protein-protein interactions 66
Table 4: genes with evolutionary properties 73
Table 5: tissue selectivity of genes 81
Table 6: gene origin in evolution 87
Table 7: gene conservation in evolution 88
Table 8: gene duplicability in evolution 90
Table 9: properties of the protein interaction networks 92
Table 10: network properties of singleton and duplicated genes 97
Table 11: relationships between origin and network properties 101
Table 12: relationships between conservation and network properties 104
Table 13: Hub conservation throughout evolution 109
Table 14: relationships between duplicability and network properties 111
Table 15: functional comparison between recent duplicated hubs and ancestral singleton hubs 114
Table 16: functional comparison between duplicated genes and singleton genes 118
Table 17: functional comparison between recent genes and ancestral genes 122
Table 18: functional comparison between dominant and recessive cancer genes 141
Table 19: integration of protein interaction networks from different sources 162
Table 20: highly conserved paralogs of cancer genes 169
Table 21: filters on cell lines 175
Table 22: candidate cell lines 176
Table 23: candidate controls 177
Table 24: experiments used to identify genetic interactions of cancer genes 179
Table 25: properties of essential genes 182

Acknowledgements

I would like to express my gratitude to Francesca Ciccarelli

For guiding me through these four years of Ph.D

For all the suggestions and discussion

And for teaching me how to be a scientist.

I would like to thank my external co-supervisor Aoife McLysaght

For her advices and criticism on my work

And for her patience to read and comment on my assays.

I would like to thank my internal co-supervisor Stefano Casola

For his support and useful discussions during these four years

I am very grateful to all the members of my group,

Anna De Grassi, Fabio Iannelli, Elena Gatti,

Matteo Cereda, Shruti Sinha, Valentina Melocchi

And Vera Pendino

For all their support and discussions

During these years of data clubs, journal clubs, retreats, lunches and breaks

And for reminding my that there is always room for a joke.

I would like to thank the NCG crew,

Vera Pendino, Adnan Syed, Shruti Sinha and, of course, Francesca Ciccarelli,

And all the IT staff

For all their help in publishing the Network of Cancer Genes.

I am grateful to all the rest of the IFOM-IEO Campus,

Who made it possible to create an atmosphere where working is not stressful,

In particular the basketball players at the Campus,

Rodrigo, Katerina, Matias, Valentina, Gabriele, Yannève, Marieta, Mattia and Aga

For all the fun we've been having all these evenings after work.

A special thank goes to all my friends,

Who remind me that work is not the only important thing.

I would like to thank my family, for believing in me and for their support.

At last, I would like to express my gratitude to Agnieszka Chronowska,

For teaching me how to face the stress of writing a Ph.D Thesis,

For all the patience she showed me after every day of work,

For not complaining about my few moments of nervousness,

For always having a word to make me smile and be relaxed.

List of abbreviations

BHLH: Basic Helix-Loop-Helix

BLAST: Basic Local Alignment Search Tool

BLAT: BLAST-like Alignment Tool

CGC: Cancer Gene Census

COG: Cluster of Orthologous Groups

COSMIC: Catalogue Of Somatic Mutations In Cancer

DDC: Duplication, Degeneration, Complementation

DIP: Database of Interacting Proteins

EAC: Escape from Adaptive Conflict

EST: Expressed Sequence Tag

FRET: Förster Resonance Energy Transfer

GEO: Gene Expression Omnibus

GO: Gene Ontology

HPRD: Human Protein Reference Database

HTMS: High-Throughput Mutational Screenings

KOG: euKaryotic Orthologous Groups

LCR: Low-Copy Repeat

LUCA: Last Universal Common Ancestor

miRNA: microRNAs

MINT: Molecular Interaction Database

NCG: Network of Cancer Genes

NCI: National Cancer Institute

ORF: Open Reading Frame

PCR: Polymerase Chain Reaction

RNAi: RNA interference

SGD: *Saccharomyces* Genome Database

shRNA: short hairpin RNA

siRNA: small interfering RNA

SNP: Single Nucleotide Polymorphism

TAP: Tandem Affinity Purification

TF: Transcription Factor

UAS_G: Upstream Activation Sequence

UTR: UnTranslated Region

WGS: Whole Genome Sequencing

Abstract

Recent studies on *S. cerevisiae* have shown that modifications in the dosage of essential genes, genes coding for protein complexes and network hubs are deleterious. These genes are intrinsically fragile toward perturbations, since dosage modifications of essential and highly connected proteins yield to alterations in protein function, causing phenotypic aberrations that may affect the whole cell function. Similar analyses on the mammalian protein interaction network have instead revealed a probable increased robustness toward dosage modifications owing to gene duplication. Unlike yeast, in mammals highly connected proteins are mostly encoded by duplicated genes, while essentiality is not correlated with duplicability. This difference suggests that dosage-sensitive genes could duplicate at a certain point in evolution, likely favoring the progressive increase in genomic complexity. In order to understand whether this hypothesis was correct, we investigated the relationships between gene properties (origin, conservation and duplicability) and network properties (connectivity and centrality) in several species from bacteria to primates. We found that all protein interaction networks maintain a core of hubs (*i.e.* highly connected proteins), which are encoded by ancient singleton genes that are involved in basic cellular functions. During vertebrate evolution, a new group of hubs emerged. These novel hubs are encoded by duplicated genes that originated with metazoans, duplicated with vertebrates, are involved in regulatory processes and in the organization of multicellular organisms. They duplicated through the two rounds of whole duplication that occurred in the early vertebrate genome and the retention of the duplication was favored by the presence of alternative mechanisms of dosage regulation.

In addition to offering novel insights into the evolution of protein interaction networks, this analysis also helped in better understanding the network properties of cancer

genes (*i.e.* gene whose mutations are causally implicated in cancer). In particular, they are representatives of the two classes of ancient singleton and recent duplicated hubs. These two groups of cancer-related hubs may reflect two ways of promoting tumorigenesis: one that interferes with basic and ancestral functions, and the other that perturbs more complex processes, such as regulation and development.

Determining the evolutionary characteristics of cancer genes and their position inside the human protein interaction network will help to understand the importance of these properties in tumorigenesis. Furthermore, we will be able to identify new putative cancer genes, given the assumption that mutations in genes that have properties similar to known cancer genes may promote tumorigenesis in a similar way.

Introduction

1. Aim of the Thesis

The control of gene dosage is crucial to regulate the expression and the function of particular categories of genes (Veitia, 2004; Veitia et al., 2008). In *S. cerevisiae*, modifications of the dosage of essential genes, genes encoding highly connected proteins and members of protein complexes are harmful (Papp et al., 2003; Prachumwat and Li, 2006; Yang et al., 2003). These genes are fragile towards dosage perturbations and modifications in their dosage may induce phenotypic aberrations (Papp et al., 2003; Veitia, 2002). Highly connected proteins and proteins that occupy central positions in the protein interaction network are essential also in multicellular eukaryotes, such as *D. melanogaster* and *C. elegans* (Hahn and Kern, 2005). Surprisingly, mammalian protein interaction networks show an increased robustness towards dosage modifications as a consequence of gene duplication. In particular, human hubs are preferentially encoded by duplicated genes (Liang and Li, 2007; Rambaldi et al., 2008), and mouse essential genes may be both singleton and duplicated (Makino et al., 2009).

The scope of this work is to determine how protein interaction networks evolved from unicellular species to mammals in order to identify the causes of the variation in the relationships between duplicability and network properties.

Understanding the evolution of the human protein interaction network is relevant also in the context of cancer. Indeed, cancer is a genetically complex disease, which may be caused by mutations in hundreds of genes (Vogelstein and Kinzler, 2004). In the past five years, several mutational studies of cancer tissues have allowed the identification of almost 1,500 genes that are actively involved in tumorigenesis. Notwithstanding the significant increase in the number of cancer genes since the first collections (Futreal et al.,

2004), they maintain the same systems-level properties. In particular, cancer genes are mostly singleton and encode highly connected proteins inside the human protein interaction network (Rambaldi et al., 2008). Therefore, their properties resemble those of yeast hubs. This may be an indirect proof of the fact that portions of the human protein interaction network that involve cancer genes have maintained their fragility towards gene duplication, despite a general increase in the robustness of the network. Perturbations of these nodes due to mutations or dosage modifications cause tumorigenesis. Understanding why cancer genes have conserved these characteristics in evolution may help in the identification of new putative candidate cancer genes, on the basis of the concept that genes that have systems-level properties similar to known cancer genes may be involved in tumorigenesis when mutated.

2. Biological networks

In the last two decades we have witnessed the exponential development of studies concerning networks as a consequence of the increase in the amount of data that have become available. Several systems, such as the World Wide Web, the Internet and the relationships between individuals, are now best described in a network-like fashion (Albert, 2005; Albert and Barabasi, 2002). The Internet reached two billion users in 2010 (Lynn, 2010), while more than one trillion distinct URLs (Alpert and Hajaj, 2008) are available in more than one hundred million web sites (DomainTools, 2011). A quick development of these networks has corresponded to an increase in the studies of network topology and network properties. The boost of these studies has favored a similar development also in the treatment of biological networks. The cell is a system of interactions between genes and gene products, which may act at different levels. Regulatory networks describe the transcriptional regulation by transcription factors, while the formation of protein complexes and physical interactions between proteins are

described by protein interaction networks and biochemical reactions are integrated into metabolic networks (Albert, 2005). Each network is not independent from the others, having each external signal that triggers reactions involving protein-protein interactions, regulatory networks and metabolic networks (Albert, 2005).

Several high-throughput techniques have been established in the last few years, which have been exploited to gather large amounts of data from a small number of experiments in a limited time. In ten years, protein-protein network data have increased from few hundreds in a single species (*S. cerevisiae*) (Uetz et al., 2000) to more than 200,000 in more than twenty species (Stark et al., 2011). In a similar way, also genetic interaction networks have been studied extensively, having the *S. cerevisiae* network that includes now more than 150,000 interactions (Stark et al., 2011). Protein and genetic interaction networks are now studied in several species in different taxonomic groups, such as bacteria (in particular *E. coli* and *M. pneumoniae*), plants (*A. thaliana*), fungi (*S. cerevisiae* and *S. pombe*), insects (*D. melanogaster*), nematodes (*C. elegans*), rodents (*M. musculus* and *R. norvegicus*) and primates (*H. sapiens*) (Kerrien et al., 2007; Stark et al., 2011). Several other representations of networks are used to study biological phenomena, such as metabolic networks (Jeong et al., 2000; Lemke et al., 2004), transcriptional regulation networks (Lee et al., 2002; Luscombe et al., 2004) and signal transduction pathways (Ma'ayan et al., 2005).

2.1. Protein-protein interactions

Protein interaction networks describe physical interactions between proteins. Several methods have been developed in the last years to identify this type of interactions. The first technique that was established to detect protein-protein interactions is yeast-two-hybrid (Cusick et al., 2005). It was first developed by Fields and Song more than 20 years ago (Fields and Song, 1989) and has been widely used to detect protein-protein

interactions in several species, from bacteria (Rain et al., 2001), to yeast (Uetz et al., 2000; Yu et al., 2008a) and higher eukaryotes (Giot et al., 2003; Li et al., 2004; Rual et al., 2005; Uetz and Pankratz, 2004). This technique exploits the properties of the *S. cerevisiae* protein GAL4, which is a potent transcriptional activator when yeast is grown on galactose-rich media (Johnston, 1987). Its two separable domains are bound to two proteins (the "bait" and the "prey", Figure 1A). If these proteins interact, GAL4 is reconstituted and activates transcription of a designated reporter gene (Figure 1B). High-throughput screenings are made using pools of hundreds of open reading frames (ORFs) fused to one of the two domains of GAL4 that are injected into the yeast cells. After incubating the cells, colonies positive for the phenotype of the reporter gene are selected (Fields and Song, 1989). The interacting proteins are amplified using primers that are specific for one of the two domains, then two PCRs are run for each positive colony, and Sanger sequencing is used to identify the two interacting proteins (Cusick et al., 2005).

A second method to detect protein-protein interactions is based on tandem affinity purification (TAP) followed by mass-spectrometry (Puig et al., 2001). A protein of interest is recognized by a TAP tag, which is fused to its N- or C- terminus. Tagged complexes containing the protein of interest can be quickly and easily purified, then the members of the complexes are identified by mass-spectrometry (Puig et al., 2001) (Figure 2).

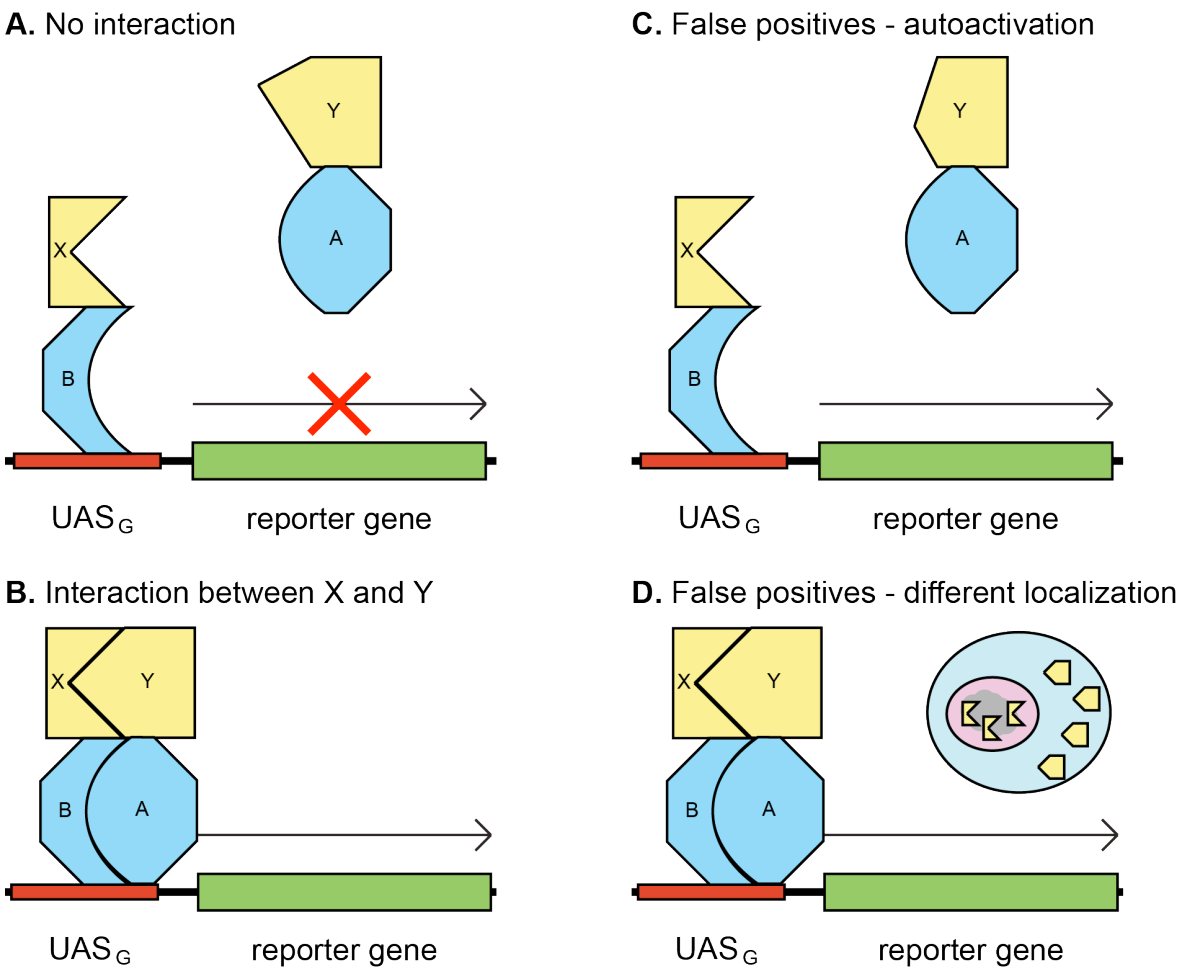


Figure 1: Detection of protein-protein interactions with in a yeast-two-hybrid system

Adapted from (Chien et al., 1991). The upstream activation sequence (UAS_G) is the promoter region that GAL4 binds. Two separable domains constitute this protein: the N-terminal domain has DNA-binding activity but cannot activate transcription, while the C-terminal domain activates transcription but does not recognize the DNA-binding site. The expression levels of this reporter gene, the interaction can be confirmed. The protein X (referred to as "bait") is translated with the GAL4 DNA-binding domain B, while the protein Y ("prey") is translated with the GAL4 activation domain A. (A) In case of no interaction between X and Y, the reporter gene will not be activated. (B) If X interacts with Y, GAL4 is reconstituted and is able to activate the transcription of a reporter gene. (C) False positives may arise if the reporter gene may be activated by only the binding between the GAL4 B-domain to UAS_G. (D) Another type of false positives may arise when X and Y are able to physically interact with each other, but they have different cellular localizations (X is nuclear, while Y is cytoplasmatic), therefore they will never interact *in vivo*.

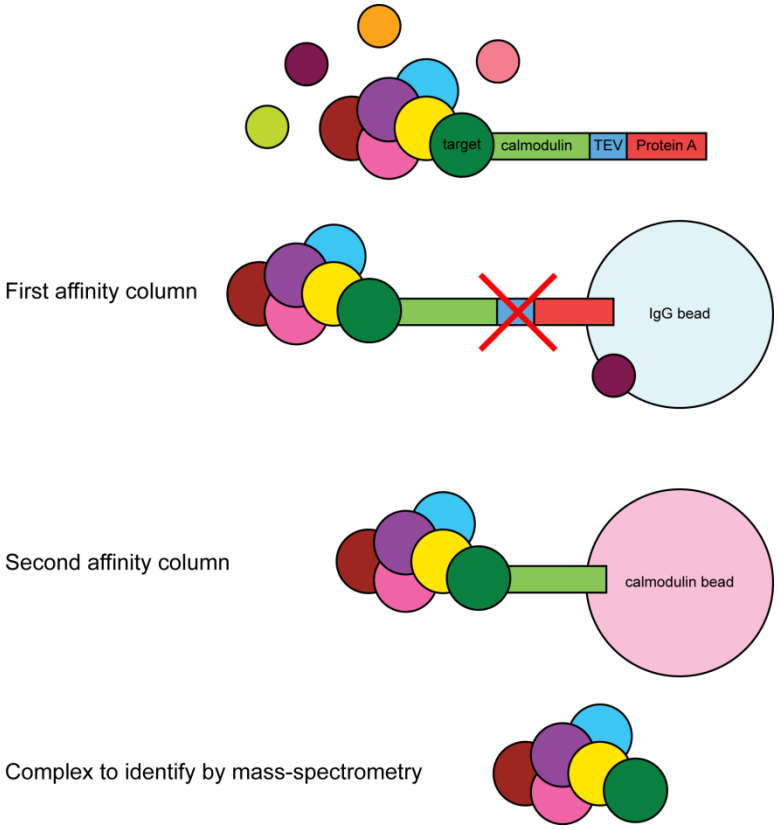


Figure 2: Detection of protein-protein interactions with TAP

Adapted from (Puig et al., 2001). The target protein is bound to a TAP tag, composed by a calmodulin-binding peptide, tobacco etch virus protease (TEV protease) cleavage site and protein A, which binds tightly to IgG. In the first affinity column, the complex to purify binds to the IgG beads. TEV protease cleaves the complex from protein A, which is then ready for the second affinity column, where the complex binds to calmodulin beads. The members of the newly purified complex are then identified by mass-spectrometry.

A last method was developed on the basis of the Förster resonance energy transfer (FRET), which exploits the ability to transfer electrons between near chromophores (<10 nm) (Truong and Ikura, 2001). The concept is similar to the yeast-two-hybrid technique: instead of utilizing the two domains of GAL4, two different mutant GFPs (the donor and the acceptor chromophores) are bound to two proteins. When the two proteins interact, the acceptor GFP becomes luminescent and the interaction may be detected directly *in vivo* (Truong and Ikura, 2001) (Figure 3). The difference with yeast-two-hybrid is that, in addition to the *in vivo* detection, the interaction is directly visible and must not be inferred from the expression of the reporter gene.

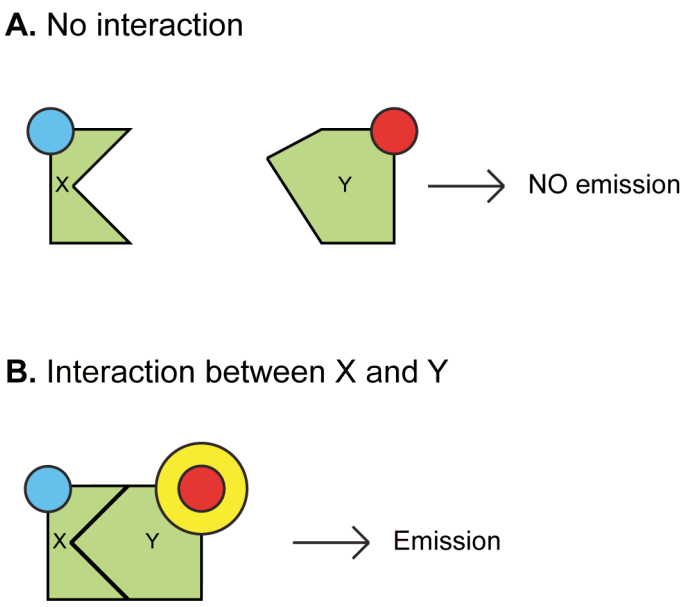


Figure 3: Detection of protein-protein interactions with FRET

Adapted from (Truong and Ikura, 2001). A chromophore is bound to protein X (cyan) and to protein Y (red). By stimulating with energy at the wavelength of the first chromophore, electrons are transferred to the second chromophore if they are less than 10 nm distant. In case X and Y interact, the second chromophore will become luminescent and will be detected. In case the two proteins do not interact, the second chromophore will not be visible.

In addition to the three methods described previously, interactions may be detected in small-scale experiments using different techniques, such as co-localization, co-purification, co-fractionation and co-crystal structure (Breitkreutz et al., 2008; Keshava Prasad et al., 2009).

2.2. False positives in the detection of protein-protein interactions

High-throughput detection of protein-protein interactions may present several false positives. The results may be misleading if the fraction of false positives is sufficiently high.

In yeast-two-hybrid experiments, false positives may arise in the detection of protein-protein interactions because of auto-activation of the reporter gene, which may

occur when the bait is able to directly activate transcription, without binding to the prey (Cusick et al., 2005) (Figure 1C). These auto-activators appear to have many interaction partners that do not share any functional similarity. They may be recognized and eliminated from the analyses by eliminating all yeast colonies that have evidence of activation of the reporter gene before the binding with the prey (Walhout and Vidal, 1999). Another type of false positives that is harder to detect is represented by proteins that are able to interact in the two-hybrid system but are never expressed at the same time or in the same cells (Figure 1D). These are nearly impossible to identify using only interaction essays, but may be detected by studying their expression levels *in vivo* (Cusick et al., 2005).

In principle, TAP should not have problems with false positives. However, the comparison of different studies of this type showed a limited overlap in the identification of the same complexes (Krause et al., 2004). This is probably due to the fact that protein complexes may involve transient interactions and a significant fraction of the members of these complexes may be hard to characterize because of their low expression levels (Cusick et al., 2005).

Intuitively, small-scale experiments have less false-positives than high-throughput screenings (Bader et al., 2004; von Mering et al., 2002). However, Yu *et al.* (Yu et al., 2008a) demonstrated that also the first yeast-two-hybrid experiments on two yeast protein interaction networks (Ito et al., 2001; Uetz et al., 2000) were of high quality in terms of false-positives rate. A problem related with small-scale experiments is that they are often based on literature curation (*i.e.* interactions are identified from text mining), which includes errors due to the difficulty to detect real interactions from a text document (Cusick et al., 2009).

These observations show that determining the reliability of a high-throughput experiment is not trivial. The best method to estimate it is based on the comparison with a reference set of high-quality interactions that were used as gold standard (Cusick et al.,

2009). This gold standard should be unbiased toward particular cell processes (*i.e.* they must include proteins from all cellular processes) and must be highly reliable and reproducible (Cusick et al., 2009).

2.3. Databases of protein interaction networks

Several databases of protein-protein interactions have been collected. Among the most complete and commonly used to analyze network data are BioGRID (Breitkreutz et al., 2008), IntAct (Kerrien et al., 2007), The Molecular Interactions Database (MINT) (Cesareni et al., 2008), the Database of Interacting Proteins (DIP) (Salwinski et al., 2004), DroID (Yu et al., 2008b) and the Human Protein Reference Database (HPRD) (Keshava Prasad et al., 2009). BioGRID, IntAct, MINT and DIP include data from several species, while HPRD focuses on human and DroID on *D. melanogaster*. However, these are all largely incomplete and the overlap between interactions derived from different databases is very low. Furthermore, until recently, there has never been a standard format to represent interactions. Therefore the integration of data from the different sources has always been challenging. To overcome these problems, the International Molecular Exchange Consortium (IMEx, http://www.imexconsortium.org/) was founded as a collaboration between several interaction data providers, with the aim of standardizing the curation rules to identify protein-protein interactions from experimental data and to determine a standard format for interaction data. However, although now a standard format of protein-protein interaction data exists, a single repository for this kind of data has not been created yet.

2.4. Genetic interactions

A genetic interaction is an unexpected phenotype that cannot be explained by combining the effects of individual genetic variants (Avery and Wasserman, 1992; Dixon

et al., 2009). Genetic interactions are detected by constructing cells with two mutated genes and analyzing their phenotype: if it deviates from the expected value, then a genetic interaction exists between the two mutated genes (Figure 4). The expected phenotype is defined as the product of the fitness of the two single mutants, although this definition is still matter of controversy (Mani et al., 2008).

A negative genetic interaction is present between two genes if the double mutant has lower fitness than expected (Figure 4). This reflects the function of two genes that are involved in parallel pathways. The deletion of one gene does not impair the cell viability because the other pathway is able to compensate its loss. Only when genes from both pathways are deleted, the function is impaired and the fitness is reduced (Dixon et al., 2009). The extreme negative genetic interaction is represented by a synthetic lethal interaction: cells with single mutants are viable, but the double mutant results in cell death.

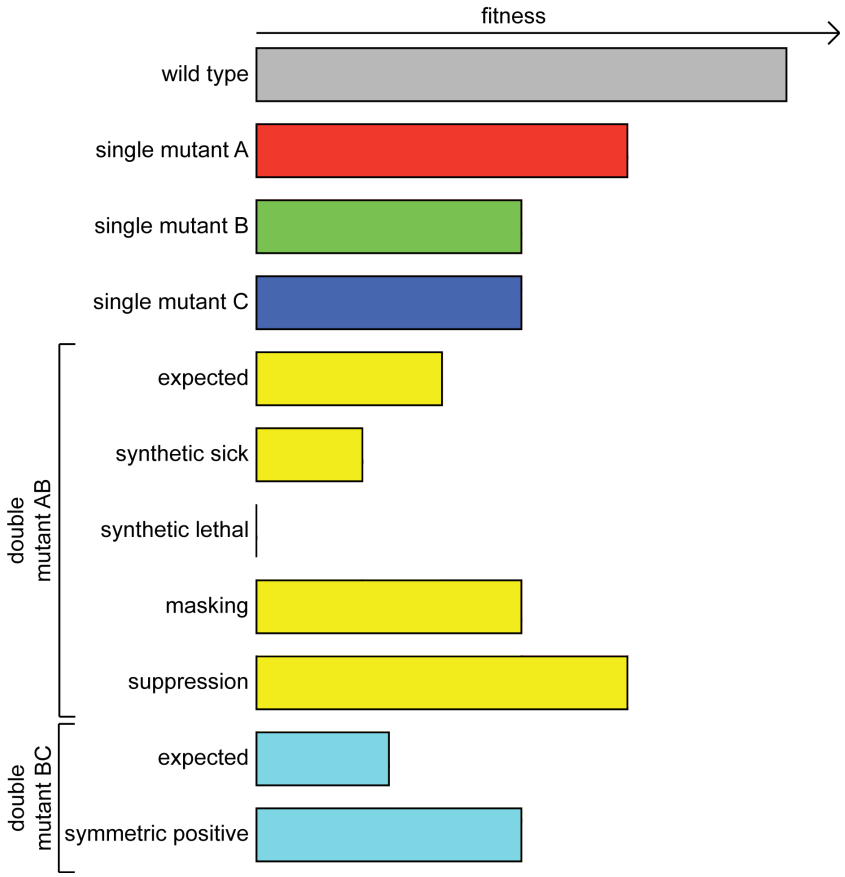


Figure 4: Detection of genetic interactions

Adapted from (Dixon et al., 2009). Genetic interactions are detected on the basis of the comparison of the fitness of the double mutant with the single mutants. Setting the wild-type fitness at 1, single mutants have lower fitness. The expected fitness of the double mutant is the product of the fitness of the single mutants. If the observed value is lower than the expected, a negative genetic interaction is established between the two mutated genes (synthetic sick interaction). The extreme fitness is present if the interaction is synthetic lethal. If the observed fitness is higher than expected, the genetic interaction is positive. In case the single mutants have the same fitness, the interaction may be symmetric positive, *i.e.* the double mutant has the same fitness as the single mutants. This may be explained by the fact that the two mutants are part of the same non-essential complex: the deletion of a single member of this complex disrupts its structure and the deletion of a second member does not reduce the fitness any further. In case of different fitness of the two mutants, the asymmetric interaction may be masking or suppression, depending on whether the fitness of the double mutant is equal to the fitness of the single mutant with lower or higher fitness.

Positive genetic interactions refer to cases where the double mutant has less severe effects than the single mutants alone. Two types of positive genetic interactions have been defined, which correspond to different biological contexts. Symmetric interactions involve genes that encode proteins from the same complex. The two single mutants have comparable fitness, because the effect is always the disruption of the protein complex. Therefore, also the double mutant will have the same fitness of the single mutants (Figure 4). Asymmetric interactions involve genes that, when mutated, give different phenotypes to the cell. The effect of the double mutant may be masking if the double mutant phenotype correspond to the sickest single mutant (*i.e.* the second mutant phenotype is masked by the first) or genetic suppression if the double mutant has better fitness than the sickest single mutant (Figure 4).

To date, *S. cerevisiae* has been the most widely used model species to study genetic interactions, because mutant strains for every gene are available and, therefore, it is not complicated to detect interactions. More than 150,000 genetic interactions have been identified among more than 5,400 yeast genes (Stark et al., 2011). In metazoans, instead, it is more complicated to detect genetic interactions. It is possible to make large-scale screenings using RNA interference (RNAi) libraries to mimic the deletion of particular

genes. *C. elegans* is particularly suited for this, because it can absorb exogenous genetic material if it is soaked in a solution (Maeda et al., 2001), or it can be fed by bacteria expressing the RNA of interest (Timmons et al., 2001). In mammals the situation is more complicated and the only approach that has given significant results consists in infecting cell lines using libraries of RNAi. This has been applied systematically to cancer cell lines in order to determine what genes are essential for the survival of tumor cells (Baldwin et al., 2010; Baldwin et al., 2008; Barbie et al., 2009; Bommi-Reddy et al., 2008; Grueneberg et al., 2008a; Grueneberg et al., 2008b). Cell line-specific essential genes likely have negative or synthetic lethal genetic interactions with the mutated genes. However, the genetic landscape of cancer cells is highly complex and these screenings are performed *ex vivo*. Therefore the identification of genetic interactions in mammals is still hard to perform.

Screenings to detect genetic interactions in metazoans may have high rates of false positives, which are caused by off target effects due to the imperfect base-pair complementarity of the small interfering RNA (siRNA) used for RNAi screenings, which results in a non-specific targeting (Echeverri et al., 2006). Finally, the introduction of alien genetic material into certain mammalian cells may induce interferone response, which interferes with the normal response to RNAi (Bridge et al., 2003).

2.5. Network classification and characteristics

In mathematics, a graph (which is used as synonym of "network") is a representation of a set of nodes connected by links (edges). Three properties are studied to determine the characteristics of each node: degree, betweenness and clustering coefficient. Degree represents the connectivity of a node, *i.e.* the number of connections that it has with other nodes in the network. Betweenness is a measure of the centrality and is calculated as the number of shortest paths that pass through a node of interest (Goh et al., 2001). Given a network V, the betweenness for a node v is calculated as:

$$C_B(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}}$$

Where σ_{st} is the number of shortest paths between *s* and, and $\sigma_{st}(v)$ is the number of shortest paths between *s* and *t* that pass through *v*. Clustering coefficient represents how the first-level neighbors (*i.e.* the nodes that interact directly with the node of interest) of a node are interconnected. It is calculated as the ratio between the number of interactions between the first neighbors of a protein and the number of all possible interactions among them:

$$C_i = \frac{2\left|\{e_{jk}\}\right|}{k_i(k_i - 1)}$$

Where k_i is the degree of the protein *i* and $\left|\{e_{jk}\}\right|$ is the number of interactions between its neighbors (Watts and Strogatz, 1998). The analysis of the distributions of these local properties (*i.e.* that are associated with each node) allows determining the global topology of the network. In particular, four types of network topologies have been identified: (1) random networks, as defined by Erdős and Rényi (Albert and Barabasi, 2002; Barabasi and Albert, 1999), (2) small-world networks, (Watts and Strogatz, 1998), (3) hierarchical networks (Ravasz et al., 2002), and (4) scale-free networks (Barabasi and Albert, 1999).

Random networks have a fixed number of nodes that are connected randomly to each other. The probability that two nodes are connected is *p*, which is constant for every pair of nodes. Following this rule, all nodes have in principle the same number of interactions and the degree follows a binomial distribution (Figure 5A) (Albert and Barabasi, 2002; Barabasi and Oltvai, 2004).

Small-world networks are random networks built by rewiring an existing ring lattice (Figure 5B) in which every node is connected with its first *k* neighbors. Each edge is rewired with probability *p*. A value of $p = 1$ represents the random network model by Erdős and Rényi (Albert and Barabasi, 2002; Watts and Strogatz, 1998).

For both these models, the probability of having highly connected nodes decreases exponentially for high degree values (Albert and Barabasi, 2002; Barabasi and Oltvai, 2004; Watts and Strogatz, 1998). A difference between these two random networks is the distance (or shortest path) between each pair of nodes, which represents the smallest number of interactions that are crossed to travel between the two nodes (Barabasi and Oltvai, 2004). While in the Erdős-Rényi model the mean distance is proportional to log$N$, where N is the number of nodes in the network, in the small-world model the mean distance is a function of p: for small values of p, it increases linearly with the number of nodes, while for higher values it behaves like the Erdős-Rényi model because a sufficient number of shortcuts is introduced in the network.

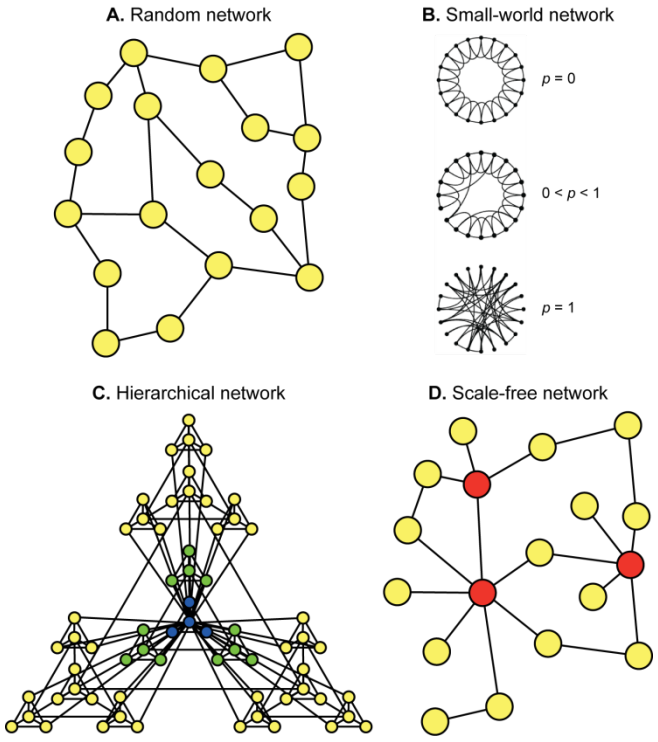


Figure 5: Network models

Adapted from (Barabasi and Oltvai, 2004; Watts and Strogatz, 1998). Four types of networks are widely studied: random, small-world, hierarchical and scale-free. (A) In a random network with N nodes, a node is connected to another with probability p. This allows the creation of a network with $\frac{pN(N-1)}{2}$ interactions. The degree follows a Poisson distribution, therefore all nodes have approximately the same number of

interactions. The clustering coefficient is constant for all levels of degree. (B) To build a small-world network, the starting point is a regular ring with n vertices, each connected to its k nearest neighbors (here, $n = 20$ and $k = 4$). Each edge is reconnected with probability p. For $p = 0$, no interactions are rewired, while randomness increases with increasing p. The maximum value of $p = 1$ represents a random network. (C) Hierarchical networks are highly modular: they are built starting from a small cluster of highly connected nodes (the four blue nodes in the center), which are then replicated and connected to the central node, producing a 16-node module (green nodes). Then this structure is repeated, producing a large 64-node network (yellow nodes). The clustering coefficient is inversely proportional to the degree, because highly interconnected nodes have low degree, while highly connected nodes link different modules, therefore their neighbors are not highly interconnected. (D) In a scale-free network, each node has a probability of having k interactions that is proportional to $k^{-\gamma}$. This allows for the presence of a small fraction of highly connected nodes, which are called hubs (nodes depicted in red). Biological networks usually have $2 < \gamma < 3$. As in random networks, the clustering coefficient is independent from the degree.

Hierarchical networks are built starting from a small cluster of highly interconnected nodes, which is replicated a fixed number of times. The central node from the resulting clusters is then connected with the central cluster and with the central node of the other resulting clusters (Figure 5C) (Albert and Barabasi, 2002; Ravasz et al., 2002).

Scale-free networks represent the fourth type of network topology. The distinguishing feature from random networks is that scale-free networks have a connectivity distribution that follows a power-law (Barabasi and Albert, 1999) (Figure 5D). The probability of having a node with *k* interactions is $P(k) \sim k^{-\gamma}$, with $2<\gamma<3$. As consequence, the probability of having nodes with high degree is significantly higher than random networks. This type of networks has a large number of nodes that have a small number of connections and a small number of nodes that are highly connected (hubs) and occupy central positions inside the network. Scale-free networks evolve following the preferential attachment theory, which states that, at each time point, a new node with *M* interactions is added to the network and connects to an already existing node *i* with probability:

$$P_i = \frac{k_i}{\sum_{j=1}^{N} k_j}$$

where k_i is the connectivity of node i and the denominator is the sum of the connectivity of all nodes inside the network. (Barabasi and Albert, 1999). Scale-free networks have also a shorter mean distance, compared with the random models: it is proportional to loglog*N*. This implies that scale-free networks are more compact than random networks.

Jeong *et al*. (Jeong et al., 2000) gave the first proof of the scale-free behavior of biological networks. They demonstrated that the metabolic networks of 43 different species are scale-free, having the vast majority of substrates involved in a small number of reactions (*i.e.* they have low connectivity), while few substrates are required for many interactions (*i.e.* they are hubs). The scale-free topology of protein-protein interaction networks has also been demonstrated in several species, such as *S. cerevisiae* (Jeong et al., 2001; Uetz et al., 2000), *D. melanogaster* (Giot et al., 2003) and *H. sapiens* (Rual et al., 2005).

The fact that real networks (both biological, social and technological, such as the World Wide Web) are scale-free is due to two processes that have an important role in the evolution of these networks (Albert and Barabasi, 2002). First, networks evolve and are the result of a growth process, with constant addition of new nodes and new interactions and rewiring of existing interactions. Random networks instead evolve randomly, with every node and interaction having the same importance compared with all the others. Second, following the preferential attachment theory, new nodes preferentially attach to hubs (Barabasi and Albert, 1999), therefore the evolution of the network maintains hubs as important nodes inside the network.

3. Orthology and paralogy

The availability of complete genomes of several different species has allowed the identification of evolutionary relationships between genes. In principle, the comparison of all gene sequences within the genome and between genomes of different species permits

the reconstruction of the history of each gene (Koonin, 2005). Five distinct events contribute to gene evolution:

1. Speciation;
2. Gene duplication, followed by divergence of duplicates;
3. Gene loss;
4. Horizontal gene transfer;
5. Gene rearrangements, such as gene fusions (Koonin, 2005).

The evolutionary relationships between genes are identified by comparing genes within the same genome and among different genomes. Orthologs are genes that evolved from a common ancestor by speciation (Koonin, 2005) (Figure 6). Paralogs are genes related via duplication (Koonin, 2005) (Figure 6).

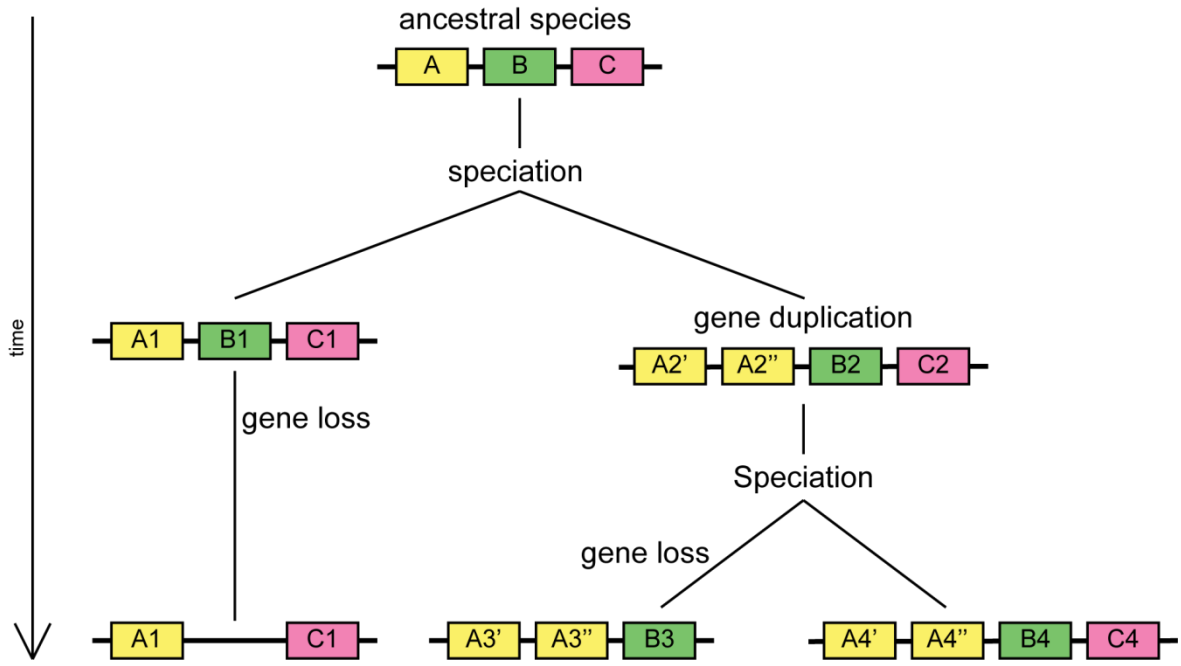


Figure 6: orthologs and paralogs

During evolution, speciation and duplication events occur, which create orthologs and paralogs. The ancestral species has 3 genes (A, B, C). A speciation event creates two species, which evolve separately. The left one undergoes gene loss and retains only A1 and C1 from the ancestor, while the right one undergoes duplication, followed by a second speciation, with the loss of the gene C in the left species. These events allow the birth of orthologous and paralogous genes. Genes from different species that share the same ancestor are orthologs: C1 and C2 are orthologs, as well as B3 and B4. Genes from the same species that share the same ancestor are orthologs: A2' and A2'' are paralogs.

3.1. Methods to detect orthology

The original method to detect orthology relationships between genes relied on phylogenetic analysis (Mirkin et al., 1995). In particular, the comparison between the topology of a gene tree and the corresponding species tree allowed the identification of paralogous genes by tree reconciliation, on the basis of the parsimony principle (*i.e.* the trees are reconciled by permitting the minimum number of gene duplications and losses) (Koonin, 2005; Mirkin et al., 1995). Several caveats do not allow the genome-wide application of this method. First, in prokaryotes and in lower eukaryotes, horizontal gene transfers undermine this method because they invalidate the relationships between the gene tree and the species tree (Koonin, 2005). Second, the reconstruction of species trees and many gene trees may not be accurate, due to the presence of artifacts and uncertainties in the tree definition (Koonin, 2005). Third, the genome-wide application of this method is computationally expensive. Several methods overcome these issues. TreeFam reconstructs orthology relationships by building phylogenetic trees of distantly related metazoans, thus eliminating the horizontal gene transfer caveat and reducing the uncertainties in the species tree definition (Ruan et al., 2008). InParanoid was developed to detect orthologs by pairwise comparisons between genomes, thus eliminating the reconciliation process (Ostlund et al., 2010). The COG database and eggNOG define orthologs as genes that are more similar to each other than to any other gene in the compared genomes (Jensen et al., 2008; Tatusov et al., 2003).

The pipeline to detect orthologs in eggNOG is as follows (Jensen et al., 2008). First, all-against-all Smith-Waterman similarities are computed for all proteins from species that are representative of different taxonomic groups (from bacteria to higher eukaryotes). Second, proteins from the same species or from tightly related species (all strains of a particular species or closely related species, such as human and chimpanzee), which are more similar to each other than to proteins from other species, are joined and considered as single entities. Orthology is assigned by joining triangles of reciprocal best

hits that involve three different species. In order to include the unassigned proteins, simple bidirectional best hits are also considered. The innovative feature of eggNOG, compared with the original KOG/COG procedure (Tatusov et al., 2003; Tatusov et al., 1997), is the construction of a hierarchy of orthologous groups to define high-resolution orthology relationships. Several lineage-specific groups are created: mammals (maNOGs), vertebrates (veNOGs), metazoans (meNOGs), fungi (fuNOGs) and insects (inNOGs) are added to the eukaryotic-specific group (KOGs) and to the COGs, which include both eukaryotes and prokaryotes.

3.2. Methods to identify paralogs

Although the relationships between genes related via duplication have been extensively studied in the last ten years, a consensus definition has not been determined and several different methods are commonly used to detect paralogous genes.

A common definition is based on a BLAST search of one gene sequence against all the other genes from the same species. Using more or less stringent cutoffs of the BLAST score, ancient and recent duplications are determined and putative paralogs are identified (Papp et al., 2003). However, this method allows identifying genes that have similar sequences, rather than paralogs (*i.e.* genes that diverged from a common ancestor).

A different method recently developed by our group (Rambaldi et al., 2008) relies on the sequence conservation on the genome. Briefly, the protein sequence of a gene is aligned to its translated genome using BLAT (Kent, 2002). A gene is duplicated if it has at least one additional hit on the genome that spans at least 60% of its length (Rambaldi et al., 2008). This definition allows identifying not only functional paralogs (*i.e.* real genes), as in the case of a BLAST search, but also pseudogenes and degenerated duplications.

A third method to detect paralogous genes relies on the reconstruction of clusters of orthologs (Jensen et al., 2008). Two genes from the same species that are included in the same cluster of orthologs are paralogs.

3.3. Evolution of paralogs

The newly duplicated genes may encounter two potential fates to become fixed in the genome: neofunctionalization or subfunctionalization (Figure 7A-C).

After duplication, one of the two paralogs may acquire a mutation that confers a new function, which can be positively selected and become fixed by genetic drift if beneficial, or be lost if detrimental. This is the classical model of neofunctionalization, which is termed mutation during non-functionality (Figure 7A) (Hughes, 1994).

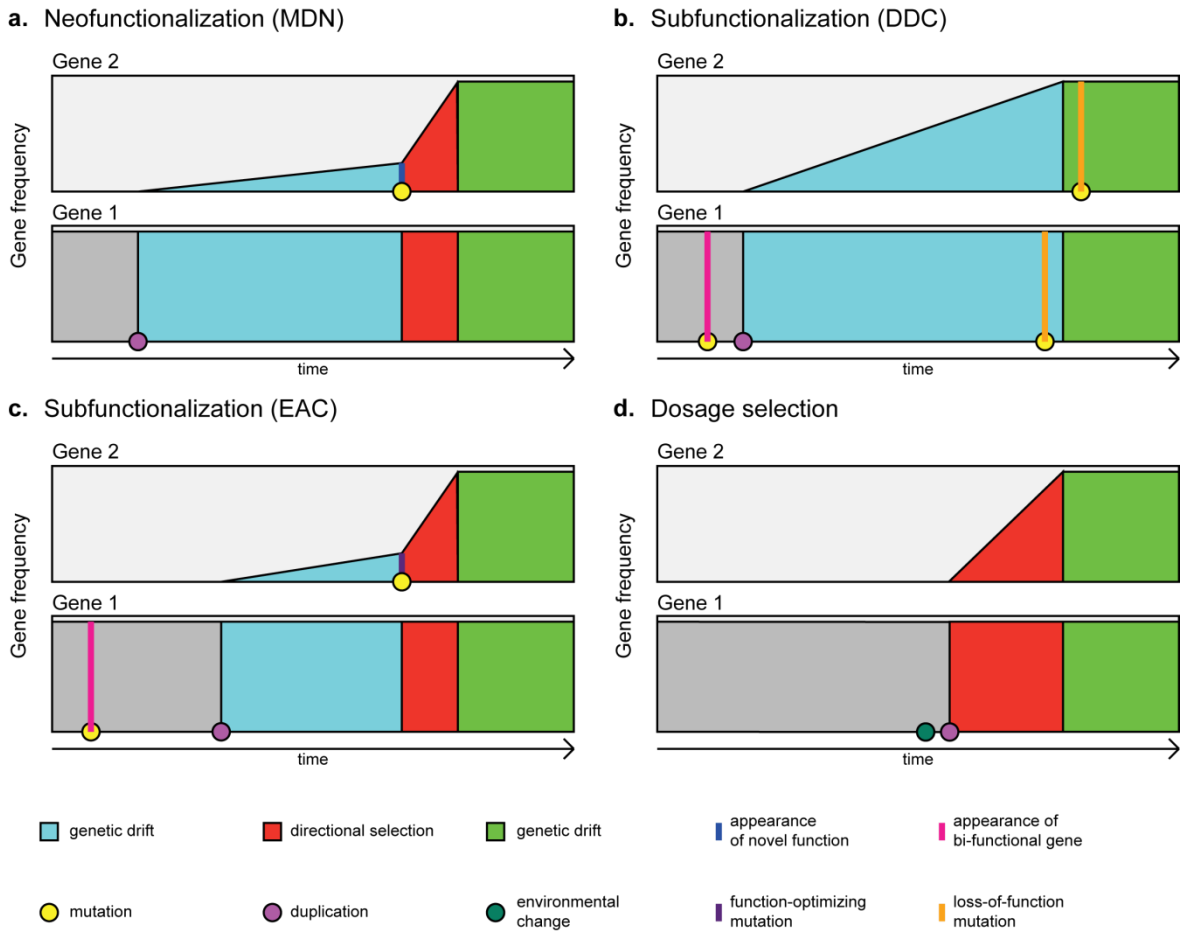


Figure 7: duplication-divergence model

Adapted from (Conant and Wolfe, 2008). Each panel shows the population frequency of both the original gene (Gene 1) and its paralog (Gene 2). Before duplication, the locus of Gene 1 is fixed in the population. (A) The mutation during non-functionality (MDN) neofunctionalization model shows that a mutation conferring a new function appears after the duplication. The paralog may become fixed in the population either because the mutation is beneficial or simply by genetic drift. (B) The duplication, degeneration, complementation model corresponds to subfunctionalization by neutral degenerative mutations. The duplication becomes fixed by genetic drift. After fixation, both copies acquire mutations that make them lose part of the original function, which is kept by the other copy. (C) In the subfunctionalization through escape from adaptive conflict (EAC) model, a mutation that optimizes the duplication occurs before fixation, which than happens by directional selection. (D) The duplication fixation through dosage selection occurs when environmental changes make an increased gene dosage beneficial.

The second model of paralog divergence implies that both paralogs evolve after duplication and tend to retain the ancestor's function only partially. Therefore they both become necessary to preserve the original function (Conant and Wolfe, 2008). Two types of subfunctionalization have been hypothesized. The first, termed duplication, degeneration, complementation (DDC) involves neutral mutations that disrupt one function of a multifunctional gene. These mutations are neutral because the original function is rescued by the paralog (Figure 7B) (Taylor and Raes, 2004). The second type, called escape from adaptive conflict (EAC), involves non-neutral mutations. After the first neutral mutation that disrupts the function of one paralog, complementary mutations are positively selected in the other paralog (Figure 7C) (Des Marais and Rausher, 2008).

Neofunctionalization and subfunctionalization do not necessarily imply the development of new functions or the disruption of existing function, but they can also refer to expression in different tissues or different expression levels (Conant and Wolfe, 2008). An example is the glutamate dehydrogenase 1 (*GLUD1*) and its paralog *GLUD2*. The latter originated by retroposition of *GLUD1* 23 million years ago (Burki and Kaessmann, 2004). While *GLUD1* is ubiquitously expressed, *GLUD2* acquired mutations after the duplication event that restricted its expression to the central nervous system and made it insensitive to inhibition by GTP (Burki and Kaessmann, 2004; Plaitakis et al., 2003). The higher

expression of glutamate dehydrogenase in the brain, due to its increased dosage, probably contributed to enhance the brain function in primates by allowing a high flux of neurotransmitter (Burki and Kaessmann, 2004).

One final mechanism of duplication fixation does not involve any mutations. An environmental change occurs and the duplication of a particular gene gives a selective advantage because of its increased dosage (Kondrashov and Kondrashov, 2006) (Figure 7D).

4. Gene duplication and genome evolution

In 1970 Susumu Ohno proposed that gene duplication is the easiest way to produce new genes, rather than creating them *de novo* (Ohno, 1970; Wolfe, 2001). His theory was based only on few known protein sequences. Recent studies demonstrated that new genes that evolved from non-coding DNA sequences are a negligible fraction of the genome (Cai et al., 2008; Knowles and McLysaght, 2009; Toll-Riera et al., 2009; Zhou et al., 2008). For example Knowles and McLysaght estimated that less than 0.1% of human genes originated in this way (Knowles and McLysaght, 2009). As Ohno hypothesized, gene duplication accounts for the vast majority of the development of new genes.

Gene duplications may arise through several mechanisms, such as chromosomal duplication or whole genome duplication (Maere et al., 2005), retroposition (Marques et al., 2005) or segmental duplication (Han et al., 2009) (Figure 8).

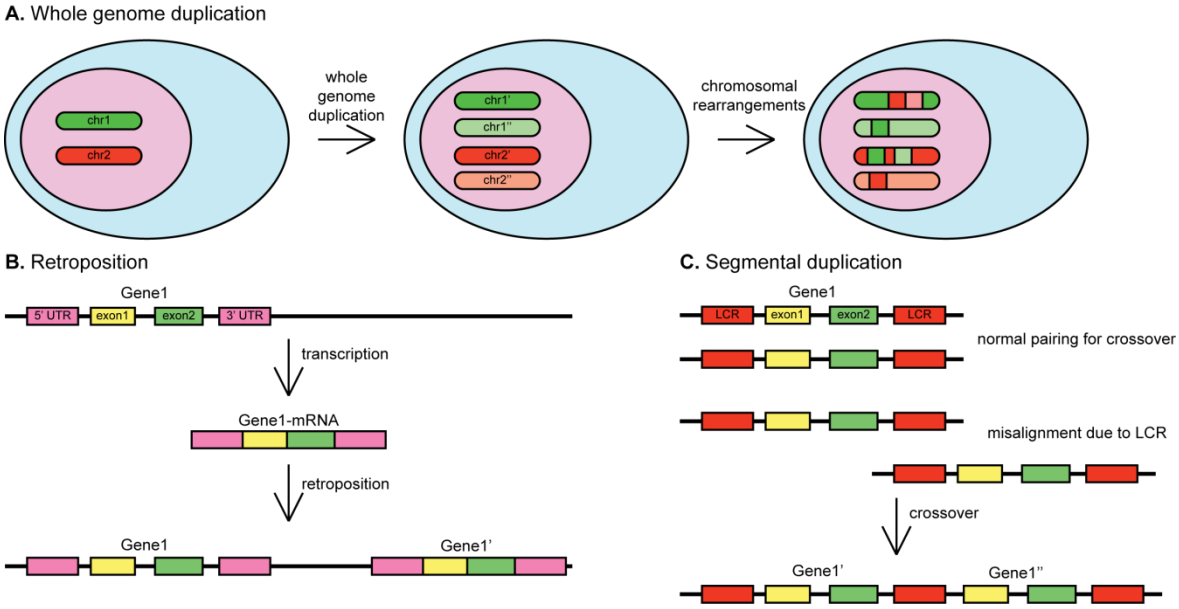


Figure 8: mechanisms of duplication

Several mechanisms allow gene duplication. (A) Whole genome duplication results in the presence of a second copy of all chromosomes. After whole genome duplication, several chromosomal rearrangements may occur (Nakatani et al., 2007). (B) After transcription, the spliced mRNA of a particular gene may be reverse transcribed into the genomic DNA sequence. This forms an intronless copy of the original gene (Marques et al., 2005). (C) Segmental duplication may arise when a genomic sequence is flanked by two highly identical sequences (LCR). Misalignment of the two LCR regions results in the duplication of the entire genomic sequence that is included between the two LCRs (Stankiewicz and Lupski, 2002).

Whole genome duplications are rare events that occurred in many eukaryotic taxa, such as plants (Proost et al.), fungi (Kellis et al., 2004) and vertebrates (Dehal and Boore, 2005). These polyploydizations are followed by major chromosomal rearrangements, which result in the retention of a fraction of duplicates (Dehal and Boore, 2005; Nakatani et al., 2007). The duplication of the entire genome allows the relaxation of the constraints that prevent or reduce gene evolution. The rate of sequence change is accelerated in both copies of the gene, independently from the long-term retention of the duplication (Dehal and Boore, 2005) (Figure 8A).

Retroposition generates intronless duplications by reverse transcription of mRNAs derived from a parental gene and integration of the resulting cDNA into the genome (Long et al., 2003). The regulatory elements of the parental copy are not duplicated and the

retroposed copy may become functional only if a new regulatory region is recruited (Long et al., 2003) (Figure 8B).

Segmental duplications are genomic regions ranging from 1 to 200 kb that are present in two or more genomic loci and share high sequence identity (90-100%) (Bailey et al., 2002; Bailey et al., 2001). This type of duplication derives from an incorrect alignment of two chromosomes before crossover, as a consequence of the presence of low-copy repeats (LCR) (Stankiewicz and Lupski, 2002). This results in the duplication of all the genomic sequence included between the two LCRs (Stankiewicz and Lupski, 2002) (Figure 8C).

4.1. Duplicability and essentiality

The stoichiometric balance between genes is strictly regulated in order to maintain the cell fitness (Veitia, 2004). The excess of one member of a protein complex may be deleterious, because it induces an imbalance with the other members (Papp et al., 2003; Veitia, 2002; Veitia, 2004; Veitia et al., 2008) (Figure 9). In case of trimeric complex A-B-A, formed by two subunits of protein A and one of protein B, overexpression of of one subunit causes an excess of the intermediate dimer A-B, because the stoichiometric balance between the two subunits is impaired (Veitia et al., 2008) (Figure 9B). Another case is represented by the heterotrimeric G-proteins complex (composed by α, β and γ subunits) that is involved in the response to pheromones. The α subunit is involved in the recognition of the pheromone signal, while β and γ activate the response (Figure 9C). Excess of β subunit induces an abnormal abundance of β-γ dimers, which activate the signal transduction even in absence of pheromone signal (Cole et al., 1990; Veitia et al., 2008) (Figure 9D).

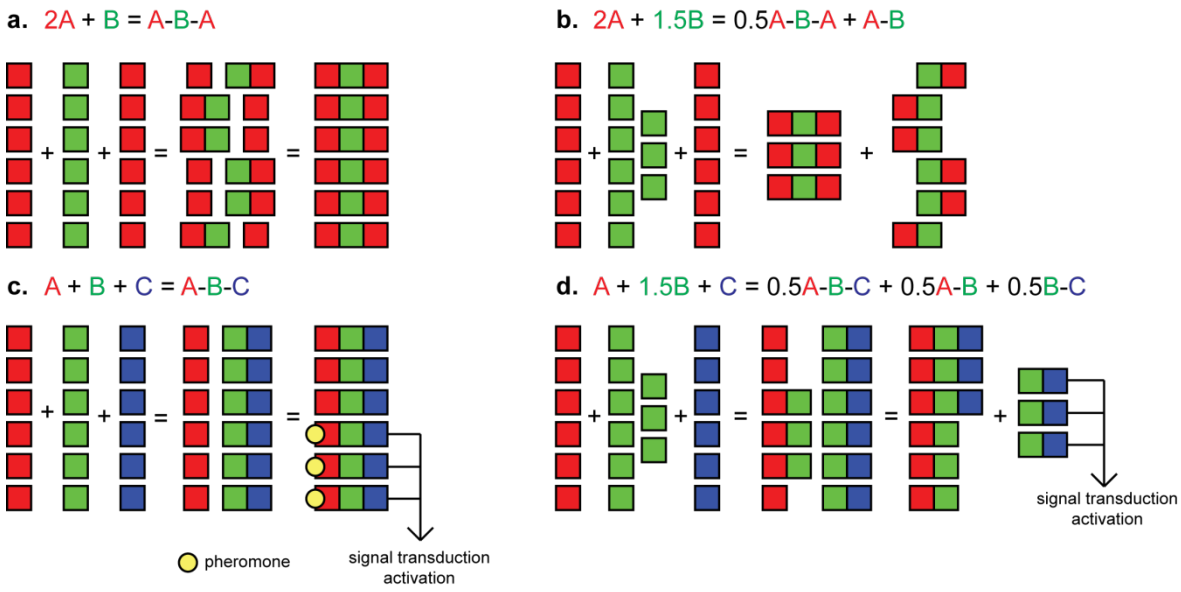


Figure 9: Dosage balance of protein complexes

Adapted from (Veitia et al., 2008). Overexpression of dosage-sensitive subunits of macromolecular complexes may be detrimental. (A) In the normal case, trimer ABA is stable and does not dissociate. (B) Overexpression of B leads to a significant decrease of ABA. (C) In the normal case of a heterotrimeric G protein complex, trimer ABC is stable and does not dissociate. The binding of pheromones to A lead to the activation of signal transduction. (D) Overexpression of B leads to abnormal excess of dimers that contain B, which trigger different signaling cascades without regulation by pheromone.

In *S. cerevisiae* and *C. elegans* singleton genes (*i.e.* genes that do not retain duplications) are essential (Conant and Wagner, 2004; Gu et al., 2003). When a duplicated gene is deleted, its paralog is able to recover, at least in part, its function, thus paralogs are able to compensate for each other's function. Only in case the second paralog is deleted, the function is lost and the fitness is significantly decreased. Paralogous genes are therefore depleted in essential genes in yeast (Papp et al., 2003; Yang et al., 2003). Deletion of singleton genes, instead, has a deleterious effect, because there cannot be any functional compensation: the function of the deleted gene is lost and the cell's fitness is impaired. In *S. cerevisiae* singleton genes also tend to encode members of protein complexes (Papp et al., 2003).

Unlike *S. cerevisiae* and *C. elegans*, mouse essential genes involved in development are preferentially duplicated (Liang and Li, 2007; Liao and Zhang, 2007;

Makino et al., 2009) and, in general, the relationships between essentiality and duplicability depend on the gene function (Makino et al., 2009). An explanation of this result about mouse developmental genes may be related with the mechanism of their duplication: these genes duplicated through whole genome duplication (Makino et al., 2009). Whole genome duplication does not modify relative dosages between genes and is the only mechanism that allows the retention of the duplication of dosage-sensitive genes (Makino et al., 2009).

5. Duplicability and protein interaction networks

Gene duplicability is tightly correlated with the number of interactions of the corresponding protein. In particular, in *S. cerevisiae*, connectivity is negatively correlated with gene duplicability and, when highly connected proteins duplicate, one copy is quickly deleted (Hughes and Friedman, 2005; Prachumwat and Li, 2006). Furthermore, highly connected and central proteins are also essential and slow-evolving (Hahn and Kern, 2005; Jeong et al., 2001). Highly connected and singleton proteins are predominantly involved in transcription, RNA metabolism, catabolism, protein synthesis and are mostly located at the ribosome or inside the nucleus, while lowly connected and duplicated proteins are located at the cell periphery (Prachumwat and Li, 2006).

The relationships between connectivity and duplicability that have been described for *S. cerevisiae* are also conserved in the *D. melanogaster* and *C. elegans* protein interaction networks (Hahn and Kern, 2005), although their incompleteness does not allow the comprehensive analyses that were performed in *S. cerevisiae*. Mammals, instead, display a peculiar behavior that cannot be associated to lower eukaryotes. Recent studies in *H. sapiens* and *M. musculus* showed that duplicability of mammalian genes positively correlates with connectivity (Liang and Li, 2007; Rambaldi et al., 2008). Furthermore, the gene family size, which is defined as the number of paralogs per gene, positively correlates with connectivity (Liang and Li, 2007). In vertebrates, gene families involved in

regulation, signal transduction, protein transport and protein modification have undergone expansion through gene duplication (Vogel and Chothia, 2006).

Although protein interaction networks of species with different levels of complexity (such as human, fly, worm and yeast) have a similar topology (Beltrao and Serrano, 2007; Hahn and Kern, 2005), these findings demonstrate that the vertebrate network developed peculiarities that might be related with the extensive expansion of the vertebrate genetic material. In particular, the observation that human hubs are preferentially duplicated (Rambaldi et al., 2008) may allow for the speculation that the evolution of the human network tolerated the retention of the duplications of hubs. This, in addition to the fact that vertebrates have a high number of tissues and cell types, may be explained by the hypothesis that a high connectivity favors the functional diversification of paralogs, in particular through tissue specialization and expression divergence (Liang and Li, 2007; Makova and Li, 2003; Vogel and Chothia, 2006).

The fact that the human genome is able to retain duplications that are deleterious in yeast may be due to several reasons, in addition to the high number of cell types. *H. sapiens* has more efficient systems to adjust the expression levels, so that duplication does not necessarily translate into doubling the expression level, and to eliminate excess of subunits of protein complexes (such as ubiquitin, chaperones, proteases) (Liang and Li, 2007). Finally, the human genome may have a higher probability for a duplication to be advantageous, if a high dosage is required for a certain function, such as the immune response (Liang and Li, 2007).

6. Cancer as a genetic disease

The fact that cancer is a genetic disease has been known for more than a century, since when, between the late nineteenth and the early twentieth century, David von Hansemann (von Hansemann, 1890) and Theodor Boveri (Boveri, 1914) identified chromosomal aberrations in dividing cancer cells. Indeed, cancer arises from genetic and

epigenetic alterations that confer selective growth advantages to the cell (Hanahan and Weinberg, 2000; Merlo et al., 2006; Vogelstein and Kinzler, 2004). All cancers are thought to develop in similar ways (Stratton et al., 2009). Tumor is the result of an evolutionary process that occurs among cell populations in a close environment delimited by the multicellular organism (Merlo et al., 2006; Stratton et al., 2009). The evolution of the cancer tissue relies on two distinct processes: the continuous acquisition of genetic alterations of single cells and the Darwinian selection acting on the genetically different cell populations (Merlo et al., 2006; Stratton et al., 2009). As a consequence of these two processes, the cancer tissue is constituted by a genetically heterogeneous population of cells that is continuously evolving (Merlo et al., 2006; Yachida et al., 2010).

Not all genomic alterations have the same effects on the cell. Indeed, some are neutral and do not affect the cell at all. These are termed passenger mutations (Stratton et al., 2009). A small fraction of mutations instead confers the growth advantage to the cell and is positively selected because they grant a selective advantage to the cell, compared with the rest of the tumor cells and the surrounding normal tissue (Merlo et al., 2006; Stratton et al., 2009). These mutations are termed "driver" and may alter the protein sequence in several different ways (Figure 10). They may be single base substitutions that impact the protein structure in different ways, by altering the protein sequence through a single residue substitution (missense mutations) or by introducing premature stop codons (nonsense mutations). Insertions and deletions of nucleotides into a coding sequence always induce non-synonymous mutations. When the insertions or deletions involve three (or its multiples) nucleotides, the effect is the insertion or the deletion of one or more aminoacid residues or a premature stop codon. If, instead, the insertion or deletion involves a different number of nucleotides (frameshift mutation), it will totally disrupt the aminoacid sequence, by changing the codon code. Other mutations may affect the splice site, inducing aberrant isoforms, or non-coding regions, the UTRs in particular, thus changing the regulation by miRNAs, or intergenic regions, with a possible impact on

promoters. However, it is harder to determine the effect of mutations that do not change the protein sequence because there is not a direct effect on the protein structure. The result is rather a change in gene expression in case of mutations in the promoter region or in the miRNA binding site.

Figure 10: non-synonymous mutations

Different types of mutations change the aminoacid sequence. Single-base substitutions may change the corresponding aminoacid residue (missense mutation) or introduce a premature STOP codon (nonsense mutation). Insertion of one nucleotide changes the downstream aminoacid sequence (frameshift mutation). Deletion of three nucleotides induces the elimination of one aminoacid residue.

Cancer may arise from the alterations of three types of genes: oncogenes, tumor suppressors and stability genes (Vogelstein and Kinzler, 2004). However, a single alteration is not sufficient to start tumorigenesis and several genes must be altered in order for the cell to develop cancer (Vogelstein and Kinzler, 2004).

The discovery of the first cancer-specific somatic mutation dates back to almost thirty years ago, when the first oncogene (*HRAS*) was found mutated in bladder carcinoma (Reddy et al., 1982). Mutated oncogenes become constitutively active when, under normal conditions, they would not be expressed. Three mechanisms may cause oncogene activation: chromosomal translocations, gene amplification or point mutations that affect residues involved in the regulation of gene activity (Vogelstein and Kinzler, 2004). One

mutated allele is sufficient to contribute to tumorigenesis (Vogelstein and Kinzler, 2004). An example of oncogenic activation is represented by the substitution of the valine at codon 599 into a glutamate in the *BRAF* gene (Davies et al., 2002). This residue is part of the activation loop of the kinase domain, which is regulated by the phosphorylation of the residues at codon 598 and 600. The presence of a glutamate at codon 599 mimics a phosphate group, thus constitutively activating the kinase (Davies et al., 2002). The constitutive activation of *BRAF* leads to aberrant cell growth by phosphoryilation of downstream targets, such as the extracellular signal-regulated kinase (ERK) (Wan et al., 2004).

Tumor suppressors need both alleles to be inactivated, in order to promote tumorigenesis (Vogelstein and Kinzler, 2004). They require mutations that impair their functions, reducing their activity (Vogelstein and Kinzler, 2004). Different types of genetic alterations impair the function of tumor suppressors genes, such as missense mutations in particular residues that are essential for their activity, nonsense mutations, insertions or deletions that result in a truncated protein or disrupt the whole protein structure, and epigenetic silencing (Vogelstein and Kinzler, 2004).

Stability genes are normally involved in activities related with DNA, such as DNA repair or mitotic recombination and chromosomal segregation. They maintain genetic alterations to a minimum level (Kinzler and Vogelstein, 1997). When inactivated, genetic alterations in all the genome increase significantly and, if these alterations involve oncogenes or tumor suppressors, the tumor will develop (Friedberg, 2003; Kinzler and Vogelstein, 1997). As in the case of tumor suppressors, stability genes require inactivation of both alleles in order to decrease the genome stability and indirectly induce cancer (Vogelstein and Kinzler, 2004). Stability genes are referred to as the "caretakers" of the genome because they are involved in the maintenance of the genome, in contrast with "gatekeepers", which are genes that directly regulate the tumor progression (Kinzler and Vogelstein, 1997).

7. The hallmarks of cancer

Notwithstanding the type and number of mutations, all cancers need eight types of alterations in the cell physiology in order to develop (Hanahan and Weinberg, 2000; Hanahan and Weinberg, 2010) (Figure 11).

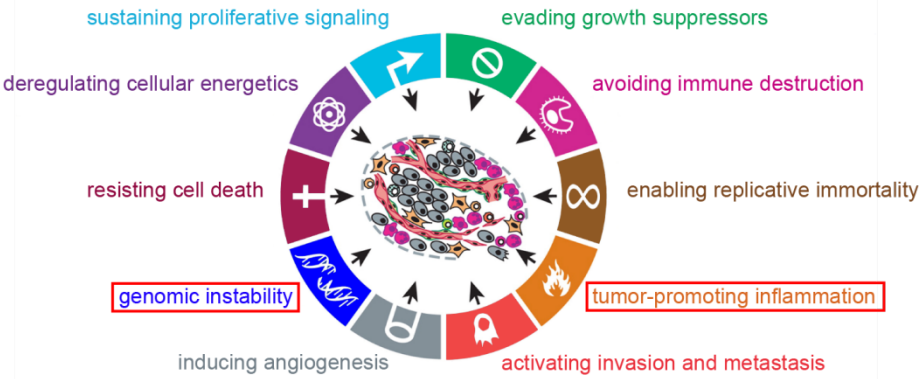


Figure 11: Hallmarks of cancer

Adapted from (Hanahan and Weinberg, 2000; Hanahan and Weinberg, 2010). All tumors arise because a cell population has acquired eight types of functional capabilities: self-sufficiency in growth signals, insensitivity to antigrowth signals, evasion from apoptosis, limitless replicative potential, sustained angiogenesis, tissue invasion and metastasis, reprogramming the energy metabolism and evading immune suppression. Furthermore, genome instability and chronic inflammation (depicted in red squares) are enabling characteristics that promote the tumor progression.

Most importantly, cancer cells must gain the ability to sustain chronic proliferation. Tumor cells are able to generate their own growth signals or their receptor signaling may be impaired: high levels of surface receptor induce hypersensitivity to their ligand or mutations in the receptor sequence may render them always active. The same result may be obtained by activating components of signaling pathways downstream of the receptor, thus separating the activation mechanisms from the receptor activation.

Cancer cells become insensible to the mechanisms that negatively regulate cell proliferation. These signals usually inhibit transcription factors that are able to activate the progression from G_1 to S phase and enter mitosis. An example of this is the retinoblastoma protein (RB), which, under normal conditions, blocks proliferation by altering the function of E2F transcription factors that control the expression of genes involved in the G_1 to S transition. Therefore, inactivation of the RB pathway renders E2F constitutively active.

A third alteration regards the resistance to cell death. This can be achieved by inhibiting the response to receptors that monitor the cell environment and should activate apoptosis in case of abnormalities, or by altering caspases and other genes that execute the cell death program.

The three alterations described so far are not sufficient for the tumor to establish, because normal cells have a limited replicative potential that is controlled by telomeres and telomerase. Telomeres protect the chromosomes from end-to-end fusions, which may cause uneven chromosomal segregation during mitosis. They shorten progressively at each replication and dictate the number of divisions that a cell is able to make. In immortalized cells, telomerase is able to add telomeres to each chromosome, increasing the replicative potential of the cell.

In order to survive, the tumor requires nutrients and oxygen and needs to eliminate wastes. Therefore it requires neovascularization. Whereas in normal tissues the endothelium and all the angiogenic material are quiescent, in the tumor tissue it is always activated because angiogenesis inducers are upregulated. However, the blood vessels generated inside the tumor tissue are aberrant and the endothelium presents abnormal levels of cell proliferation and apoptosis, which may also lead to microhemorrhaging.

Another acquired capability of the tumor is the ability of cancer cells to leave the tumor tissue, invade the surrounding tissues and establish metastases. These cells enter into blood or lymphatic vessels and reach distant tissues where they are able to exit the vessels and establish a new tumor colony.

Cancer cells develop a deregulated cellular metabolism. In particular, even in presence of oxygen, the glucose metabolism is reprogrammed to only glycolysis, which brings to a decreased efficiency of ATP production. This is balanced by the upregulation of glucose transporters, which increase the glucose import into the cytoplasm. Since the low oxygen intake may be a problem for cancer cells because the vascularization is aberrant, glycolysis and increased glucose intake may be positively selected, also as a response to the hypoxia due to the aberrant vascularization of the tumor.

The last hallmark of cancer is the escape of cancer cells from the immune system control. It is demonstrated that cancer cells are not targeted and destroyed by the immune system. However, it is still matter of debate why and how cancer cells are not targeted and eliminated by the patient's immune system.

7.1. Enabling characteristics of cancer

In addition to these eight hallmarks of cancer, two enabling characteristics emerge in all tumors and facilitate the acquisition of the hallmark (Figure 11).

First, genomic instability increases the genomic alterations of a cell. This is caused by impairments in the systems that detect and repair DNA defects and normally keep a low rate of mutations during each cell generation. Cancer cells often increase their mutation rate (Negrini et al., 2010). The causes of this may be numerous. The sensitivity to mutagenic agents may be increased because of mutations in genes that target these agents before their interaction with DNA. Increased mutability may be also caused by the impairment of the caretaker genes, which directly repair DNA damage or detect DNA damage and activate the repair mechanisms or force the damaged cell to enter apoptosis or senescence (Hanahan and Weinberg; Negrini et al.).

Tumor progression is favored also by inflammation. Every tumor contains immune cells that can contribute to many of the hallmarks described so far by supplying growth

factors, survival factors and inductive signals to the tumor environment during the immune response (Hanahan and Weinberg).

8. Known and candidate cancer genes

Cancer is a genetic disease that is caused by one or more mutations in oncogenes, tumor suppressors or stability genes. The patophysiological diversity and complexity of this disease are a consequence of the high heterogeneity at the genetic level. Indeed, in the last few years a massive effort has been made to identify the cancer genes, *i.e.* the mutated genes that are causally implicated in tumorigenesis (Futreal et al., 2004). In particular, two ways have been travelled:

1. The identification of cancer genes from a literature search, in order to collect genes that are already known to be involved in tumorigenesis. This strategy has allowed the publication of the Cancer Gene Census (Futreal et al., 2004) and the census of amplified genes (Santarius et al., 2010);
2. The mutational screenings of cancer tissues in order to identify novel candidate cancer genes. To date, more than 25 experiments of this type have been performed on several different cancer types.

8.1. Cancer Gene Census

The Cancer Gene Census at the Wellcome Trust Sanger Institute (http://www.sanger.ac.uk/genetics/CGP/Census/) represents the biggest effort to catalogue cancer genes. A gene is included in the census if driver mutations in primary patient material have been reported independently by at least two studies (Futreal et al., 2004).

The number of known cancer genes has been continuously increasing, from 291 identified in 2004 (Futreal et al., 2004) to 457 genes in the latest release (March 22nd 2011). Mutations are nucleotide substitutions that lead to aminoacid changes, premature stop codons or alterations at the splice site, insertions and deletions in coding sequences, chromosomal rearrangements that lead to chimerical transcripts or gene deregulation due to alterations in the promoter regions, or copy-number modifications. Most of the genes in the census (313, 68.5% of all cancer genes) undergo genomic translocations (Figure 12). This corresponds to a bias in the census towards genes that are mutated in leukemia or lymphomas (229, 50.1%) (Figure 13). Almost 80% of all cancer genes are dominant and mutations in one allele are sufficient to contribute to tumorigenesis (oncogenes), while recessive genes (tumor suppressors) are one fifth of all cancer genes (Figure 14). They need mutations in both alleles in order to promote cancer formation.

The Cancer Gene Census identifies cancer genes by analyzing their mutations (Futreal et al., 2004). However, oncogenes may be activated and contribute to tumorigenesis also if their dosage is significantly increased by overexpression or genomic amplification. In order to identify genes that are activated with these mechanisms, a census of amplified genes in cancer was recently published (Santarius et al., 2010). It includes 77 genes that present evidence of being both amplified at the genomic level and overexpressed at the RNA level (Santarius et al., 2010).

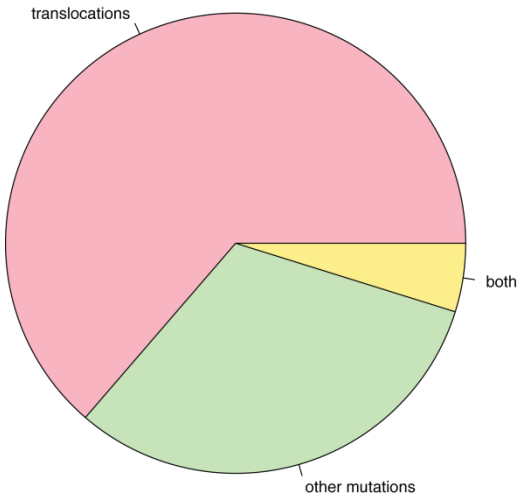


Figure 12: translocations in the Cancer Gene Census

Adapted from (Futreal et al., 2004). The majority of cancer genes in the Cancer Gene census is represented by translocation, while less than a third involves other mutation types.

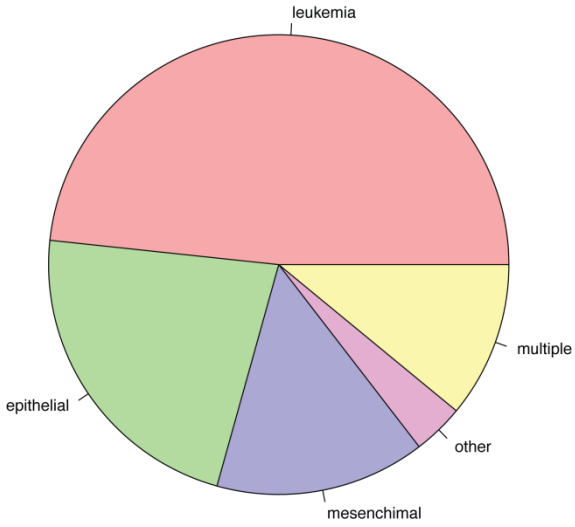


Figure 13: tumor types in the Cancer Gene Census

Adapted from (Futreal et al., 2004). Almost half of the cancer genes from the Cancer Gene Census harbor mutations that drive leukemia. Less than 15% of the genes are involved in more than one cancer type. “multiple” refers to genes that are involved in at least two cancer types.

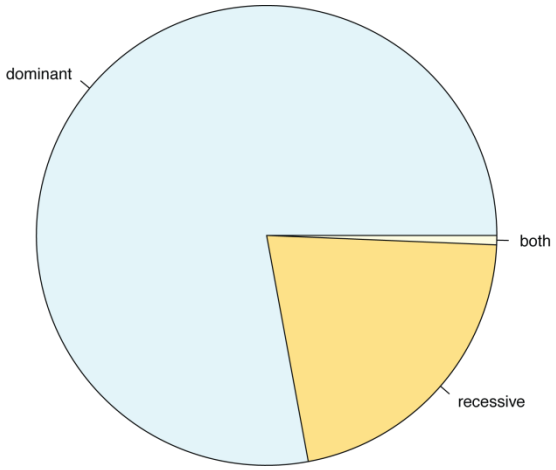


Figure 14: dominant and recessive genes from the Cancer Gene Census

Adapted from (Futreal et al., 2004). Dominant genes represent almost 80% of all cancer genes. “both” represents three genes (CREBBP, CBL, PRKAR1A) that may have both dominant and recessive behavior.

8.2. Novel candidate cancer genes

A further level of the mutational complexity of cancer has been added by the several high-throughput screenings of the cancer genome that were performed in the last five years. Starting from the early work by Sjoblom *et al.* (Sjoblom et al., 2006) on breast and colorectal cancers, more than twenty experiments have been published, which identified more than 1,000 candidate cancer genes. These experiments may be of two types: high-throughput mutational screenings that sequence part or all human genes in order to detect genes that are frequently mutated in cancer, and whole genome sequencing experiments that identify the complete mutational landscape of a single cancer patient.

In the last four years, eighteen high-throughput mutational screenings were published, which identified almost 20,000 mutations in more than 7,000 genes. The identification of candidate cancer genes was performed on the basis of the mutation type (*i.e.* synonymous or non-synonymous) and the mutation frequency (*i.e.* how often a gene was found mutated among the analyzed samples). The high-throughput mutational screenings of cancer tissues may be divided into two categories, depending on the number of sequenced genes: whole exome sequencing studies and sequencing of a smaller subset of genes. Twelve screens were performed on the whole exome of one or more patients: the number of screened genes varied between studies, but it ranged between 18,000 and more than 20,000 genes. Five experiments were performed on a selected set of genes, which were chosen because either they were already known to be involved in cancer or mutations in their sequence might be related to cancer because of their function. This latter category of experiments is biased towards already known mutated genes or genes whose mutations are likely to be involved in tumorigenesis because they are functionally related to known cancer genes. A single experiment (Greif et al.) represents an intermediate situation, having 10,000 screened genes.

The experiments of whole genome sequencing do not present any biases towards already studied genes and allow the identification of mutations in intergenic regions.

Different sequencing techniques were used to detect mutations in 39 patients from nine distinct experiments and in two cancer cell lines (Clark et al.; Pleasance et al.). From these eleven screenings, candidate cancer genes were identified as the genes that presented non-synonymous mutations, which were validated further with orthogonal methods, such as Sanger sequencing.

8.3. Methods to identify driver mutations

Since the vast majority of the mutations identified in the high-throughput mutational screenings are passenger, statistical methods must be used to identify possible candidate cancer genes on the basis of their mutation frequency. The pipeline to detect driver mutations in whole exome sequencing experiments is usually supported by two screenings: discovery and validation (Sjoblom et al., 2006) (Figure 15). The discovery screening is used to identify mutations in the cancer patients in an unbiased way. First, the initial dataset of all human genes to screen for mutations is defined. Second, the corresponding genomic regions are amplified and sequenced. The analysis of the sequencing results identifies putative nucleotide changes, which must be filtered to eliminate synonymous mutations and known polymorphisms. In order to remove artifacts due to the amplification or the sequencing steps, the regions containing mutations are re-sequenced and only mutations that are found also in this second run are considered as true. Finally the comparison with the matched normal sample allows the filtering of all germline variants and the retention of the real mutations (Figure 15). This pipeline was originally published in 2006 (Sjoblom et al., 2006) and more recent works used a modified version, which normally does not involve re-sequencing of the mutated regions for confirmation.

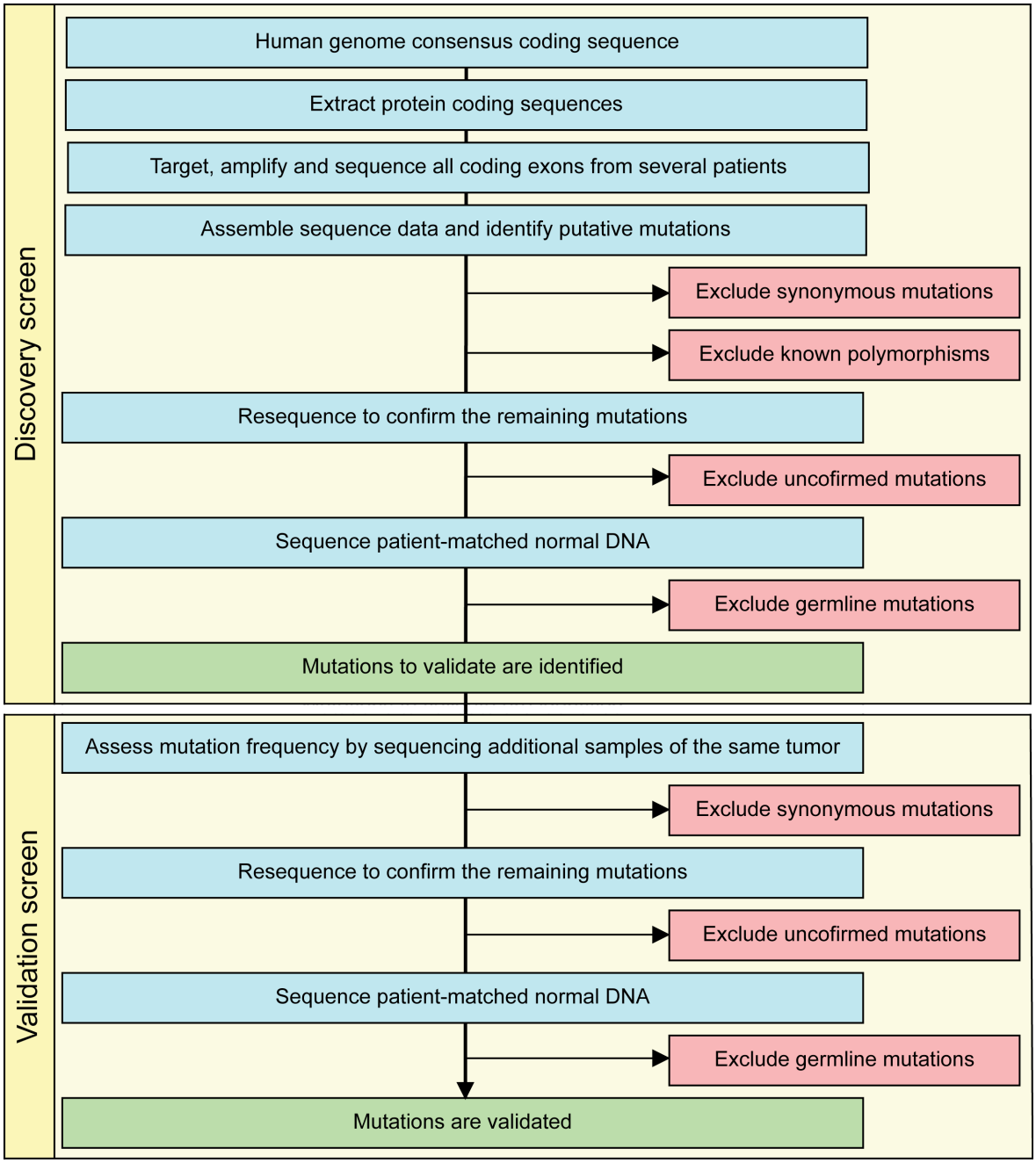


Figure 15: Pipeline to detect candidate cancer genes in whole-exome mutational screenings

Adapted from (Sjoblom et al., 2006). The scheme describes the steps to detect and validate candidate cancer genes on the basis of their mutations. In the discovery screen, all the human genes are identified and sequenced. The putative mutations are detected and filtered to eliminated synonymous mutations and known polymorphisms. A step that is eliminated from later experiments consists in the resequencing of all the mutations that have passed the two filters. The *bona fide* real mutations are then compared with the corresponding normal sample in order to exclude germline mutations. A second screening is then performed for all the mutated genes on a larger set of samples from the same tumor, in order to validate the mutations identified in the first screening. All the mutations that are present also in this validation screen are considered as “driver”.

In order to discriminate between driver and passenger mutations, all mutated regions are sequenced in a larger set of patients with the same cancer (Figure 15). Candidate cancer genes are mutated in both the discovery screen and this second screen, which is referred to as the validation screen, while mutations that occur only in the discovery screen are expected to be passenger (Sjoblom et al., 2006).

The five experiments that were performed on selected genes use an alternative method. Instead of screening all human genes for mutations, a subset of human genes is selected for mutation detection (Barretina et al., 2010; Dalgliesh et al., 2010; Ding et al., 2008; Kan et al., 2010; McLendon et al., 2008). These genes are usually known cancer genes or genes that, because of their function, are likely to be involved in cancer, such as kinases, genes involved in protein degradation or associated with receptor signaling. These genes are screened for mutations in a high number of cancer samples and the distinction between candidate cancer genes and genes with passenger mutations is made upon the mutation frequency among the cancer samples.

High-throughput mutational screenings that focus on the entire human exome are the first unbiased studies to identify candidate cancer genes. This approach is completely different from the Cancer Gene Census, which instead collects data from several small-scale experiments. These experiments are normally hypothesis-driven, therefore only genes that are likely to be involved in cancer are studied. This likelihood is related with their function or with previous reports of the same genes or pathways to be modified in cancer.

Whole exome mutational screenings are not the only high-throughput screenings to identify putative cancer genes. First Davies *et al*. (Davies et al., 2005), then Greenman *et al.* (Greenman et al., 2007) investigated the presence of mutations among 518 kinases in several different cancer samples. Driver mutations were discriminated from passenger on the basis of the ratio between non-synonymous and synonymous substitutions (Davies et al., 2005; Greenman et al., 2007). Kinases were selected for these screenings because

previous reports had detected the kinase domain as the most frequently mutated in cancer (Futreal et al., 2004). Therefore these experiments were biased towards genes that were already known to be enriched in driver mutations.

Another unbiased approach to identify mutated genes is whole genome sequencing, although this method allows the identification of mutations in only one or few patients. The pipeline to identify mutations was first proposed by Ley *et al*. (Ley et al., 2008) (Figure 16) and envisages several filters. First, single nucleotide variants (SNVs) are identified inside the tumoral DNA. All the variants that are present also in the respective normal DNA are eliminated because they are *bona fide* SNPs. This filter eliminates the vast majority of variants, reducing their number from hundreds of thousands or millions to tens of thousands. From all the tumor-specific variants, those that are known SNPs are excluded from further analyses. In order to analyze only mutations in the coding sequence, all those that fall inside intergenic regions or introns are eliminated. One last filter eliminates synonymous substitutions and what remains must be validated with orthogonal methods, usually with Sanger sequencing, in order to eliminate false positives (Ley et al., 2008) (Figure 16).

Whole genome sequencing of the cancer genome does not allow the identification of driver and passenger mutations. However, a detailed study of the mutations that are identified allows the confirmation of already known cancer genes or the discovery of putative cancer genes, because of their function or because they interact with already known cancer genes.

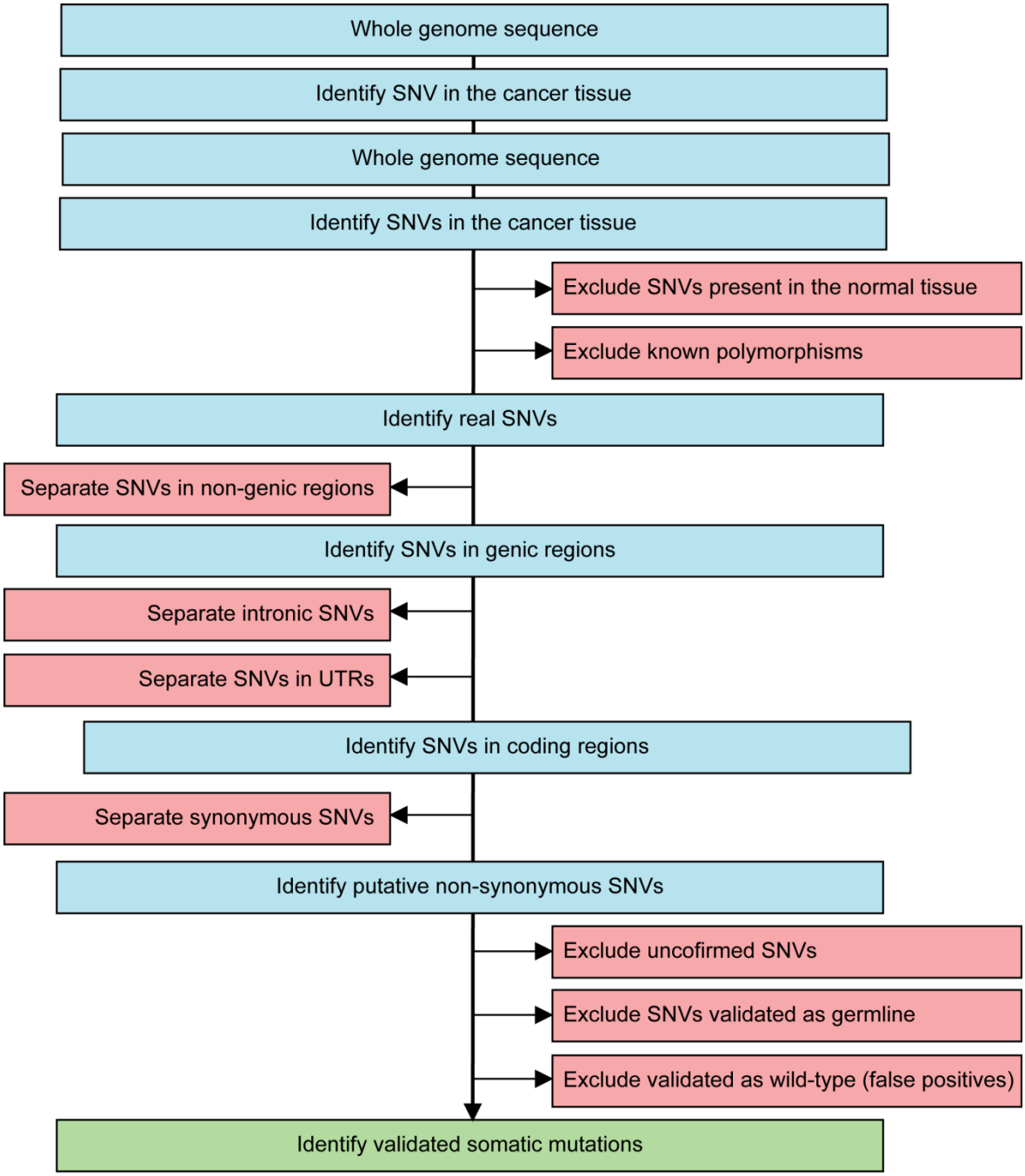


Figure 16: Pipeline to detect true mutations in whole genome sequencing studies

Adapted from (Ley et al., 2008). Of all the SNVs detected by whole genome sequencing, all known SNPs and germline variants are eliminated. All SNVs are validated with orthogonal methods, such as Sanger sequencing, in order to identify real mutations.

9. Systems-level properties of cancer genes

High-throughput mutational screenings and whole genome sequencing experiments of the cancer genome, together with the Cancer Gene Census, identified in the last few years ~1,500 genes that are actively involved in cancer, which account for at least 7% of all human genes. Cancer is a highly heterogeneous disease, which may be caused by an enormous amount of genomic alterations, considering also the fact that only few genes are mutated in many cancer types (Figure 17) (Ciccarelli, 2010). A means to reduce this complexity is to focus on mutated pathways, rather than single genes (Vogelstein and Kinzler, 2004). Two examples were extensively studied in the past few years: RB and p53 pathways (Figure 18). The RB pathway controls the cell's transition from a resting stage (G_0 or G_1) to replication (S phase). Several oncogenes (CDK, cyclin D1, TAL1 and TFE3) and tumor suppressors (RB, p16, CDKN2A) are involved in this pathway (Figure 18A). p53 is a transcription factor that inhibits cell growth and promotes apoptosis. It is a tumor suppressor that is found mutated frequently in most tumor types (Vogelstein et al., 2000). However, also mutations in many upstream and downstream genes were found to be involved in tumorigenesis. In particular, amplification of MDM2 induces a faster degradation of p53 and, consequently, blocks apoptosis (Figure 18B) (Vogelstein and Kinzler, 2004). In addition to these, genes involved in several other pathways were found mutated in multiple tumors, such as hypoxia-inducible factor (HIF), WNT, phosphoinositide 3-kinae (PI3K), small mother against decapentaplegic (SMADs) and receptor tyrosine kinases (RTKs) (Vogelstein and Kinzler, 2004). The large number of genes that are mutated in each pathway implies that impairments in these pathways are associated with tumorigenesis. However, mutations are mutually exclusive: mutations in only one gene involved in a single pathway are sufficient to promote tumorigenesis, since the functional effect of all mutation is similar (Vogelstein and Kinzler, 2004).

Even considering mutated pathways instead of mutated genes does not significantly reduce the complexity and the heterogeneity of cancer. The only common feature shared

by all cancers seems to be the uncontrolled proliferation of tumor cells. Nevertheless, several groups in the last few years have investigated whether cancer genes have common properties, in order to understand whether the complexity of cancer may be reduced. These systems-level features may help identifying new putative cancer genes or discriminating between driver and passenger mutations.

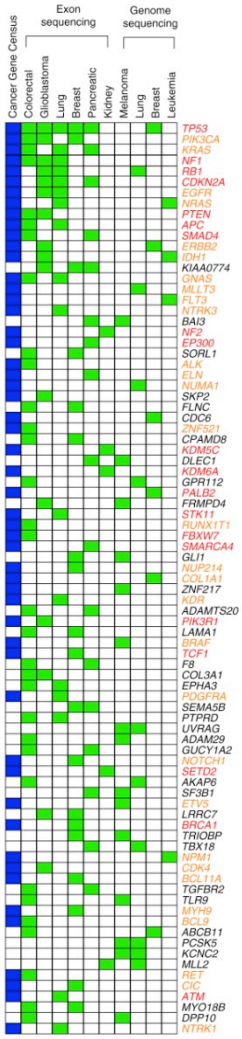


Figure 17: Heterogeneity of cancer genes identified in different cancer types

Adapted from (Ciccarelli, 2010). Of all candidate cancer genes identified in several different studies, only 85 were found mutated in at least two cancer types. The gene names are colored in red if they are recessive genes from the Cancer Gene Census, orange if they are dominant, black if they are not included in the Cancer Gene Census.

A large-scale approach to infer the effects of mutations is the analysis of protein-protein interactions: this allows identifying whether the position of a gene inside the

human network is indicative of its role in cancer (Ciccarelli, 2010). Indeed it was demonstrated that cancer genes are highly connected inside the human protein interaction network (Jonsson and Bates, 2006; Rambaldi et al., 2008).

This is indicative of an intrinsic fragility of the network: modifications of genes that encode proteins with few connections are less detrimental than highly connected proteins.

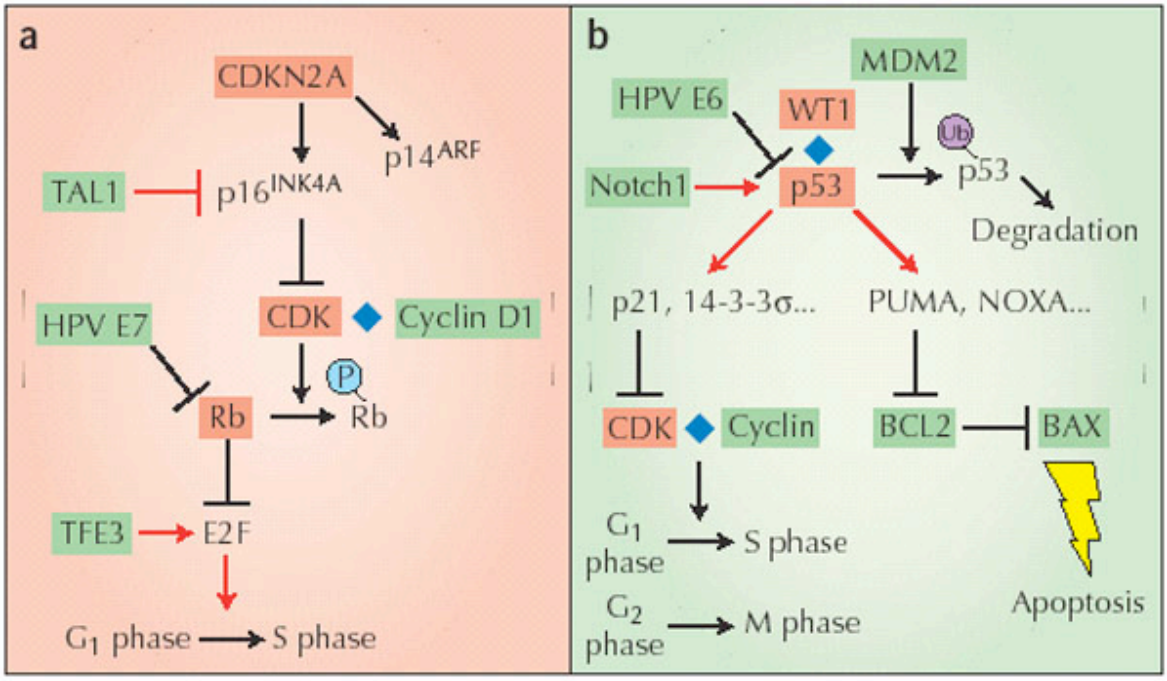


Figure 18: Cancer genes in Rb and p53 pathways

Adapted from (Vogelstein and Kinzler, 2004). Two examples of pathways that are frequently altered in cancer are displayed: (A) Rb and (B) p53. Green boxes indicate genes that are frequently somatically mutated in cancer, while red boxes show genes that harbor also germline mutations. Diamonds indicate protein-protein interactions. Red arrows and T-bars indicate transcriptional induction and repression.

Methods

1. Identification of unique gene sets

In order to determine how protein interaction networks evolve, we analyzed the genomic and network properties of four species: *E. coli*, *S. cerevisiae*, *D. melanogaster* and *H. sapiens*. As explained later, these species are the only whose protein interaction network includes at least 50% of their proteins.

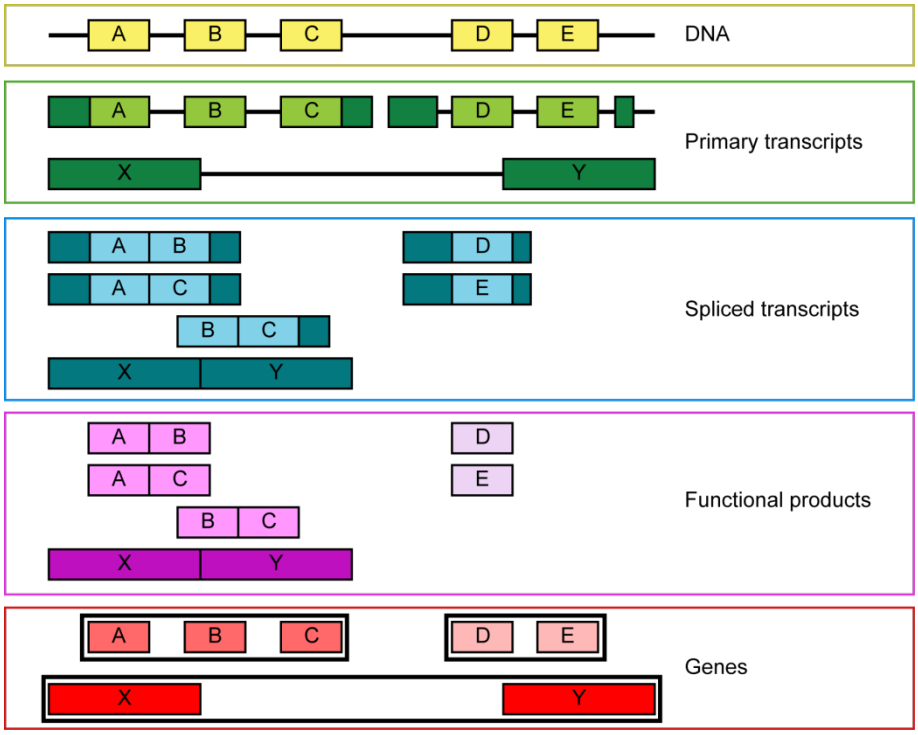


Figure 19: definition of gene

Adapted from (Gerstein et al., 2007). The difficulty to define a gene on the basis of overlapping transcripts is exemplified in the figure. A genomic region produces three primary transcripts (ABC, DE, XY). After alternative splicing ABC results in three different spliced transcript, DE in two, while XY in one, which does not translate into proteins. The three transcripts from ABC are overlapping, while D and E share only UTRs. Considering both coding and non-coding sequences to identify genes, this region includes three genes: ABC, DE and XY.

The identification how many unique genes are present are present in a eukaryotic species is not a trivial issue, given the complexity of the genome (Gerstein et al., 2007). Even trying to determine the definition of a gene is tricky, since many features must be taken into account, such as alternative splicing, non-coding RNAs, pseudogenes, overlaps between different transcripts (Gerstein et al., 2007) (Figure 19). When analyzing genes and gene properties, this fact is not taken into consideration, and the number of genes from the species of interest is retrieved from one of the available databases, such as Ensembl (Flicek et al.) or Entrez (Maglott et al.). We instead developed our method to determine the total number of human genes. Given that the central point of our analyses was protein-protein interaction networks, we did not consider non-coding genes. In order to identify a set of unique human genes, we aligned all protein sequences to the human genome, in order to identify the corresponding transcripts. We defined a gene as the union of all overlapping transcripts and we took the longest isoform as representative transcript.

Table 1: Identification of unique genes

Genes	***H. sapiens***	***D. melanogaster***	***S. cerevisiae***	***E. coli***
Initial protein sequences	38,015	14,134		
Initial genes	25,635	14,134	**6.752**	**4.497**
Proteins with BLAT hit on the genome	37,752 (99.3%)	14,111 (99.8%)		
Removed isoforms	15.506	266		
Non-overlapping Genes	22,160 (86.4%)	13,819 (97.7%)		
Spurious hits	139	36		
Unique genes with best hit >60%	**22,020 (99.4%)**	**13,783 (99.7%)**		

For *S. cerevisiae* and *E. coli*, the unique genes are directly derived from SGD and EcoCyc. For *H. sapiens* and *D. melanogaster*, the initial dataset of genes must be filtered in order to eliminate overlapping genes and retain only one isoform per gene locus. The initial protein sequences are BLATted to the corresponding genomes and only the longest isoform is retained for each locus. A second filter is then applied to eliminate all spurious hits, *i.e.* all hits below 60% of their protein length.

As starting datasets of genes, we used RefSeq v. 37 for human (Pruitt et al., 2007), FlyBase FB2009_01 for *D. melanogaster* (Drysdale, 2008), the Saccharomyces Genome

Database (SGD) for *S. cerevisiae* (frozen at January 5th 2010) (Engel et al., 2010) and EcoCyc v.14.0 for *E. coli* (Keseler et al., 2009) (Table 1). *S. cerevisiae* has 6,752 genes and *E. coli* has 4,497. For the two metazoans we applied a pipeline to retain one only isoform per gene and to identify the unique genes, since many isoforms may be present in the same locus. In order to verify the presence of a single gene per locus, we mapped the protein sequence to the corresponding genomic sequence as reported in Rambaldi *et al.* (Rambaldi et al., 2008) (Figure 20). First, we retrieved all protein sequences from the respective database (Table 1). Second, we aligned all these protein sequences to their genome reference assembly using the BLAST-like Alignment Tool (BLAT) (Kent, 2002) (Table 1). All the hits on alternate haplotype regions (*chr22_h2_hap1*, *chr5_h2_hap1*, *chr6_cox_hap1* and *chr6_qbl_hap2* on the human genome), heterochromatic sequences (*chr2LHet*, *chr2RHet*, *chr3LHet*, *chr3RHet*, *chrXHet*, and *chrYHet* on the fly genome) or on random chromosomes and plasmids were eliminated. 37,752 human sequences and 14,111 fly sequences passed this first filter (Table 1). Then we applied a second filter in order to retain only one isoform for each gene. All the proteins with overlapping best hits were clustered together and only the longest protein was retained as representative for the cluster. This allowed us to remove 15,506 human and 266 fly redundant isoforms (Table 1). Afterward, a third filter was applied. All the genes with their best hit shorter than 60% of the original protein length were removed. This was done to eliminate spurious hits, which do not correspond to the original locus of the gene. Following the results of the application of these three filters, we concluded that *H. sapiens* has 22,020 unique non-overlapping genes, while *D. melanogaster* has 13,783 (Table 1).

In addition to these four species, we analyzed the genomic properties of seven other species, in order to have a more comprehensive view of the evolution of these properties. We added *M. musculus*, *G. gallus*, *D. rerio*, *A. millifera*, *C. elegans*, *S. pombe* and *B. subtilis*. In order to determine the number of unique genes, we used RefSeq v. 37 for the

three vertebrates (Pruitt et al., 2007) and Ensembl v. 46 for the other species (Flicek et al., 2011).

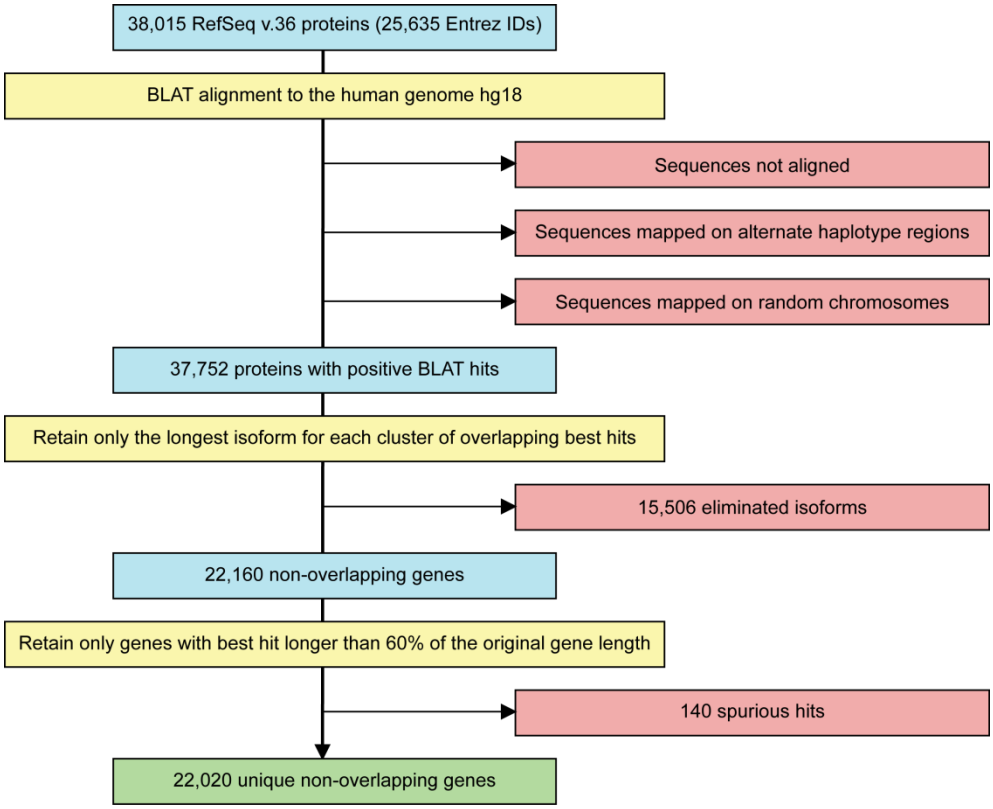


Figure 20: Pipeline to identify unique human genes

In order to identify a unique set of non-overlapping human genes, 38,015 RefSeq protein sequences were aligned to the human genome using BLAT. Once eliminated all the sequences that did not map to real chromosomes, all the sequences with overlapping hits were clustered together and only the longest was retained. A last filter eliminated all best hits shorter than 60% of the original protein length.

2. Identification of cancer genes

Cancer genes are genes whose mutations are causally implicated in oncogenesis (Futreal et al., 2004). On the basis of this definition and the available data, we identified three subsets of cancer genes, which differ for the methods of identification: known cancer genes from the Cancer Gene Census and from the census of amplified genes in cancer (Futreal et al., 2004; Santarius et al., 2010), candidate cancer genes from high-throughput

mutational screenings and genes with non-synonymous mutations in whole genome sequencing experiments of cancer tissues.

We used the 427 genes included in the Cancer Gene Census version of March 30th 2010 (Futreal et al., 2004). A first comparison with the 22,020 unique human genes allowed us to eliminate 20 cancer genes, because they were not included in the set of unique genes, since they were either discontinued or not associated with RefSeq proteins. In addition to these genes, we retrieved the 77 genes amplified in cancer from Santarius *et al.* (Santarius et al., 2010), which were all included in the 22,020 unique human genes. In total, we were able to retrieve 480 known cancer genes, of which 460 (96%) were also included in the dataset of unique human genes. For the latest analyses, we retrieved a more recent version of the Cancer Gene Census (March 22nd 2011), which included 447 genes, thus bringing the total known cancer genes to 501 (Table 2).

To identify the candidate cancer genes from high-throughput mutational screenings, we gathered the mutational information from 18 experiments, which included 12 whole-exome sequencing studies, 5 experiments on selected gene lists and one that sequenced 10,000 genes (Table 2). More than one third of all human genes were found mutated in at least one patient. In order to detect candidate cancer genes in this vast mutational landscape, only genes that were mutated at high frequency or that were also found mutated in validation screenings were considered. This allowed us to identify 699 candidate cancer genes, of which 686 (98%) were included in the list of unique human genes. The number of cancer genes from each experiment is highly heterogeneous, ranging from 5 to 140. Also when considering the same tumor type, differences are significant. For example, two experiments on pancreatic cancer identified 7 and 82 candidate cancer genes, respectively (Jiao et al., 2011; Jones et al., 2008).

Table 2: experiments to identify cancer genes

Type	Screen	Candidates	Candidates in NCG 3.0	Mutated genes	Mutations	Screening type	Cancer type
CGC	Futreal, Nature 2004 (March 30th 2010)	427	NA	0	0	literature-based	leukemia, mesenchimal, epithelial, dominant, recessive
	Futreal, Nature 2004 (March 22nd 2011)	447	444	0	0	literature-based	leukemia, mesenchimal, epithelial, dominant, recessive
	Santarius, Nature 2010	77	77	0	0	literature-based	amplified
	All	**501**	**498**	**0**	**0**		
HTMS	Agrawal, Science 2011	6	6	725	911	whole-exome sequencing (18,000 genes)	head and neck squamous cell carcinoma
	Barretina, Nature Genetics 2010	21	21	21	46	722 genes sequenced	sarcoma
	Cancer Genome Atlas, Nature 2008	8	8	223	453	601 genes sequenced	glioblastoma
	Chapman, Nature 2011	10	10	498	560	whole-exome sequencing (164,687 exons)	myeloma
	Dalgliesh, Nature 2010	5	5	398	722	3,544 genes sequenced	kidney
	Ding, Nature 2008	26	26	357	1,013	623 genes sequenced	lung
	Greif, Leukemia 2011	5	5	5	5	10,000 genese sequenced	acute myeloid leukemia
	Gui, Nature Genetics 2011	21	21	328	465	whole-exome sequencing (18,000 genes)	bladder
	Jiao, Science 2011	7	7	150	231	whole-exome sequencing (18,000 genes)	pancreas
	Jones, Science 2008	82	81	1,253	1,823	whole-exome sequencing (20,661 genes)	pancreas
	Kan, Nature 2010	112	112	967	2,576	1,507 genes sequenced	breast, lung, ovarian, pancreas, prostate
	Li, Nature Genetics 2011	5	5	411	429	whole-exome sequencing (18,000 genes)	liver
	Parsons, Science 2008	42	42	1,940	2,449	whole-exome sequencing (20,661 genes)	glioblastoma
	Parsons, Science 2010	9	9	218	225	whole-exome sequencing (225,752 exons)	medulloblastoma
	Pasqualucci, Nature Genetics 2011	54	54	93	96	whole-exome sequencing (180,000 exons)	large B-cell lymphoma
	Stransky, Science 2011	76	76	NA	NA	whole-exome sequencing (188,260 exons)	head and neck squamous cell carcinoma
	Wei, Nature Genetics 2011	68	68	3,026	4,226	whole-exome sequencing (20,000 genes)	melanoma
	Wood, Science 2007	272	272	1,881	2,693	whole-exome sequencing (18,191 genes)	breast, colorectal
	All	**699**	**698**	**7,439**	**16,797**		
WGS	Berger, Nature 2011	88	88	156	162	WGS for 7 patients	prostate
	Chapman, Nature 2011	10	10	1,036	1,418	WGS for 23 patients	myeloma
	Clark, PLOS Genetics 2010	60	60	61	62	WGS for U87MG cell line	glioblastoma
	Ding, Nature 2010	49	49	49	49	WGS for one patient	breast
	Lee, Nature 2010	16	16	344	373	WGS for one patient	lung
	Ley, Nature 2008	10	10	10	10	WGS for one patient	acute myeloid leukemia
	Mardis, NEJM 2009	12	12	12	12	WGS for one patient	acute myeloid leukemia
	Pleasance, Nature 2010	60	59	133	137	WGS for NCI-H209 cell line	lung
	Pleasance, Nature 2010	62	61	276	299	WGS for one patient	melanoma
	Shah, Nature 2009	31	30	563	776	WGS for one patient	breast
	Totoki, Nature Genetics 2011	71	71	71	72	WGS for one patient	liver
	All	**457**	**454**	**2,439**	**3,370**		
All	**All**	**1,499**	**1,494**	**8,531**	**20,167**		

Experiments to detect cancer genes are grouped into three categories, depending on the method of identification: Cancer Gene Census (CGC), High-Throughput (HTMS) and Whole Genome Sequencing (WGS). For HTMS and WGS, the screening type is derived from the Methods section of each experiment. For HTMS the number of genes or, as an alternative, exons that are screened for mutations is reported, while for WGS the number of screened patients or the number name of the screened cell line is reported. The columns "Mutated genes" and "Mutations" are derived from the Supplementary Information from each experiments. For Stransky *et al.* (Stransky et al., 2011) it was not possible to derive this type of information. The comumn "Candidates in NCG 3.0" represents the number of candidates that could be successfully mapped to up-to-date Entrez IDs and were included in NCG 3.0 (http://bio.ifom-ieo-campus.it/ncg).

We identified 456 cancer genes from nine different experiments of whole genome sequencing of 41 cancer genomes (Table 2). Of these, 453 (99%) were included in the list of unique human genes. In opposition to the high-throughput mutational screenings, the heterogeneity in the number of detected genes is lower, having 10 and 12 genes mutated in acute myeloid leukemia (Ley et al., 2008; Mardis et al., 2009) and 31 and 49 in breast cancer (Ding et al., 2010; Shah et al., 2009).

The total number of cancer genes detected with the three different methods is 1,499, of which 1,464 (98%) were included in the list of unique human genes. This is the most complete collection of cancer genes to date, and includes 6.6% of all human genes.

3. Reconstruction of protein interaction networks

We integrated protein-protein interaction data from several sources, in order to build the most complete, to our knowledge, protein interaction networks of model species (Table 3).

Table 3: sources of protein-protein interactions

Species	Network	Database	Version	Proteins	Interactions	Experiments
E. coli	All	IntAct	Jan 23rd 2009	2,819	13,663	186
		DIP	Jan 26th 2009	1,456	5,646	418
			Total	**2,884**	**15,888**	**445**
	Gold Set	IntAct	Jan 23rd 2009	481	590	186
		DIP	Jan 26th 2009	529	675	418
			Total	**703**	**1,004**	**445**
S. cerevisiae	All	BioGRID	2.0.49 (Feb 1st 2009)	5,005	41,401	4,230
		IntAct	Jan 23rd 2009	5,563	45,580	476
		MINT	Feb 5th 2009	5,219	30,904	174
		DIP	Jan 26th 2009	4,898	17,222	1,278
		Yu	Oct 2008	1,228	1,636	1
			Total	**5,937**	**91,652**	**4523**
	Gold Set	BioGRID	2.0.49 (Feb 1st 2009)	3,671	16,425	4,228
		IntAct	Jan 23rd 2009	3,138	9,357	476
		MINT	Feb 5th 2009	2,880	7,920	174
		DIP	Jan 26th 2009	3,032	7,671	1,278
		Yu	Oct 2008	700	545	1
			Total	**3,930**	**21,731**	**4,521**
D. melanogaster	All	BioGRID	2.0.49 (Feb 1st 2009)	6,925	21,775	145
		IntAct	Jan 23rd 2009	7,729	23,671	148
		MINT	Feb 5th 2009	7,129	21,540	45
		DIP	Jan 26th 2009	7,015	21,630	35
		DroID	4.0 (Jul 2008)	7,139	22,872	282
			Total	**10,563**	**61,014**	**363**
	Gold Set	BioGRID	2.0.49 (Feb 1st 2009)	445	418	145
		IntAct	Jan 23rd 2009	831	1,083	148
		MINT	Feb 5th 2009	305	242	45
		DIP	Jan 26th 2009	229	183	35
		DroID	4.0 (Jul 2008)	940	1,186	282
			Total	**1,392**	**2,236**	**363**
H. sapiens	All	BioGRID	2.0.49 (Feb 1st 2009)	7,163	23,588	8,815
		IntAct	Jan 23rd 2009	7,066	22,119	1,374
		MINT	Feb 5th 2009	5,151	12,653	1,210
		DIP	Jan 26th 2009	1,108	1,326	739
		HPRD	Sep 1st 2007	8,697	34,938	17,770
			Total	**11,988**	**68,498**	**19,886**
	Gold Set	BioGRID	2.0.49 (Feb 1st 2009)	5,550	16,153	8,815
		IntAct	Jan 23rd 2009	2,936	4,887	1,374
		MINT	Feb 5th 2009	2,278	3,591	1,209
		DIP	Jan 26th 2009	1,058	1,207	739
		HPRD	Sep 1st 2007	7,296	26,910	17,770
			Total	**9,127**	**39,868**	**19,885**

For each database and each species, the number of proteins, interactions and Pubmed IDs supporting the interactions ("Experiments") are reported. The gold set includes all the interactions that are supported by single-gene experiments or by more more than one high-throughput experiment.

The first filter that we applied to the raw data gathered from each single database consists in the retention of only primary physical interactions: only direct evidence for the presence of interactions was considered, while all interactions inferred from orthology were discarded. Second, in order to integrate the data from the different sources, we chose a unique protein identifier and converted all protein identifiers to this. For human we used Entrez IDs, for *D. melanogaster* FlyBase IDs, for *S. cerevisiae* SGD IDs and for *E. coli* EcoCyc IDs. The mappings to these identifiers were done by querying BioMart (Haider et al., 2009) and the Protein Identifier Cross-Reference (PICR) Service (Cote et al., 2007) with the lists of protein IDs that were not in the chosen format yet. Although some genes were lost because they could not be mapped to the chosen identifiers, this allowed us to eliminate redundancies between the different databases and to have a unique type of protein identifiers that may be easily converted into other types in order to integrate network data with other sources, such as orthology.

The integration of data from different sources allowed us to have a significantly more complete view of the interactions that each protein undertakes, compared to using single databases. For example, the human protein TP53 has between 27 (DIP) and 239 (HPRD) interactions, while the integration of all databases allowed us to detect interactions with 409 distinct proteins. Of these, 209 (51%) were detected in more than one database.

After building the protein interaction networks, we divided all the 25,217 publications into two categories: high-throughput and single-gene. Since a manual characterization of the experimental type could not be made, we set an arbitrary cutoff to 100 interactions. Each experiment associated with at least 100 interactions was labeled as "high-throughput", single-gene otherwise. This division allowed us to distinguish between four categories of interactions:

1. Supported by single-gene experiments;
2. Supported by high-throughput and single-gene experiments;
3. Supported by more than one high-throughput experiment;
4. Supported by only one high-throughput experiment.

From this division, we identified a gold set of high-confidence interactions, which included all the interactions associated with one of the former three categories. We eliminated the interactions supported by only one high-throughput experiment from the gold set, because this category is the most enriched in false-positives (Bader et al., 2004; von Mering et al., 2002).

In order to determine how networks evolve, for each protein in each network we measured three properties: degree, clustering coefficient and betweenness. Degree is a measure of the connectivity of a protein, *i.e.* the number of interactions that a protein has in the protein interaction network. Clustering coefficient is a measure of the local interconnectivity and is calculated as the ration between the number of interactions among the first-level neighbors of the protein of interest and the total possible interactions between them (Watts and Strogatz, 1998). Betweenness is a measure of the centrality of a protein and is calculated as the number of shortest paths that pass trough that protein.

4. Orthology and paralogy assignment

We derived the orthology relationships from eggNOG 1.0 (Jensen et al., 2008).

On the basis of the lineage-specific orthologous groups defined in eggNOG, we derived a simplified version of the tree of life, with seven internal nodes that correspond to major transitions in evolution (last universal common ancestor, eukaryotes, opisthokonts, metazoans, vertebrates, mammals, and group-specific transition) (Figure 21).

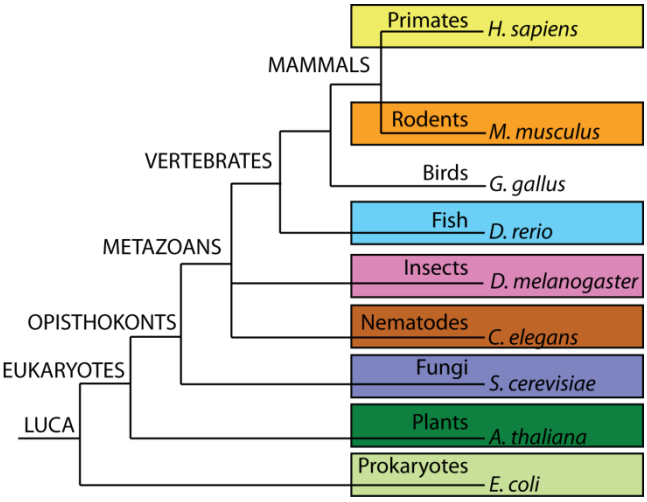


Figure 21: Tree of life

In order to detect origin and conservation of genes, we built a simplified version of the tree of life, with six internal nodes. LUCA represents Last Universal Common Ancestor, *i.e.* the last ancestor before the split between prokaryotes and eukaryotes.

We identified the 373 species present in eggNOG and assigned each to the most specific internal node. In particular, we identified 338 prokaryotes, 4 plants, 8 fungi, one nematode, 3 insects, 6 other invertebrates, 3 fish, 2 rodents, 3 primates, 3 other mammals and 2 other vertebrates. For example, *H. sapiens* is representative of primates, *D. melanogaster* of insects, *S. cerevisiae* of fungi and *E. coli* of bacteria. These group-specific nodes do not reflect comparable evolutionary transitions. Indeed, for human we were able to obtain a better resolution than for the other species. In particular, the presence of three primate species (*H. sapiens*, *P. troglodytes* and *M. mulatta*) in addition to five other mammals (*M. domestica*, *B. taurus*, C. *familiaris*, *M. musculus* and *R. norvegicus*) allowed us to discriminate between human genes that originated with primates and those that originated with mammals. The presence of only three distantly related insects (*D. melanogaster*, *Apis mellifera* and *Anopheles gambiae*), instead, allowed a maximum resolution of insect-specific genes. For *S. cerevisiae* and *E. coli* the resolution was even lower, and we were able to detect fungi-specific and prokaryote-specific genes.

The presence of different levels of resolution did not introduce biases to the analyses, given a very low fraction of group-specific genes in all species. Other than identifying the group-specific genes, we performed the analysis of the orthology relationships to identify a common evolutionary trend among species from different taxa. Since the number of group-specific genes is very low, it does not affect the overall trend in a significant way.

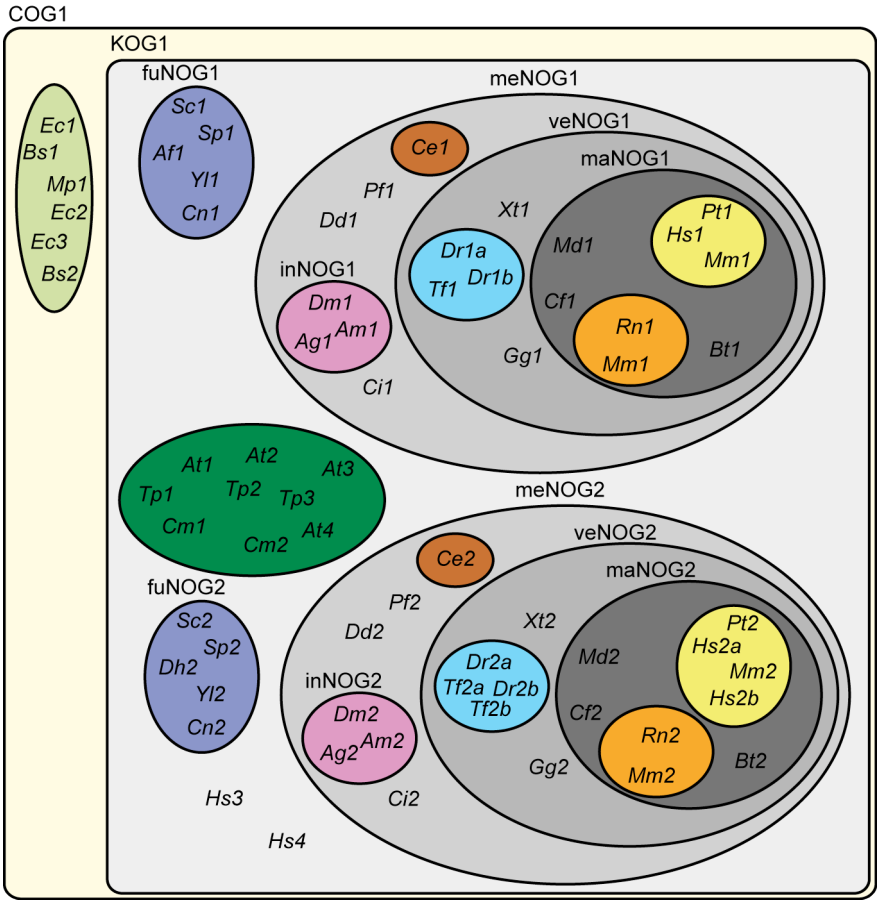


Figure 22: resolution of the clusters of orthologs

eggNOG builds clusters of orthologs with different levels of resolution. The algorithm to assign the orthology starts to allocate orthology relationships between highly similar genes, and creates the first clusters of orthologs maNOG1 and maNOG2. By relaxing its parameters, more distantly related orthologs are assigned to the genes inside these clusters, and the vertebrate-specific clusters (veNOG1 and veNOG2) are built. With the same parameters, also the insect-specific clusters (inNOG1 and inNOG2) are built. A further relaxation allows the identification of metazoan-specific orthologs in meNOG1 and meNOG2. These two clusters have a significant similarity and they are included in the same KOG (KOG1), together with plant-specific genes and the two fungi-specific clusters (fuNOG1 and fuNOG2). The relaxation of the parameters to assign orthology allows also the insertion into the most inclusive clusters (KOGs and COGs) of other genes that could not be assigned to any specific cluster (for example *Hs3* and *Hs4*). COG1 includes KOG1

and six bacterial genes. In addition to the original clusters of orthologs defined by eggNOG, we identified further clusters: fish-specific (fiNOG, depicted in cyan), worm-specific (woNOG, brown) and bacteria-specific (baNOG, light green).
Abbreviations:
Primates: *Hs: H. sapiens; Pt: P. troglodytes; Mm: M. mulatta*.
Rodents: *Mm: M. musculus; Rn: R. norvegicus.*
Other mammals: *Md: M. domestica; Cf: C. familiaris; Bt: B. Taurus.*
Fishes: *Dr: D. rerio; Tf: T. rubripes*.
Other vertebrates: *Xt: X. tropicalis; Gg: G. gallus.*
Insects: Dm*: D. melanogaster; Am: A. millifera; Ag: A. gambiae;*
Nematodes: *Ce: C. elegans*.
Other metazoans: *Pf: P. falciparum; Dd: D. discoideum; Ci: C. intestinalis*.
Fungi: *Sc: S. cerevisiae; Sp: S. pombe; Af: A. fumigatus; Yl: Y. lipolytica; Cn: C. neoformans*.
Plants: *At: A. thaliana; Tp: T. pseudonana; Cm: C. merolae*.
Bacteria: *Ec: E. coli; Bs: B. subtilis; Mp: M. pneumoniae*.

We exploited the different levels of resolution of the orthologous clusters in order to check for the presence of orthologs of each gene in every internal node of the tree of life (Figure 22). For example, for human genes, we investigated the presence of non-primate mammalian orthologs in maNOGs, non-mammalian vertebrate orthologs in veNOGs and so on. For species that do not have a group-specific cluster of orthologs, we extrapolated it from the most specific group to which their genes could be assigned. For example, *D. rerio* does not have e fish-specific cluster of orthologs, therefore we extracted all fish-specific genes from each vertebrate-specific cluster containing *D. rerio* genes, and created the corresponding fish-specific orthologous groups (fiNOGs). We repeated the same procedure for worms and bacteria, starting from metazoan-specific clusters of orthologs and from COGs, respectively (Figure 22).

5. Evolutionary origin, conservation and duplicability

Given the presence/absence of each gene in clusters of orthologs with different resolution and the presence/absence of their orthologs in all internal nodes of the tree of

life, we derived three evolutionary properties for each gene: origin, conservation and duplicability (Figure 23).

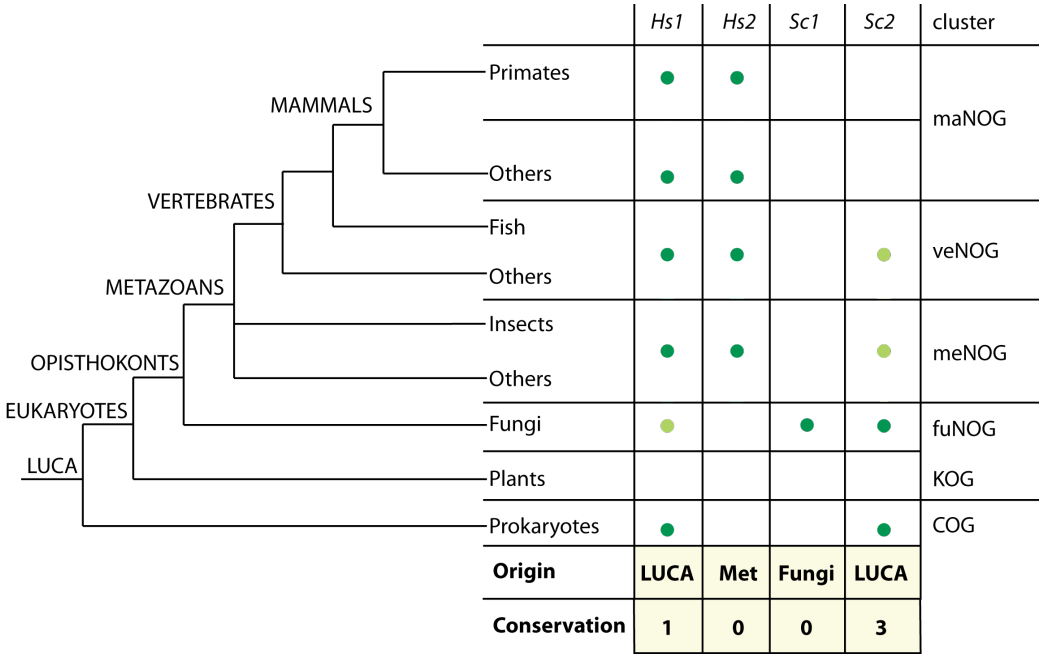


Figure 23: Origin and conservation assignment

Origin and conservation are calculated depending on the presence or absence of orthologs at the each branch of the tree of life. The origin of a gene is calculated as the farthest branch where an ortholog of the gene of interest can be found. *Hs1* and *Sc2* originated with LUCA because they have orthologs in prokaryotes. *Hs2* originated with metazoans because it has orthologs in metazoans but not in other eukaryotes or prokaryotes. *Sc1* originated with fungi because it is not present in any other clusters of orthologs. Conservation reflects the number of lineages where orthologs of the gene of interest cannot be found. *Hs2* and *Sc1* have conservation 0, because no lineages have lost orthologs since its origin. *Hs1* has conservation 1 because it is lost in plants, while *Sc2* has conservation 3 because it is lost in primates, mammals and plants.
Light green shows where, instead of using the specific cluster shown in the last column (dark green), KOG is used to detect orthologs of the gene of interest.

The origin of a gene corresponds to the most ancient internal node of the tree of life where an ortholog is found (Figure 23). We could not assign the origin to a small number of genes (120 in *H. sapiens*, 270 in *D. melanogaster* and 5 in *S. cerevisiae*) because they were included in clusters that did not include representative orthologs (Table 4). For example, there are human genes that are only included in a KOG, but this cluster contains only metazoan-specific genes. We could not assign the origin of these genes to metazoans,

because they are not associated to any metazoan-specific cluster. Therefore, these genes were eliminated from further analyses.

Table 4: genes with evolutionary properties

Genes	*H. sapiens*	*D. melanogaster*	*S. cerevisiae*	*E. coli*
Unique genes in EggNOG 1.0 (Jensen et al., 2008)	18,205	10,543	5,411	4,196
Genes with traceable origin and conservation	18,085	10,273	5,406	4,196
Genes with duplicability information	18,074	10,227	5,400	4,196
Duplicated genes (%)	11,826 (65.4)	6,020 (58.9)	2,260 (41.9)	2,153 (51.3)

From all the unique genes present in eggNOG v. 1.0 (Jensen et al., 2008), all genes included only in clusters without representative orthologs are eliminated, because the origin cannot be assigned. All eukaryotic genes that are not included in KOG are filtered out because duplicability cannot be assigned.

We defined conservation as the number of internal nodes between the origin of a gene and the group-specific cluster where no orthologs could be associated to the gene of interest. The maximum conservation corresponds to presence of orthologs in all nodes, and its value is 0. If orthologs are lost in one lineage, then the conservation is 1, and so on (Figure 23). This method to calculate conservation, together with the fact that we considered the same number of internal nodes (seven) for all the species, makes the measure of conservation independent from the origin of the genes. Our definition of conservation is different from the typical definitions of conservation in evolution, which usually measures the sequence divergence among orthologs. We defined it in this way because we were interested in the rate retention and loss of genes through evolution, rather than simple sequence conservation. Our definition measures the importance of a gene for an organism: the more it is important, the more it is conserved in evolution and cannot be lost in any lineages. The standard definition, instead, is used in order to determine how fast a gene family evolves.

Duplicability was defined upon the presence or absence of more than one gene from the same species in a cluster of orthologs. A gene is duplicated if at least one other gene from the same species is included in its cluster of orthologs, otherwise it is singleton. Since the resolution of the most specific orthologous groups is different for every species, we defined a gene as duplicated if an ortholog was present in the same KOG (Table 4). This definition allowed us to have comparable measures of duplicability between distantly related species. The only exceptions were prokaryotes: since they were not included in KOGs, their duplicability was based on COGs, which could not be used for all the other species, because a small number of eukaryotic genes are included in these clusters (on average, less than 50%). This method did not allow us to date the time of duplication, but rather the rate of duplication. Duplication is a random event and, if it is not deleterious for the fitness of the organism, it can be retained. The use of this definition of duplicability shows whether this duplication was selected in evolution. However, using the same data, in principle it is possible to calculate when the duplication occurred on the basis of the presence or absence of paralogs in other species. For a small fraction of the eukaryotic genes (63 in total), we were not able to assign duplicability, because they were not included in KOG. Therefore these genes were excluded from further analysis.

With the previously described methods, we were able to assign origin, conservation and duplicability to a high number of genes: 18,074 in H. sapiens (82% of all human genes), 10,227 in D. melanogaster (74%), 5,400 in S. cerevisiae (80%) and 4,196 in E. coli (93%) (Table 4).

6. Comparison of gene and network properties

In order to determine how protein interaction networks evolve, we compared the network properties of *H. sapiens*, *D. melanogaster*, *S. cerevisiae* and *E. coli* proteins with the evolutionary properties of the corresponding genes. We first grouped all the genes on the basis of their origin and compared the degree and betweenness distributions of proteins

with the same origin with the corresponding distributions of older and younger proteins. Similarly, we compared the distributions of network properties of proteins with a given conservation with more and less conserved proteins. To analyze duplicability, we proceeded in an analogous way. Among all genes with the same origin, we analyzed the differences in the network properties between singleton and duplicated proteins. To make all these comparisons, we used Wilcoxon test. This non-parametric statistical test is used in presence of continuous non-normal distributions in order to determine whether two independent samples derive from a single homogeneous population.

Since the dimensions of the samples are highly variable (more than two orders of magnitude of difference between the biggest and the smallest dataset), we made a second analysis in order to exclude a possible bias related to this fact. Therefore we developed a randomization test. For each level of evolutionary origin, we extracted 500 random proteins and derived their median degree. We then extracted 500 proteins that originated later in evolution, calculated their median degree and the difference compared to the proteins with the selected origin. We repeated this operation 100,000 times and derived the distributions of the median degree differences. Finally, we computed a *z*-score as the ratio between the number of comparisons with a difference <0 and the total number of iterations. If this value is next to zero, then it is reasonable to conclude that the two distributions are different and, in particular, that younger proteins are less connected. We repeated the same procedure for the comparisons with older proteins. The only difference here is that the numerator of the *z*-score was calculated as the number of differences >0. We repeated the computation of these iterations at all levels of origin and conservation for both degree and betweenness. In case a group included less than 500 genes, we selected the lowest number of genes and computed the 100,000 iterations picking this number of genes. For example, only 84 primate-specific human genes have network information, therefore they were compared with random sets of 84 older genes.

We developed a method to visualize the results of both the analyses with Wilcoxon tests and the randomization tests, which is derived from the visualization of microarray data: we created heatmaps whose colors represent the level of significance of *p*-values and *z*-values. First, we log-transformed these values. Then we assigned different color-scales on the basis of whether the considered category was enriched or depleted. Red boxes were associated with significantly higher values of degree and betweenness, while green boxes were associated with lower values. Black represented not significant values.

7. Functional analysis

We used the Biological Process branch of the Gene Ontology (GO) to perform the faunctional analyses (Ashburner et al., 2000). We exploited the tree-like structure of GO in order to analyze the enrichment in particular functional categories of each class of genes. A GO level refers to the number of branching points that separate a GO term from the root, which is set as level 1 by default. Depending on the type of analysis, three separate trees may be used, which correspond to three different roots: Biological Process (GO: 0008150), Molecular Function (GO:0003674) and Cellular Component (GO:0005575). We focused only on Biological Process, in order to determine the functional differences between the different classes of human hubs.

We developed a *R* pipeline that relies on the *GO.db* and *org.Hs.eg.db* packages in order to compute the functional enrichment analysis between two classes of genes (Figure 24). The input is represented by two lists of Entrez IDs that correspond to the two classes to compare. All the GO terms that are associated to Biological Process are extracted using the *GO.db* function *GOBPCHILDREN* and are subsequently divided into their level. Since each node may have several parents, and therefore it may appear at different levels, it is assigned to its most inclusive level. Then all genes that are associated to each GO term are extracted using the *org.Hs.eg.db* function *org.Hs.egGO2ALLEGS*. The statistical analyses are made level by level. First, all the genes from the two lists that are associated at each

level are counted. Then a Fisher's exact test is computed for all the GO terms that include at least one gene from one or both lists, in order to determine the enrichment of one gene list in that particular term. Given the high number of statistical tests that are computed, a correction for the false-discovery rate must be made. This is done for each level separately using the Benjamini-Hochberg method.

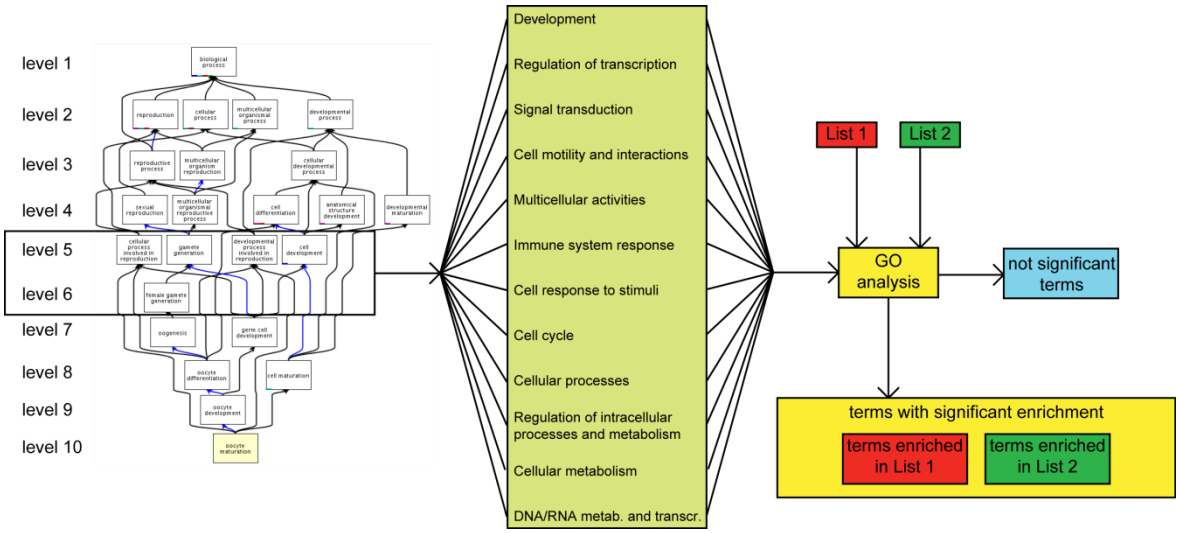


Figure 24: Pipeline to assess enrichment in GO terms

The tree-like structure of GO is exploited to extract all terms at level 5 and 6, which are then divided into 12 different functional categories. Having two lists of genes as input, the enrichment is assessed separately for all the terms with Fisher's exact test followed by adjustment for the false discovery rate (GO analysis). The terms with *p*-value >0.05 are eliminated, while the others are divided in two categories, on the basis of whether list 1 or list 2 is enriched.

Since the levels go from highly inclusive, with thousands of genes that are associated to a single term (*e.g.* the term GO:0009987, "cellular process", has 210,383 associated gene products from several species), to highly specific, with only few genes associated to the terms at a specific level (*e.g.* GO:0007092, "activation of mitotic anaphase-promoting complex activity", has 14 associated gene products), we decided to focus only on GO terms at level 5 and 6 (Figure 24). This is a good compromise between analyzing specific terms and handling a fair number of genes. We further grouped the 1,213 terms at these levels into twelve functional categories:

- Cell cycle,
- Cell motility and interactions,
- Cell response to stimuli,
- Cellular metabolism,
- Cellular processes,
- Development,
- DNA/RNA metabolism and transcription,
- Immune system response,
- Multicellular activities,
- Regulation of intracellular processes and metabolism,
- Regulation of transcription,
- Signal transduction.

We then performed four comparisons:

1. Ancient singleton hubs and recent duplicated hubs;
2. Genes that originated with the last universal common ancestor and eukaryotes (ancient) and genes that originated with metazoans and vertebrates (recent);
3. Singletons and duplicated genes
4. Dominant and recessive cancer genes.

All the terms were grouped by their functional category and the involvement of a class of genes in a particular category was highlighted not only by the *p*-value, but also by the number of significantly enriched terms, compared with the other class (Figure 24).

8. Identification of ohnologs

Ohnologs are genes that duplicated via whole genome duplication (Wolfe, 2000). This term was first coined by Ken Wolfe in 2000, in honor of the late Susumu Ohno who,

thirty years before with only a handful of known protein sequences, first proposed that the easiest way to produce new genes is through duplication, rather than creating them *de novo*, and that whole genome duplications may allow the duplication of entire pathways. He also proposed that two or three rounds of whole genome duplications occurred in the early evolution of the vertebrate genome (Ohno, 1970; Wolfe, 2001). The final evidence that the early vertebrate genome underwent two rounds of whole duplications was given first by Dehal and Boore in 2005 (Dehal and Boore, 2005) and by Nakatani *et al.* (Nakatani et al., 2007), who were able to reconstruct the ancestral vertebrate chromosomes. From this experiment, we derived the definition of the 4,174 ohnologs in the human genome (Nakatani et al., 2007). We intersected these genes with the 22,020 unique human genes in our dataset and 307 ohnologs were discarded because they could not be mapped to any of these genes. We then eliminated all singleton genes from this list and retain 3,618, which were both ohnologs and with evidence of duplication from our pipeline. The 249 singletons were likely false positives and therefore they were discarded from further analyses. 62 other genes were eliminated because they were in the ohnologs dataset but they originated in mammals or primates. Therefore, these were also likely false positives. Finally we were able to detect 3,556 ohnologs.

9. miRNA targets

miRNAs are post-transcriptional regulators of gene expression (Bartel, 2004). They are short sequences of RNA that promote the degradation of mature RNAs through an imperfect pairing with the 3' UTR. miRNAs emerged in evolution with higher eukaryotes, indeed they are present in the genome of plants and metazoans, but not in yeast (Wheeler et al., 2009). They are continuously acquired in evolution, but they are rarely lost (Sempere et al., 2006).

With the help of Vera Pendino, we derived the interactions between miRNAs and their targets from two sources: Tarbase v.5 (June 2008), which includes 1,051 interactions

between 101 miRNAs and their 808 target genes (Papadopoulos et al., 2009), and miRecords v.1 (August 15th 2008), which include 1,311 interactions that involve 112 miRNAs and 613 human genes (Xiao et al., 2009). We chose these two databases because they collect only experimentally validated interactions. Each miRNA-target gene interaction may be validated by three types of experiment: microarrays, mass-spectrometry or single-gene. In total 986 genes are target of miRNAs.

10. Tissue-selectivity of human genes

In order to determine the value of gene expression in different human tissues, we used the data from two experiments that analyzed microarray data in 36 (Ge et al., 2005) and 79 (Su et al., 2004) human tissues. From the latter study, we eliminated the data about six cancer tissues, in order to avoid the influence of the disease conditions on our analysis.

We downloaded the expression data for these two experiments from the Gene Expression Omnibus (GEO, http://www.ncbi.nlm.nih.gov/geo/) (Barrett et al., 2011). They both are based on the Affymetrix Human Genome U133A Array (GPL96), which include 22,283 distinct probes that target 13,789 distinct genes. The Su experiment (Su et al., 2004) used also the GNF1H platform (GPL1074), which was discarded from the analysis in order to have only compatible gene sets for the tissue-selectivity analysis. We were able to assign the origin to 10,060 genes (73% of all human genes) (Table 5).

We defined a gene as tissue-specific if it is expressed in less than 25% of the analyzed tissues in at least one of the two studies. Therefore a gene is tissue-specific if it is expressed in less than 8 or 17 tissues, respectively. Housekeeping genes are expressed in more than 97% of the tissues (*i.e.* 35 and 71, respectively) (Table 5).

Table 5: tissue selectivity of genes

Experiment	N tissues	N samples	Type	N probes	N genes	genes in our dataset
Su et al. 2004	73	146	Total	22,283	13,787	10,060
			tissue-selective	6,308	4,241	3,622
			housekeeping	4,493	3,750	3,319
Ge et al. 2005	36	36	Total	22,283	13,787	10,060
			tissue-selective	2,876	2,514	2,192
			housekeeping	1,759	1,532	1,321
Total	109	182	Total	22,283	13,787	10,060
			tissue-selective	NA	4,988	4,616
			housekeeping	NA	3,765	3,500

Tissue selectivity of genes is derived from two experiments, which determined the expression levels of 13,787 genes in 182 samples from 109 non-cancer tissues. The last column represents the intersection with the 18,074 that have origin and duplicability information. Tissue-selective genes are expressed in less than 25% of tissues, while housekeeping genes are expressed in more than 97% of tissues.

The definition of whether a gene is expressed in a tissue was different for the two experiments. The Ge experiment (Ge et al., 2005) defined a gene as expressed in a particular tissue if it had a significant detection *p*-value (<0.05). The definition of tissue-specificity given by the authors differed from ours, since they considered as tissue-specific all gene expressed only in one tissue. In particular, a gene must have a detection *p*-value lower than 0.02 and the second highest expression score must be less than half of the highest. The Su experiment (Su et al., 2004) defined a gene as expressed if its expression level is higher than 200.

The results from the two experiments were similar, in terms of consensus tissue-specific and housekeeping genes: 1,717 genes were labeled as tissue-specific and 1,517 as housekeeping in both experiments. However, we considered the union between the two experiments and we were able to assign the origin to 4,616 tissue-specific and 3,500 housekeeping genes (Table 5).

11. Analysis of the mechanisms that control duplicated hubs

The retention of duplication after whole genome duplication, the regulation by miRNAs and the tissue selectivity of a gene are three mechanisms that made it possible to retain the duplication of hubs. We intersected singleton and duplicated hubs with the lists of genes that are involved in these three mechanisms, and repeated this at all levels of gene origin. We then compared, using Fisher's exact test, the fraction of singleton and duplicated hubs that are associated to at least one of these mechanisms, in order to determine whether duplicated genes that originated with metazoans and vertebrates are enriched in ohnologs, miRNA targets and tissue-selective genes. We made the same analysis also for housekeeping genes.

12. Identification of recent paralogs of cancer genes

We used the results of the alignment of RefSeq proteins to the human genome (Figure 20) not only to determine a non-redundant set of unique human genes, but also to detect duplications. The BLAT algorithm is preferred for this kind of analysis because it is fast and it is able to align short sequences with high levels of identity. In order to determine what the unique genes in the human genome are, we focused only on the best hit for each sequence. We instead used all the other hits in order to determine duplicability. Previous works in our lab (Rambaldi et al., 2008; Syed et al., 2010) defined as duplicated all genes that have additional hits on the genome above 60% of their length. Here, instead, we focused on all additional hits, in order to analyze in detail the differences between cancer genes and the rest of human genes. We set different thresholds of duplicability (from 0 to 100%): at 0% we defined as singletons all those genes that have no additional hits on the genome, at 100% we identified genes that had a perfect copy of their coding sequence on the genome. For each level of duplicability, we compared the fraction of

duplicated genes between cancer genes (i.e. genes included in the Cancer Gene Census) and the rest of human genes, using chi-squared test. We repeated the same analysis for dominant and recessive cancer genes.

In order to determine whether the additional hits for each gene overlapped already known genes or intergenic regions, we controlled whether each hit intersected exons from one of the other 22,019 genes or from mRNA sequences extracted from the UCSC Genome Browser (*all_mrna* table from "*mRNA and EST tracks*") (Kent et al., 2002). If the results were negative, the hit was labeled as "genomic". We repeated the analyses for both types of hits separately.

The normalization by the gene length was performed by counting the number of duplicated bases, instead of the number of duplicated genes. At each level of conservation the normalized duplicability was calculated in the following way:

$$\frac{\frac{x}{100}}{\frac{N}{1}} = \frac{\frac{x}{100}}{\frac{N}{1}}$$

Where $\frac{x}{100}$ is the length of a gene, $\frac{x}{100}$ is the number of duplicated genes, N is the total number of genes and x is the conservation level (between 0 and 100%).

13. Tools

We used scripts written in Perl version 5.8.8 (http://www.perl.org/) to integrate all data from the different sources and to prepare them for all the statistical analyses.

All statistical analyses were performed using R version 2.10.1 (http://www.r-project.org/). For the network analyses, we used the R package *igraph* v. 0.5.2 (http://igraph.sourceforge.net/), while *gplots* version 2.8.0 (http://cran.r-project.org/web/packages/gplots/index.html) was used to create all the heatmaps, using the command *heatmap.2*. To make the functional analyses we used two additional packages: *GO.db* version 2.5.0 (http://www.bioconductor.org/packages/2.8/data/annotation/html/GO.db.html) and *org.Hs.eg.db* version 2.5.0

(http://www.bioconductor.org/packages/2.8/data/annotation/html/org.Hs.eg.db.html). These packages are included in R Bioconductor version 2.8 (http://www.bioconductor.org/).

In order to store all the data that we produced and to easily access and query it, we built a database, formed by 20 tables. In order to access to it, we created a website, called the Network of Cancer Genes (NCG), which is publicly available at http://bio.ifom-ieo-campus.it/ncg (Appendix 1). The website is cancer-gene centered and may be used to analyze genomic and network properties of cancer genes (see Results). The database may be queried using MySQL version 14.12. inside the Network of Cancer Genes, we built a network visualization, in order to visually inspect the interactions of each cancer gene, using Cytoscape Web version 0.7.3 (Lopes et al., 2010).

Results

1. Gene and network properties undergo modifications during evolution

1.1. Origin distribution, conservation and duplicability change in evolution

For the analysis of gene properties, we considered seven species (*M. musculus*, *G. gallus*, *D. rerio*, *A. millifera*, *C. elegans*, *S. pombe* and *B. subtilis*), in addition to the four species that have comprehensive network information (*H. sapiens*, *D. melanogaster*, *S. cerevisiae* and *E. coli*), in order to have representatives of all nodes of the tree of life, and to have a limited evolutionary distance between species (Figure 25).

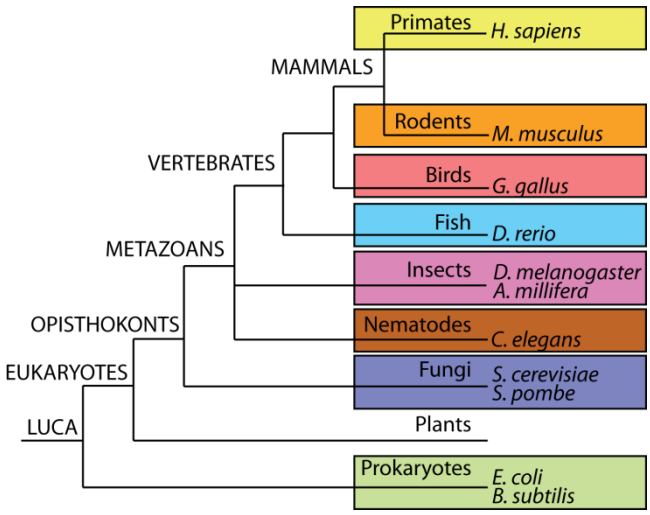


Figure 25: Representative species used in the analysis

The species that are used for the analysis of origin, conservation and duplicability are highlighted on the tree of life.

Overall, we noticed a high variability for these gene properties across the different species. Around 60% of mammalian genes originated early in evolution (*i.e.* they share orthologs with prokaryotes or early eukaryotes) (Figure 26, Table 6), as it was previously

reported for human (Domazet-Loso and Tautz, 2008). The fraction of ancient genes is tightly correlated with the complexity of the organisms, since unicellular eukaryotes have more than 90% of their genes that originated early in evolution, while three quarters of insect genes are ancient. The high fraction of vertebrate-specific genes (between 14.6% and 20.8%) is likely related with the two rounds of whole genome duplications that occurred in the early vertebrate genome (Dehal and Boore, 2005; Nakatani et al., 2007). The absence of group-specific genes for *M. musculus* and *G. gallus* is due to the low number of species in eggNOG that are associated with the corresponding group-specific nodes. Hence, since no orthologs could be associated with these genes, they were not included in any cluster of orthologs.

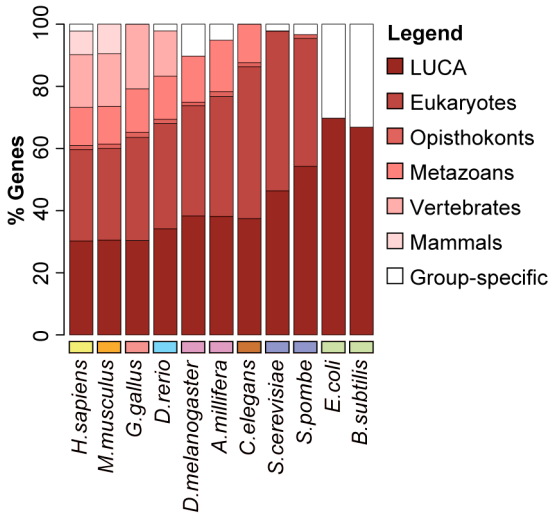


Figure 26: gene origin in evolution

The percentage of genes that originated in every internal node of the tree of life is shown for eleven species: a primate (*H. sapiens*), a rodent (*M. musculus*), a bird (*G. gallus*), a fish (*D. rerio*), two insects (*D. melanogaster* and *A. mellifera*), a nematode (*C. elegans*), two fungi (*S. cerevisiae* and *S. pombe*), and two bacteria (*E. coli* and *B. subtilis*). The color code refers to the tree of life in Figure 25.

Table 6: gene origin in evolution

Vertebrates	*H. sapiens*		*M.musculus*		*G. gallus*		*D. rerio*	
	Genes	**%**	**Genes**	**%**	**Genes**	**%**	**Genes**	**%**
LUCA	5,466	30.2	5,473	30.5	4,237	30.3	3,160	34.1
Eukaryotes	5,321	29.4	5,282	29.4	4,635	33.2	3,134	33.9
Opisthokonts	234	1.3	260	1.4	226	1.6	128	1.4
Metazoans	2,217	12.3	2,180	12.2	1,955	14.0	1,283	13.9
Vertebrates	3,062	16.9	3,042	17.0	2,909	20.8	1,350	14.6
Mammals	1,377	7.6	1,699	9.5	NA	NA	NA	NA
Group-specific	397	2.2	0	0.0	0	0.0	199	2.2
TOTAL	**18,074**	**100**	**17,936**	**100**	**13,962**	**100**	**9,254**	**100**

Invertebrates	*D. melanogaster*		*A. mellifera*		*C. elegans*	
	Genes	**%**	**Genes**	**%**	**Genes**	**%**
LUCA	3,918	38.3	3,003	38.1	3,593	37.4
Eukaryotes	3,626	35.5	3,035	38.5	4,691	48.9
Opisthokonts	112	1.1	127	1.6	124	1.3
Metazoans	1,523	14.9	1,302	16.5	1,185	12.3
Group-specific	1,048	10.2	407	5.2	5	0.1
TOTAL	**10,227**	**100**	**7,874**	**100**	**9,598**	**100**

Unicellular species	*S. cerevisiae*		*S. pombe*		*E. coli*		*B. subtilis*	
	Genes	**%**	**Genes**	**%**	**Genes**	**%**	**Genes**	**%**
LUCA	2,505	46.4	2,321	54.2	2,926	69.7	2,462	66.9
Eukaryotes	2,773	51.4	1,762	41.2	NA	NA	NA	NA
Opisthokonts	7	0.1	55	1.3	NA	NA	NA	NA
Group-specific	115	2.1	142	3.3	1,270	30.3	1,219	33.1
TOTAL	**5,400**	**100**	**4,280**	**100**	**4,196**	**100**	**3,681**	**100**

Origin is defined as the most ancient internal node of the tree of life where an ortholog of the gene of interest is found. For all the 11 species, the number of genes and the percentage of all the genes with an assigned origin is shown. "NA" represents lineages where the origin could no be calculated, such as mammalian-specific genes for *G. gallus* or *D. rerio*. The total number of genes refers to all the genes that have both origin and duplicability information.

The analysis of conservation also showed substantial differences between vertebrates and the other species (Figure 27, Table 7). Vertebrates have more than 90% of the genes that are highly conserved (conservation between 0 and 2), while invertebrates, *S. cerevisiae* and bacteria have only between 63 and 75%. These findings are related with the origin of the genes: vertebrates have a high fraction of young genes, therefore a smaller number of internal nodes may show a loss of orthologs, compared with ancient genes. An additional explanation of the differences between vertebrates and invertebrates may be the

fact that invertebrates retain a high fraction of ancient genes, which are lost in other lineages.

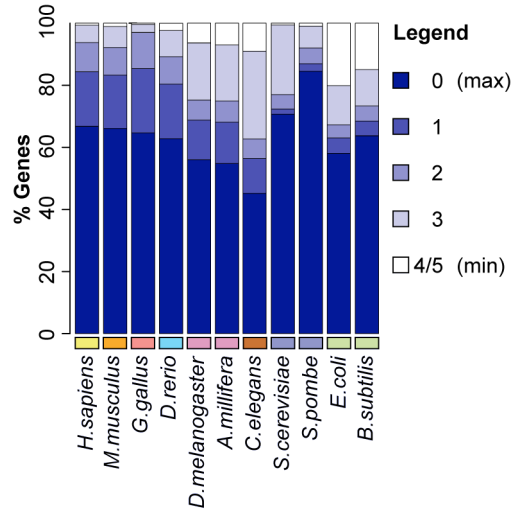


Figure 27: gene conservation in evolution

The percentage of genes that have the same level of conservation is shown for the eleven species. Zero represents the maximum level of conservation, *i.e.* at least one ortholog of the gene of interest is found in all nodes of the tree of life. Higher values represent the number of internal nodes of the tree of life where no orthologs could be associated with the gene of interest. The color code refers to the tree of life in Figure 25.

Table 7: gene conservation in evolution

Vertebrates	***H. sapiens***		***M.musculus***		***G. gallus***		***D. rerio***	
	Genes	**%**	**Genes**	**%**	**Genes**	**%**	**Genes**	**%**
0	12,071	66.8	11,850	66.1	9,027	64.7	5,808	62.8
1	3,178	17.6	3,084	17.2	2,896	20.7	1,630	17.6
2	1,697	9.4	1,593	8.9	1,620	11.6	816	8.8
3	1,012	5.6	1,222	6.8	364	2.6	783	8.5
4 and 5	116	0.6	187	1.0	55	0.4	217	2.3
TOTAL	**18,074**	**100**	**17,936**	**100**	**13,962**	**100**	**9,254**	**100**

Invertebrates	***D. melanogaster***		***A. mellifera***		***C. elegans***	
	Genes	**%**	**Genes**	**%**	**Genes**	**%**
0	5,725	56.0	4,315	54.8	4,335	45.2
1	1,311	12.8	1,048	13.3	1,079	11.2
2	661	6.5	536	6.8	605	6.3
3	1,879	18.4	1,422	18.1	2,712	28.3
4 and 5	651	6.4	553	7.0	867	9.0
TOTAL	**10,227**	**100**	**7,874**	**100**	**9,598**	**100**

Unicellular species	*S. cerevisiae*		*S. pombe*		*E. coli*		*B. subtilis*	
	Genes	**%**	**Genes**	**%**	**Genes**	**%**	**Genes**	**%**
0	3,816	70.7	3,616	84.5	2,436	58.1	2,346	63.7
1	90	1.7	104	2.4	209	5.0	175	4.8
2	253	4.7	217	5.1	177	4.2	179	4.9
3	1,209	22.4	302	7.1	530	12.6	431	11.7
4 and 5	32	0.6	41	1.0	844	20.1	550	14.9
TOTAL	**5,400**	**100**	**4,280**	**100**	**4,196**	**100**	**3,681**	**100**

Conservation is defined as the number of branches of the tree of life where no orthologs of the gene of interest are found, since its origin. Zero represents the most conserved genes, since its orthologs are present in all branches of the tree of life, while 4 and 5 represent the least conserved genes, which are lost in a high number of nodes. The total number of genes is as in Table 6.

Duplicability adds a new level of complexity to this scenario. Among eukaryotes, duplicability is tightly related to organism complexity (Figure 28, Table 8) and fungi are less duplicated than metazoans, as expected (Yang et al., 2003). The higher fraction of duplicated genes in bacteria compared to fungi is due to the higher level of inclusiveness of the clusters of orthologs used to define duplicated genes among prokaryotes (COGs intead of KOGs). Among metazoans, two thirds of the genes are duplicated, with two exceptions: *D. rerio* and insects. *D. rerio* has 77% of duplicated genes: this fact is explained by the recent duplication that occurred in the fish ancestral genome (Christoffels et al., 2004; Jaillon et al., 2004; Meyer and Van de Peer, 2005; Taylor et al., 2001). The other exception is represented by insects, which are less duplicated than other metazoans. This is a proof of the compactness of their genome (Petrov and Hartl, 1998), as was previously demonstrated in two independent ways: insects have a high rate of DNA loss (Petrov, 2002) and a low rate of fixed transposable elements (Gonzalez et al., 2008).

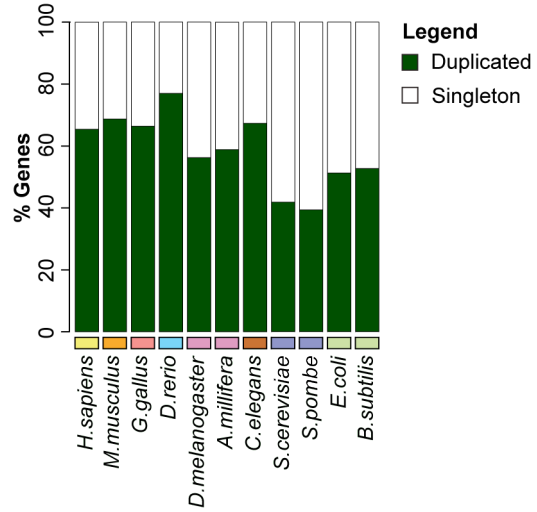


Figure 28: gene duplicability in evolution

The percentage of duplicated genes is shown for the eleven species. For eukaryotes, a gene is duplicated if it has a paralog in its eukaryotic-specific cluster of orthologs (KOG), for bacteria duplicability is based on the presence of orthologs in the corresponding COG. The color code refers to the tree of life in Figure 25.

Table 8: gene duplicability in evolution

Vertebrates	*H. sapiens*		*M.musculus*		*G. gallus*		*D. rerio*	
	Genes	**%**	**Genes**	**%**	**Genes**	**%**	**Genes**	**%**
Duplicated genes	11,826	65.4	12,331	68.7	9,338	66.4	7,130	77.0
Singletons	6,248	34.6	5,609	31.3	4,728	33.6	2,130	23.0
TOTAL	**18,074**	**100**	**17,940**	**100**	**14,066**	**100**	**9,260**	**100**

Invertebrates	*D. melanogaster*		*A. mellifera*		*C. elegans*	
	Genes	**%**	**Genes**	**%**	**Genes**	**%**
Duplicated genes	6,020	58.9	4,461	56.3	6,463	67.3
Singletons	4,207	41.1	3,466	43.7	3,135	32.7
TOTAL	**10,227**	**100**	**7,927**	**100**	**9,598**	**100**

Unicellular species	*S. cerevisiae*		*S. pombe*		*E. coli*		*B. subtilis*	
	Genes	**%**	**Genes**	**%**	**Genes**	**%**	**Genes**	**%**
Duplicated genes	2,260	41.9	1,690	39.4	2,153	51.3	1,942	52.8
Singletons	3,140	58.1	2,602	60.6	2,043	48.7	1,739	47.2
TOTAL	**5,400**	**100**	**4,292**	**100**	**4,196**	**100**	**3,681**	**100**

A gene is duplicated if another gene of the same species is present in its KOG (for eukaryotes) or COG (for prokaryotes). The total number of genes is as in Table 6.

1.2. Networks have different levels of completeness

We chose four species that are representative of different levels of complexity and have a defined protein interaction network: a prokaryote (*E. coli*), a unicellular eukaryote (*S. cerevisiae*) and two multicellular eukaryotes (*D. melanogaster* and *H. sapiens*). Publicly available databases of protein-protein interactions include data from several other species. However, only four of these include more than 1,000 interactions: *M. pneumoniae*, A. *thaliana*, *C. elegans* and *M. musculus*. These networks were discarded from our analyses because they are highly incomplete. The network of *C. elegans* has only 4,040 proteins (20% of the total proteins), *M. musculus* has 2,587 proteins (12%) and *A. thaliana* has 2,586 proteins (10%). The network of *M. pneumoniae* was discarded because it includes 1,054 interactions among 409 proteins (60%) that were derived from only one single high-throughput experiment (Kuhner et al., 2009).

We rebuilt the protein interaction networks for *E. coli*, *S. cerevisiae*, *D. melanogaster* and *H. sapiens*. Since several resources that store protein-protein interactions data were publicly available and the overlap between them is rather low (Table 3), their integration resulted in the most complete network that is currently possible (Table 9), based only on primary data (*i.e.* no orthology-inferred interactions). The advantage of using the integration of all resources was noteworthy for all the four chosen species. For *E. coli* we gathered network data from two sources, which have 13,663 and 5,646 interactions, while the total protein interaction network includes 15,888 interactions. The improvement was even more relevant for the other species. The biggest network for *S.cerevisiae* includes 45,580 interactions, while the total network that we were able to reconstruct has 91,652 interactions. Also for *H. sapiens* the biggest network that was available covers only half of the total interactions that we could define, while the *D. melanogaster* biggest network has only 22,872 interactions, while we reconstructed 61,014.

Table 9: properties of the protein interaction networks

Network properties			*H. sapiens*	*D. melanogaster*	*S. cerevisiae*	*E. coli*
Total Network	Nodes (% total proteins)		11,988 (54%)	10,563 (77%)	5,937 (88%)	2,884 (64%)
	Interactions		68,498	61,014	91,541	15,888
	High-Throughput (%)		29,023 (42%)	58,921 (97%)	77,615 (85%)	15,078 (95%)
	Single-Gene Experiments (%)		39,475 (58%)	2,093 (3%)	13,926 (15%)	810 (5%)
	Degree	Median	5	5	15	5
		Mean	11.4	11.5	30.9	11.0
	Betweeness	Median	898	1.011	930	287
		Mean	16,885	16,888	6,014	3,222
Gold Set	Nodes (% total nodes)		9,127 (41%)	1,392 (10%)	3,921 (58%)	703 (16%)
	Interactions (% total interactions)		39,868 (58%)	2,236 (4%)	21,721 (24%)	1,004 (6%)
	Degree	Median	4	2	5:05	2
		Mean	8.7	3.2	11.1	2.8
	Betweeness	Median	682	0	932	0
		Mean	14,208	2,633	6,107	618

Network properties are described for the four species. High-throughput studies include at least 100 interactions. The gold set includes all interactions that are supported by single experiments or more than one high-throughput experiment.

The available data did not allow the reconstruction of a fully connected network for each species. However, a very small number of nodes could not be linked to the biggest subnetwork (315 in total, 137 in *H. sapiens*, 151 in *D. melanogaster*, 3 in *S. cerevisiae* and 24 in *E. coli*). This corresponds to 1% of all the nodes, with a minimum of 0.05% in *S. cerevisiae* and a maximum of 1.43% in *D. melanogaster*. In the eliminated species, instead, this fraction is significantly higher. For example the *M. musculus* protein interaction network has 2,586 proteins but 313 form 121 additional small networks (2.6 proteins/network), which account for 12.1% of the entire mouse protein interaction network.

The four networks are different in terms of number of nodes and interactions, coverage of all genes and type of interactions (Table 9). Nevertheless, they are all scale-free and their power-laws all have similar γ values, between 2.09 and 2.21 (Figure 29,

Figure 30, Figure 31, Figure 32). The most complete network in terms of nodes is *S. cerevisiae*, with 88% of the proteins that have at least one interaction, while the least complete is human, with only 54% of the proteins (Table 9). *S. cerevisiae* has also the most complete network in terms of interactions, since the average degree is three times higher than that of the other species. *D. melanogaster* and *E. coli* have more than 95% of the interactions derived from high-throughput experiments (Table 3). This might imply the presence of a substantial fraction of false-positives, although the matter is still controversial (Bader et al., 2004; von Mering et al., 2002; Yu et al., 2008a). In *H. sapiens* instead most of the interactions are derived from single-gene experiments.

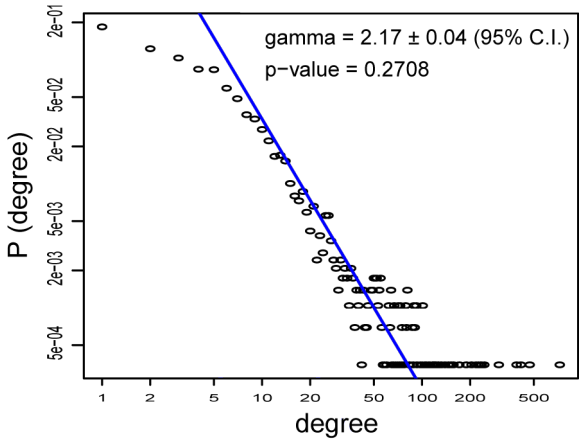


Figure 29: degree distribution of the *E. coli* network

The degree is the count of the interactions that each node has, while P represents the probability of a node to have a certain value of degree. The blue line represents the interpolation of the nodes with degree > 10. The power law is 2.17, which corresponds to a scale-free network. The *p*-value from Kolmogorov-Smirnov test is calculated to determine whether the line fits the data. Since it is not significant, the null hypothesis cannot be rejected and the calculated power-law describes adequately the degree distribution of the network.

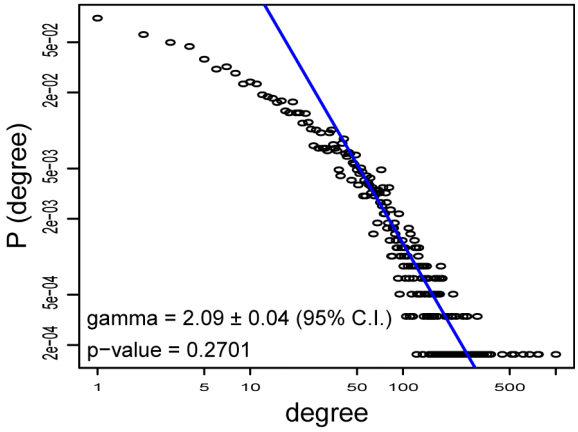


Figure 30: degree distribution of the *S. cerevisiae* network

The blue line is interpolated as in the previous figure. The power-law is 2.09, which corresponds to a scale-free network.

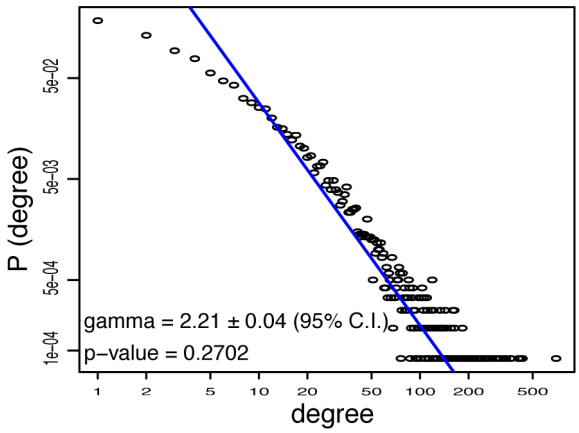


Figure 31: degree distribution of the *D. melanogaster* network

The blue line is interpolated as in the previous figures. Network is scale-free, because the power-law is 2.21.

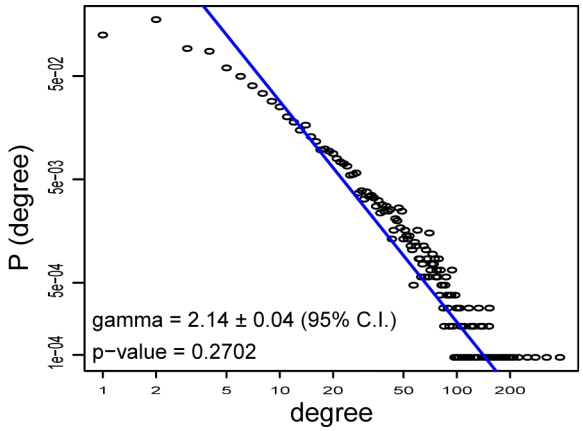


Figure 32: degree distribution of the *H. sapiens* network

The blue line is interpolated as in the previous figure. The power-law is 2.14, which corresponds to a scale-free network.

Given the different incidence of false positives in the four networks due to different fractions of interactions derived from high-throughput experiments, we defined a "gold set" of interactions, supported by single-gene experiments or by more than one high-throughput screening. This latter condition was chosen in order to minimize the effects of false positives derived from high-throughput experiments. Since almost all interactions of

D. melanogaster and *E. coli* were derived from large-scale screenings, the gold set for these two species is too small and cannot be used for any statistical analysis, because it includes less than 20% of all proteins. Instead the gold set networks of *H. sapiens* and *S. cerevisiae* are rather similar: they include 42% and 58% of total proteins, with an average degree of 9 and 11, respectively (Table 3, Table 9).

1.3. Human duplicated genes encode highly connected and central proteins

We measured connectivity as the degree of a protein, *i.e.* the number of interactions that it makes inside the network. To measure centrality, we used betweenness, which counts the number of shortest paths that pass through a protein. In this case, the higher the number of shortest paths, the more central is the protein.

It was previously shown that gene duplicability affects the protein network properties of human and yeast genes in different ways (Hughes and Friedman, 2005; Liang and Li, 2007; Prachumwat and Li, 2006; Rambaldi et al., 2008). In particular, it was found that yeast singleton genes encode highly connected proteins, while duplicated proteins are less connected than singletons (Hughes and Friedman, 2005; Prachumwat and Li, 2006). In human, instead, duplicated genes encode more connected proteins than singletons (Liang and Li, 2007; Rambaldi et al., 2008). These results were found with smaller networks than the ones we reconstructed and were performed only on two distant species. We therefore repeated this analysis for all the four networks and we found the same relationships in the *H. sapiens* and *S. cerevisiae* network, both considering the whole network and the gold set (Figure 33, Table 10). The trend of centrality is very similar to connectivity: in *H. sapiens*, duplicated proteins are more central than singletons, while in *S. cerevisiae* singletons are more central than duplicated proteins. Furthermore, we also found that also *E. coli* and *D. melanogaster* have properties similar to *S. cerevisiae*. The only exception is represented by

the centrality of yeast proteins, which does not present any differences between singleton and duplicated proteins.

These findings seem to support the hypothesis that the modification of the relationships between network properties and duplicability must have occurred in the ancestral vertebrate genome or later in evolution.

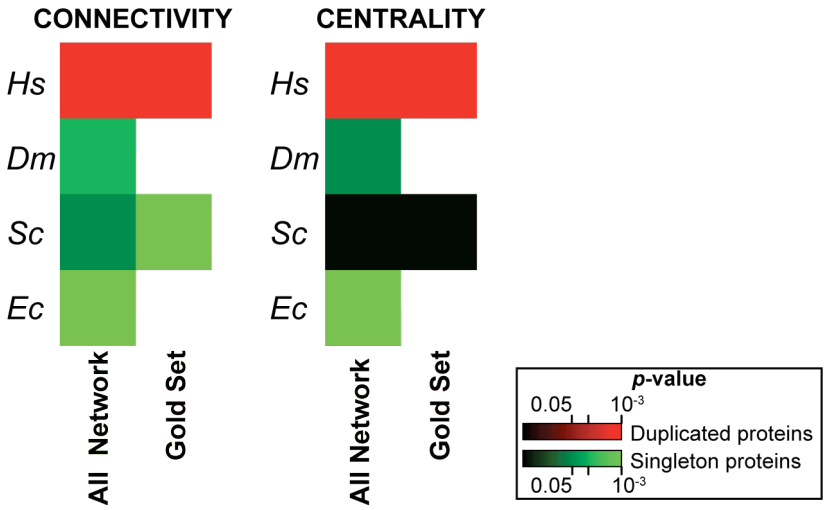


Figure 33: network properties of singleton and duplicated genes

Wilcoxon test is used to compare degree and betweenness of singleton and duplicated genes in all four species. The resulting *p*-values are transformed into heatmaps in order to visually detect differences between duplicated and singleton genes. Red indicates that duplicated genes encode significantly more connected or more central proteins than singletons, green indicates that singleton proteins are significantly more connected or central, black indicates no enrichment, while analyses that are not performed (such as connectivity and centrality in the gold set for *D. melanogaster* and *E. coli*) are depicted in white.

Table 10: network properties of singleton and duplicated genes

Network	Species	N genes	Duplicated Genes					Singleton Genes					*p*-value degree	*p*-value betweenness
			N genes	Degree		Betweenness		N genes	Degree		Betweenness			
				Mean	Median	Mean	Median		Mean	Median	Mean	Median		
Total network	*H. sapiens*	10,373	6,963	12.5	5.0	19,620	1,098	3,410	10.2	5.0	13,299	817	**3.50E-04**	**4.08E-04**
	D. melanogaster	6,298	3,429	12.9	5.0	19,762	1,213	2,869	13.2	6.0	20,922	1734	**9.20E-03**	**3.33E-02**
	S. cerevisiae	5,232	2,162	35.3	17.0	7,841	1,121	3,070	33.2	19.0	5,892	1318	**3.44E-02**	1.51E-01
	E. coli	2,839	1,448	8.4	4.0	2,141	198	1,391	13.9	5.0	4,427	423	**1.52E-09**	**2.63E-05**
Gold set	*H. sapiens*	8,051	5,573	9.4	4.0	16,270	787	2,478	7.1	4.0	9,605	405	**3.06E-07**	**1.48E-06**
	S. cerevisiae	3,758	1,449	9.5	5.0	6,748	846	2,309	12.6	7.0	6,079	1183	**2.62E-14**	4.50E-01

The degree and betweenness distributions of duplicated proteins are compared with singletons using Wilcoxon test. Significant *p*-values (<0.05) are colored in red if duplicated proteins are more connected or more central, in green if singletons are more connected or more central.

2. An ancient network core is conserved in all species

The preferential attachment theory of network evolution (Barabasi and Albert, 1999) describes the laws that regulate the expansion of a network in time. In particular, new nodes attach preferentially to already highly connected nodes. In order to verify whether this theory of network evolution is applicable to protein interaction networks and to understand how protein interaction network evolve, we analyzed connectivity and centrality of each protein in respect to the origin of the corresponding gene.

2.1. Old proteins are highly connected and central

In all species, we discovered that older proteins are more connected and more central than younger proteins (Figure 34, Figure 35, Table 11). The signal is evident in all four species and in both the gold sets, with some exceptions, which may be due to the incompleteness of the protein interaction network data. *D. melanogaster* proteins that originated with the last universal common ancestor are less connected and central than those that appeared with eukaryotes. This tendency is explained by the fact that the network is far from being complete and that more than 95% of the *D. melanogaster* interactions are supported by only one high-throughput experiment. Therefore, there may be a significant fraction of false positives that confound the signal. A similar trend is present also in the gold set of *S. cerevisiae*, which includes less than 60% of all proteins, hence the incompleteness of the network may explain this peculiar trend. One last exception is represented by human proteins that originated with metazoans, which, in the gold set, are more connected than older proteins. Although this may be an artifact due to the incompleteness of the human network (only 42% of the total proteins), with the analysis of duplicability it will become evident how this is a real biological signal.

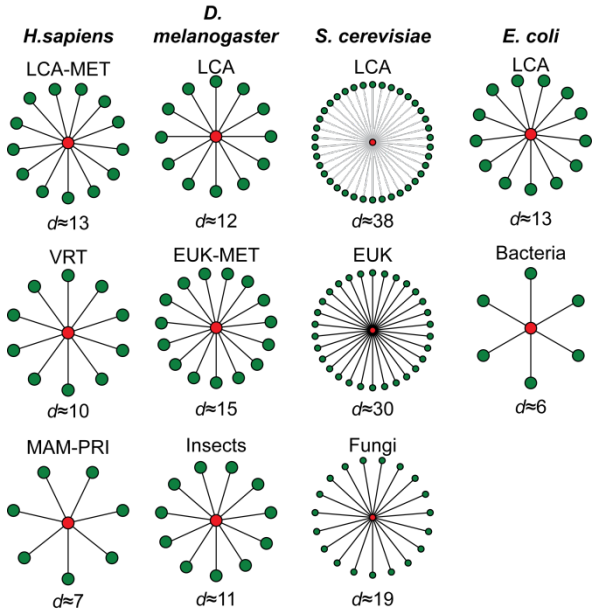


Figure 34: degree of proteins with different origins

The visual analysis of proteins that are representative of a certain level of origin is shown. For all species, old genes encode highly connected and central proteins. The only exception is *D. melanogaster*, for which proteins that originated with the last universal common ancestor (LCA) have lower degree than those that originated with eukaryotes and metazoans. The degree always refers to the red central node.

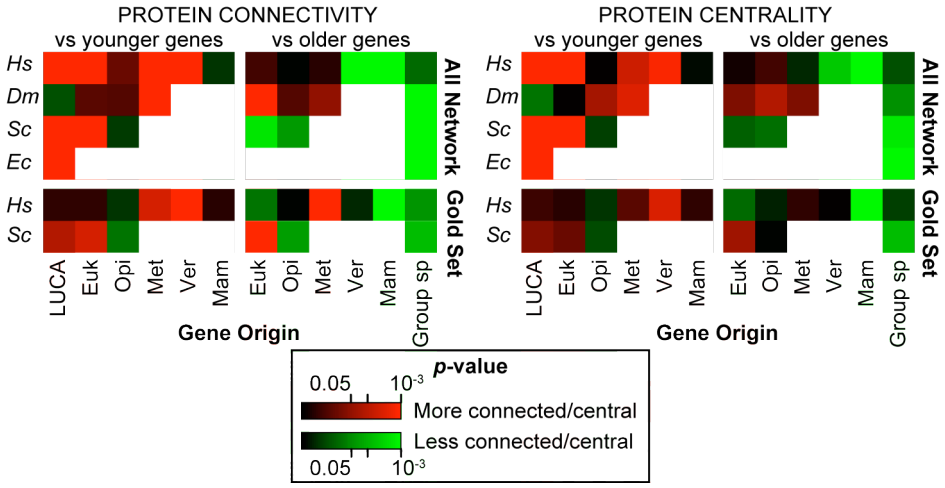


Figure 35: relationships between origin and network properties

Degree (left) and betweenness (right) are compared between proteins that originated at a certain node and older or younger proteins. The analysis is performed using Wilcoxon test and the resulting *p*-values are transformed into heatmaps. Red represents the fact that proteins that originated at a certain node have significantly higher degree or betweenness, while green represents lower degree or betweenness. Black

indicates no significant differences, while white represents analysis that cannot be performed, such as the analysis of *D. melanogaster* proteins that originated with vertebrates or mammals.

The gene origin does not influence the clustering coefficient of the corresponding protein (Table 11). This is somewhat expected, given the network topology. Indeed, in a scale-free network that has a degree distribution that follows a power-law with 2<gamma<3, the clustering coefficient does not depend on the value of the degree (Barabasi and Oltvai, 2004).

Table 11: relationships between origin and network properties

Network	Species	Considered Category								Comparison with genes that originated later												
				Degree		Clustering coefficient		Betweenness			Degree		Clustering coefficient		Betweenness		Wilcoxon Test			Randomization	
		Origin	N genes	Mean	Median	Mean	Median	Mean	Median	N genes	Mean	Median	Mean	Median	Mean	Median	*p*-value degree	*p*-value clustering coefficient	*p*-value betweenness	*z*-score degree	*z*-score betweenness
Total Network	*H. sapiens*	LUCA	3,248	12,22	5	0,1392	0,0543	17,898	1,163	3,878	10,62	4	0,1486	0,0552	15,357	780	1,84E-04	1,72E-01	8,20E-05	1,29E-01	2,87E-01
		Eukaryotes	3,247	12,58	5	0,1393	0,0514	19,795	1,206	3,878	10,62	4	0,1566	0,0585	15,357	780	4,74E-07	8,83E-02	9,44E-04	9,85E-02	2,37E-01
		Opisthokonts	178	9,44	5	0,1111	0,0460	9,352	1,189	3,700	10,68	4	0,1591	0,0606	15,646	758	9,11E-02	4,80E-02	8,00E-01	7,33E-01	8,68E-01
		Metazoans	1,399	13,26	5	0,1541	0,0636	21,443	1,027	2,301	9,11	4	0,1623	0,0581	12,121	605	1,53E-10	9,08E-01	5,59E-03	1,94E-03	3,07E-02
		Vertebrates	1,683	10,03	4	0,1547	0,0606	14,037	786	618	6,60	3	0,1843	0,0513	6,903	296	4,21E-05	9,53E-01	1,33E-03	5,00E-05	2,85E-03
		Mammals	534	6,52	3	0,1790	0,0529	7,035	270	84	7,07	4	0,2160	0,0474	6,064	462	3,10E-01	6,76E-01	6,78E-01	6,59E-01	4,03E-01
		Primates	84	7,07	4	0,2160	0,0474	6,064	462	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
	D. melanogaster	LUCA	2,425	11,93	5	0,0259	0,0000	18,324	1,259	1,618	13,62	5,5	0,0257	0,0000	23,075	1,418	1,53E-01	3,60E-02	5,80E-02	8,99E-01	8,53E-01
		Eukaryotes	2,255	13,68	6	0,0255	0,0000	20,408	1,809	1,618	13,62	5,5	0,0259	0,0000	23,075	1,418	1,21E-01	7,61E-01	8,00E-01	4,80E-01	7,12E-01
		Opisthokonts	84	17,00	6,5	0,0234	0,0000	39,329	3,221	1,534	13,44	5	0,0260	0,0000	22,185	1,370	1,38E-01	9,62E-02	1,97E-02	1,85E-01	1,54E-01
		Metazoans	1,001	14,43	6	0,0302	0,0000	24,752	1,833	533	11,59	5	0,0172	0,0000	17,365	810	5,19E-06	9,28E-02	3,03E-03	2,12E-02	5,65E-02
		Insects	533	11,59	5	0,0172	0,0000	17,365	810	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
	S. cerevisiae	LUCA	2,433	38,52	20	0,1962	0,1429	7,759	1,333	111	17,32	8	0,2177	0,1662	2,584	441	3,06E-09	4,49E-05	2,69E-04	0,00E+00	4,61E-03
		Eukaryotes	2,688	30,69	18	0,2193	0,1667	5,906	1,188	111	17,32	8	0,1790	0,1039	2,584	441	9,69E-08	2,19E-03	9,45E-04	6,60E-04	3,51E-02
		Opisthokonts	6	8,00	5	0,0867	0,0000	538	92	105	17,85	8	0,1837	0,1071	2,701	464	2,61E-01	2,05E-01	2,29E-01	8,47E-01	8,84E-01
		Fungi	105	17,85	8	0,1837	0,1071	2,701	464	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
	E. coli	LUCA	2,136	12,70	5	0,1275	0,0714	3,955	385	703	6,28	4	0,1298	0,0545	1154	116	9,40E-12	2,23E-01	1,82E-08	0,00E+00	0,00E+00
		Bacteria	703	6,28	4	0,1298	0,0545	1,154	116	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Golden set	*H. sapiens*	LUCA	2,482	8,87	4	0,1392	0,0543	15,948	710	3,099	8,41	4	0,1486	0,0552	13,015	708	1,59E-01	1,72E-01	1,46E-01	3,31E-01	2,40E-01
		Eukaryotes	2,470	8,87	4	0,1393	0,0514	13,992	592	3,099	8,41	4	0,1566	0,0585	13,015	708	5,62E-01	8,83E-02	6,86E-01	3,20E-01	3,80E-01
		Opisthokonts	146	7,45	4	0,1111	0,0460	9,561	567	2,953	8,45	4	0,1591	0,0606	13,186	713	8,00E-01	4,80E-02	5,54E-01	7,22E-01	7,19E-01
		Metazoans	1,115	10,19	5	0,1541	0,0636	15,940	896	1,838	7,40	4	0,1623	0,0581	11,515	576	1,21E-06	9,08E-01	8,00E-02	4,37E-03	1,30E-01
		Vertebrates	1,355	8,10	4	0,1547	0,0606	13,331	782	483	5,42	3	0,1843	0,0513	6,420	182	6,91E-05	9,53E-01	9,80E-04	9,00E-05	3,80E-03
		Mammals	414	5,46	3	0,1790	0,0529	6,604	182	69	5,17	3	0,2160	0,0474	5,319	190	5,45E-01	6,76E-01	9,72E-01	4,04E-01	3,54E-01
		Primates	69	5,17	3	0,2160	0,0474	5,319	190	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
	S. cerevisiae	LUCA	1,690	11,09	5	0,1962	0,1429	6,708	1,016	70	6,80	3	0,2177	0,1662	3,225	81	2,08E-02	4,49E-05	3,13E-03	1,27E-02	4,22E-02
		Eukaryotes	1,998	11,83	7	0,2193	0,1667	6,131	1,124	70	6,80	3	0,1790	0,1039	3,225	81	1,09E-04	2,19E-03	1,49E-02	3,97E-03	8,12E-02
		Opisthokonts	4	1,75	1	0,0867	0,0000	431	0	66	7,11	3	0,1837	0,1071	3,395	133	5,64E-02	2,05E-01	6,99E-01	9,40E-01	8,30E-01
		Fungi	66	7,11	3	0,1837	0,1071	3,395	133	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA

Network	Species	Considered Category								Comparison with genes that originated earlier												
				Degree		Clustering coefficient		Betweenness			Degree		Clustering coefficient		Betweenness		Wilcoxon Test			Randomization	
		Origin	N genes	Mean	Median	Mean	Median	Mean	Median	N genes	Mean	Median	Mean	Median	Mean	Median	*p*-value degree	*p*-value clustering coefficient	*p*-value betweenness	*z*-score degree	*z*-score betweenness
Total Network	*H. sapiens*	LUCA	3,248	12,22	5	0,1392	0,0543	17,898	1,163	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
		Eukaryotes	3,247	12,58	5	0,1393	0,0514	19,795	1,206	3,248	12,22	5	0,1392	0,0543	17,898	1,163	2,25E-01	7,87E-01	5,52E-01	5,85E-01	5,89E-01
		Opisthokonts	178	9,44	5	0,1111	0,0460	9,352	1,189	6,495	12,40	5	0,1393	0,0524	18,846	1,180	8,00E-01	2,05E-01	2,31E-01	6,56E-02	4,99E-02
		Metazoans	1,399	13,26	5	0,1541	0,0636	21,443	1,027	6,673	12,32	5	0,1385	0,0523	18,593	1,180	4,03E-01	4,72E-02	3,79E-01	7,22E-01	7,03E-01
		Vertebrates	1,683	10,03	4	0,1547	0,0606	14,037	786	8,072	12,48	5	0,1412	0,0545	19,087	1,154	5,62E-08	1,14E-01	4,59E-03	4,70E-02	1,66E-01
		Mammals	534	6,52	3	0,1790	0,0529	7,035	270	9,755	12,06	5	0,1434	0,0549	18,216	1,066	2,10E-13	5,22E-01	1,11E-06	1,00E-02	5,98E-02
		Primates	84	7,07	4	0,2160	0,0474	6,064	462	10,289	11,77	5	0,1451	0,0549	17,635	1,000	7,16E-02	5,70E-01	1,41E-01	1,99E-02	1,96E-02
	D. melanogaster	LUCA	2,425	11,93	5	0,0259	0,0000	18,324	1,259	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
		Eukaryotes	2,255	13,68	6	0,0255	0,0000	20,408	1,809	2,425	11,93	5	0,0259	0,0000	18,324	1,259	8,72E-04	4,52E-02	4,70E-02	9,22E-01	7,33E-01
		Opisthokonts	84	17,00	6,5	0,0234	0,0000	39,329	3,221	4,680	12,77	6	0,0257	0,0000	19,328	1,470	1,54E-01	6,80E-02	1,28E-02	8,74E-01	9,13E-01
		Metazoans	1,001	14,43	6	0,0302	0,0000	24,752	1,833	4,764	12,85	6	0,0257	0,0000	19,681	1,479	3,18E-02	2,90E-01	5,26E-02	8,68E-01	8,50E-01
		Insects	533	11,59	5	0,0172	0,0000	17,365	810	5,765	13,12	6	0,0265	0,0000	20,561	1,529	3,68E-05	1,84E-01	2,69E-02	1,14E-01	1,96E-01
	S. cerevisiae	LUCA	2,433	38,52	20	0,1962	0,1429	7,759	1,333	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
		Eukaryotes	2,688	30,69	18	0,2193	0,1667	5,906	1,188	2,433	38,52	20	0,1962	0,1429	7,759	1,333	1,89E-03	7,26E-06	1,03E-01	1,11E-01	3,00E-01
		Opisthokonts	6	8,00	5	0,0867	0,0000	538	92	5,121	34,41	19	0,2082	0,1556	6,786	1,250	2,05E-02	7,27E-02	6,16E-02	1,86E-02	2,88E-02
		Fungi	105	17,85	8	0,1837	0,1071	2,701	464	5,127	34,38	19	0,2081	0,1556	6,779	1,248	1,30E-07	2,65E-02	1,57E-03	3,20E-04	2,07E-02
	E. coli	LUCA	2,136	12,70	5	0,1275	0,0714	3,955	385	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
		Bacteria	703	6,28	4	0,1298	0,0545	1,154	116	2,136	12,70	5	0,1275	0,0714	3,955	385	9,40E-12	2,23E-01	1,82E-08	0,00E+00	0,00E+00
Golden set	*H. sapiens*	LUCA	2,482	8,87	4	0,1392	0,0543	15,948	710	NA	NA	NA	0,1486	0,0552	NA	NA	NA	NA	NA	NA	NA
		Eukaryotes	2,470	8,87	4	0,1393	0,0514	13,992	592	2,482	8,87	4	0,1566	0,0585	15,948	710	6,03E-02	7,87E-01	7,08E-02	5,02E-01	3,21E-01
		Opisthokonts	146	7,45	4	0,1111	0,0460	9,561	567	4,952	8,87	4	0,1591	0,0606	14,972	647	8,00E-01	2,05E-01	4,65E-01	2,04E-01	1,91E-01
		Metazoans	1,115	10,19	5	0,1541	0,0636	15,940	896	5,098	8,83	4	0,1623	0,0581	14,817	645	1,91E-04	4,72E-02	3,58E-01	9,00E-01	6,28E-01
		Vertebrates	1,355	8,10	4	0,1547	0,0606	13,331	782	6,213	9,07	4	0,1843	0,0513	15,019	685	3,89E-01	1,14E-01	9,20E-01	1,83E-01	3,34E-01
		Mammals	414	5,46	3	0,1790	0,0529	6,604	182	7,568	8,90	4	0,2160	0,0474	14,717	702	1,67E-05	5,22E-01	7,64E-04	3,69E-02	1,06E-01
		Primates	69	5,17	3	0,2160	0,0474	5,319	190	7,982	8,72	4	NA	NA	14,296	664	2,36E-02	5,70E-01	2,20E-01	2,36E-02	3,97E-02
	S. cerevisiae	LUCA	1,690	11,09	5	0,1962	0,1429	6,708	1,016	NA	NA	NA	0,2177	0,1662	NA	NA	NA	NA	NA	NA	NA
		Eukaryotes	1,998	11,83	7	0,2193	0,1667	6,131	1,124	1,690	11,09	5	0,1790	0,1039	6,708	1,016	3,11E-06	7,26E-06	2,06E-02	6,17E-01	4,08E-01
		Opisthokonts	4	1,75	1	0,0867	0,0000	431	0	3,688	11,49	6	0,1837	0,1071	6,396	1,071	1,97E-02	7,27E-02	8,00E-01	1,23E-02	6,94E-02
		Fungi	66	7,11	3	0,1837	0,1071	3,395	133	3,692	11,48	6	NA	NA	6,389	1,068	6,93E-03	2,65E-02	7,16E-03	1,36E-02	8,15E-02

The network properties of genes born at each level of evolution are compared with those of genes that originated later and earlier. The comparisons are made using the Wilcoxon test and a randomization test (only for degree and betweenness), which is made to determine whether the significance is due only to the different numbers of genes for each category. 100,000 randomizations of 500 proteins are made in order to determine whether the significant differences determined using the Wilcoxon test are due to eventual biases in the number of genes for each age category. Significant *p*-values and *Z*-scores are depicted in red for enrichment and in green for depletion. NA: not available.

2.2. Conserved proteins are highly connected and central

The analysis of conservation showed that highly conserved proteins are more connected and more central than less conserved proteins (Figure 36, Table 12). This demonstrates that the tendency of gene loss in particular lineages is influenced by the number of interactions and the position inside the protein interaction network of the corresponding protein. Central proteins with many interactions tend to be conserved throughout evolution, because their loss would disrupt the network in a more detrimental way, compared to lowly connected and peripheral proteins.

The signal for conservation is strong and consistent, with only two exceptions in *E. coli* and in the golden set of *S. cerevisiae*, which may be due to the incompleteness of the networks, rather than a real biological signal, given their apparent randomness.

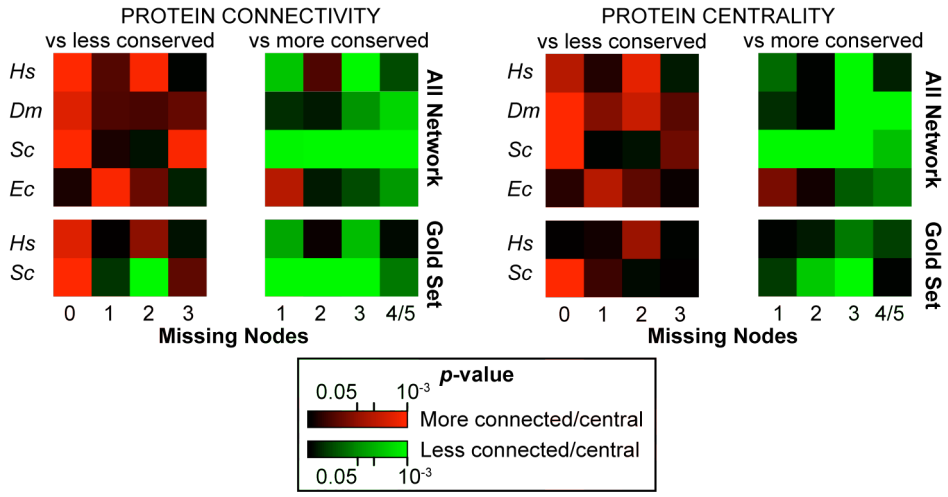


Figure 36: Relationships between conservation and network properties

Degree (left) and betweenness (right) are compared between proteins with a certain level of conservation and more and less conserved proteins. Conservation is calculated as the number of internal nodes of the tree of life where no orthologs can be found. 0 represents the maximum. The analysis is performed as explained in Figure 33. Red is associated with higher connectivity and centrality, green with lower. Black represents no difference.

Table 12: relationships between conservation and network properties

Network	Species	Considered Category						Comparison with less conserved genes								
		Conservation	N genes	Degree		Betweenness		N genes	Degree		Betweenness		Wilcoxon Test		Randomization	
				Mean	Median	Mean	Median		Mean	Median	Mean	Median	*p*-value degree	*p*-value betweenness	*z*-score degree	*z*-score betweenness
Total Network	*H. sapiens*	0	7,248	11,91	5	17,594	1,093	3,125	11,32	4	17,421	837	**2,59E-05**	**1,06E-02**	3,45E-01	4,91E-01
		1	1,926	10,85	5	16,348	951	1,199	12,08	4	19,144	614	1,44E-01	4,36E-01	7,73E-01	6,84E-01
		2	873	13,15	5	22,832	921	326	9,22	3	9,267	258	**1,47E-03**	**2,60E-03**	**1,64E-02**	**6,15E-03**
		3	287	9,26	3	9,395	246	39	8,95	4	8,327	349	8,23E-01	4,92E-01	4,54E-01	4,23E-01
		4	39	8,95	4	8,327	349	NA	NA	NA	NA	NA	NA	NA	NA	NA
	D. melanogaster	0	3,861	13,41	6	21,988	1,654	2,437	12,32	5	17,602	1,148	**3,28E-03**	**3,60E-04**	1,96E-01	1,28E-01
		1	843	12,46	6	17,837	1,404	1,594	12,25	5	17,478	1,011	1,63E-01	**4,37E-02**	4,33E-01	4,63E-01
		2	418	12,76	6	19,139	1,294	1,176	12,07	5	16,887	889	1,91E-01	**6,11E-03**	3,06E-01	2,69E-01
		3	923	12,35	5	16,995	1,041	253	11,04	4	16,494	476	1,00E-01	1,33E-01	2,16E-01	4,53E-01
		4	253	11,04	4	16,494	476	NA	NA	NA	NA	NA	NA	NA	NA	NA
	S. cerevisiae	0	3,702	39,17	22	7,945	1,576	1,530	21,66	12	3,677	627	**1,24E-46**	**3,53E-31**	**0,00E+00**	**4,80E-03**
		1	82	24,95	12,5	3,800	559	1,448	21,48	12	3,670	627	5,33E-01	7,12E-01	2,23E-01	3,88E-01
		2	245	25,28	11	6,874	572	1,203	20,70	12	3,018	633	5,72E-01	6,14E-01	6,38E-02	**3,90E-02**
		3	1,176	20,84	13	3,030	650	27	14,52	4	2,474	141	**1,32E-03**	**7,73E-02**	1,71E-01	4,23E-01
		4	27	14,52	4	2,474	141	NA	NA	NA	NA	NA	NA	NA	NA	NA
	E. coli	0	1,584	12,48	5	4,207	314	1,255	9,38	5	2,067	260	4,97E-01	3,69E-01	**2,61E-02**	**2,54E-02**
		1	153	11,65	6	2,475	731	1,102	9,07	4	2,010	217	**1,37E-03**	**1,12E-02**	9,82E-02	1,97E-01
		2	134	11,90	5	2,800	384	968	8,68	4	1,901	183	8,81E-02	1,08E-01	6,00E-02	9,33E-02
		3	401	8,52	4	1,762	250	567	8,78	4	1,999	141	4,40E-01	6,56E-01	6,01E-01	7,69E-01
		4	567	8,78	4	1,999	141	NA	NA	NA	NA	NA	NA	NA	NA	NA
Gold Set	*H. sapiens*	0	5,689	8,75	4	14,237	711	2,362	8,55	3,5	14,176	580	**3,21E-03**	9,88E-01	4,22E-01	4,89E-01
		1	1,473	7,97	4	12,443	695	889	9,49	3	17,047	468	8,92E-01	5,56E-01	9,08E-01	8,37E-01
		2	663	10,05	4	19,559	566	226	7,86	3	9,678	222	**3,90E-02**	**2,45E-02**	1,06E-01	6,18E-02
		3	201	8,17	3	10,405	200	25	5,40	4	3,834	531	5,68E-01	7,41E-01	2,01E-01	1,56E-01
		4	25	5,40	4	3,834	531	NA	NA	NA	NA	NA	NA	NA	NA	NA
	S. cerevisiae	0	2,783	12,86	7	7,407	1,469	975	7,25	4	3,280	339	**1,33E-26**	**2,56E-18**	**0,00E+00**	**0,00E+00**
		1	48	6,75	3	4,381	434	927	7,28	4	3,223	334	2,94E-01	2,47E-01	6,09E-01	2,82E-01
		2	137	5,76	2	4,313	27	790	7,54	4	3,034	387	**1,58E-04**	6,72E-01	9,43E-01	1,37E-01
		3	781	7,54	4	2,973	396	9	7,78	2	8,321	0	1,10E-01	9,72E-01	4,99E-01	7,45E-01
		4	9	7,78	2	8,321	0	NA	NA	NA	NA	NA	NA	NA	NA	NA

Network	Species	Considered Category						Comparison with less conserved genes								
		Conservation	N genes	Degree		Betweenness		N genes	Degree		Betweenness		Wilcoxon Test		Randomization	
				Mean	Median	Mean	Median		Mean	Median	Mean	Median	*p*-value degree	*p*-value betweenness	*z*-score degree	*z*-score betweenness
Total Network	*H. sapiens*	0	7,248	11,91	5	17,594	1,093	NA	NA	NA	NA	NA	NA	NA	NA	NA
		1	1,926	10,85	5	16,348	951	7,248	11,91	5	17,594	1,093	**5,48E-03**	7,82E-02	2,35E-01	4,09E-01
		2	873	13,15	5	22,832	921	9,174	11,69	5	17,332	1,062	1,69E-01	7,75E-01	7,48E-01	7,66E-01
		3	287	9,26	3	9,395	246	10,047	11,81	5	17,810	1,040	**1,50E-05**	**3,98E-04**	2,90E-01	2,62E-01
		4	39	8,95	4	8,327	349	10,334	11,74	5	17,577	998	1,63E-01	4,20E-01	2,73E-01	2,23E-01
	D. melanogaster	0	3,861	13,41	6	21,988	1,654	NA	NA	NA	NA	NA	NA	NA	NA	NA
		1	843	12,46	6	17,837	1,404	3,861	13,41	6	21,988	1,654	3,26E-01	3,27E-01	2,28E-01	1,46E-01
		2	418	12,76	6	19,139	1,294	4,704	13,24	6	21,244	1,583	4,87E-01	9,47E-01	3,68E-01	3,14E-01
		3	923	12,35	5	16,995	1,041	5,122	13,20	6	21,072	1,565	**2,51E-02**	**2,56E-04**	3,20E-01	2,16E-01
		4	253	11,04	4	16,494	476	6,045	13,07	6	20,450	1,498	**3,18E-03**	**6,11E-04**	1,18E-01	2,21E-01
	S. cerevisiae	0	3,702	39,17	22	7,945	1,576	NA	NA	NA	NA	NA	NA	NA	NA	NA
		1	82	24,95	12,5	3,800	559	3,702	39,17	22	7,945	1,576	**1,22E-03**	**9,53E-04**	**1,21E-02**	9,17E-02
		2	245	25,28	11	6,874	572	3,784	38,86	22	7,855	1,541	**2,02E-11**	**7,05E-08**	**6,10E-04**	3,94E-01
		3	1,176	20,84	13	3,030	650	4,029	38,03	21	7,796	1,476	**5,27E-32**	**4,19E-21**	**3,77E-02**	1,50E-01
		4	27	14,52	4	2,474	141	5,205	34,15	19	6,719	1,239	**1,74E-05**	**6,46E-03**	**1,69E-02**	1,46E-01
	E. coli	0	1,584	12,48	5	4,207	314	NA	NA	NA	NA	NA	NA	NA	NA	NA
		1	153	11,65	6	2,475	731	1,584	12,48	5	4,207	314	**1,02E-02**	**5,86E-02**	4,26E-01	3,01E-01
		2	134	11,90	5	2,800	384	1,737	12,40	5	4,055	344	5,11E-01	5,91E-01	4,73E-01	4,36E-01
		3	401	8,52	4	1,762	250	1,871	12,37	5	3,965	358	1,67E-01	1,17E-01	**8,20E-03**	**7,20E-03**
		4	567	8,78	4	1,999	141	2,272	11,69	5	3,576	334	**2,18E-02**	**4,92E-02**	**1,86E-02**	**4,84E-02**
Gold Set	*H. sapiens*	0	5,689	8,75	4	14,237	711	NA	NA	NA	NA	NA	NA	NA	NA	NA
		1	1,473	7,97	4	12,443	695	5,689	8,75	4	14,237	711	**1,52E-02**	7,62E-01	2,15E-01	3,10E-01
		2	663	10,05	4	19,559	566	7,162	8,59	4	13,868	709	6,56E-01	4,94E-01	7,87E-01	7,58E-01
		3	201	8,17	3	10,405	200	7,825	8,71	4	14,350	685	**7,32E-03**	**5,12E-02**	4,53E-01	4,18E-01
		4	25	5,40	4	3,834	531	8,026	8,70	4	14,251	659	6,58E-01	2,19E-01	1,33E-01	7,89E-02
	S. cerevisiae	0	2,783	12,86	7	7,407	1,469	NA	NA	NA	NA	NA	NA	NA	NA	NA
		1	48	6,75	3	4,381	434	2,783	12,86	7	7,407	1,469	**4,68E-04**	2,60E-01	**6,57E-03**	1,69E-01
		2	137	5,76	2	4,313	27	2,831	12,75	7	7,356	1,452	**5,94E-13**	**4,97E-03**	**0,00E+00**	**5,11E-02**
		3	781	7,54	4	2,973	396	2,968	12,43	6	7,216	1,327	**5,00E-14**	**1,38E-16**	2,01E-01	2,25E-01
		4	9	7,78	2	8,321	0	3,749	11,41	6	6,332	1,050	**4,65E-02**	7,21E-01	2,84E-01	5,97E-01

The network properties of genes with a certain level of conservation are compared with those of genes that have higher or lower conservation. Conservation is calculated as the number of internal nodes where no orthologs for the gene of interest are found. The comparisons are made using the Wilcoxon test and a randomization test, which is made as explained in Table 11. Significant *p*-values and *Z*-scores are depicted in red for enrichment and in green for depletion. NA: not available.

2.3. A randomization test confirms the relationships between evolutionary and network properties

We observed a high variability in terms of the number of genes that were included in each category of origin and conservation: it spans three orders of magnitude, from less than ten to more than 7,000 proteins (Table 11, Table 12). This may introduce a bias in the results: in particular, when comparing categories that include a high number of proteins, even small differences may appear as significant, that are not significant if at least one category contains a small number of genes. To overcome this potential bias, we performed a randomization test for each category of origin and conservation. We executed 100,000 iterations and at each iteration we measured the difference in the average degree and betweenness between 500 randomly selected proteins that originated at a given evolutionary time point and 500 younger or older proteins. The randomization was repeated for each evolutionary time point and in each species separately. If one group contained less than 500 proteins, we selected an equal number also from the older/younger groups (for example, since there are only 84 human genes originated with primates, these were compared with 84 randomly picked older genes). At the end of all randomizations, we obtained two distributions of differences between the average degree of proteins originated at a given evolutionary time point and younger and older proteins. From these distributions a *Z*-score was measured as the fraction of randomizations with a difference <0 when comparing with younger proteins, and >0 when comparing with older proteins. A similar analysis was made also for centrality and both connectivity and centrality with respect to conservation.

The randomization analysis confirmed that ancestral and highly conserved genes encode more connected and more central proteins, compared with younger and less conserved genes (Table 11, Table 12, Figure 37, Figure 38).

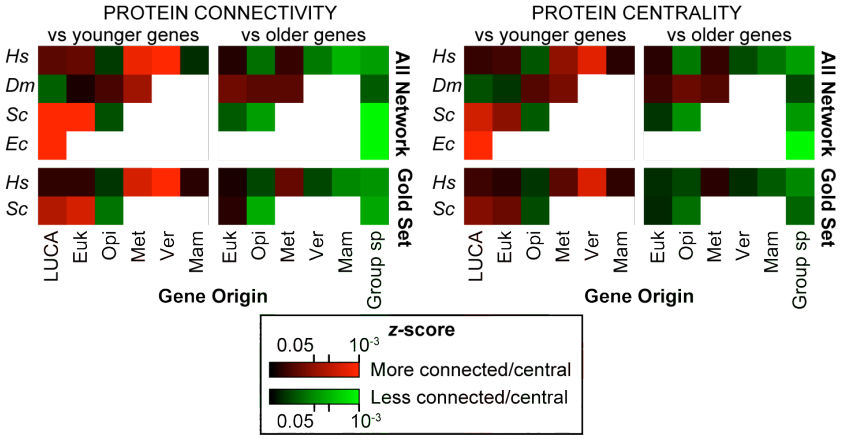


Figure 37: Relationships between origin and network properties using a randomization test

Degree (left) and betweenness (right) are compared between proteins that originated at a certain node and older or younger proteins. The analysis is performed using a randomization test and the resulting *z*-scores are transformed into heatmaps. The test is performed by randomly selecting 500 proteins with the same origin, determining the mean degree and betweenness and comparing these values with those of 500 randomly selected proteins that originated before and later in evolution. The procedure is repeated 100,000 times and *z*-score is calculated as the fraction of randomizations with a negative difference when comparing with younger proteins and with a positive difference when comparing with older proteins. The color code is as described in Figure 33.

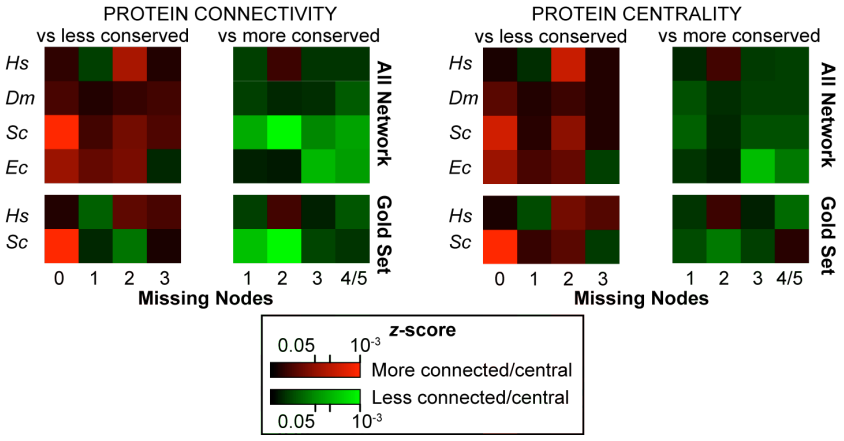


Figure 38: Relationships between conservation and network properties using a randomization test

The connectivity (left) and centrality (right) of proteins with a certain level of conservation are compared with more and less conserved proteins using a randomization test as described in Figure 37.

2.4. A core of singleton hubs is conserved in evolution

The results described so far show that protein interaction networks in different species seem to evolve in similar ways, following the preferential attachment theory. Ancestral proteins are highly connected and central, while young proteins are at the network periphery. Furthermore, highly connected and central proteins are retained in all lineages, since they have higher conservation compared to lowly connected and peripheral proteins. These properties are similar in all four species, although the networks are highly heterogeneous in terms of completeness and type of interactions. The conservation of the relationships between evolutionary and network properties is the consequence of the conservation of ancestral genes in evolution. In order to demonstrate this, we examined whether the network properties are conserved among orthologs and, in particular, whether the most connected proteins in one species have orthologs that are also highly connected. In absence of a consensus definition of hubs (*i.e.* the most highly connected nodes inside a network) (Vallabhajosyula et al., 2009), we identified them as the top 25% most connected nodes in each protein interaction network (degree ≥ 12). Half of the singleton hubs that originated early in evolution have orthologs that are also hubs in at least one of the other species (Table 13). Therefore protein interaction networks have a core of ancestral proteins, which are highly connected and central. These proteins are highly conserved in evolution and preserve their properties in all species.

Table 13: Hub conservation throughout evolution

Origin	Hub type	*Hs*	*Dm*	*Sc*	*Ec*
LUCA	Singleton Hubs	226	258	386	348
	Hubs with Orthologs in Networks	169	194	287	188
	Orthologous Hubs (%)	76 (45)	99 (51)	113 (39)	93 (49)
Eukaryotes	Singleton Hubs	235	305	500	NA
	Hubs with Orthologs in Networks	212	227	277	NA
	Orthologous Hubs (%)	117 (55)	113 (50)	137 (49)	NA
Total	**Singleton Hubs**	**461**	**563**	**886**	**348**
	Hubs with Orthologs in Networks	**381**	**427**	**564**	**188**
	Orthologous Hubs (%)	**193 (51)**	**212 (50)**	**250 (44)**	**93 (49)**

For each species, the number of singleton hubs that originated in the last universal common ancestor (LUCA) and in eukaryotes is extracted. Of these proteins, only those that have orthologs in at least one of the other three species are considered and the percentage of ancestral singleton hubs that have orthologs that are also hubs is calculated.

3. A novel group of duplicated hubs is acquired in the human network

So far, we have demonstrated that the origin and the level of conservation of a gene influence the network properties of its encoded protein and the netework core is highly conserved throughout evolution. Furthermore, the relationships duplicability, connectivity and centrality are different between different species. In particular, while human hubs are mostly duplicated, in the other species singleton genes encode more connected proteins than duplicated genes. In order to understand the evolutionary reasons for this change, we studied how network properties of singleton and duplicated proteins are affected by their origin. We compared the network properties of singleton and duplicated proteins that have the same origin. In case the origin of a gene influences the relationships between network properties and duplicability, this analysis should show significant differences between human and the other species.

The analysis of ancient genes for all four species showed the same relationships between origin, duplicability and network properties: among genes that originated between

the last universal common ancestor and eukaryotes, singletons are more connected and more central than duplicated proteins, supporting the presence of a core of ancestral proteins that are highly conserved in all networks (Figure 39, Table 14). Surprisingly, this was evident also in the *H. sapiens* protein interaction network, although the general trend is opposite, *i.e.* duplicated genes encode more connected and central proteins than singletons. The differences between human and the other species became evident in the analysis of younger genes. Indeed, while *D. melanogaster*, *S. cerevisiae* and *E. coli* showed a slight enrichment of singletons in hubs, the *H. sapiens* network presented a significant inversion of the trend. Duplicated genes that originated with metazoans and, with less statistical evidence, vertebrates and mammals, encode proteins that are more connected than singletons. This signal was evident also when analyzing the gold set. The difference between duplicated and singleton genes was so strong that the overall signal appeared different for *H. sapiens*, compared to the other species.

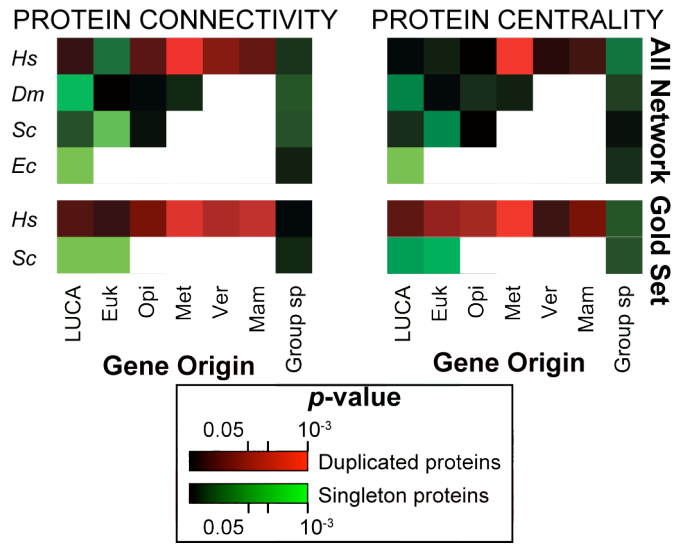


Figure 39: Relationships between conservation and network properties

Wilcoxon test is used to compare degree and betweenness of singleton and duplicated genes at different levels of origin in all four species. The resulting *p*-values are transformed into heatmaps as described in Figure 33.

Table 14: relationships between duplicability and network properties

Network	Species	Origin	N genes	Duplicated Genes					Singleton Genes					*p*-value degree	*p*-value betweenness
				N genes	Degree		Betweenness		N genes	Degree		Betweenness			
					Mean	Median	Mean	Median		Mean	Median	Mean	Median		
Total Network	*H. sapiens*	LUCA	3248	2468	12,5	5,0	18945	1088	780	11,3	5,0	14586	1340	3,08E-01	8,00E-01
		Eukaryotes	3247	2474	12,9	5,0	21615	1131	773	11,7	6,0	13970	1359	7,15E-02	5,20E-01
		Opisthokonts	178	148	9,8	5,5	9916	1221	30	7,6	3,5	6572	896	1,12E-01	8,00E-01
		Metazoans	1399	901	14,2	6,0	23900	1737	498	11,5	4,0	16996	478	1,18E-03	5,39E-07
		Vertebrates	1683	765	11,1	4,0	15110	872	918	9,2	4,0	13143	681	3,83E-02	3,88E-01
		Mammals	534	187	7,7	4,0	9236	500	347	5,9	3,0	5849	197	8,86E-02	2,01E-01
		Primates	84	20	5,8	2,5	4550	103	64	7,5	4,0	6537	981	2,73E-01	6,23E-02
	D. melanogaster	LUCA	2425	1493	11,7	5,0	17804	1012	932	12,2	6,0	19157	1648	7,05E-03	4,22E-02
		Eukaryotes	2255	1279	14,1	6,0	21247	1629	976	13,2	6,0	19308	1935	8,00E-01	8,00E-01
		Opisthokonts	84	40	14,1	5,5	23371	1014	44	19,7	9,0	53836	4433	8,00E-01	3,30E-01
		Metazoans	1001	396	14,0	6,0	23344	1586	605	14,7	7,0	25674	2199	4,15E-01	5,39E-01
		Insects	533	221	11,3	4,0	17328	418	312	11,8	5,0	17391	934	1,34E-01	2,08E-01
	S. cerevisiae	LUCA	2433	1284	39,1	19,0	9084	1296	1149	37,9	21,0	6278	1366	1,41E-01	3,19E-01
		Eukaryotes	2688	863	30,0	16,0	6080	970	1825	31,0	19,0	5824	1334	2,04E-03	3,52E-02
		Opisthokonts	6	1	2,0	2,0	156	156	5	9,2	7,0	614	27	6,67E-01	8,00E-01
		Fungi	105	14	14,7	5,0	2919	258	91	18,3	9,0	2667	493	1,42E-01	6,58E-01
	E. coli	LUCA	2136	1283	8,8	4,0	2287	212	853	18,5	7,0	6464	696	2,32E-20	4,36E-10
		Bacteria	703	165	5,5	3,0	1005	106	538	6,5	4,0	1199	124	4,80E-01	3,16E-01
Golden Set	*H. sapiens*	LUCA	2482	1935	9,4	4,0	17942	776	547	6,9	4,0	8894	524	1,34E-01	1,04E-01
		Eukaryotes	2470	1918	9,2	4,0	15392	600	552	7,5	4,0	9128	542	2,76E-01	2,42E-02
		Opisthokonts	146	130	8,0	4,0	10525	753	16	3,4	2,0	1735	4	5,31E-02	1,80E-02
		Metazoans	1115	775	11,0	5,0	18390	1281	340	8,3	4,0	10356	286	3,24E-03	4,92E-04
		Vertebrates	1355	650	8,9	4,0	14635	1124	705	7,3	4,0	12130	510	1,40E-02	2,59E-01
		Mammals	414	147	6,4	4,0	8388	948	267	4,9	3,0	5621	91	8,61E-03	5,79E-02
		Primates	69	18	5,2	2,0	3882	55	51	5,2	3,0	5826	225	8,00E-01	1,23E-01
	S. cerevisiae	LUCA	1690	856	9,0	4,0	6782	716	834	13,2	6,0	6633	1355	4,14E-07	1,88E-02
		Eukaryotes	1998	586	10,2	5,0	6767	1001	1412	12,5	7,0	5868	1167	2,58E-06	1,04E-02
		Opisthokonts	4	0	NA	NA	NA	NA	4	1,8	1,0	431	0	1,00E+00	1,00E+00
		Fungi	66	7	7,7	2,0	962	0	59	7,0	3,0	3683	193	4,12E-01	1,55E-01

For each level of origin, the network properties of duplicated and singleton genes are compared. Duplicability is defined as the presence of paralogous genes inside the same KOG (for eukaryotes) or COG (fo *E. coli*). The comparisons are made using the Wilcoxon test. Significant *p*-values are depicted in red for enrichment in duplicated genes and in green for enrichment in singletons. NA: not available.

Since *D. melanogaster* genes that originated with metazoans did not show the same properties of *H. sapiens*, it is likely that the genes that originated with metazoans gained the "hub" status after the speciation of insects.

These findings support the hypothesis that the evolution of protein interaction networks has allowed the birth of two classes of hubs, one ancestral, which is conserved in all species, and one that has arisen later in evolution. The first constitutes the core of the network and is composed of proteins encoded by ancient and highly conserved singleton genes. The second originated with metazoans and, to a lower extent, vertebrates and mammals, and is composed of young duplicated genes. These hubs appeared in the ancestor of vertebrates and their duplication has been retained in evolution because their sensitivity to dosage modifications has likely been reduced by other factors.

4. Ancient and recent human hubs are involved in different functions

The human protein interaction network shares a core of ancestral singleton hubs with all other species, and acquired a class of young duplicated hubs that is instead absent in other species. Given the different origin and duplicability, these two classes of hubs are likely to have different biological functions. Of the 2,573 human hubs, 461 are ancestral (*i.e.* originated with the last universal common ancestor or eukaryotes) and singleton, while 468 are recent (*i.e.* originated with metazoans or vertebrates) and duplicated.

We designed a pipeline to analyze the functional differences between ancestral singleton and recent duplicated hubs that relies on the Biological Process branch of the Gene Ontology (GO) (Ashburner et al., 2000). Briefly, all the GO terms at level 5 and 6 are mapped to twelve macro-categories to facilitate the understanding of the functional

differences between the two classes. Then all genes from the two classes are mapped to each GO term and the enrichment in one of the two classes is calculated using Fisher's exact test. Depending on the number and the level of significance of the GO terms and the number of significant GO terms in each macro-category, the functional differences between the two classes become evident.

This analysis showed indeed that there are significant functional differences between the two classes of hubs (Figure 40, Table 15). Ancient singleton hubs are involved in basic biological processes that are related to the cell survival. In particular, this class is enriched in GO terms that are associated to the categories "DNA and RNA metabolism and transcription" and "cellular metabolism". Recent duplicated hubs, instead, are involved in the organization of the multicellular organism, in cell communication and in regulatory functions. Indeed, they are associated with the development of organs and tissues, cell motility and interactions with the environment, response to external stimuli and immune response. Furthermore, they are involved in the cell homeostasis and in the regulation of several pathways, transcription and cell cycle.

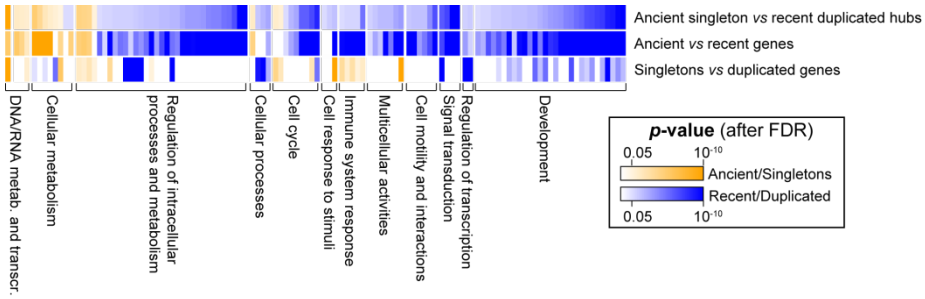


Figure 40: Functional differences between the two classes of human hubs

Functional differences are analyzed between ancient singleton hubs and recent duplicated hubs, ancestral and recent human genes, and singleton and duplicated human genes. "Ancestral" refers to as genes that originated between the last universal common ancestor and Opisthokonts, while "recent" identifies genes that originated with Metazoans or later. For each GO term, Fisher's exact test is performed for all the three analyses and the resulting *p*-values are adjusted for the false discovery rate. Each vertical bar corresponds to a single GO term. All GO terms are grouped into 12 functional categories. Blue bars represent significant enrichment for duplicated and recent genes or hubs, while orange represents enrichment in singleton and recent genes or hubs. White represents no enrichment.

Table 15: functional comparison between recent duplicated hubs and ancestral singleton hubs

Process	GO level	GO description	GO ID	N genes (list 1)	N genes (list 2)	% of total genes (list 1)	% of total genes (list 2)	p-value	adjusted p-value	Enriched list
Cell cycle	5	cell cycle arrest	GO:0007050	12	2	2,75	0,48	1,25E-02	3,27E-02	1
Cell cycle	5	apoptosis	GO:0006915	88	28	20,18	6,71	6,80E-09	9,42E-08	1
Cell cycle	5	cell development	GO:0048468	60	9	13,76	2,16	1,19E-10	2,69E-09	1
Cell cycle	5	negative regulation of growth	GO:0045926	16	1	3,67	0,24	2,55E-04	1,28E-03	1
Cell cycle	5	regulation of cell proliferation	GO:0042127	58	13	13,30	3,12	3,99E-08	4,22E-07	1
Cell cycle	5	cell growth	GO:0016049	21	2	4,82	0,48	5,85E-05	3,29E-04	1
Cell cycle	6	M phase	GO:0000279	8	25	1,83	6,00	2,09E-03	8,77E-03	2
Cell cycle	6	regulation of programmed cell death	GO:0043067	72	20	16,51	4,80	1,86E-08	4,43E-07	1
Cell cycle	6	cell cycle checkpoint	GO:0000075	2	12	0,46	2,88	5,94E-03	2,12E-02	2
Cell motility and interactions	5	taxis	GO:0042330	14	1	3,21	0,24	9,25E-04	3,47E-03	1
Cell motility and interactions	5	regulation of cellular component movement	GO:0051270	25	2	5,73	0,48	4,72E-06	3,40E-05	1
Cell motility and interactions	5	regulation of cell adhesion	GO:0030155	14	1	3,21	0,24	9,25E-04	3,47E-03	1
Cell motility and interactions	6	chemotaxis	GO:0006935	14	1	3,21	0,24	9,25E-04	4,71E-03	1
Cell motility and interactions	6	cell migration	GO:0016477	39	3	8,94	0,72	3,52E-09	1,08E-07	1
Cell motility and interactions	6	positive regulation of cellular component movement	GO:0051272	15	1	3,44	0,24	4,87E-04	2,82E-03	1
Cell response to stimuli	6	sensory perception	GO:0007600	20	3	4,59	0,72	4,44E-04	2,79E-03	1
Cell response to stimuli	6	cellular response to hormone stimulus	GO:0032870	15	2	3,44	0,48	2,22E-03	8,98E-03	1
Cell response to stimuli	6	response to peptide hormone stimulus	GO:0043434	14	3	3,21	0,72	1,22E-02	3,89E-02	1
Cellular metabolism	5	cellular nitrogen compound biosynthetic process	GO:0044271	1	8	0,23	1,92	1,84E-02	4,74E-02	2
Cellular metabolism	5	macromolecule catabolic process	GO:0043285	29	58	6,65	13,91	6,18E-04	2,47E-03	2
Cellular metabolism	5	cellular macromolecule catabolic process	GO:0044265	17	51	3,90	12,23	6,60E-06	4,57E-05	2
Cellular metabolism	5	macromolecule biosynthetic process	GO:0043284	136	167	31,19	40,05	8,06E-03	2,27E-02	2
Cellular metabolism	5	cellular lipid metabolic process	GO:0044255	23	8	5,28	1,92	9,88E-03	2,74E-02	1
Cellular metabolism	6	modification-dependent macromolecule catabolic process	GO:0043632	12	32	2,75	7,67	1,66E-03	7,64E-03	2
Cellular metabolism	6	cellular protein metabolic process	GO:0044267	69	129	15,83	30,94	1,80E-07	3,85E-06	2
Cellular metabolism	6	steroid metabolic process	GO:0008202	12	2	2,75	0,48	1,25E-02	3,89E-02	1
Cellular processes (not DNA/RNA)	5	establishment of RNA localization	GO:0051236	1	12	0,23	2,88	1,39E-03	4,81E-03	2
Cellular processes (not DNA/RNA)	5	ion transport	GO:0006811	29	7	6,65	1,68	2,64E-04	1,29E-03	1
Cellular processes (not DNA/RNA)	6	endocytosis	GO:0006897	29	5	6,65	1,20	3,16E-05	2,70E-04	1
Cellular processes (not DNA/RNA)	6	cation transport	GO:0006812	26	7	5,96	1,68	1,17E-03	5,80E-03	1
Development	5	cell fate commitment	GO:0045165	18	1	4,13	0,24	6,94E-05	3,67E-04	1
Development	5	regulation of cell differentiation	GO:0045595	45	6	10,32	1,44	1,04E-08	1,34E-07	1
Development	5	embryonic morphogenesis	GO:0048598	18	2	4,13	0,48	3,70E-04	1,62E-03	1
Development	5	skeletal system development	GO:0001501	31	2	7,11	0,48	9,75E-08	9,75E-07	1
Development	5	urogenital system development	GO:0001655	10	1	2,29	0,24	1,14E-02	3,07E-02	1
Development	5	immune system development	GO:0002520	26	9	5,96	2,16	5,42E-03	1,63E-02	1
Development	5	nervous system development	GO:0007399	73	15	16,74	3,60	7,91E-11	2,03E-09	1
Development	5	organ development	GO:0048513	110	21	25,23	5,04	2,84E-17	1,70E-15	1
Development	5	cellular component morphogenesis	GO:0032989	38	2	8,72	0,48	9,36E-10	1,68E-08	1
Development	5	anatomical structure formation involved in morphogenesis	GO:0048646	30	5	6,88	1,20	1,80E-05	1,16E-04	1
Development	5	positive regulation of developmental process	GO:0051094	55	13	12,61	3,12	1,99E-07	1,88E-06	1
Development	5	positive regulation of multicellular organismal process	GO:0051240	24	1	5,50	0,24	1,30E-06	1,01E-05	1
Development	5	negative regulation of developmental process	GO:0051093	55	9	12,61	2,16	2,42E-09	3,63E-08	1
Development	5	regulation of tissue remodeling	GO:0034103	13	1	2,98	0,24	1,75E-03	5,83E-03	1
Development	6	negative regulation of cell differentiation	GO:0045596	20	2	4,59	0,48	1,09E-04	8,95E-04	1
Development	6	positive regulation of cell differentiation	GO:0045597	28	3	6,42	0,72	3,75E-06	4,42E-05	1

Development	6	neurogenesis	GO:0022008	39	6	8,94	1,44	3,66E-07	7,12E-06	1
Development	6	central nervous system development	GO:0007417	29	8	6,65	1,92	6,46E-04	3,64E-03	1
Development	6	regulation of nervous system development	GO:0051960	13	1	2,98	0,24	1,75E-03	7,64E-03	1
Development	6	vasculature development	GO:0001944	21	3	4,82	0,72	2,49E-04	1,71E-03	1
Development	6	organ morphogenesis	GO:0009887	59	6	13,53	1,44	2,72E-12	1,94E-10	1
Development	6	tissue development	GO:0009888	52	5	11,93	1,20	2,75E-11	1,47E-09	1
Development	6	sensory organ development	GO:0007423	14	1	3,21	0,24	9,25E-04	4,71E-03	1
Development	6	heart development	GO:0007507	21	6	4,82	1,44	5,49E-03	1,99E-02	1
Development	6	muscle organ development	GO:0007517	17	2	3,90	0,48	6,77E-04	3,71E-03	1
Development	6	gland development	GO:0048732	15	1	3,44	0,24	4,87E-04	2,82E-03	1
Development	6	cell morphogenesis	GO:0000902	37	2	8,49	0,48	1,83E-09	6,53E-08	1
Development	6	cell part morphogenesis	GO:0032990	21	1	4,82	0,24	9,60E-06	9,78E-05	1
Development	6	regulation of bone remodeling	GO:0046850	13	1	2,98	0,24	1,75E-03	7,64E-03	1
DNA/RNA metabolism and transcription	5	nucleobase, nucleoside, nucleotide and nucleic acid transport	GO:0015931	1	12	0,23	2,88	1,39E-03	4,81E-03	2
DNA/RNA metabolism and transcription	6	DNA metabolic process	GO:0006259	7	76	1,61	18,23	4,54E-18	4,86E-16	2
DNA/RNA metabolism and transcription	6	ribonucleoprotein complex assembly	GO:0022618	1	11	0,23	2,64	2,68E-03	1,04E-02	2
DNA/RNA metabolism and transcription	6	nucleic acid transport	GO:0050657	1	12	0,23	2,88	1,39E-03	6,75E-03	2
Immune system response	5	inflammatory response	GO:0006954	26	1	5,96	0,24	3,37E-07	2,76E-06	1
Immune system response	5	positive regulation of immune system process	GO:0002684	13	2	2,98	0,48	7,11E-03	2,06E-02	1
Immune system response	6	T cell activation	GO:0042110	12	2	2,75	0,48	1,25E-02	3,89E-02	1
Immune system response	6	B cell activation	GO:0042113	13	2	2,98	0,48	7,11E-03	2,41E-02	1
Immune system response	6	regulation of leukocyte activation	GO:0002694	11	1	2,52	0,24	6,15E-03	2,16E-02	1
Multicellular activities	5	muscle contraction	GO:0006936	15	1	3,44	0,24	4,87E-04	1,99E-03	1
Multicellular activities	5	blood circulation	GO:0008015	16	2	3,67	0,48	1,23E-03	4,43E-03	1
Multicellular activities	5	transmission of nerve impulse	GO:0019226	28	5	6,42	1,20	5,52E-05	3,20E-04	1
Multicellular activities	5	cognition	GO:0050890	25	3	5,73	0,72	2,33E-05	1,40E-04	1
Multicellular activities	5	wound healing	GO:0042060	18	2	4,13	0,48	3,70E-04	1,62E-03	1
Multicellular activities	5	homeostasis of number of cells	GO:0048872	14	2	3,21	0,48	3,99E-03	1,24E-02	1
Multicellular activities	6	synaptic transmission	GO:0007268	25	5	5,73	1,20	2,84E-04	1,84E-03	1
Regulation of intracellular processes and metabolism	5	regulation of protein localization	GO:0032880	17	2	3,90	0,48	6,77E-04	2,65E-03	1
Regulation of intracellular processes and metabolism	5	regulation of transport	GO:0051049	37	3	8,49	0,72	1,29E-08	1,45E-07	1
Regulation of intracellular processes and metabolism	5	regulation of biosynthetic process	GO:0009889	145	99	33,26	23,74	2,41E-03	7,75E-03	1
Regulation of intracellular processes and metabolism	5	negative regulation of metabolic process	GO:0009892	54	23	12,39	5,52	4,82E-04	1,99E-03	1
Regulation of intracellular processes and metabolism	5	positive regulation of metabolic process	GO:0009893	72	23	16,51	5,52	2,17E-07	1,95E-06	1
Regulation of intracellular processes and metabolism	5	regulation of cellular metabolic process	GO:0031323	176	119	40,37	28,54	3,14E-04	1,49E-03	1
Regulation of intracellular processes and metabolism	5	regulation of nitrogen compound metabolic process	GO:0051171	141	94	32,34	22,54	1,64E-03	5,58E-03	1
Regulation of intracellular processes and metabolism	5	regulation of macromolecule metabolic process	GO:0060255	164	117	37,61	28,06	3,53E-03	1,11E-02	1
Regulation of intracellular processes and metabolism	5	positive regulation of cellular process	GO:0048522	125	40	28,67	9,59	9,21E-13	3,32E-11	1
Regulation of intracellular processes and metabolism	5	negative regulation of cellular process	GO:0048523	116	36	26,61	8,63	4,26E-12	1,28E-10	1
Regulation of intracellular processes and metabolism	5	negative regulation of cellular component organization	GO:0051129	15	3	3,44	0,72	7,19E-03	2,06E-02	1
Regulation of intracellular processes and metabolism	5	regulation of cell activation	GO:0050865	11	1	2,52	0,24	6,15E-03	1,82E-02	1
Regulation of intracellular processes and metabolism	5	cellular homeostasis	GO:0019725	32	6	7,34	1,44	1,92E-05	1,19E-04	1
Regulation of intracellular processes and metabolism	5	chemical homeostasis	GO:0048878	37	5	8,49	1,20	3,07E-07	2,64E-06	1
Regulation of intracellular processes and metabolism	5	negative regulation of catalytic activity	GO:0043086	11	26	2,52	6,24	1,08E-02	2,94E-02	2
Regulation of intracellular processes and metabolism	5	regulation of ligase activity	GO:0051340	1	22	0,23	5,28	1,51E-06	1,13E-05	2
Regulation of intracellular processes and metabolism	6	positive regulation of transport	GO:0051050	16	1	3,67	0,24	2,55E-04	1,71E-03	1
Regulation of intracellular processes and metabolism	6	negative regulation of transport	GO:0051051	17	1	3,90	0,24	1,33E-04	1,02E-03	1
Regulation of intracellular processes and metabolism	6	negative regulation of biosynthetic process	GO:0009890	38	18	8,72	4,32	1,22E-02	3,89E-02	1
Regulation of intracellular processes and metabolism	6	positive regulation of biosynthetic process	GO:0009891	58	17	13,30	4,08	1,52E-06	2,04E-05	1
Regulation of intracellular processes and metabolism	6	regulation of cellular biosynthetic process	GO:0031326	143	99	32,80	23,74	3,86E-03	1,47E-02	1
Regulation of intracellular processes and metabolism	6	negative regulation of macromolecule metabolic process	GO:0010605	53	23	12,16	5,52	7,02E-04	3,75E-03	1
Regulation of intracellular processes and metabolism	6	negative regulation of cellular metabolic process	GO:0031324	48	23	11,01	5,52	4,16E-03	1,54E-02	1
Regulation of intracellular processes and metabolism	6	positive regulation of macromolecule metabolic process	GO:0010604	68	22	15,60	5,28	6,62E-07	1,18E-05	1
Regulation of intracellular processes and metabolism	6	positive regulation of cellular metabolic process	GO:0031325	69	23	15,83	5,52	8,76E-07	1,44E-05	1
Regulation of intracellular processes and metabolism	6	positive regulation of nitrogen compound metabolic process	GO:0051173	55	18	12,61	4,32	1,34E-05	1,25E-04	1

Regulation of intracellular processes and metabolism	6	positive regulation of cell proliferation	GO:0008284	26	6	5,96	1,44	4,67E-04	2,82E-03	1
Regulation of intracellular processes and metabolism	6	negative regulation of cell proliferation	GO:0008285	33	6	7,57	1,44	1,11E-05	1,08E-04	1
Regulation of intracellular processes and metabolism	6	cellular chemical homeostasis	GO:0055082	29	4	6,65	0,96	8,80E-06	9,42E-05	1
Regulation of intracellular processes and metabolism	6	ion homeostasis	GO:0050801	29	5	6,65	1,20	3,16E-05	2,70E-04	1
Regulation of intracellular processes and metabolism	6	negative regulation of ligase activity	GO:0051352	1	21	0,23	5,04	3,04E-06	3,82E-05	2
Regulation of intracellular processes and metabolism	6	positive regulation of transferase activity	GO:0051347	16	3	3,67	0,72	4,19E-03	1,54E-02	1
Regulation of intracellular processes and metabolism	6	regulation of ubiquitin-protein ligase activity	GO:0051438	1	22	0,23	5,28	1,51E-06	2,04E-05	2
Regulation of transcription	6	regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	GO:0019219	139	94	31,88	22,54	2,68E-03	1,04E-02	1
Regulation of transcription	6	regulation of gene expression	GO:0010468	143	96	32,80	23,02	1,74E-03	7,64E-03	1
Signal transduction	5	signal transduction	GO:0007165	205	54	47,02	12,95	2,64E-28	4,75E-26	1
Signal transduction	6	cell surface receptor linked signaling pathway	GO:0007166	128	13	29,36	3,12	8,50E-28	1,82E-25	1
Signal transduction	6	intracellular signaling pathway	GO:0007242	92	31	21,10	7,43	9,46E-09	2,53E-07	1
Signal transduction	6	regulation of signal transduction	GO:0009966	56	9	12,84	2,16	1,34E-09	5,72E-08	1

The pipeline to determine the functional enrichment between two lists of genes is described in Figure 24. Fisher's exact test is calculated for each GO term and all GO terms at level 5 and 6 are grouped into 12 functional categories. *List 1* represents recent (*i.e.* originated with metazoans and vertebrates) duplicated hubs, while *list 2* represents ancestral (*i.e.* originated with the last universal common ancestor) singleton hubs.

In order to understand the causes of the functional differences between the two classes of human hubs, we conducted the same analysis on duplicated and singleton genes (Figure 40, Table 16) and on ancestral and recent genes (Figure 40, Table 17). The investigation of the significant GO terms and categories demonstrated that the biological processes that involve recent duplicated hubs are the same that are associated to young genes, while ancient singleton hubs are representative of the old genes, in terms of function. Similar functional differences between old and young genes had already been found also in yeast, where basic cellular processes, such as transcription and replication, are typical of old genes, while young genes are involved in genetic, transcriptional and posttranslational regulation (Kunin et al., 2004).

The signal between duplicated and singleton genes showed less statistical support and presented some interesting exceptions. The most evident was represented by the genes involved in the immune system response. Recent duplicated hubs are enriched in this category, and so are recent genes. However, the comparison of singleton and duplicated genes showed that singletons are enriched in the same terms. Cellular metabolism instead showed the opposite behavior, although with less statistical support: both ancient singleton hubs and ancient genes are involved in this process, but duplicated genes are enriched when compared with singletons.

These results demonstrate that the two classes of human hubs are involved in the biological processes that are typical of genes with their same origin and, at least in part with the same duplicability.

Table 16: functional comparison between duplicated genes and singleton genes

Process	GO level	GO description	GO ID	N genes (list 1)	N genes (list 2)	% of total genes (list 1)	% of total genes (list 2)	p-value	adjusted p-value	Enriched list
Cell cycle	5	sister chromatid segregation	GO:0000819	12	16	0,13	0,43	2,90E-03	1,92E-02	2
Cell cycle	5	cell cycle phase	GO:0022403	251	162	2,82	4,31	2,96E-05	3,41E-04	2
Cell cycle	5	apoptosis	GO:0006915	553	305	6,21	8,11	1,26E-04	1,23E-03	2
Cell cycle	5	cell development	GO:0048468	473	122	5,31	3,25	2,41E-07	3,94E-06	1
Cell cycle	5	regulation of cell cycle	GO:0051726	153	107	1,72	2,85	8,49E-05	8,83E-04	2
Cell cycle	6	M phase	GO:0000279	192	141	2,16	3,75	7,29E-07	2,23E-05	2
Cell cycle	6	cell cycle checkpoint	GO:0000075	37	38	0,42	1,01	1,82E-04	2,45E-03	2
Cell cycle	6	regulation of mitotic cell cycle	GO:0007346	71	62	0,80	1,65	3,52E-05	7,19E-04	2
Cell cycle	6	regulation of cell cycle process	GO:0010564	47	39	0,53	1,04	2,00E-03	1,72E-02	2
Cell motility and interactions	5	homophilic cell adhesion	GO:0007156	94	4	1,06	0,11	2,00E-10	5,61E-09	1
Cell motility and interactions	5	calcium-dependent cell-cell adhesion	GO:0016339	21	1	0,24	0,03	8,30E-03	4,34E-02	1
Cell motility and interactions	5	calcium-independent cell-cell adhesion	GO:0016338	1	21	0,01	0,56	1,28E-10	3,85E-09	2
Cell motility and interactions	5	cell-matrix adhesion	GO:0007160	81	16	0,91	0,43	3,55E-03	2,21E-02	1
Cell motility and interactions	5	cell-substrate junction assembly	GO:0007044	21	1	0,24	0,03	8,30E-03	4,34E-02	1
Cell response to stimuli	5	response to virus	GO:0009615	49	51	0,55	1,36	7,95E-06	1,01E-04	2
Cell response to stimuli	5	response to bacterium	GO:0009617	58	52	0,65	1,38	9,32E-05	9,37E-04	2
Cell response to stimuli	5	response to fungus	GO:0009620	5	11	0,06	0,29	1,41E-03	1,01E-02	2
Cell response to stimuli	6	sensory perception	GO:0007600	362	296	4,07	7,87	1,97E-17	1,82E-15	2
Cell response to stimuli	6	defense response to bacterium	GO:0042742	34	44	0,38	1,17	1,13E-06	3,27E-05	2
Cell response to stimuli	6	defense response to fungus	GO:0050832	2	8	0,02	0,21	1,47E-03	1,42E-02	2
Cell response to stimuli	6	response to gamma radiation	GO:0010332	3	9	0,03	0,24	1,55E-03	1,45E-02	2
Cellular metabolism	5	cellular nitrogen compound biosynthetic process	GO:0044271	77	79	0,86	2,10	4,33E-08	8,93E-07	2
Cellular metabolism	5	macromolecule catabolic process	GO:0043285	818	291	9,19	7,74	8,90E-03	4,59E-02	1
Cellular metabolism	5	aromatic compound catabolic process	GO:0019439	6	14	0,07	0,37	2,28E-04	2,08E-03	2
Cellular metabolism	5	macromolecule biosynthetic process	GO:0043284	2180	758	24,49	20,16	1,15E-07	2,04E-06	1
Cellular metabolism	5	heterocycle biosynthetic process	GO:0018130	16	28	0,18	0,74	3,33E-	4,51E-05	2

								06		
Cellular metabolism	5	aromatic compound biosynthetic process	GO:0019438	6	11	0,07	0,29	2,93E-03	1,92E-02	2
Cellular metabolism	5	cofactor biosynthetic process	GO:0051188	33	52	0,37	1,38	2,33E-09	5,71E-08	2
Cellular metabolism	5	macromolecule modification	GO:0043412	1151	348	12,93	9,26	2,84E-09	6,54E-08	1
Cellular metabolism	5	collagen metabolic process	GO:0032963	32	3	0,36	0,08	4,70E-03	2,71E-02	1
Cellular metabolism	5	cytokine metabolic process	GO:0042107	43	35	0,48	0,93	5,79E-03	3,20E-02	2
Cellular metabolism	5	lipoprotein metabolic process	GO:0042157	23	54	0,26	1,44	2,47E-13	9,68E-12	2
Cellular metabolism	5	energy derivation by oxidation of organic compounds	GO:0015980	77	68	0,86	1,81	1,37E-05	1,68E-04	2
Cellular metabolism	5	phosphate metabolic process	GO:0006796	886	234	9,95	6,23	3,91E-12	1,40E-10	1
Cellular metabolism	5	tetrapyrrole metabolic process	GO:0033013	10	15	0,11	0,40	1,71E-03	1,18E-02	2
Cellular metabolism	5	coenzyme metabolic process	GO:0006732	64	68	0,72	1,81	1,42E-07	2,42E-06	2
Cellular metabolism	6	tetrapyrrole biosynthetic process	GO:0033014	5	14	0,06	0,37	9,44E-05	1,41E-03	2
Cellular metabolism	6	pteridine and derivative biosynthetic process	GO:0042559	4	9	0,04	0,24	3,71E-03	2,81E-02	2
Cellular metabolism	6	protein catabolic process	GO:0030163	756	218	8,49	5,80	1,14E-07	4,19E-06	1
Cellular metabolism	6	lipoprotein biosynthetic process	GO:0042158	14	42	0,16	1,12	3,04E-12	1,86E-10	2
Cellular metabolism	6	coenzyme biosynthetic process	GO:0009108	26	34	0,29	0,90	1,37E-05	3,02E-04	2
Cellular metabolism	6	aminoglycan metabolic process	GO:0006022	36	31	0,40	0,82	4,49E-03	3,22E-02	2
Cellular metabolism	6	cellular protein metabolic process	GO:0044267	1732	654	19,46	17,40	6,71E-03	4,41E-02	1
Cellular metabolism	6	carboxylic acid metabolic process	GO:0019752	334	180	3,75	4,79	7,75E-03	4,86E-02	2
Cellular metabolism	6	phosphorylation	GO:0016310	744	216	8,36	5,75	2,53E-07	8,21E-06	1
Cellular metabolism	6	dephosphorylation	GO:0016311	140	17	1,57	0,45	1,82E-08	8,36E-07	1
Cellular metabolism	6	porphyrin metabolic process	GO:0006778	10	15	0,11	0,40	1,71E-03	1,56E-02	2
Cellular metabolism	6	oxidoreduction coenzyme metabolic process	GO:0006733	21	22	0,24	0,59	3,82E-03	2,85E-02	2
Cellular metabolism	6	glycerolipid metabolic process	GO:0046486	77	59	0,86	1,57	6,48E-04	7,30E-03	2
Cellular metabolism	6	immunoglobulin production	GO:0002377	14	20	0,16	0,53	4,78E-04	5,69E-03	2
Cellular processes (not DNA/RNA)	5	electron transport chain	GO:0022900	43	62	0,48	1,65	4,62E-10	1,21E-08	2
Cellular processes (not DNA/RNA)	5	cellular protein localization	GO:0034613	271	154	3,04	4,10	3,49E-03	2,21E-02	2
Cellular processes (not DNA/RNA)	5	vesicle-mediated transport	GO:0016192	402	122	4,52	3,25	8,89E-04	7,11E-03	1
Cellular processes (not DNA/RNA)	5	ion transport	GO:0006811	624	114	7,01	3,03	2,93E-20	1,91E-18	1
Cellular processes (not DNA/RNA)	5	neurotransmitter transport	GO:0006836	67	10	0,75	0,27	9,83E-04	7,56E-03	1
Cellular processes (not DNA/RNA)	5	hormone transport	GO:0009914	33	30	0,37	0,80	3,28E-03	2,11E-02	2

Cellular processes (not DNA/RNA)	5	organic acid transport	GO:0015849	108	19	1,21	0,51	1,79E-04	1,67E-03	1
Cellular processes (not DNA/RNA)	6	peroxisomal transport	GO:0043574	2	12	0,02	0,32	2,23E-05	4,73E-04	2
Cellular processes (not DNA/RNA)	6	endocytosis	GO:0006897	185	40	2,08	1,06	4,65E-05	8,33E-04	1
Cellular processes (not DNA/RNA)	6	cation transport	GO:0006812	447	93	5,02	2,47	1,18E-11	6,50E-10	1
Cellular processes (not DNA/RNA)	6	anion transport	GO:0006820	110	15	1,24	0,40	4,29E-06	1,08E-04	1
Cellular processes (not DNA/RNA)	6	carboxylic acid transport	GO:0046942	107	19	1,20	0,51	1,74E-04	2,40E-03	1
Development	5	cell fate commitment	GO:0045165	95	8	1,07	0,21	9,34E-08	1,74E-06	1
Development	5	regulation of cell differentiation	GO:0045595	262	74	2,94	1,97	1,62E-03	1,13E-02	1
Development	5	embryonic development ending in birth or egg hatching	GO:0009792	173	46	1,94	1,22	4,46E-03	2,61E-02	1
Development	5	embryonic morphogenesis	GO:0048598	178	28	2,00	0,74	7,68E-08	1,50E-06	1
Development	5	regionalization	GO:0003002	127	18	1,43	0,48	1,39E-06	2,02E-05	1
Development	5	skeletal system development	GO:0001501	224	56	2,52	1,49	2,60E-04	2,27E-03	1
Development	5	nervous system development	GO:0007399	660	147	7,41	3,91	1,97E-14	8,59E-13	1
Development	5	organ development	GO:0048513	1052	360	11,82	9,58	2,34E-04	2,08E-03	1
Development	5	morphogenesis of a branching structure	GO:0001763	42	6	0,47	0,16	6,92E-03	3,77E-02	1
Development	5	cellular component morphogenesis	GO:0032989	258	64	2,90	1,70	7,15E-05	7,99E-04	1
Development	6	regulation of smooth muscle cell proliferation	GO:0048660	11	15	0,12	0,40	3,94E-03	2,90E-02	2
Development	6	cell fate specification	GO:0001708	37	1	0,42	0,03	4,50E-05	8,33E-04	1
Development	6	cell fate determination	GO:0001709	26	2	0,29	0,05	6,38E-03	4,24E-02	1
Development	6	negative regulation of cell differentiation	GO:0045596	120	23	1,35	0,61	2,09E-04	2,75E-03	1
Development	6	neurogenesis	GO:0022008	324	65	3,64	1,73	2,51E-09	1,26E-07	1
Development	6	chordate embryonic development	GO:0043009	172	45	1,93	1,20	3,37E-03	2,66E-02	1
Development	6	anterior/posterior pattern formation	GO:0009952	87	7	0,98	0,19	1,74E-07	5,99E-06	1
Development	6	somatic diversification of immune receptors	GO:0002200	10	16	0,11	0,43	8,45E-04	8,97E-03	2
Development	6	synapse assembly	GO:0007416	26	2	0,29	0,05	6,38E-03	4,24E-02	1
Development	6	central nervous system development	GO:0007417	246	60	2,76	1,60	6,18E-05	1,03E-03	1
Development	6	regulation of nervous system development	GO:0051960	97	21	1,09	0,56	4,35E-03	3,16E-02	1
Development	6	organ morphogenesis	GO:0009887	464	124	5,21	3,30	1,79E-06	4,94E-05	1
Development	6	tissue development	GO:0009888	418	125	4,70	3,33	4,47E-04	5,61E-03	1
Development	6	sensory organ development	GO:0007423	138	26	1,55	0,69	4,68E-05	8,33E-04	1
Development	6	muscle organ development	GO:0007517	161	40	1,81	1,06	1,80E-03	1,58E-02	1

Development	6	respiratory tube development	GO:0030323	57	10	0,64	0,27	6,97E-03	4,49E-02	1
Development	6	embryonic organ development	GO:0048568	32	2	0,36	0,05	1,10E-03	1,12E-02	1
Development	6	cell morphogenesis	GO:0000902	241	52	2,71	1,38	2,78E-06	7,30E-05	1
Development	6	cell part morphogenesis	GO:0032990	146	28	1,64	0,74	3,77E-05	7,43E-04	1
DNA/RNA metabolism and transcription	5	chromatin organization	GO:0006325	279	73	3,13	1,94	1,46E-04	1,39E-03	1
DNA/RNA metabolism and transcription	5	chromosome condensation	GO:0030261	4	18	0,04	0,48	6,14E-07	9,63E-06	2
DNA/RNA metabolism and transcription	6	DNA metabolic process	GO:0006259	249	249	2,80	6,62	3,94E-22	4,35E-20	2
DNA/RNA metabolism and transcription	6	RNA metabolic process	GO:0016070	1634	597	18,36	15,88	8,23E-04	8,92E-03	1
DNA/RNA metabolism and transcription	6	ribonucleoprotein complex assembly	GO:0022618	22	24	0,25	0,64	1,73E-03	1,56E-02	2
DNA/RNA metabolism and transcription	6	protein-DNA complex assembly	GO:0065004	72	9	0,81	0,24	1,20E-04	1,74E-03	1
DNA/RNA metabolism and transcription	6	regulation of helicase activity	GO:0051095	1	6	0,01	0,16	3,57E-03	2,77E-02	2
Immune system response	5	production of molecular mediator of immune response	GO:0002440	21	24	0,24	0,64	9,18E-04	7,19E-03	2
Immune system response	5	inflammatory response	GO:0006954	185	118	2,08	3,14	5,64E-04	4,62E-03	2
Immune system response	5	positive regulation of immune system process	GO:0002684	103	68	1,16	1,81	5,24E-03	2,98E-02	2
Immune system response	6	natural killer cell activation	GO:0030101	7	14	0,08	0,37	4,98E-04	5,73E-03	2
Immune system response	6	T cell activation	GO:0042110	79	71	0,89	1,89	5,37E-06	1,29E-04	2
Immune system response	6	B cell activation	GO:0042113	47	40	0,53	1,06	1,35E-03	1,33E-02	2
Immune system response	6	lymphocyte proliferation	GO:0046651	37	39	0,42	1,04	7,60E-05	1,20E-03	2
Immune system response	6	regulation of leukocyte activation	GO:0002694	64	60	0,72	1,60	1,11E-05	2,55E-04	2
Multicellular activities	5	smooth muscle cell proliferation	GO:0048659	12	16	0,13	0,43	2,90E-03	1,92E-02	2
Multicellular activities	5	cognition	GO:0050890	406	311	4,56	8,27	1,55E-15	7,59E-14	2
Multicellular activities	6	striated muscle contraction	GO:0006941	50	6	0,56	0,16	1,09E-03	1,12E-02	1
Regulation of intracellular processes and metabolism	5	regulation of biosynthetic process	GO:0009889	2069	544	23,24	14,47	3,82E-30	3,75E-28	1
Regulation of intracellular processes and metabolism	5	regulation of cellular metabolic process	GO:0031323	2383	654	26,77	17,40	1,12E-30	1,47E-28	1
Regulation of intracellular processes and metabolism	5	regulation of nitrogen compound metabolic process	GO:0051171	2000	500	22,47	13,30	4,49E-34	1,76E-31	1
Regulation of intracellular processes and metabolism	5	regulation of macromolecule metabolic process	GO:0060255	2260	642	25,39	17,08	4,33E-25	3,40E-23	1
Regulation of intracellular processes and metabolism	5	regulation of cell activation	GO:0050865	67	63	0,75	1,68	6,85E-06	8,94E-05	2
Regulation of intracellular processes and metabolism	5	negative regulation of catalytic activity	GO:0043086	135	84	1,52	2,23	5,68E-03	3,18E-02	2
Regulation of intracellular processes and metabolism	5	regulation of lyase activity	GO:0051339	63	8	0,71	0,21	3,57E-04	3,04E-03	1
Regulation of intracellular processes and metabolism	5	regulation of ligase activity	GO:0051340	39	37	0,44	0,98	5,66E-04	4,62E-03	2
Regulation of intracellular processes and metabolism	6	regulation of protein complex assembly	GO:0043254	54	8	0,61	0,21	3,07E-03	2,53E-02	1

Regulation of intracellular processes and metabolism	6	regulation of cellular biosynthetic process	GO:0031326	2061	542	23,15	14,42	5,53E-30	1,52E-27	1
Regulation of intracellular processes and metabolism	6	positive regulation of cell proliferation	GO:0008284	190	112	2,13	2,98	5,02E-03	3,51E-02	2
Regulation of intracellular processes and metabolism	6	positive regulation of cell activation	GO:0050867	41	42	0,46	1,12	8,61E-05	1,32E-03	2
Regulation of intracellular processes and metabolism	6	regulation of endothelial cell proliferation	GO:0001936	12	15	0,13	0,40	5,32E-03	3,63E-02	2
Regulation of intracellular processes and metabolism	6	regulation of fibroblast proliferation	GO:0048145	12	15	0,13	0,40	5,32E-03	3,63E-02	2
Regulation of intracellular processes and metabolism	6	negative regulation of homeostatic process	GO:0032845	5	10	0,06	0,27	3,34E-03	2,66E-02	2
Regulation of intracellular processes and metabolism	6	positive regulation of homeostatic process	GO:0032846	8	13	0,09	0,35	2,79E-03	2,33E-02	2
Regulation of intracellular processes and metabolism	6	negative regulation of lyase activity	GO:0051350	29	1	0,33	0,03	4,85E-04	5,69E-03	1
Regulation of intracellular processes and metabolism	6	negative regulation of ligase activity	GO:0051352	29	34	0,33	0,90	7,31E-05	1,19E-03	2
Regulation of intracellular processes and metabolism	6	positive regulation of ligase activity	GO:0051351	33	35	0,37	0,93	1,62E-04	2,29E-03	2
Regulation of intracellular processes and metabolism	6	regulation of lipase activity	GO:0060191	63	11	0,71	0,29	4,62E-03	3,27E-02	1
Regulation of intracellular processes and metabolism	6	regulation of ubiquitin-protein ligase activity	GO:0051438	39	35	0,44	0,93	1,34E-03	1,33E-02	2
Regulation of transcription	6	regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	GO:0019219	1984	495	22,29	13,17	5,21E-34	2,88E-31	1
Regulation of transcription	6	regulation of gene expression	GO:0010468	2012	530	22,60	14,10	4,94E-29	9,09E-27	1
Signal transduction	6	intracellular signaling pathway	GO:0007242	1101	303	12,37	8,06	5,70E-13	4,50E-11	1

The functional analysis is made between duplicated (*list 1*) and singleton genes (*list 2*), as explained in Table 15.

Table 17: functional comparison between recent genes and ancestral genes

Process	GO level	GO description	GO ID	N genes (list 1)	N genes (list 2)	% of total genes (list 1)	% of total genes (list 2)	p-value	adjusted p-value	Enriched list
Development	5	sex differentiation	GO:0007548	42	51	1,21	0,63	2,94E-03	8,16E-03	1
Development	5	reproductive structure development	GO:0048608	38	39	1,09	0,49	4,29E-04	1,34E-03	1
Multicellular activities	5	ovulation cycle	GO:0042698	20	14	0,57	0,17	5,70E-04	1,77E-03	1
DNA/RNA metabolism and transcription	5	nucleobase, nucleoside and nucleotide metabolic process	GO:0055086	77	271	2,21	3,37	7,10E-04	2,15E-03	2
Cellular metabolism	5	cellular nitrogen compound catabolic process	GO:0044270	5	76	0,14	0,95	1,46E-07	8,40E-07	2
Cellular metabolism	5	cellular nitrogen compound biosynthetic process	GO:0044271	22	130	0,63	1,62	7,25E-06	3,06E-05	2
Cellular metabolism	5	macromolecule catabolic process	GO:0043285	165	897	4,74	11,16	7,55E-31	3,75E-29	2
Cellular metabolism	5	cellular macromolecule catabolic process	GO:0044265	107	547	3,07	6,80	6,35E-17	8,69E-16	2

Cellular metabolism	5	organic acid catabolic process	GO:0016054	10	90	0,29	1,12	1,72E-06	8,14E-06	2
Cellular metabolism	5	cofactor catabolic process	GO:0051187	1	27	0,03	0,34	7,47E-04	2,23E-03	2
Cellular metabolism	5	vitamin biosynthetic process	GO:0009110	2	24	0,06	0,30	9,55E-03	2,31E-02	2
Cellular metabolism	5	organic acid biosynthetic process	GO:0016053	25	122	0,72	1,52	2,78E-04	8,90E-04	2
Cellular metabolism	5	heterocycle biosynthetic process	GO:0018130	2	42	0,06	0,52	4,38E-05	1,69E-04	2
Cellular metabolism	5	cofactor biosynthetic process	GO:0051188	1	84	0,03	1,04	2,22E-12	2,01E-11	2
Cellular metabolism	5	macromolecule modification	GO:0043412	233	121 6	6,69	15,12	1,53E-39	1,01E-37	2
Cellular metabolism	5	cytokine metabolic process	GO:0042107	46	24	1,32	0,30	1,20E-09	8,83E-09	1
Cellular metabolism	5	cellular amino acid derivative metabolic process	GO:0006575	28	105	0,80	1,31	2,23E-02	4,85E-02	2
Cellular metabolism	5	pteridine and derivative metabolic process	GO:0042558	1	20	0,03	0,25	7,85E-03	1,91E-02	2
Cellular metabolism	5	water-soluble vitamin metabolic process	GO:0006767	3	51	0,09	0,63	1,69E-05	7,00E-05	2
Cellular metabolism	5	phosphate metabolic process	GO:0006796	201	868	5,77	10,79	7,87E-19	1,56E-17	2
Cellular metabolism	5	hydrogen peroxide metabolic process	GO:0042743	1	19	0,03	0,24	1,27E-02	2,89E-02	2
Cellular metabolism	5	tetrapyrrole metabolic process	GO:0033013	2	23	0,06	0,29	1,48E-02	3,31E-02	2
Cellular metabolism	5	coenzyme metabolic process	GO:0006732	1	130	0,03	1,62	2,26E-19	4,72E-18	2
Cellular metabolism	5	monosaccharide metabolic process	GO:0005996	36	168	1,03	2,09	4,19E-05	1,63E-04	2
Cellular metabolism	5	carbohydrate catabolic process	GO:0016052	14	93	0,40	1,16	4,68E-05	1,75E-04	2
Cellular metabolism	5	cellular carbohydrate metabolic process	GO:0044262	41	196	1,18	2,44	5,76E-06	2,51E-05	2
Cellular metabolism	5	cellular lipid metabolic process	GO:0044255	134	480	3,85	5,97	1,98E-06	9,25E-06	2
Immune system response	5	adaptive immune response based on somatic recombination of immune receptors built from immunoglobulin superfamily domains	GO:0002460	48	29	1,38	0,36	5,68E-09	3,88E-08	1
Immune system response	5	leukocyte mediated cytotoxicity	GO:0001909	10	5	0,29	0,06	3,85E-03	1,05E-02	1
Immune system response	5	production of molecular mediator of immune response	GO:0002440	21	17	0,60	0,21	1,29E-03	3,81E-03	1
Immune system response	5	leukocyte mediated immunity	GO:0002443	44	35	1,26	0,44	3,00E-06	1,37E-05	1
Immune system response	5	antigen processing and presentation of peptide antigen via MHC class I	GO:0002474	7	4	0,20	0,05	2,25E-02	4,85E-02	1
Immune system response	5	myeloid leukocyte activation	GO:0002274	20	19	0,57	0,24	7,70E-03	1,89E-02	1
Cell motility and interactions	5	heterophilic cell-cell adhesion	GO:0007157	9	6	0,26	0,07	2,05E-02	4,49E-02	1
Cell motility and interactions	5	homophilic cell adhesion	GO:0007156	72	9	2,07	0,11	2,58E-28	1,14E-26	1
Cell motility and interactions	5	neuron cell-cell adhesion	GO:0007158	6	1	0,17	0,01	3,94E-03	1,05E-02	1
Cell motility and interactions	5	calcium-dependent cell-cell adhesion	GO:0016339	10	2	0,29	0,02	2,19E-04	7,31E-04	1
Cell motility and interactions	5	cell-matrix adhesion	GO:0007160	44	44	1,26	0,55	1,00E-04	3,54E-04	1

DNA/RNA metabolism and transcription	5	chromatin organization	GO:0006325	36	303	1,03	3,77	3,50E-18	6,03E-17	2
Multicellular activities	5	pigment granule organization	GO:0048753	7	2	0,20	0,02	4,48E-03	1,16E-02	1
Cell cycle	5	cell cycle arrest	GO:0007050	38	51	1,09	0,63	1,44E-02	3,26E-02	1
Signal transduction	5	generation of a signal involved in cell-cell signaling	GO:0003001	61	30	1,75	0,37	6,19E-13	5,86E-12	1
Cell cycle	5	apoptosis	GO:0006915	348	437	9,99	5,43	5,52E-18	8,98E-17	1
Multicellular activities	5	striated muscle cell proliferation	GO:0014855	9	3	0,26	0,04	1,78E-03	5,05E-03	1
Multicellular activities	5	smooth muscle cell proliferation	GO:0048659	16	11	0,46	0,14	2,38E-03	6,71E-03	1
Development	5	muscle cell differentiation	GO:0042692	32	40	0,92	0,50	1,00E-02	2,40E-02	1
Development	5	cell fate commitment	GO:0045165	67	34	1,92	0,42	6,83E-14	7,32E-13	1
Development	5	fat cell differentiation	GO:0045444	20	22	0,57	0,27	1,79E-02	3,98E-02	1
Development	5	regulation of cell differentiation	GO:0045595	170	150	4,88	1,87	5,65E-18	8,98E-17	1
Cell cycle	5	cell development	GO:0048468	292	272	8,39	3,38	8,41E-28	3,34E-26	1
Cell cycle	5	stem cell differentiation	GO:0048863	11	8	0,32	0,10	1,22E-02	2,80E-02	1
Development	5	pigment cell differentiation	GO:0050931	10	3	0,29	0,04	6,90E-04	2,11E-03	1
Cellular processes (not DNA/RNA)	5	intracellular transport	GO:0046907	119	482	3,42	5,99	3,44E-09	2,44E-08	2
Cellular processes (not DNA/RNA)	5	cellular protein localization	GO:0034613	101	311	2,90	3,87	1,02E-02	2,42E-02	2
Multicellular activities	5	muscle contraction	GO:0006936	68	94	1,95	1,17	1,39E-03	4,05E-03	1
Multicellular activities	5	muscle adaptation	GO:0043500	8	4	0,23	0,05	9,92E-03	2,39E-02	1
Multicellular activities	5	vascular process in circulatory system	GO:0003018	32	24	0,92	0,30	2,93E-05	1,16E-04	1
Multicellular activities	5	blood circulation	GO:0008015	79	99	2,27	1,23	6,90E-05	2,54E-04	1
Multicellular activities	5	transmission of nerve impulse	GO:0019226	163	132	4,68	1,64	1,71E-19	3,78E-18	1
Multicellular activities	5	cognition	GO:0050890	224	192	6,43	2,39	2,28E-24	7,53E-23	1
Multicellular activities	5	neuromuscular process	GO:0050905	26	21	0,75	0,26	3,58E-04	1,14E-03	1
Development	5	embryonic pattern specification	GO:0009880	20	5	0,57	0,06	3,82E-07	2,11E-06	1
Development	5	embryonic development ending in birth or egg hatching	GO:0009792	110	106	3,16	1,32	1,64E-10	1,33E-09	1
Development	5	embryonic morphogenesis	GO:0048598	121	80	3,48	0,99	8,71E-19	1,65E-17	1
Development	5	regionalization	GO:0003002	101	42	2,90	0,52	1,56E-23	4,41E-22	1
Development	5	tube morphogenesis	GO:0035239	38	28	1,09	0,35	4,80E-06	2,12E-05	1
Development	5	skeletal system development	GO:0001501	149	127	4,28	1,58	1,17E-16	1,50E-15	1
Development	5	urogenital system development	GO:0001655	40	35	1,15	0,44	3,71E-05	1,46E-04	1

Development	5	immune system development	GO:0002520	122	140	3,50	1,74	2,51E-08	1,56E-07	1
Development	5	nervous system development	GO:0007399	388	366	11,14	4,55	2,71E-36	1,54E-34	1
Development	5	endocrine system development	GO:0035270	37	14	1,06	0,17	5,33E-10	4,12E-09	1
Development	5	organ development	GO:0048513	640	674	18,38	8,38	1,52E-50	2,01E-48	1
Development	5	appendage morphogenesis	GO:0035107	33	29	0,95	0,36	2,32E-04	7,66E-04	1
Development	5	limb development	GO:0060173	34	30	0,98	0,37	1,67E-04	5,73E-04	1
Multicellular activities	5	hair cycle process	GO:0022405	15	12	0,43	0,15	6,06E-03	1,52E-02	1
Development	5	morphogenesis of a branching structure	GO:0001763	27	21	0,78	0,26	2,03E-04	6,89E-04	1
Development	5	establishment of tissue polarity	GO:0007164	6	1	0,17	0,01	3,94E-03	1,05E-02	1
Development	5	regulation of anatomical structure morphogenesis	GO:0022603	85	86	2,44	1,07	8,34E-08	4,94E-07	1
Development	5	cellular component morphogenesis	GO:0032989	151	150	4,34	1,87	2,78E-13	2,76E-12	1
Development	5	anatomical structure formation involved in morphogenesis	GO:0048646	127	137	3,65	1,70	9,73E-10	7,29E-09	1
Cell response to stimuli	5	behavioral defense response	GO:0002209	8	2	0,23	0,02	1,67E-03	4,77E-03	1
Immune system response	5	inflammatory response	GO:0006954	144	97	4,14	1,21	1,13E-21	3,00E-20	1
Immune system response	5	innate immune response	GO:0045087	56	63	1,61	0,78	1,17E-04	4,10E-04	1
Multicellular activities	5	wound healing	GO:0042060	73	75	2,10	0,93	1,26E-06	6,24E-06	1
Cell response to stimuli	5	fear response	GO:0042596	11	2	0,32	0,02	7,76E-05	2,82E-04	1
Multicellular activities	5	adult locomotory behavior	GO:0008344	17	15	0,49	0,19	6,69E-03	1,65E-02	1
Cell motility and interactions	5	taxis	GO:0042330	74	30	2,13	0,37	6,94E-18	1,06E-16	1
Multicellular activities	5	eating behavior	GO:0042755	6	1	0,17	0,01	3,94E-03	1,05E-02	1
Immune system response	5	immune response to tumor cell	GO:0002418	5	1	0,14	0,01	1,13E-02	2,61E-02	1
Cell response to stimuli	5	response to virus	GO:0009615	36	43	1,03	0,53	4,37E-03	1,14E-02	1
Cell response to stimuli	5	response to organic cyclic substance	GO:0014070	20	9	0,57	0,11	1,88E-05	7,71E-05	1
Cell response to stimuli	5	response to ATP	GO:0033198	7	1	0,20	0,01	1,35E-03	3,97E-03	1
Cell response to stimuli	5	response to alkaloid	GO:0043279	17	5	0,49	0,06	7,05E-06	3,01E-05	1
Cell response to stimuli	5	response to ethanol	GO:0045471	15	7	0,43	0,09	2,65E-04	8,62E-04	1
Regulation of intracellular processes and metabolism	5	regulation of protein localization	GO:0032880	47	47	1,35	0,58	6,24E-05	2,32E-04	1
Cellular processes (not DNA/RNA)	5	establishment of protein localization	GO:0045184	164	572	4,71	7,11	7,49E-07	3,97E-06	2
Cellular processes (not DNA/RNA)	5	establishment of RNA localization	GO:0051236	7	76	0,20	0,95	2,95E-06	1,36E-05	2
Cellular processes (not DNA/RNA)	5	lipid storage	GO:0019915	21	13	0,60	0,16	2,12E-04	7,12E-04	1

Cellular processes (not DNA/RNA)	5	hydrogen transport	GO:0006818	8	53	0,23	0,66	2,96E-03	8,16E-03	2
Cellular processes (not DNA/RNA)	5	lipid transport	GO:0006869	55	81	1,58	1,01	1,10E-02	2,57E-02	1
Cellular processes (not DNA/RNA)	5	hormone transport	GO:0009914	38	21	1,09	0,26	7,28E-08	4,38E-07	1
Cellular processes (not DNA/RNA)	5	peptide transport	GO:0015833	27	29	0,78	0,36	5,10E-03	1,32E-02	1
Cellular processes (not DNA/RNA)	5	vitamin transport	GO:0051180	10	7	0,29	0,09	1,55E-02	3,46E-02	1
DNA/RNA metabolism and transcription	5	nucleobase, nucleoside, nucleotide and nucleic acid transport	GO:0015931	8	90	0,23	1,12	1,86E-07	1,06E-06	2
Regulation of intracellular processes and metabolism	5	regulation of transport	GO:0051049	150	134	4,31	1,67	1,10E-15	1,32E-14	1
Cellular processes (not DNA/RNA)	5	transmembrane transport	GO:0055085	35	191	1,01	2,38	3,42E-07	1,91E-06	2
Cell motility and interactions	5	regulation of cellular component movement	GO:0051270	71	60	2,04	0,75	1,32E-08	8,56E-08	1
Immune system response	5	negative regulation of immune system process	GO:0002683	23	17	0,66	0,21	3,92E-04	1,24E-03	1
Immune system response	5	positive regulation of immune system process	GO:0002684	86	57	2,47	0,71	1,16E-13	1,21E-12	1
Immune system response	5	regulation of immune response	GO:0050776	69	57	1,98	0,71	1,19E-08	8,00E-08	1
Regulation of intracellular processes and metabolism	5	regulation of biosynthetic process	GO:0009889	851	163 4	24,44	20,32	1,03E-06	5,23E-06	1
Regulation of intracellular processes and metabolism	5	negative regulation of metabolic process	GO:0009892	234	334	6,72	4,15	1,24E-08	8,21E-08	1
Regulation of intracellular processes and metabolism	5	positive regulation of metabolic process	GO:0009893	294	355	8,44	4,41	9,99E-17	1,32E-15	1
Regulation of intracellular processes and metabolism	5	regulation of cellular metabolic process	GO:0031323	979	190 2	28,12	23,65	4,69E-07	2,52E-06	1
Regulation of intracellular processes and metabolism	5	regulation of multicellular organismal metabolic process	GO:0044246	7	4	0,20	0,05	2,25E-02	4,85E-02	1
Regulation of intracellular processes and metabolism	5	regulation of nitrogen compound metabolic process	GO:0051171	808	157 5	23,21	19,59	1,30E-05	5,44E-05	1
Regulation of intracellular processes and metabolism	5	regulation of macromolecule metabolic process	GO:0060255	894	186 1	25,67	23,14	3,71E-03	1,02E-02	1
Cell response to stimuli	5	positive regulation of cell killing	GO:0031343	9	2	0,26	0,02	6,10E-04	1,88E-03	1
Development	5	regulation of multicellular organism growth	GO:0040014	26	14	0,75	0,17	6,23E-06	2,69E-05	1
Cell cycle	5	negative regulation of growth	GO:0045926	38	35	1,09	0,44	9,87E-05	3,53E-04	1
Cell cycle	5	positive regulation of growth	GO:0045927	22	21	0,63	0,26	4,23E-03	1,11E-02	1
Regulation of intracellular processes and metabolism	5	positive regulation of cellular process	GO:0048522	605	673	17,38	8,37	1,71E-42	1,66E-40	1
Cell response to stimuli	5	positive regulation of response to stimulus	GO:0048584	83	69	2,38	0,86	3,44E-10	2,73E-09	1
Development	5	positive regulation of developmental process	GO:0051094	227	219	6,52	2,72	2,04E-20	5,06E-19	1
Regulation of intracellular processes and metabolism	5	positive regulation of cellular component organization	GO:0051130	67	56	1,92	0,70	2,64E-08	1,61E-07	1
Development	5	positive regulation of multicellular organismal process	GO:0051240	82	52	2,35	0,65	1,37E-13	1,40E-12	1
Regulation of intracellular processes and metabolism	5	negative regulation of cellular process	GO:0048523	512	669	14,70	8,32	7,47E-24	2,28E-22	1
Development	5	negative regulation of developmental process	GO:0051093	188	221	5,40	2,75	9,81E-12	8,66E-11	1

Regulation of intracellular processes and metabolism	5	negative regulation of cellular component organization	GO:0051129	44	58	1,26	0,72	6,44E-03	1,60E-02	1
Development	5	negative regulation of multicellular organismal process	GO:0051241	54	47	1,55	0,58	1,07E-06	5,40E-06	1
Immune system response	5	regulation of response to external stimulus	GO:0032101	54	46	1,55	0,57	8,46E-07	4,36E-06	1
Multicellular activities	5	regulation of behavior	GO:0050795	21	6	0,60	0,07	4,55E-07	2,48E-06	1
Signal transduction	5	signal transduction	GO:0007165	1342	143 4	38,54	17,83	1,21E-119	4,81E-117	1
Cell motility and interactions	5	regulation of cell adhesion	GO:0030155	45	41	1,29	0,51	2,58E-05	1,04E-04	1
Cell cycle	5	regulation of cell proliferation	GO:0042127	279	263	8,01	3,27	5,37E-26	1,94E-24	1
Regulation of intracellular processes and metabolism	5	regulation of cell activation	GO:0050865	71	37	2,04	0,46	3,47E-14	3,82E-13	1
Regulation of intracellular processes and metabolism	5	regulation of cytokine production	GO:0001817	77	47	2,21	0,58	2,98E-13	2,88E-12	1
Development	5	regulation of tissue remodeling	GO:0034103	35	22	1,01	0,27	1,42E-06	6,81E-06	1
Cell cycle	5	cell growth	GO:0016049	78	87	2,24	1,08	4,67E-06	2,08E-05	1
Regulation of intracellular processes and metabolism	5	negative regulation of cell size	GO:0045792	33	30	0,95	0,37	2,70E-04	8,72E-04	1
Regulation of intracellular processes and metabolism	5	temperature homeostasis	GO:0001659	9	1	0,26	0,01	1,52E-04	5,29E-04	1
Regulation of intracellular processes and metabolism	5	cellular homeostasis	GO:0019725	160	178	4,60	2,21	2,24E-11	1,93E-10	1
Regulation of intracellular processes and metabolism	5	regulation of homeostatic process	GO:0032844	44	26	1,26	0,32	2,08E-08	1,31E-07	1
Multicellular activities	5	multicellular organismal homeostasis	GO:0048871	18	19	0,52	0,24	1,91E-02	4,20E-02	1
Multicellular activities	5	homeostasis of number of cells	GO:0048872	35	46	1,01	0,57	1,46E-02	3,29E-02	1
Regulation of intracellular processes and metabolism	5	chemical homeostasis	GO:0048878	187	181	5,37	2,25	5,10E-17	7,23E-16	1
Regulation of intracellular processes and metabolism	5	anatomical structure homeostasis	GO:0060249	44	61	1,26	0,76	1,04E-02	2,45E-02	1
Multicellular activities	5	hemostasis	GO:0007599	52	51	1,49	0,63	1,93E-05	7,82E-05	1
Regulation of intracellular processes and metabolism	5	positive regulation of catalytic activity	GO:0043085	172	217	4,94	2,70	3,91E-09	2,72E-08	1
Regulation of intracellular processes and metabolism	5	regulation of hydrolase activity	GO:0051336	114	143	3,27	1,78	1,31E-06	6,36E-06	1
Regulation of intracellular processes and metabolism	5	regulation of transferase activity	GO:0051338	98	158	2,81	1,96	5,79E-03	1,46E-02	1
Regulation of intracellular processes and metabolism	5	regulation of lyase activity	GO:0051339	49	19	1,41	0,24	1,24E-12	1,15E-11	1
Regulation of intracellular processes and metabolism	5	regulation of ligase activity	GO:0051340	5	70	0,14	0,87	8,41E-07	4,36E-06	2
Development	6	development of primary sexual characteristics	GO:0045137	36	43	1,03	0,53	4,37E-03	1,40E-02	1
Development	6	female sex differentiation	GO:0046660	21	18	0,60	0,22	2,45E-03	8,42E-03	1
Multicellular activities	6	ovulation cycle process	GO:0022602	19	14	0,55	0,17	1,77E-03	6,32E-03	1
Cellular metabolism	6	nucleoside phosphate metabolic process	GO:0006753	77	253	2,21	3,15	5,13E-03	1,58E-02	2
Cellular metabolism	6	amine catabolic process	GO:0009310	4	62	0,11	0,77	2,40E-06	1,55E-05	2

Cellular metabolism	6	amine biosynthetic process	GO:0009309	5	68	0,14	0,85	2,10E-06	1,37E-05	2
Cellular metabolism	6	protein catabolic process	GO:0030163	142	793	4,08	9,86	2,39E-28	3,25E-26	2
Cellular metabolism	6	modification-dependent macromolecule catabolic process	GO:0043632	78	446	2,24	5,55	1,28E-16	3,67E-15	2
Cellular metabolism	6	cytokine biosynthetic process	GO:0042089	45	24	1,29	0,30	2,39E-09	2,76E-08	1
Cellular metabolism	6	lipoprotein biosynthetic process	GO:0042158	8	47	0,23	0,58	1,15E-02	3,23E-02	2
Cellular metabolism	6	proteoglycan metabolic process	GO:0006029	21	15	0,60	0,19	7,34E-04	3,07E-03	1
Cellular metabolism	6	cellular protein metabolic process	GO:0044267	391	192 1	11,23	23,89	2,59E-59	7,02E-57	2
DNA/RNA metabolism and transcription	6	DNA metabolic process	GO:0006259	95	386	2,73	4,80	1,52E-07	1,33E-06	2
Cellular metabolism	6	macromolecule glycosylation	GO:0043413	20	97	0,57	1,21	1,58E-03	5,97E-03	2
Cellular metabolism	6	carboxylic acid metabolic process	GO:0019752	65	435	1,87	5,41	3,74E-20	2,90E-18	2
Cellular metabolism	6	cellular amino acid derivative biosynthetic process	GO:0042398	6	39	0,17	0,49	1,37E-02	3,72E-02	2
Cellular metabolism	6	phosphorylation	GO:0016310	191	732	5,49	9,10	1,45E-11	2,01E-10	2
Cellular metabolism	6	dephosphorylation	GO:0016311	12	129	0,34	1,60	4,72E-10	5,69E-09	2
Cellular metabolism	6	porphyrin metabolic process	GO:0006778	2	23	0,06	0,29	1,48E-02	3,97E-02	2
Cellular metabolism	6	group transfer coenzyme metabolic process	GO:0006752	1	24	0,03	0,30	1,94E-03	6,90E-03	2
Cellular metabolism	6	membrane lipid metabolic process	GO:0006643	10	61	0,29	0,76	2,57E-03	8,73E-03	2
Immune system response	6	T-helper 1 type immune response	GO:0042088	7	2	0,20	0,02	4,48E-03	1,41E-02	1
Cellular metabolism	6	immunoglobulin production	GO:0002377	17	12	0,49	0,15	1,77E-03	6,32E-03	1
Immune system response	6	lymphocyte mediated immunity	GO:0002449	40	29	1,15	0,36	2,05E-06	1,36E-05	1
Immune system response	6	lymphocyte activation during immune response	GO:0002285	10	6	0,29	0,07	1,07E-02	3,09E-02	1
Immune system response	6	natural killer cell activation	GO:0030101	12	2	0,34	0,02	2,71E-05	1,49E-04	1
Immune system response	6	T cell activation	GO:0042110	75	51	2,15	0,63	1,20E-11	1,71E-10	1
Immune system response	6	B cell activation	GO:0042113	52	22	1,49	0,27	1,27E-12	2,23E-11	1
Immune system response	6	lymphocyte proliferation	GO:0046651	43	21	1,23	0,26	1,17E-09	1,38E-08	1
Development	6	regulation of smooth muscle cell proliferation	GO:0048660	16	9	0,46	0,11	6,12E-04	2,64E-03	1
Multicellular activities	6	regulation of muscle cell differentiation	GO:0051147	12	4	0,34	0,05	2,84E-04	1,32E-03	1
Development	6	muscle cell development	GO:0055001	12	10	0,34	0,12	1,86E-02	4,82E-02	1
Development	6	cell fate specification	GO:0001708	27	11	0,78	0,14	2,44E-07	1,95E-06	1
Development	6	cell fate determination	GO:0001709	16	12	0,46	0,15	3,28E-03	1,09E-02	1
Development	6	negative regulation of cell differentiation	GO:0045596	68	68	1,95	0,85	1,23E-06	8,77E-06	1

Development	6	positive regulation of cell differentiation	GO:0045597	91	66	2,61	0,82	4,94E-13	9,25E-12	1
Cell cycle	6	developmental programmed cell death	GO:0010623	6	2	0,17	0,02	1,17E-02	3,23E-02	1
Development	6	neurogenesis	GO:0022008	195	178	5,60	2,21	2,14E-19	1,45E-17	1
Development	6	cell maturation	GO:0048469	27	24	0,78	0,30	6,85E-04	2,90E-03	1
Development	6	melanocyte differentiation	GO:0030318	10	2	0,29	0,02	2,19E-04	1,03E-03	1
Cellular processes (not DNA/RNA)	6	cytoskeleton-dependent intracellular transport	GO:0030705	5	37	0,14	0,46	1,04E-02	3,05E-02	2
Cellular processes (not DNA/RNA)	6	Golgi vesicle transport	GO:0048193	6	82	0,17	1,02	1,26E-07	1,14E-06	2
DNA/RNA metabolism and transcription	6	ribonucleoprotein complex assembly	GO:0022618	4	41	0,11	0,51	9,48E-04	3,81E-03	2
DNA/RNA metabolism and transcription	6	protein-DNA complex assembly	GO:0065004	4	75	0,11	0,93	4,22E-08	4,02E-07	2
Multicellular activities	6	plasma lipoprotein particle remodeling	GO:0034369	11	6	0,32	0,07	5,62E-03	1,70E-02	1
Multicellular activities	6	regulation of muscle contraction	GO:0006937	23	26	0,66	0,32	1,81E-02	4,74E-02	1
Multicellular activities	6	smooth muscle contraction	GO:0006939	27	15	0,78	0,19	5,40E-06	3,26E-05	1
Multicellular activities	6	striated muscle adaptation	GO:0014888	5	1	0,14	0,01	1,13E-02	3,18E-02	1
Multicellular activities	6	muscle hypertrophy	GO:0014896	7	1	0,20	0,01	1,35E-03	5,16E-03	1
Multicellular activities	6	regulation of blood pressure	GO:0008217	32	37	0,92	0,46	5,32E-03	1,63E-02	1
Multicellular activities	6	synaptic transmission	GO:0007268	142	115	4,08	1,43	4,73E-17	1,51E-15	1
Cell response to stimuli	6	sensory perception	GO:0007600	199	160	5,72	1,99	7,58E-24	6,86E-22	1
Multicellular activities	6	startle response	GO:0001964	9	1	0,26	0,01	1,52E-04	7,37E-04	1
Development	6	embryonic axis specification	GO:0000578	7	2	0,20	0,02	4,48E-03	1,41E-02	1
Development	6	chordate embryonic development	GO:0043009	109	105	3,13	1,31	2,19E-10	2,76E-09	1
Development	6	embryonic appendage morphogenesis	GO:0035113	31	22	0,89	0,27	2,86E-05	1,55E-04	1
Development	6	anterior/posterior pattern formation	GO:0009952	67	25	1,92	0,31	3,56E-17	1,29E-15	1
Development	6	dorsal/ventral pattern formation	GO:0009953	28	18	0,80	0,22	1,71E-05	9,89E-05	1
Development	6	segmentation	GO:0035282	22	11	0,63	0,14	2,34E-05	1,31E-04	1
Development	6	branching morphogenesis of a tube	GO:0048754	23	18	0,66	0,22	8,92E-04	3,67E-03	1
Development	6	kidney development	GO:0001822	32	32	0,92	0,40	9,28E-04	3,79E-03	1
Development	6	synapse assembly	GO:0007416	13	7	0,37	0,09	2,11E-03	7,34E-03	1
Development	6	central nervous system development	GO:0007417	142	150	4,08	1,87	2,58E-11	3,51E-10	1
Development	6	nerve development	GO:0021675	13	7	0,37	0,09	2,11E-03	7,34E-03	1
Development	6	regulation of nervous system development	GO:0051960	64	51	1,84	0,63	1,53E-08	1,60E-07	1

Development	6	vasculature development	GO:0001944	96	111	2,76	1,38	8,21E-07	6,03E-06	1
Development	6	organ morphogenesis	GO:0009887	295	256	8,47	3,18	1,83E-31	3,31E-29	1
Development	6	tissue development	GO:0009888	260	245	7,47	3,05	2,12E-24	2,30E-22	1
Development	6	sensory organ development	GO:0007423	89	66	2,56	0,82	1,65E-12	2,80E-11	1
Development	6	heart development	GO:0007507	68	85	1,95	1,06	1,83E-04	8,79E-04	1
Development	6	muscle organ development	GO:0007517	79	112	2,27	1,39	1,07E-03	4,23E-03	1
Development	6	pancreas development	GO:0031016	18	6	0,52	0,07	7,73E-06	4,56E-05	1
Development	6	gland development	GO:0048732	58	31	1,67	0,39	9,70E-12	1,42E-10	1
Development	6	limb morphogenesis	GO:0035108	33	29	0,95	0,36	2,32E-04	1,08E-03	1
Development	6	cell morphogenesis	GO:0000902	139	134	3,99	1,67	6,21E-13	1,12E-11	1
Development	6	cell part morphogenesis	GO:0032990	88	76	2,53	0,95	3,21E-10	3,96E-09	1
Cell response to stimuli	6	behavioral fear response	GO:0001662	8	2	0,23	0,02	1,67E-03	6,16E-03	1
Cell response to stimuli	6	acute inflammatory response	GO:0002526	36	28	1,03	0,35	1,52E-05	8,89E-05	1
Multicellular activities	6	blood coagulation	GO:0007596	49	50	1,41	0,62	6,07E-05	3,14E-04	1
Multicellular activities	6	circadian sleep/wake cycle	GO:0042745	5	1	0,14	0,01	1,13E-02	3,18E-02	1
Cell motility and interactions	6	chemotaxis	GO:0006935	74	30	2,13	0,37	6,94E-18	2,90E-16	1
Cell response to stimuli	6	response to food	GO:0032094	6	1	0,17	0,01	3,94E-03	1,28E-02	1
Cell response to stimuli	6	detection of light stimulus	GO:0009583	17	14	0,49	0,17	5,09E-03	1,58E-02	1
Cell response to stimuli	6	response to steroid hormone stimulus	GO:0048545	27	30	0,78	0,37	8,49E-03	2,52E-02	1
Cell response to stimuli	6	response to nicotine	GO:0035094	9	1	0,26	0,01	1,52E-04	7,37E-04	1
Cell response to stimuli	6	response to calcium ion	GO:0051592	19	16	0,55	0,20	2,95E-03	9,94E-03	1
Cellular processes (not DNA/RNA)	6	apical protein localization	GO:0045176	6	2	0,17	0,02	1,17E-02	3,23E-02	1
Cellular processes (not DNA/RNA)	6	protein transport	GO:0015031	161	567	4,62	7,05	5,25E-07	3,96E-06	2
Cellular processes (not DNA/RNA)	6	endocytosis	GO:0006897	88	117	2,53	1,46	1,13E-04	5,59E-04	1
Cellular processes (not DNA/RNA)	6	peptide secretion	GO:0002790	25	20	0,72	0,25	4,69E-04	2,09E-03	1
Cellular processes (not DNA/RNA)	6	secretion by cell	GO:0032940	120	109	3,45	1,36	1,87E-12	3,07E-11	1
Cellular processes (not DNA/RNA)	6	acid secretion	GO:0046717	7	3	0,20	0,04	1,10E-02	3,18E-02	1
Regulation of intracellular processes and metabolism	6	regulation of secretion	GO:0051046	68	45	1,95	0,56	3,92E-11	5,07E-10	1
Cellular processes (not DNA/RNA)	6	ion transmembrane transport	GO:0034220	5	38	0,14	0,47	7,00E-03	2,11E-02	2
Regulation of intracellular processes and metabolism	6	regulation of ion transport	GO:0043269	35	16	1,01	0,20	2,69E-08	2,66E-07	1

Regulation of intracellular processes and metabolism	6	regulation of neurotransmitter transport	GO:0051588	10	5	0,29	0,06	3,85E-03	1,26E-02	1
Cellular processes (not DNA/RNA)	6	monoamine transport	GO:0015844	17	11	0,49	0,14	1,34E-03	5,15E-03	1
Regulation of intracellular processes and metabolism	6	regulation of amine transport	GO:0051952	17	1	0,49	0,01	1,83E-08	1,84E-07	1
Regulation of intracellular processes and metabolism	6	regulation of organic acid transport	GO:0032890	8	2	0,23	0,02	1,67E-03	6,16E-03	1
DNA/RNA metabolism and transcription	6	nucleic acid transport	GO:0050657	7	76	0,20	0,95	2,95E-06	1,84E-05	2
Regulation of intracellular processes and metabolism	6	positive regulation of transport	GO:0051050	84	54	2,41	0,67	8,21E-14	1,71E-12	1
Regulation of intracellular processes and metabolism	6	negative regulation of transport	GO:0051051	50	42	1,44	0,52	1,71E-06	1,16E-05	1
Cell motility and interactions	6	cell migration	GO:0016477	147	122	4,22	1,52	4,40E-17	1,49E-15	1
Cell motility and interactions	6	negative regulation of cellular component movement	GO:0051271	26	25	0,75	0,31	1,99E-03	7,00E-03	1
Cell motility and interactions	6	positive regulation of cellular component movement	GO:0051272	36	22	1,03	0,27	7,75E-07	5,77E-06	1
Immune system response	6	positive regulation of immune response	GO:0050778	46	39	1,32	0,49	4,39E-06	2,71E-05	1
Immune system response	6	regulation of immune effector process	GO:0002697	28	24	0,80	0,30	3,99E-04	1,82E-03	1
Immune system response	6	regulation of adaptive immune response	GO:0002819	19	12	0,55	0,15	5,32E-04	2,31E-03	1
Regulation of intracellular processes and metabolism	6	negative regulation of biosynthetic process	GO:0009890	181	248	5,20	3,08	9,59E-08	8,98E-07	1
Regulation of intracellular processes and metabolism	6	positive regulation of biosynthetic process	GO:0009891	248	268	7,12	3,33	3,59E-18	1,77E-16	1
Regulation of intracellular processes and metabolism	6	regulation of cellular biosynthetic process	GO:0031326	847	163 0	24,33	20,27	1,45E-06	9,98E-06	1
Regulation of intracellular processes and metabolism	6	negative regulation of macromolecule metabolic process	GO:0010605	219	321	6,29	3,99	1,93E-07	1,61E-06	1
Regulation of intracellular processes and metabolism	6	negative regulation of cellular metabolic process	GO:0031324	213	314	6,12	3,90	3,96E-07	3,07E-06	1
Regulation of intracellular processes and metabolism	6	negative regulation of nitrogen compound metabolic process	GO:0051172	160	231	4,60	2,87	5,17E-06	3,16E-05	1
Regulation of intracellular processes and metabolism	6	positive regulation of catabolic process	GO:0009896	20	22	0,57	0,27	1,79E-02	4,70E-02	1
Regulation of intracellular processes and metabolism	6	positive regulation of macromolecule metabolic process	GO:0010604	279	335	8,01	4,17	3,41E-16	8,42E-15	1
Regulation of intracellular processes and metabolism	6	positive regulation of cellular metabolic process	GO:0031325	285	341	8,18	4,24	1,35E-16	3,67E-15	1
Regulation of intracellular processes and metabolism	6	positive regulation of nitrogen compound metabolic process	GO:0051173	227	245	6,52	3,05	1,08E-16	3,25E-15	1
Regulation of transcription	6	regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	GO:0019219	799	156 4	22,95	19,45	2,41E-05	1,34E-04	1
Regulation of intracellular processes and metabolism	6	regulation of phosphorus metabolic process	GO:0051174	141	204	4,05	2,54	2,19E-05	1,24E-04	1
Regulation of transcription	6	regulation of gene expression	GO:0010468	791	162 8	22,72	20,25	3,03E-03	1,02E-02	1
Regulation of intracellular processes and metabolism	6	regulation of lipid metabolic process	GO:0019216	37	37	1,06	0,46	4,97E-04	2,19E-03	1
Development	6	positive regulation of multicellular organism growth	GO:0040018	11	5	0,32	0,06	1,66E-03	6,16E-03	1
Regulation of intracellular processes and metabolism	6	positive regulation of cell proliferation	GO:0008284	149	124	4,28	1,54	2,61E-17	1,01E-15	1
Cell motility and interactions	6	positive regulation of cell adhesion	GO:0045785	22	12	0,63	0,15	3,83E-05	2,06E-04	1
Cell cycle	6	positive regulation of cell cycle	GO:0045787	28	14	0,80	0,17	1,42E-06	9,87E-06	1

Regulation of intracellular processes and metabolism	6	positive regulation of cell activation	GO:0050867	51	23	1,46	0,29	9,30E-12	1,40E-10	1
Cell response to stimuli	6	positive regulation of response to biotic stimulus	GO:0002833	5	1	0,14	0,01	1,13E-02	3,18E-02	1
Cell response to stimuli	6	positive regulation of response to external stimulus	GO:0032103	30	11	0,86	0,14	1,71E-08	1,75E-07	1
Multicellular activities	6	positive regulation of behavior	GO:0048520	18	5	0,52	0,06	2,70E-06	1,72E-05	1
Regulation of intracellular processes and metabolism	6	positive regulation of cytokine production	GO:0001819	28	20	0,80	0,25	5,48E-05	2,86E-04	1
Regulation of intracellular processes and metabolism	6	positive regulation of tissue remodeling	GO:0034105	14	8	0,40	0,10	1,57E-03	5,96E-03	1
Regulation of intracellular processes and metabolism	6	negative regulation of cell proliferation	GO:0008285	136	122	3,91	1,52	3,95E-14	8,58E-13	1
Cell cycle	6	negative regulation of cell cycle	GO:0045786	25	28	0,72	0,35	1,02E-02	3,00E-02	1
Regulation of intracellular processes and metabolism	6	negative regulation of cell activation	GO:0050866	17	12	0,49	0,15	1,77E-03	6,32E-03	1
Immune system response	6	negative regulation of cytokine production	GO:0001818	12	9	0,34	0,11	1,45E-02	3,92E-02	1
Regulation of intracellular processes and metabolism	6	negative regulation of tissue remodeling	GO:0034104	11	2	0,32	0,02	7,76E-05	3,94E-04	1
Immune system response	6	regulation of response to tumor cell	GO:0002834	5	1	0,14	0,01	1,13E-02	3,18E-02	1
Cell response to stimuli	6	regulation of defense response	GO:0031347	40	40	1,15	0,50	2,09E-04	9,97E-04	1
Cell cycle	6	regulation of programmed cell death	GO:0043067	257	309	7,38	3,84	7,80E-15	1,84E-13	1
Cell cycle	6	regulation of mitotic cell cycle	GO:0007346	52	74	1,49	0,92	8,27E-03	2,47E-02	1
Signal transduction	6	cell surface receptor linked signaling pathway	GO:0007166	806	438	23,15	5,45	4,72E-158	2,56E-155	1
Signal transduction	6	intracellular signaling pathway	GO:0007242	489	819	14,04	10,19	3,76E-09	4,09E-08	1
Signal transduction	6	regulation of signal transduction	GO:0009966	290	352	8,33	4,38	2,23E-16	5,77E-15	1
Regulation of intracellular processes and metabolism	6	regulation of endothelial cell proliferation	GO:0001936	16	8	0,46	0,10	4,21E-04	1,90E-03	1
Regulation of intracellular processes and metabolism	6	regulation of epithelial cell proliferation	GO:0050678	27	14	0,78	0,17	2,96E-06	1,84E-05	1
Immune system response	6	regulation of leukocyte activation	GO:0002694	67	35	1,92	0,44	2,26E-13	4,39E-12	1
Regulation of intracellular processes and metabolism	6	regulation of chemokine production	GO:0032642	10	4	0,29	0,05	1,76E-03	6,32E-03	1
Immune system response	6	regulation of interferon-gamma production	GO:0032649	11	8	0,32	0,10	1,22E-02	3,32E-02	1
Immune system response	6	regulation of interleukin-2 production	GO:0032663	13	6	0,37	0,07	6,60E-04	2,82E-03	1
Immune system response	6	regulation of interleukin-6 production	GO:0032675	17	12	0,49	0,15	1,77E-03	6,32E-03	1
Development	6	regulation of bone remodeling	GO:0046850	34	21	0,98	0,26	1,92E-06	1,29E-05	1
Development	6	regulation of neurological system process	GO:0031644	36	35	1,03	0,44	3,73E-04	1,71E-03	1
Regulation of intracellular processes and metabolism	6	regulation of cell growth	GO:0001558	64	62	1,84	0,77	1,25E-06	8,84E-06	1
Regulation of intracellular processes and metabolism	6	cell redox homeostasis	GO:0045454	5	47	0,14	0,58	6,89E-04	2,90E-03	2
Regulation of intracellular processes and metabolism	6	cellular chemical homeostasis	GO:0055082	151	121	4,34	1,50	2,84E-18	1,54E-16	1
Regulation of intracellular processes and metabolism	6	negative regulation of homeostatic process	GO:0032845	9	4	0,26	0,05	4,24E-03	1,37E-02	1
Regulation of intracellular processes and metabolism	6	positive regulation of homeostatic process	GO:0032846	15	5	0,43	0,06	4,65E-05	2,45E-04	1
Regulation of intracellular processes and metabolism	6	tissue homeostasis	GO:0001894	18	18	0,52	0,22	1,65E-02	4,37E-02	1
Multicellular activities	6	leukocyte homeostasis	GO:0001776	17	8	0,49	0,10	1,07E-04	5,34E-04	1
Regulation of intracellular processes and metabolism	6	carbohydrate homeostasis	GO:0033500	20	15	0,57	0,19	1,26E-03	4,90E-03	1

Regulation of intracellular processes and metabolism	6	ion homeostasis	GO:0050801	152	144	4,37	1,79	2,13E-14	4,82E-13	1
Regulation of intracellular processes and metabolism	6	lipid homeostasis	GO:0055088	21	17	0,60	0,21	1,29E-03	4,99E-03	1
Regulation of intracellular processes and metabolism	6	negative regulation of hydrolase activity	GO:0051346	16	15	0,46	0,19	1,68E-02	4,43E-02	1
Regulation of intracellular processes and metabolism	6	negative regulation of lyase activity	GO:0051350	23	7	0,66	0,09	2,00E-07	1,65E-06	1
Regulation of intracellular processes and metabolism	6	negative regulation of ligase activity	GO:0051352	3	59	0,09	0,73	1,06E-06	7,66E-06	2
Regulation of intracellular processes and metabolism	6	positive regulation of hydrolase activity	GO:0051345	83	53	2,38	0,66	1,04E-13	2,09E-12	1
Regulation of intracellular processes and metabolism	6	positive regulation of transferase activity	GO:0051347	71	85	2,04	1,06	6,68E-05	3,42E-04	1
Regulation of intracellular processes and metabolism	6	positive regulation of lyase activity	GO:0051349	28	11	0,80	0,14	1,01E-07	9,33E-07	1
Regulation of intracellular processes and metabolism	6	positive regulation of ligase activity	GO:0051351	3	65	0,09	0,81	1,70E-07	1,44E-06	2
Regulation of intracellular processes and metabolism	6	positive regulation of oxidoreductase activity	GO:0051353	13	10	0,37	0,12	1,05E-02	3,05E-02	1
Regulation of intracellular processes and metabolism	6	regulation of GTPase activity	GO:0043087	19	80	0,55	0,99	1,55E-02	4,13E-02	2
Regulation of intracellular processes and metabolism	6	regulation of peptidase activity	GO:0052547	32	30	0,92	0,37	4,41E-04	1,98E-03	1
Regulation of intracellular processes and metabolism	6	regulation of lipase activity	GO:0060191	55	11	1,58	0,14	4,53E-19	2,73E-17	1
Regulation of intracellular processes and metabolism	6	regulation of ubiquitin-protein ligase activity	GO:0051438	4	69	0,11	0,86	2,55E-07	2,01E-06	2
Regulation of intracellular processes and metabolism	6	regulation of monooxygenase activity	GO:0032768	12	10	0,34	0,12	1,86E-02	4,82E-02	1
Regulation of intracellular processes and metabolism	6	regulation of protein homodimerization activity	GO:0043496	5	1	0,14	0,01	1,13E-02	3,18E-02	1

The functional analysis is made between recent (*i.e.* genes that originated with metazoans or vertebrates, *list 1*) and ancestral genes (*i.e.*, genes that originated with the last universal common ancestor or eukaryotes *list 2*), as explained in Table 15.

5. Gene dosage of human duplicated hubs is regulated through three different mechanisms

The duplication of a gene entails dosage imbalance between itself and its interactors. In principle, the more connections the protein encoded by this gene has, the more deleterious will be the effect of the dosage imbalance caused by its duplication. Therefore, the presence of a class of duplicated hubs in the human protein interaction network is counterintuitive. The fact the human duplicated hubs appeared at a particular time in evolution suggests that an unexpected event occurred in the ancestor of vertebrates. It is now demonstrated that the early vertebrate genome underwent two rounds of whole genome duplication (Dehal and Boore, 2005; Nakatani et al., 2007) and that this particular mechanism of large-scale gene duplication preserves the dosage balance between genes, hence dosage-sensitive genes may retain duplications. This has been demonstrated in both yeast (Qian and Zhang, 2008) and vertebrates (Makino and McLysaght, 2010) and may explain why *H. sapiens* has a class of duplicated hubs that is not conserved in non-vertebrate species.

Other mechanisms may allow the retention of duplicates of dosage-sensitive genes. First, enhanced post-transcriptional regulation may act as a buffer to control the gene dosage. Therefore, if a gene is subjected to this type of regulation, its duplication does not affect the fitness of the organism and may be retained in the genome. microRNAs are an example of such regulation. Tissue selectivity is another mechanism that may allow the retention of the duplicates of dosage-sensitive genes. Indeed, if subfunctionalization occurs after duplication, then the paralogs may become tissue-specific and be expressed in different tissues, therefore they would not interfere with each other (Fernandez and Chen, 2009; Semon and Wolfe, 2007).

To understand whether the effect of the combination of these three mechanisms influences the retention of hub duplications, we analyzed the enrichment of young duplicated hubs in genes that are involved in at least one of these mechanisms. We found that 61.4% of duplicated hubs are also ohnologs, *i.e.* that duplicated via whole genome duplication (Wolfe, 2000), miRNA targets, or tissue-selective genes, a fraction that is significantly higher compared with singleton hubs (33.9%, *p*-value 4e-40 from Fisher's exact test) (Figure 41). All three mechanisms, separately, show a similar trend (Figure 42, Figure 43, Figure 44), but the most significant is represented by ohnologs. The effect of miRNA targets is weak, because we could analyze only less than 1,000 genes. Therefore, the signal is not significant, but it contributes to the general trend, when added to that of ohnologs and tissue-specific genes.

It is remarkable to note that housekeeping genes (*i.e.* genes that are expressed in all human tissues) preferentially encode for ancestral hubs, independently from their duplicability (Figure 45). This is yet another characteristic that is peculiar of the hubs that are conserved throughout evolution. Not only they highly conserved throughout evolution, but they are needed to be expressed in all cells of a multicellular organism.

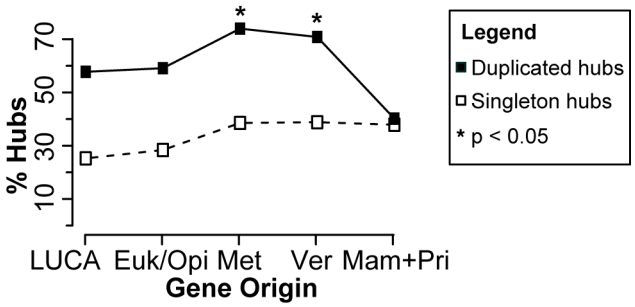


Figure 41: dosage regulation of hubs

The fraction of genes that are ohnologs, miRNA targets or tissue-specific is compared between singleton and duplicated hubs. Given the small number of hubs that originated with opisthokonts and primates (43 and 17, respectively), genes that originated at these two levels are grouped with eukaryotes and mammals, respectively. "*" represents significant enrichment when compared with more ancient genes (Fisher's exact test).

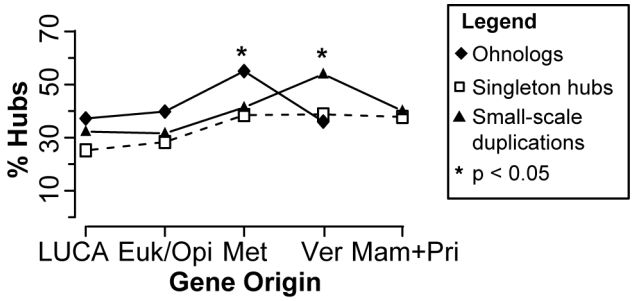


Figure 42: relationships between ohnologs and singleton hubs

The origin distribution of hubs that are ohnologs is compared with singleton hubs and hubs duplicated via small-scale duplications. "*" represents significant enrichment when compared with more ancient genes (Fisher's exact test).

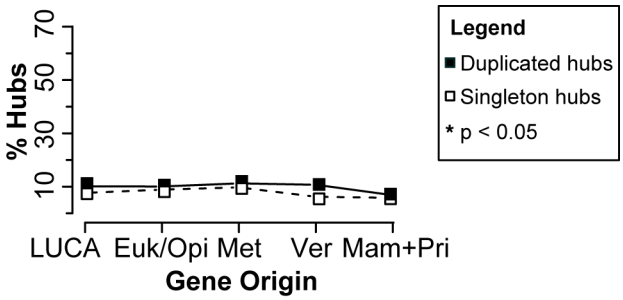


Figure 43: miRNA regulation of singleton and duplicated hubs

The origin distribution of duplicated hubs that are targets of miRNAs is compared with singleton hubs.

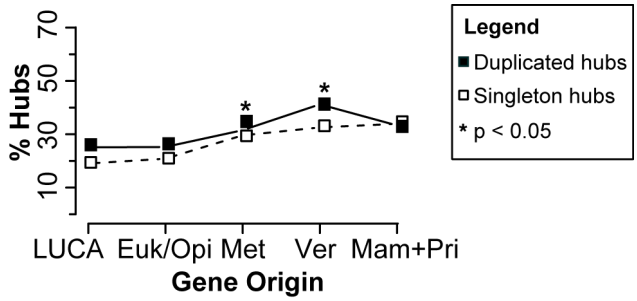
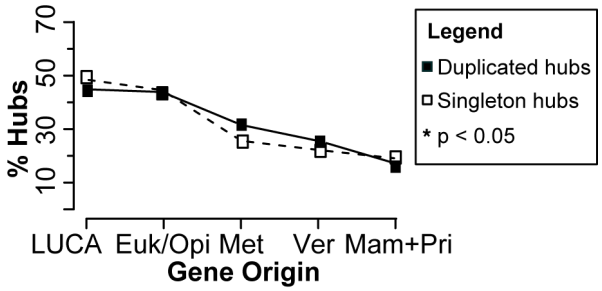


Figure 44: tissue-selectivity of singleton and duplicated hubs

The origin distribution of duplicated tissue-specific hubs with singleton hubs. Tissue-selective genes are expressed in less than 25% of human tissues. "*" represents significant enrichment when compared with more ancient genes (Fisher's exact test).

Figure 45: relationships between housekeeping genes and duplicability of hubs



The origin distribution of housekeeping hubs is compared between singleton and duplicated hubs. Housekeeping genes are expressed in at least 97% of human tissues.

6. Dominant and recessive cancer genes are representative of ancient and recent human hubs

We demonstrated that all protein interaction networks conserve a core of highly connected, central proteins, which are preferentially encoded by singleton genes that are expressed in all tissues of the multicellular organism (at least in *H. sapiens*). The human protein interaction network gained a new set of hubs, which originated later in evolution and duplicated in the two rounds of whole genome duplications in the ancestral vertebrate. The presence of this new class of hubs in the human protein interaction network would imply an increased robustness towards dosage perturbations, since even genes that encode highly connected proteins are allowed to retain their paralogs. However, this is not the case. The dosage of young duplicated hubs is tightly regulated by miRNAs or through tissue-selective expression and mutations in their corresponding genes are often associated to cancer (Jonsson and Bates, 2006; Rambaldi et al., 2008; Syed et al., 2010). Actually, the relationships between hubs and their involvement in cancer are highly complex.

Cancer genes are genes whose mutations are associated to the development of cancer (Futreal et al., 2004). As it was previously shown (Rambaldi et al., 2008; Syed et al., 2010), cancer genes encode highly connected proteins in the human protein interaction

network. Their origin differs from that of the rest of the human genes: they are enriched in genes born with metazoans and are depleted in genes that originated with mammals and primates (Figure 46). Furthermore, the analysis of dominant and recessive cancer genes (Futreal et al., 2004) showed that they have different evolutionary properties. In particular, while the origin of dominant genes reflects that of all cancer genes, recessive genes are ancient: they are enriched only in genes that originated with the last universal common ancestor (Figure 46).

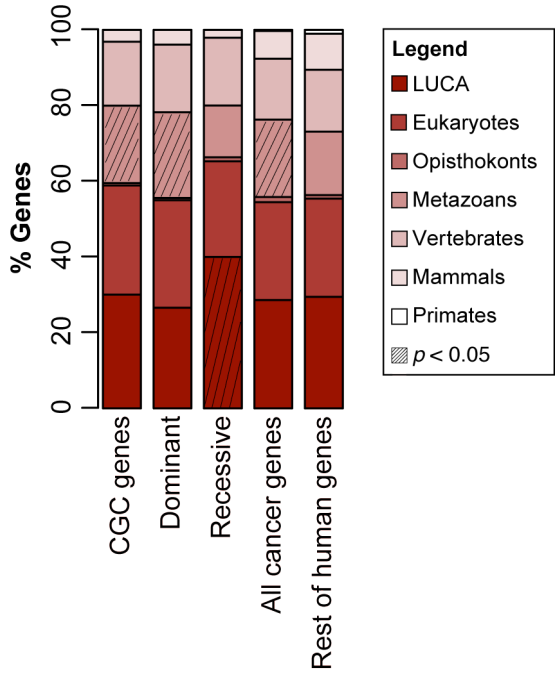


Figure 46: origin of cancer genes

The fraction of genes that originated at each level of evolution is plotted for different categories of cancer genes. "CGC" represents the cancer genes from the cancer gene census (Futreal et al., 2004), dominant and recessive genes are defined as in (Futreal et al., 2004), while "all cancer genes" includes all the candidate cancer genes from high-throughput mutational screenings and whole genome sequencing of cancer samples. *P*-values are calculated with Fisher's exact test. LUCA, last universal common ancestor.

Overall, cancer genes have less highly conserved duplications than the rest of human genes, although the duplicability of a gene is tightly related to its time of appearance. Hence, we repeated the analysis of duplicability of cancer genes at each level of origin. Among old genes, all cancer genes are less duplicated than the rest of human

genes. Younger genes, instead showed striking differences. Recessive genes that originated recently in evolution are all singleton, while dominant genes that originated with metazoans and vertebrates are enriched in duplicated genes (18.5%), compared with the rest of human genes (11.0%) (Figure 47).

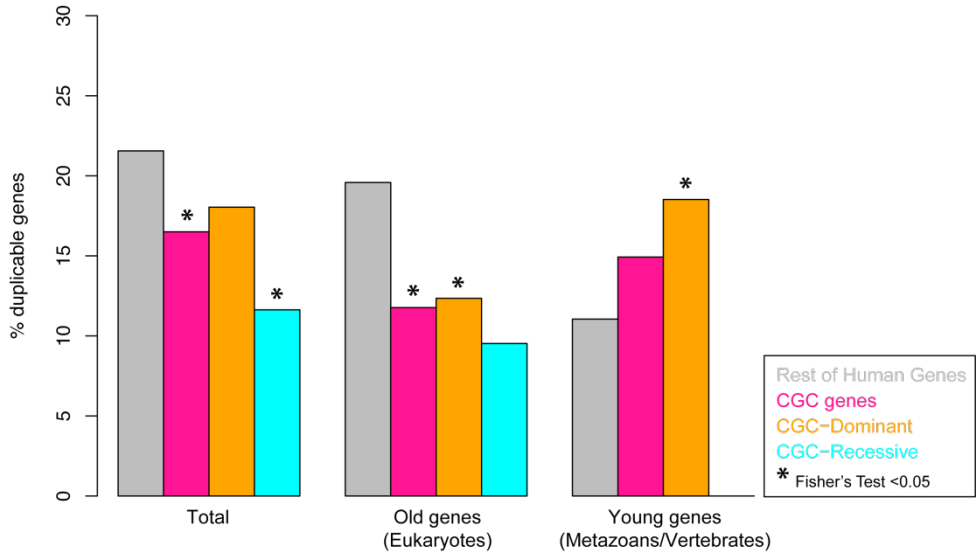


Figure 47: relationships between origin and duplicability of cancer genes

The fraction of duplicated genes is calculated for genes that originated at different times in evolution. Eukaryotes are taken as representative of ancient genes, while metazoans and vertebrates are representative of recent genes. Duplicability is calculated as the presence of additional hits longer than 60% of the original length of the gene of interest. Fisher's exact test is calculated with respect to the rest of the human genes.

The analysis of gene and network properties of cancer genes showed that there are two evolutionarily distinct classes of cancer genes that correspond to the two distinct classes of human hubs. Recessive genes are representative of old hubs: they originated early in evolution and are mostly singleton. Dominant genes are instead representative of young hubs: they mostly originated with metazoans and retain duplications. The analysis of the biological processes that involve cancer genes showed less striking results compared with the analyses of hubs. This is mostly due to the small number of dominant (276) and recessive (80) genes that are associated to GO terms at level 5 and 6. Dominant genes are enriched in regulation of metabolism, while recessive genes are particularly enriched in

processes related with cell cycle, metabolism and DNA metabolism (Table 18). This roughly reflects the functional differences between the two classes of hubs: recessive genes are representative of ancestral singleton hubs and are involved in basic cellular functions, while dominant are mostly involved in complex cellular functions that are related to the regulation of metabolism.

7. Cancer genes are neighbors in the human protein interaction network

Given the fact that cancer genes are highly connected and central in the human protein interaction network, we analyzed whether this reflects a different distance between cancer proteins and the rest of human proteins. We calculated the distance between each couple of proteins inside the human protein interaction network, thus creating a 11,988-by-11,988 distance matrix with each row and column representing a protein. We then calculated the average distance between each cancer protein and the other cancer proteins that have network information by extracting the corresponding rows and columns. We found that both dominant and recessive proteins are closer to each other than the rest of human genes (*p*-value 7e-94 and 2e-34, respectively, from Wilcoxon test, Figure 48).

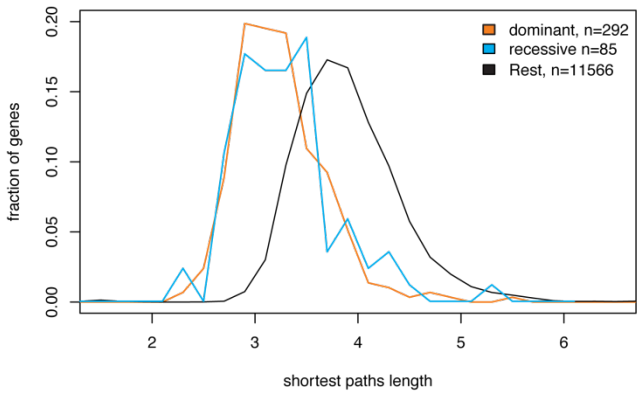


Figure 48: distance between cancer proteins

The distribution of mean distance between dominant proteins, recessive proteins and the rest of the human proteins. The distance is defined as the smallest number of interactions that must be crossed to join two nodes.

Table 18: functional comparison between dominant and recessive cancer genes

Process	GO level	GO description	GO ID	N genes (list 1)	N genes (list 2)	% of total genes (list 1)	% of total genes (list 2)	p-value	adjusted p-value	Enriched list
Regulation of intracellular processes and metabolism	6	regulation of cellular biosynthetic process	GO:0031326	156	29	53,42465753	35,36585366	4,09E-03	3,90E-02	1
Regulation of intracellular processes and metabolism	5	regulation of biosynthetic process	GO:0009889	156	29	53,42465753	35,36585366	4,09E-03	4,93E-02	1
Cell cycle	6	cell cycle checkpoint	GO:0000075	4	16	1,369863014	19,51219512	1,86E-08	1,68E-06	2
Cell cycle	5	regulation of cell cycle	GO:0051726	15	20	5,136986301	24,3902439	1,77E-06	7,18E-05	2
Cell cycle	6	regulation of mitotic cell cycle	GO:0007346	8	11	2,739726027	13,41463415	5,01E-04	1,30E-02	2
Cell cycle	6	regulation of cell cycle process	GO:0010564	4	7	1,369863014	8,536585366	2,98E-03	2,99E-02	2
Cell motility and interactions	5	regulation of cell adhesion	GO:0030155	4	7	1,369863014	8,536585366	2,98E-03	4,82E-02	2
Cell response to stimuli	6	response to UV	GO:0009411	6	12	2,054794521	14,63414634	3,60E-05	1,63E-03	2
Cell response to stimuli	5	response to light stimulus	GO:0009416	9	13	3,082191781	15,85365854	1,04E-04	2,80E-03	2
Cell response to stimuli	5	response to ionizing radiation	GO:0010212	2	8	0,684931507	9,756097561	1,23E-04	2,84E-03	2
Cellular metabolism	5	macromolecule catabolic process	GO:0043285	16	21	5,479452055	25,6097561	1,06E-06	5,70E-05	2
Cellular metabolism	5	cellular macromolecule catabolic process	GO:0044265	12	19	4,109589041	23,17073171	7,45E-07	5,70E-05	2
Cellular metabolism	6	modification-dependent macromolecule catabolic process	GO:0043632	9	11	3,082191781	13,41463415	9,13E-04	2,06E-02	2
Cellular metabolism	6	protein catabolic process	GO:0030163	14	13	4,794520548	15,85365854	2,50E-03	2,66E-02	2
Development	6	gastrulation	GO:0007369	3	7	1,02739726	8,536585366	1,32E-03	2,65E-02	2
Development	6	tissue development	GO:0009888	27	19	9,246575342	23,17073171	1,84E-03	2,66E-02	2
Development	5	embryonic morphogenesis	GO:0048598	12	12	4,109589041	14,63414634	1,59E-03	2,87E-02	2
Development	5	positive regulation of developmental process	GO:0051094	27	18	9,246575342	21,95121951	3,45E-03	4,93E-02	2
DNA/RNA metabolism and transcription	6	DNA metabolic process	GO:0006259	16	32	5,479452055	39,02439024	5,27E-13	9,54E-11	2
Regulation of intracellular processes and metabolism	5	negative regulation of catalytic activity	GO:0043086	2	10	0,684931507	12,19512195	7,27E-06	2,35E-04	2
Regulation of intracellular processes and metabolism	6	negative regulation of transferase activity	GO:0051348	2	8	0,684931507	9,756097561	1,23E-04	4,44E-03	2
Regulation of intracellular processes and metabolism	6	negative regulation of macromolecule metabolic process	GO:0010605	28	19	9,589041096	23,17073171	2,17E-03	2,66E-02	2
Regulation of intracellular processes and metabolism	6	negative regulation of cellular metabolic process	GO:0031324	27	19	9,246575342	23,17073171	1,84E-03	2,66E-02	2

Regulation of intracellular processes and metabolism	6	regulation of phosphorus metabolic process	GO:0051174	19	15	6,506849315	18,29268293	2,14E-03	2,66E-02	2
Regulation of intracellular processes and metabolism	6	negative regulation of cell proliferation	GO:0008285	15	14	5,136986301	17,07317073	1,51E-03	2,66E-02	2
Regulation of intracellular processes and metabolism	6	positive regulation of hydrolase activity	GO:0051345	7	9	2,397260274	10,97560976	2,33E-03	2,66E-02	2
Regulation of intracellular processes and metabolism	5	negative regulation of metabolic process	GO:0009892	29	20	9,931506849	24,3902439	1,35E-03	2,74E-02	2
Regulation of intracellular processes and metabolism	5	negative regulation of cellular process	GO:0048523	66	32	22,60273973	39,02439024	4,26E-03	4,93E-02	2
Regulation of intracellular processes and metabolism	5	anatomical structure homeostasis	GO:0060249	8	9	2,739726027	10,97560976	4,07E-03	4,93E-02	2

The functional analysis is made between dominant (*list 1*) and recessive genes (*list 2*), as explained in Table 15.

8. Cancer genes are depleted in highly conserved paralogs

It was previously demonstrated that cancer genes are depleted in highly conserved paralogs, compared with the rest of human genes (Rambaldi et al., 2008). However, the previous paragraphs showed that gene origin influences the duplicability of cancer genes (Figure 47). Furthermore, dominant and recessive cancer genes have different behaviors in terms of duplicability. These differences are highly complex and depend also on the level of the conservation of duplication. Therefore, after analyzing the role of the gene origin, we investigated the level of conservation of the paralogs.

Instead of considering as duplicated all genes that have additional hits above 60% of their length as in Rambaldi *et al.* (Rambaldi et al., 2008), we set variable thresholds and investigated whether cancer genes behave differently than the rest of human genes. We discovered that cancer genes are more duplicated when considering low levels of paralog conservation (less than 30%), while they are less duplicated at higher levels of conservation (more than 40%) (Figure 49). However, not all cancer genes have the same behavior: recessive genes are less duplicated than the rest of human genes independently on the level of conservation. We performed the same analyses considering two types of duplications separately: duplications that overlap known genes and those that fall in intronic or intergenic regions. The relationships between both types of cancer genes and the rest of human genes did not change when analyzing duplications that overlap known genes (Figure 50). Duplications that do not overlap known genes instead always involve less than 20% of human genes and are relatively constant between 10 and 80% of sequence conservation (Figure 51). Therefore, the differences in duplicability between cancer genes and the rest of human genes are due to functional duplications, not to spurious hits on the genome, which do not overlap known genes.

The different duplicability of cancer genes and the rest of human genes might be due to their different length. The median length of cancer genes (3,979 bp) is significantly different from the rest of human genes (1,941 bp, *p*-value from Wilcoxon test 8.5e-82) (Figure 52). In principle, longer genes have a higher low-coverage duplicability, because the likelihood to find a highly identical sequence on another chromosomal location is proportional to the gene length. At high levels of coverage, instead, long genes have a lower probability to find highly identical sequences. In order to eliminate this potential bias, we repeated the analysis normalizing by the gene length. Instead of considering the number of duplicated genes, we used the sum of the length of all duplicated genes and divided this by the total gene length. The results remain consistent with the previous findings also when normalizing by the gene length (Figure 53, Figure 54, Figure 55).

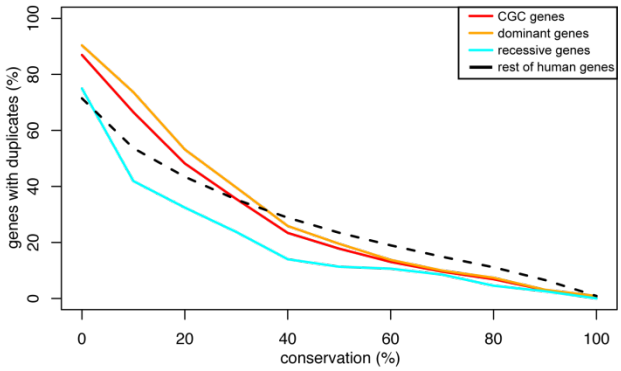


Figure 49: conservation of duplicates of cancer genes

Duplicability as a function of the conservation level (*i.e.* the length of the longest additional hit on the genome) is calculated for all cancer genes (CGC genes, *i.e.* all genes that are included in the cancer gene census), dominant genes, recessive genes and the rest of human genes.

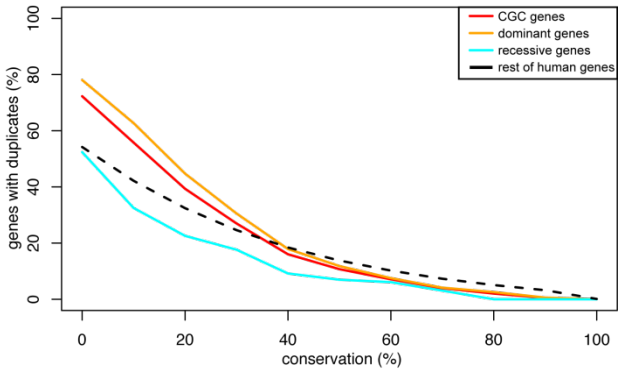


Figure 50: conservation of genic duplicates of cancer genes

Duplicability is calculated as in Figure 49. Only hits that overlap exons of known genes are considered.

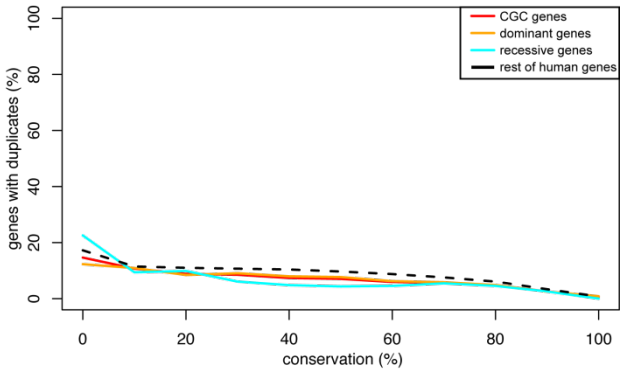


Figure 51: conservation of genomic duplicates of cancer genes

Duplicability is calculated as in Figure 49. Only hits that do not overlap exons of known genes are considered.

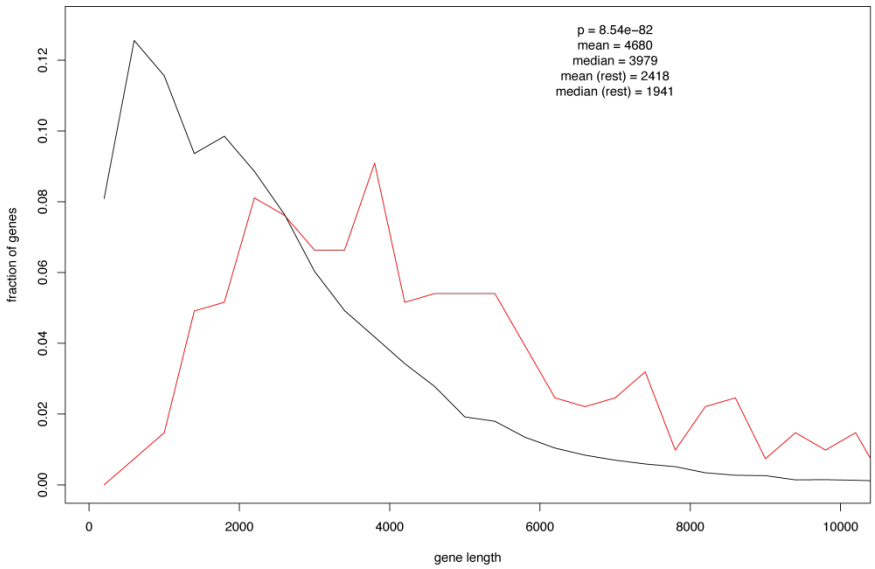


Figure 52: length of cancer genes

The distribution of the length of cancer genes (red) is compared with the length distribution of the rest of human genes (black). To calculate the gene length only the coding sequence of the longest isoform of each gene is considered.

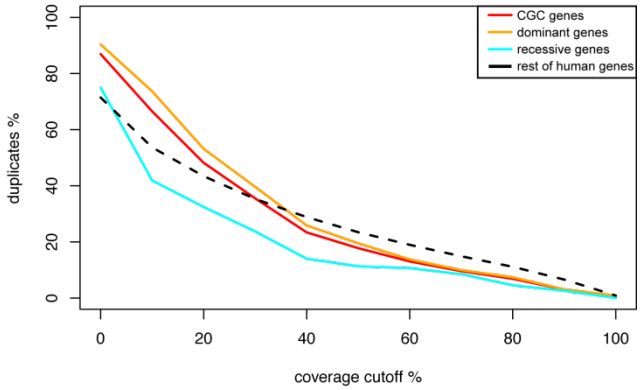


Figure 53: conservation of normalized duplicates of cancer genes

Duplicability is calculated as in Figure 49. In order to normalize by the gene length, the fraction of duplicates is calculated as ratio between the sum of the length of all duplicated genes and the sum of the length of all genes, for each level of conservation.

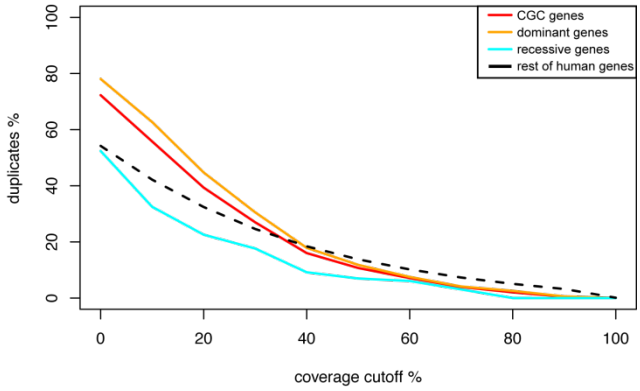


Figure 54: conservation of normalized genic duplicates of cancer genes

Duplicability is calculated and normalized as described in Figure 49. Only hits that overlap exons of known genes are considered.

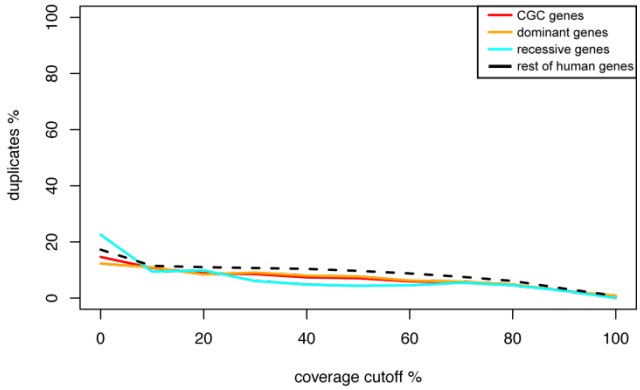


Figure 55: conservation of normalized genomic duplicates of cancer genes

Duplicability is calculated and normalized as described in Figure 49. Only hits that do not overlap exons of known genes are considered.

Discussion

In the past few years, several technical innovations allowed extensive studies of biological networks. In particular, the development of high-throughput techniques such as yeast-two-hybrid and TAP increased the protein interaction networks data exponentially and protein-protein interactions analysis is now feasible in several species with different levels of complexity (Breitkreutz et al., 2008; Kerrien et al., 2007). However, we are far from having a comprehensive picture of protein interaction networks. Yu *et al.* (Yu et al., 2008a) estimated that the protein interaction network of *S. cerevisiae* consists of 18,000 ± 4,500 interactions, but we identified almost 100,000 interactions between yeast proteins. Furthermore, it is still matter of debate whether high-throughput data include more or less false positives than literature-based data (Gandhi et al., 2006; Hart et al., 2006; Venkatesan et al., 2009; Yu et al., 2008a). Given the high rates of both false negatives and false positives, it would seem unfeasible to compare networks from different species, especially taking into consideration the fact that different networks have different levels of completeness and are based on different combinations of high-throughput and literature-based primary data (Table 8). Notwithstanding these pivotal issues, we identified several properties that are common in the available protein interaction networks. From any analyses, we first eliminated all the networks that include too few proteins and interactions, such as *M. musculus*, *C. elegans*, *A. thaliana* and *M. pneumoniae*, while we retained *H. sapiens*, *D. melanogaster*, *S. cerevisiae* and *E. coli*. Although the number of species is small, they well represent different levels of complexity. Indeed we were able to analyze one prokaryote (*E. coli*), one unicellular eukaryote (*S. cerevisiae*), one lower metazoan (*D, melanogaster*) and one vertebrate (*H. sapiens*).

Gene properties presented less issues for these four species. Only for *H. sapiens* we needed to convert the identifiers, in order to have the same type of gene IDs between

network and gene properties. For other species not all data were available. In particular, for *M. musculus*, *G. gallus* and *C. elegans* we were able to identify none or very few group-specific genes. This is due to the fact that we could use only one or very few species to define the corresponding group-specific nodes. However, this did not affect the results in a significant way, because metazoans, and vertebrates in particular, have a very low fraction of group-specific genes.

Having identified the duplicability and the origin of genes from several different species, we could hypothesize how protein interaction networks evolved. In particular, we discovered the role that gene duplications have in the evolution of protein interaction networks. We identified a network core that is common to all the analyzed species. It is composed of ancient central hubs whose singleton status is retained throughout evolution. These genes evolve slowly, hence we were able to identify their orthologs in distantly related species, such as vertebrates and prokaryotes. Our studies on the function and the expression of human ancient hubs, together with previous findings in *S. cerevisiae* (Kunin et al., 2004), showed that these are ubiquitously expressed and are involved in basic cellular processes, related with the survival of the single cell. Functions like transcription, replication and cellular metabolism are typical of these proteins. Although the proteins inside the network core have similar properties, they are not a homogeneous set. Indeed, it was demonstrated that yeast hubs might be divided into two categories: date and party hubs (Han et al., 2004). This distinction was made on the basis of the expression of each hub and its interactors. The expression of the interactors of party hubs is highly correlated, therefore party hubs may interact simultaneously with their interactors. Date hubs, instead, display a low level of co-expression, which may be explained by the fact that their interactions occur at different times and in different locations (Figure 56) (Han et al., 2004).

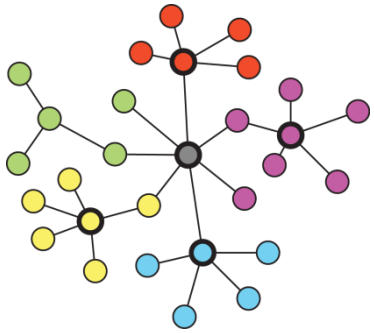


Figure 56: date and party hubs

Adapted from (Han et al., 2004). Nodes with thick borders are hubs. Different colors represent different locations and/or different time of expression. The grey hub is a date hub, because it interacts with different proteins at different times and localizations, while the other four hubs are party hubs because they are co-localized with their interactors.

In *S. cerevisiae*, *E. coli*, *D. melanogaster* and *H. sapiens*, connectivity and centrality negatively correlate with the age of the corresponding genes, *i.e.* younger proteins are less connected and more peripheral than older proteins. Furthermore, ancient hubs are preferentially singleton. Studies on the *S. cerevisiae* protein interaction network showed that essential genes are preferentially singleton (Papp et al., 2003; Yang et al., 2003) and occupy central positions inside the yeast protein interaction network (Gandhi et al., 2006), while peripheral proteins are involved in ongoing adaptive evolution (Kim et al., 2007). Our results support these findings, showing that all protein interaction networks conserve of backbone of essential singleton nodes that are ancient, slow-evolving and are involved in basic cellular processes. The network periphery, instead, includes non-essential young proteins that are under positive selection. Therefore, their duplications are more likely to be retained in the genome.

Ancient human hubs, independently from their duplicability status, are expressed in several tissues and many of them are housekeeping (*i.e.* they are expressed in all tissues), while younger hubs are expressed in fewer tissues (Figure 44, Figure 45). This strengthens

the hypothesis that the network core is composed of genes that are important for the survival of the cell and, therefore, must be expressed in all the cells of an organism.

The protein interaction networks of *H. sapiens*, *D. melanogaster*, *S. cerevisiae* and *E. coli* are scale-free (Figure 29, Figure 30, Figure 31, Figure 32). Barabasi and Albert (Barabasi and Albert, 1999) described the "preferential attachment" theory of evolution of scale-free networks starting from two assumptions:

1. The models that describe random networks (Erdős-Rényi and Watts-Strogatz) start with a fixed number of nodes, which are either randomly connected or reconnected, without implying any expansion of the network (Barabasi and Albert, 1999). In biological networks and, in general, real world networks, the probability to gain new nodes is non-zero;
2. In random networks, the probability of a connection between two nodes is random and uniform. However, nodes in real networks establish connections in a non-random way. In particular, new nodes tend to attach to already existing nodes that are highly connected, with a probability that is proportional to their degree (Barabasi and Albert, 1999).

The preferential attachment theory seems to be supported by our results. Indeed, we found that, in all four networks, older proteins are highly connected and central, while younger proteins have fewer connections and are located at the network periphery.

The preferential attachment theory does not take into account the mechanisms of gene evolution. The evolution of new genes from non-coding DNA sequences accounts for a negligible fraction of the genome (Cai et al., 2008; Knowles and McLysaght, 2009; Toll-Riera et al., 2009; Zhou et al., 2008), while gene duplication is responsible for the vast majority of new genes (Wolfe, 2001). Therefore, while the preferential attachment mechanism may be applied to genes that are created *de novo* from non-coding sequences (Figure 57), it cannot be valid to explain the evolution of paralogous genes.

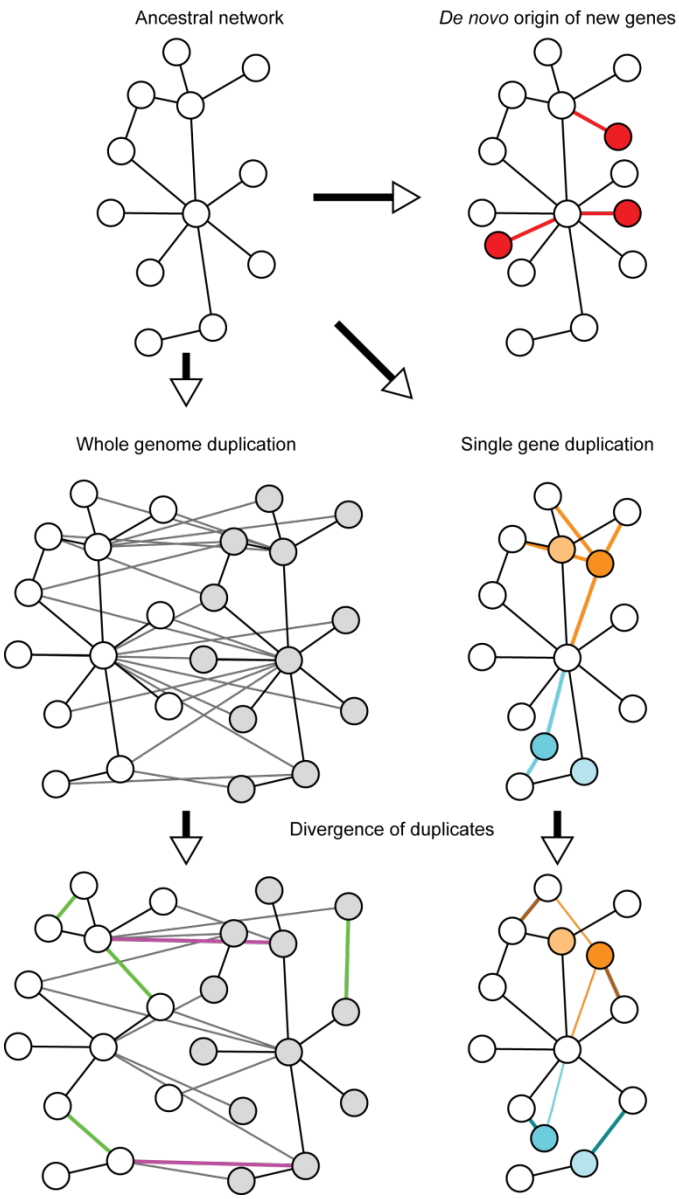


Figure 57: three models of network evolution

The network evolves in different ways, depending on mechanisms of appearance of new genes. Genes that originate *de novo* from non-coding DNA sequences (red) preferentially attach to existing hubs. In case of single-gene duplication (orange and cyan), the new gene will have the same interactions of its paralog. Through neo- and subfunctionalization the two paralogs diverge, gaining new interactions and losing others. After whole genome duplication, all interactions are duplicated, both lowly connected peripheral and highly connected central nodes. The divergence of duplicates allows the rewiring of interactions (green and violet).

The "duplication and divergence" model takes into account the fact that the principal source of new genes is the duplication of existing genes (Evlampiev and Isambert, 2008). Right after the duplication, the two paralogs have identical interactions.

Both paralogs undergo positive selection and diverge, losing part of their interactions and gaining new ones, following the mechanisms of sub- and neofunctionalization. Peripheral nodes are more likely to retain duplications because they have fewer connections and include a smaller fraction of essential proteins than central nodes. The presence of a second copy creates smaller perturbations in nodes that have few connections. Therefore the network periphery evolves more quickly than the network core (Figure 57).

A further level of complexity to the evolution of protein interaction networks is given by the duplication of the entire genome. Several taxonomic groups underwent one or more rounds of whole genome duplications, such as plants (up to three rounds) (Proost et al.), fungi (Kellis et al., 2004) and vertebrates (two rounds, with a further round in fish) (Dehal and Boore, 2005; Nakatani et al., 2007). As a consequence of whole genome duplication, also protein-protein interactions are duplicated. Then neo- and subfunctionalization occur and some interactions are lost, while new ones are created (Figure 57). However, the difference compared to single-gene duplications is that through whole genome duplication also the duplication of hubs may be retained. Furthermore, also dosage-sensitive genes may retain their duplicates, because, having all their interactors duplicated, the dosage balance remains unchanged. Another consequence of the two rounds of whole genome duplication in the ancestor of vertebrates may be the increase of organismal complexity. Two facts indirectly support this hypothesis. First, the number of genes that constitute gene families positively correlates with the organismal complexity (Vogel and Chothia, 2006). Second, the expansions of vertebrate gene families correlate with each other (Vogel and Chothia, 2006), suggesting that they followed similar evolutionary paths.

Human recent hubs duplicated preferentially through the two rounds of whole duplications of the early vertebrate genome. We have no evidence to understand whether they were already singleton hubs or they gained the “hub” status after duplication, although we are able to make some speculations. Following the “duplication and

divergence" theory of network evolution, the vertebrate protein interaction network underwent massive rearrangements after whole genome duplication, with partial loss of redundant interactions and formation of *de novo* interactions. Given that the network periphery is under positive selection (Kim et al., 2007), young proteins (*i.e.* proteins that originated with metazoans or later) underwent substantial rewiring and gained many interactions. The network core, instead, was able to resist to these rearrangements and ancestral hubs retained their singleton status. However the signal of ancestral singleton proteins to be more connected than duplicated proteins is weaker in human than in the other protein interaction networks (Figure 39, Table 8). This may be due to the fact that a fraction of human ancestral hubs was able to retain duplications.

Genes that duplicated through whole genome duplication do not undergo further small-scale duplications or copy number variations, while genes duplicated via small-scale duplications are more prone to copy number variations (Makino and McLysaght, 2010). Therefore whole genome duplications allow the retention of the duplication of dosage-sensitive genes, which could not duplicate otherwise. The fact that the dosage of human duplicated hubs is tightly regulated is evident also for other reasons. These genes are regulated post-transcriptionally by miRNAs. These short RNA strands control the dosage of a gene by coupling with its 3' UTR, thus allowing mRNA degradation. miRNAs control the dosage of duplicated hubs: although the signal is not statistically significant, independently from the origin, the fraction of duplicated hubs that are miRNA targets is higher than singletons (Figure 43).

An example that explains how the dosage of duplicated hubs is accurately regulated by miRNAs is represented by PTEN and its highly conserved paralog PTENP1. Although it is an ancient gene (it has orthologs in prokaryotes), PTEN is a duplicated hub that occupies a central position in the human protein interaction network. Its paralog is a processed pseudogene located on chromosome 9, which has 97% of sequence conservation with PTEN. Also part of the PTENP1 3' UTR is highly identical to PTEN: the 5' region

has 95% conservation, the 3' region has 50% conservation, but PTENP1 3' UTR lacks the 3'-most 1 kilobase. PTENP1 does not encode a functional protein because of a missense mutation in the initiatior methionine codon (Fujii et al., 1999). A recent study by Poliseno *et al*. (Poliseno et al.) showed that the binding sites for five miRNAs are highly conserved between PTEN and PTENP1 3'UTRs (miR-17, miR-19, miR-21, miR-26 and miR-214) (Figure 58). The gene expression levels of the two paralogs are highly similar in normal tissues and their mRNA abundance is lowered upon the expression of miR-19b and miR-20a. Furthermore, the overexpression of the 3' UTR of one of the two paralogs results in the downregulation of both PTEN and PTENP1 (Poliseno et al.). These observations show that the expression of both paralogs is regulated by the same elements. In particular PTENP1 functions as a decoy of PTEN-targeting miRNAs, allowing a higher expression level of PTEN by sequestering the miRNAs that should target it (Poliseno et al.).

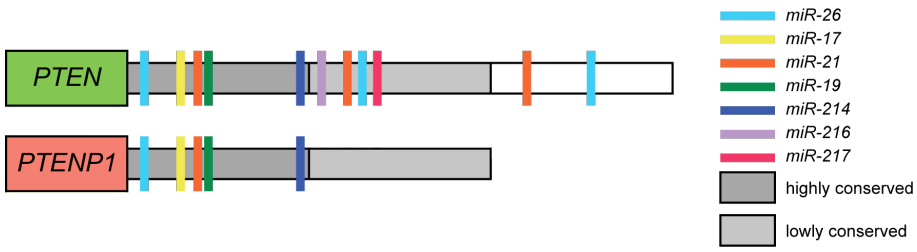


Figure 58: miRNA regulation of PTEN and PTENP1

Adapted from (Poliseno et al., 2010). The 3'-UTR of PTEN and PTENP1 includes three regions: highly conserved (dark gray), lowly conserved (light gray) and a PTEN-specific region. miRNA binding sites are highlighted by the vertical colored lines.

These results show that the duplication of PTEN did not cause catastrophic perturbations in the human protein interaction network. Although its paralog does not encode functional proteins, its expression helps the regulation of PTEN, because it acts as a decoy for the miRNAs that regulate PTEN, thus increasing the dosage of PTEN (Salmena et al., 2011). A similar example is given by KRAS and its processed pseudogene paralog KRAS1P. The overexpression of KRAS1P *in vitro* results in increased KRAS

mRNA abundance, because of this dosage imbalance induced between the two paralogs (Poliseno et al., 2010).

In addition to being duplicated through whole genome duplication and being tightly regulated by miRNAs, human duplicated hubs that appeared recently in evolution are mostly tissue-specific. Therefore the retention of more than one copy of a single gene does not imply an increased dosage, but subfunctionalization, intended as diversification of the gene expression in different tissues. Hence the two paralogs are duplicated at the genomic level but their function retains a "singleton" status. The expansion of tissues and cell types with vertebrates (Vogel and Chothia, 2006) may have helped this particular type of subfunctionalization.

Being singleton and essential, deletion of hubs brings to cell death in *S. cerevisiae* (Hughes and Friedman, 2005; Prachumwat and Li, 2006). This demonstrates that the network is not robust towards dosage modifications of hubs, since they cannot retain duplicates and cannot be lost by the cell. In *H. sapiens*, different evidence suggests the same conclusion, taking also into consideration the presence of the new class of duplicated hubs. First, genes that duplicated through whole genome duplication do not undergo further duplications (Makino and McLysaght, 2010). Second, the dosage of duplicated hubs is controlled by alternative mechanisms, such as miRNAs or tissue-selective expression. Third, mutations in the sequence of hubs are often associated with disease. Germline mutations that are associated with disease do not affect hubs (Goh et al., 2007), while somatic mutations of hubs are often involved in tumorigenesis (Jonsson and Bates, 2006; Rambaldi et al., 2008).

Previous analyses showed that cancer genes are mostly singleton and encode proteins with higher connectivity than the rest of human genes (Jonsson and Bates, 2006; Rambaldi et al., 2008). This holds true when analyzing cancer genes as a unique entity. However, we were able to distinguish between two categories of cancer genes, which are significantly different: dominant and recessive cancer genes (Futreal et al., 2004).

Dominant genes require only mutations in one allele to develop cancer and are associated with gain-of-function mutations (Vogelstein and Kinzler, 2004). Recessive genes, instead, need loss-of-function mutations in both alleles to start tumorigenesis (Vogelstein and Kinzler, 2004). Both dominant and recessive genes are more connected and more central than the rest of human genes: hubs represent the 25% most connected nodes inside the human protein interaction network, but they include more than half of dominant genes and three quarters of recessive genes (Figure 59). Dominant and recessive genes represent two subgroups of cancer genes with strikingly different characteristics. Recessive genes are ancient and associated with basic cellular processes, while dominant genes are more recent (*i.e.* they are enriched in genes that originated with metazoans) and involved in regulatory functions and processes related to multicellularity (Figure 46, Table 18). A further difference is represented by duplicability. Although all cancer genes are less duplicated than the rest of human genes, their duplicability depends on the origin of the genes. Ancient cancer genes are less duplicated than the rest of human genes, while, among younger genes, dominant and recessive genes have different behaviors: recessive genes that originated with metazoans or later are never duplicated, while dominant genes with the same origin are enriched in duplicated genes, compared with the rest of the human genome (Figure 47).

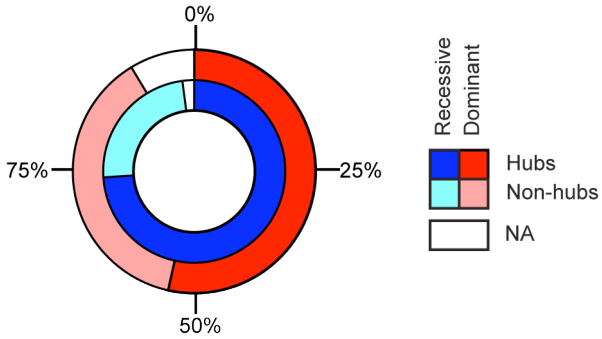


Figure 59: cancer hubs

The pie charts show the distribution of hubs among recessive and dominant genes. The white part represents the genes that are not included in the protein interaction network. Hubs are defined as the top 25% most connected nodes in the human protein interaction network.

Recessive and dominant cancer genes roughly correspond to caretakers and gatekeepers, which promote two distinct mechanisms of tumorigenesis (Domazet-Loso and Tautz, 2010; Kinzler and Vogelstein, 1997). Caretakers are responsible of the maintenance of genome stability and are involved in DNA repair (recessive genes), while gatekeepers are mainly involved in regulatory processes and directly control cell proliferation (dominant genes) (Hanahan and Weinberg; Kinzler and Vogelstein, 1997). Therefore, while in the first case mutations impair mechanisms that are directly related to proteins that interact with DNA and increase the mutation rate of the cell's DNA, mutations in gatekeeper genes spoil regulatory processes, hence perturbing normal interactions that are necessary for the correct cell functioning. However, tumorigenesis arises because the function of hubs, which are fragile components of the human protein interaction network, is impaired. This promotes the perturbation of a number of proteins and cellular processes that is proportional to the number of interactions that involve the mutated hub.

Having identified the complex mechanisms that regulate the evolution of protein interaction networks, we investigated what rules control other types of biological networks. In particular, we analyzed the human genetic interaction network, with the aim at identifying the network structure and understanding the characteristics of cancer genes. Given the absence of databases that collect genetic interactions in human (only less than 300 interactions are present in BioGRID), we developed a method to identify genetic interactions in human. We reconstructed the human genetic interaction network and exploited both protein interaction and genetic interaction network data to develop a method to detect new putative cancer genes (Appendix 2).

Appendix 1 – Network of Cancer Genes: a web resource for integration and analysis of gene and network properties of cancer genes

In order to collect and organize the information about evolutionary and protein interaction network properties of cancer genes, with the help of Adnan Syed, we built the Network of Cancer Genes (NCG, http://bio.ifom-ieo-campus.it/ncg), a web application that stores information on several systems-level properties of cancer genes. The first version of NCG was published in 2008 (Rambaldi et al., 2008), and we updated in 2010 (Syed et al., 2010). Together with Vera Pendino and Shruti Sinha, we have now completed a second updated version, with several improvements.

1. Database description

NCG collects information about origin, orthology relationships, duplicability and function of cancer genes, together with their interactions with miRNAs and the network properties of their encoded proteins.

The newest version of NCG (NCG 3.0) includes data for 1,494 cancer genes, gathered from several different sources (Table 2):

- 498 genes from the Cancer Gene Census (Futreal et al., 2004) (updated at March 22nd 2011) and from the census of amplified genes in cancer (Santarius et al.). These are known cancer genes, *i.e.* genes whose involvement in cancer has already been experimentally demonstrated;
- 698 candidate cancer genes from 18 high-throughput mutational screenings. These genes were selected among the 7,439 that were found mutated because of their high mutation frequency among different samples;

- 454 mutated genes in 11 whole genome sequencing studies of 39 cancer patients.

We were not able to include all genes from the original data as they were extracted from the corresponding experiments, because for five we could not associated an up-to-date Entrez IDs.

The definition of duplicability was explained in the Methods section. Briefly, a gene is duplicated if it has a second hit on the genome that spans at least 60% of its length. In addition to this threshold, on the website the user may choose different levels of sequence conservation in order to analyze more or less conserved paralogs of the gene of interest.

We modified the dataset of unique human genes, in order to have the most recent updates. Instead of using RefSeq, we chose GenCode v.7.0 (Harrow et al., 2006). This is a highly curated set of unique human genes that is used as reference for the ENCODE project, the 1000 Genomes Project and to capture baits of the whole human exome (Coffey et al., 2011). It is therefore likely to contain all genes that are found mutated in current and future high-throughput mutational screenings of cancer exomes. As with RefSeq proteins, we aligned all the 84,408 protein sequences (20,700 genes) to the human genome build hg18, using BLAT (Figure 20). We were able to positively align the 99.2% (83,769) of these sequences. After the application of all the filters described in Figure 20, we retrieved 19,560 genes. The loss of many of the 1,140 genes upon the application of the filter for isoforms was due to the presence of many sequences that span two consecutive genes. These sequences seem to be errors in the Ensembl database, since their only supports are Ensembl transcripts that include exons from clearly different genes (Figure 60). For these sequences, there is no evidence of other mRNAs or ESTs. In these cases, the clusters of sequences with overlapping best hits span both genes and only the longest isoform is retained. In order to overcome this problem, we made the union between the coordinates of the 19,960 genes from GenCode and the results from the application of the pipeline to the

most recent version of RefSeq (version 46, 19,356 genes after the removal of isoforms) (Figure 60). We were able to add 971 genes, bringing the number of unique genes to 20,531.

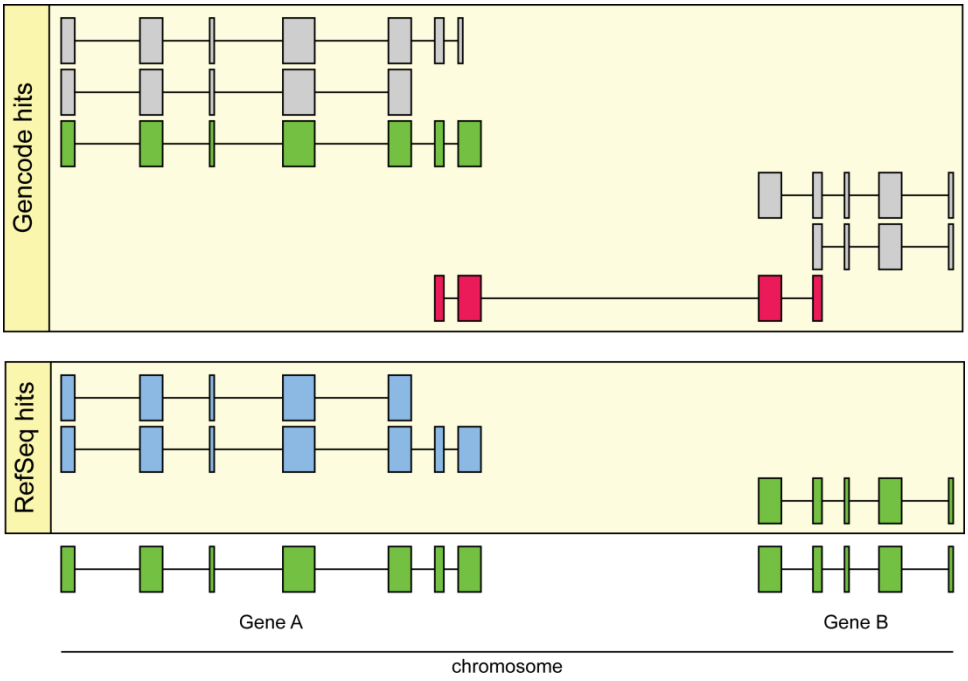


Figure 60: Loss of genes in the genome alignment pipeline

The figure shows an example of a misidentification of genes due to the presence of an isoform (magenta) that spans two consecutive genes. When considering only the Gencode hits, the six isoforms of Gene A and Gene B are overlapping, therefore only the longest isoform (green) is retained and Gene B is completely lost. When aligning RefSeq sequences to the genome, both genes are retained, because the spurious isoform is not present. The final union of the results from both Gencode and RefSeq isoform filters allows the retrieval of both Gene A and Gene B.

We updated the orthology information to the most recent release of eggNOG (Muller et al., 2010) (version 2.0), which include 20,122 genes. The new protein-protein interaction data was retrieved from the most recent releases of the databases described in Table 3. This allowed us to define 98,492 interactions between 13,531 proteins (Table 19).

Table 19: integration of protein interaction networks from different sources

Dataset	Version	Nodes	Interactions	Publications
BioGRID	3.1.75 (Apr 1st 2011)	8,940	35,893	11,491
IntAct	138 (Mar 2nd 2011)	8,785	34,677	2,090
MINT	Dec 15th 2010	5,428	13,309	2,709
DIP	Oct 10th 2010	4,669	12,422	2,263
HPRD	9 (Apr 13th 2010)	8,897	37,026	18,837
Total		**13,531**	**98,492**	**25,915**

Primary protein interaction network data are gathered from five sources. The number of nodes is calculated as the number of Entrez IDs that it is possible to associate to the original data from each database. Interactions are non-redundant.

The integration of the newest data in terms of duplicability, orthology and protein interaction networks allowed us to have information of 23,535 genes (Figure 61). For 12.161 genes (51.7%) we were able to retrieve duplicability, orthology and network information (Figure 61).

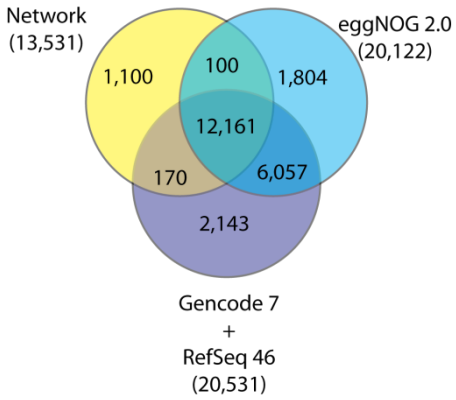


Figure 61: Orthology visualization in NCG

The figure shows the number of genes that have information from the three sources. The total number of genes that have information from at least one dataset is 23,535.

The new version of NCG allows the user to gather information about the interactions between cancer genes and miRNAs. Tarbase v.5 (June 2008) (Papadopoulos et

al., 2009) and miRecords v.1 (August 15th 2008) (Xiao et al., 2009) were used to gather these interactions and the information about what genes are hosts of miRNAs (see Methods section).

We also exploited the functional analysis of hubs to create a section of functional information of cancer genes. For each gene the user may retrieve the functional classes that are associated to it and the GO terms from Biological Process at level 5 and 6. The data collected in NCG are stored in a MySQL database. The web interface to interrogate the database is built in Perl. We adopted Cytoscape Web (Lopes et al., 2010) to visualize both the protein interaction networks and the miRNAs-cancer genes networks.

2. Web interface

The user may retrieve information on cancer genes in three ways:

1. Search for a specific gene or a list of genes by using one of the supported identifiers (Entrez ID, RefSeq ID, Ensembl protein ID or gene symbol);
2. Select a pre-compiled list of cancer genes, which includes all the genes from a single study or from one of the three types of cancer genes;
3. Select all genes that have similar properties. The user may choose all cancer genes that have peculiar characteristics, such as duplicability, function, origin, network properties or tissues where the genes were found mutated.

The primary output of the query is divided into six sections:

1. The summary table that provides several links to external databases, in addition to the gene information and its involvement in cancer. If the user is interested in the involvement of a gene in cancer, a link opens a new page with the description of all the experiments that found it mutated;
2. The duplicability report, which describes whether the gene has duplications above 60% of its length;
3. The description of its origin;

4. The description of its network properties;
5. The report of its interactions with miRNAs, *i.e.* whether it is target or host of miRNAs;
6. The functional report, with the list of all functional classes that are associated to it.

Each of these sections allows the user to open a new page, which describes the gene properties in detail. The duplicability page displays all the duplications of the gene of interest, with information about the duplicated loci (*i.e.* whether they overlap a real gene or an intergenic region). The user may set different thresholds of duplicability in order to study more or less conserved paralogs. The orthology page shows the tree of life with information about the presence or absence of orthologs and a table with all the orthologs of the gene of interest (Figure 62). The network page shows all the interactions of the protein of interest and between its primary interactors (Figure 63), in addition to a brief report that describes the properties of its interactors. The miRNA page is very similar to the network page: it shows all the interactions that involve the miRNAs that target the gene of interest (Figure 64). The function page displays all the Biological Process GO terms at level 5 and 6 that are associated to the gene of interest.

The user may also search directly for miRNAs and their cancer targets, in several ways:

1. Search for a particular miRNA or cancer gene that is a target of miRNAs;
2. Retrieve the list of all the 118 cancer genes that are targets of miRNAs;
3. Retrieve the list of all the 55 cancer genes that are host of miRNAs.

In the first case the miRNAs page will be displayed, while in the latter two cases the result page will be shown with the properties of all cancer genes that are target or host of miRNAs.

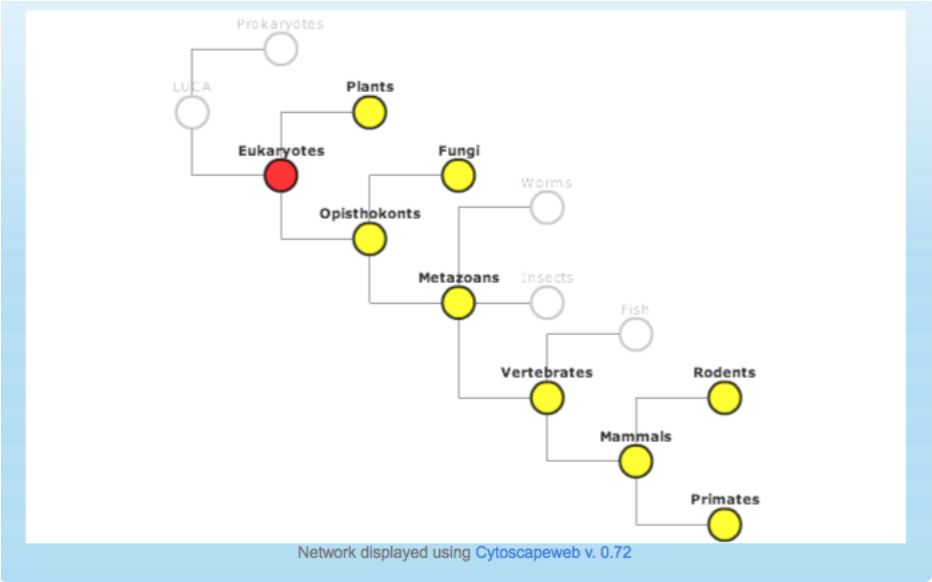


Figure 62: Orthology visualization in NCG

The figure shows the orthology relationships of ADAM metallopeptidase domain 29 (ADAM29). The tree of life shows the presence (yellow) or absence (white) of orthologs for a gene of interest. The red node represents the origin of the gene.

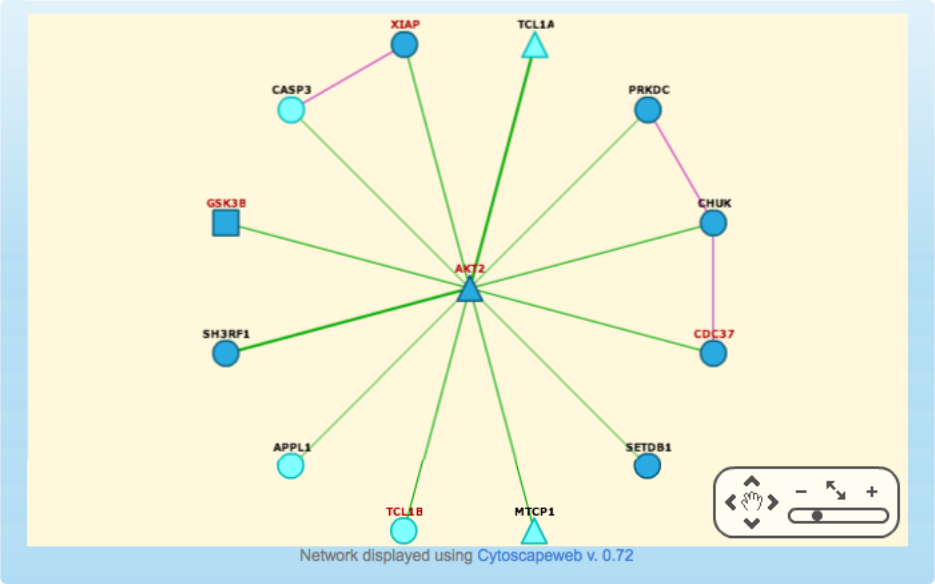


Figure 63: protein interaction network visualization in NCG

The figure shows the 12 interactors of c-abl oncogene 1, non-receptor tyrosine kinase (AKT2). Triangles represents genes from the cancer gene census, squares are candidate cancer genes. Dark blue nodes represent ancient proteins, while cyan represents recent proteins. Primary interactions (*i.e.* interactions that involve AKT2) are depicted in green, while secondary interactions (*i.e.* interactions between the interactors of AKT2) are depicted in violet. Thick lines correspond to interactions detected by more than one expreriment, while

thin lines are detected only in one experiment. Red proteins are duplicated (*i.e.* they have an additional hit that covers than 60% of their length), while black proteins are singleton.

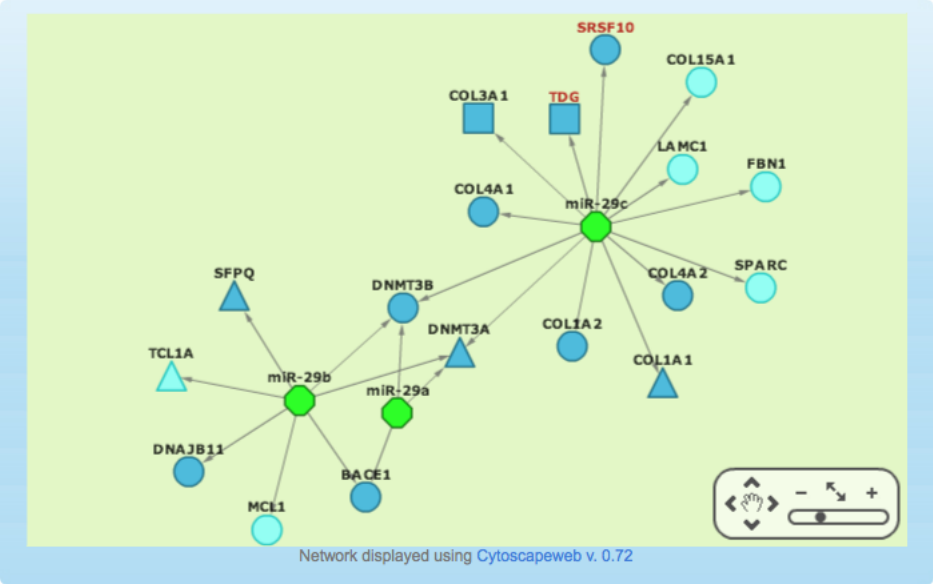


Figure 64: miRNA-gene interactions in NCG

The figure shows the interactions between miRNAs and DNA (cytosine-5-)-methyltransferase 3 alpha (DNMT3A). miRNAs are depicted in green, while the gene properties are shown as explained in Figure 63.

Appendix 2 – The human genetic interaction network

1. Introduction

Genetic interactions were used in *S. cerevisiae* to identify the presence of functional relationships between genes and discriminate between two distinct types of paralogs (VanderSluis et al.). In particular, negative genetic interactions may imply functional redundancy between two genes (VanderSluis et al.). The application of this concept to paralogs entails the identification of two definite classes of duplicated genes: functional paralogs and dosage paralogs (VanderSluis et al.). Functional paralogs have at least partially redundant functions. They have negative genetic interactions between them and few negative genetic interactions with other genes because they functionally buffer one another (VanderSluis et al.). Therefore the deletion of one paralog is not lethal because the other is able to at least partially restore the original function. Dosage paralogs instead do not functionally buffer one another, are highly identical and are involved in the same function (VanderSluis et al.). They have similar genetic interactions and, in order to correctly accomplish their function and maintain the dosage balance, both functional copies are needed, therefore the deletion of a single copy has a severe effect (VanderSluis et al.).

2. Genetic interactions between cancer genes and their paralogs

Assuming the hypothesis that the distinction between functional and dosage paralogs exists also in human, duplicated cancer genes should be depleted only in dosage duplicates, because both copies must be functional in order to accomplish their function correctly. Indeed we found that, only considering highly conserved duplications, cancer

genes are less duplicated than the rest of human genes, while they are enriched in lowly conserved duplications (less than 30% sequence identity) (Figure 50).

We divided the 35 cancer genes from the Cancer Gene Census with highly conserved duplicates into three possible categories of duplicates (Table 20). Since some duplicated cancer genes are reciprocal paralogs, we identified 27 distinct clusters of paralogs. By literature search, we identified the majority (16, 63%) as diverging paralogs: nine have different expression (*i.e.* their expression levels are different or they are expressed in different tissues) and eight have different function. Only the lysine (K)-specific demethylase 5C (KDM5C) and its paralog KDM5D are dosage duplicates. KDM5C is a recessive gene that is involved in clear cell renal carcinoma, while its paralog has never been associated with cancer. They are both histone demethylases that repress SMAD3 activity (Kim et al., 2008). They both may act as monomers, heterodimers and KDM5C also as homodimer, activating different transcription factors (Kim et al., 2008) (Figure 65). The dosage of both paralogs must be tightly regulated, in order to activate the right transcription factors (Kim et al., 2008). When KDM5C is mutated, its function is impaired and it cannot activate any transcription factor (Kim et al., 2008). KDM5D, instead, is wild type and can activate its targets as monomer, thus producing aberrant transcription factor activation.

For five clusters of paralogs (seven cancer genes) we were not able to determine the paralog type, while we identified the last five genes as functional duplicates: clathrin heavy chain (CLTC), pre-B-cell leukemia homeobox 1 (PBX1), guanine nucleotide binding protein q polypeptide (GNAQ), SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily a, member 4 (SMARCA4) and Von Hippel-Lindau tumor suppressor (VHL).

Table 20: highly conserved paralogs of cancer genes

Cancer Gene(s)	Paralog(s)	Coverage (%)	Identity (%)	Mutation type (CGC)	Cancer Type (CGC)	Paralog type
LCP1	PLS3	62,84	84,27	T (BCL6)	NHL	different expression
MSN	RDX	82,05	86,7	T (ALK)	ALCL	different expression
PAX3	PAX7	86,73	85,84	T (FOXO1A, NCOA1)	alveolar rhabdomyosarcoma	different expression
POU5F1	POU5F1B	73,49	95,82	T (EWSR1)	sarcoma	different expression
PRKAR1A	PRKAR1B	95,83	87,51	T (RET), M, N, F, S	papillary thyroid	different expression
SEPT6	SEPT11	65,58	85,25	T (MLL)	AML	different expression
SH3GL1	SH3GL2	78,8	80,29	T (MLL)	AL	different expression
TPM3, TPM4	TPM1, TPM2, TPM3, TPM4	66,49	88,46	T (Alk, NTRK1)	papillary thyroid, ALCL	different expression
UTX	UTY	65,02	87,42	D, N, F, S	renal, oesophageal, myoepithelioma	different expression
AKT1, AKT2	AKT1, AKT2, AKT3	62,79	89,38	M, A	breast, colorectal, ovarian, NSCLC, pancreatic	different function
FCGR2B	FCGR2A, FCGR2C	68,39	93,36	T (?)	ALL	different function
HRAS, KRAS, NRAS	HRAS, KRAS, NRAS	83,6	86,1	M	many	different function
KLK2	KLK3	75,52	84,38	T (ETV4)	prostate	different function
MYH9, MYH11	MYH9, MYH11, MYH10	63,43	80,34	T (ALK, CBFB)	ALCL, AML	different function
NPM1	CLEC2D	62,5	92,67	T (ALK, RARA, MLF1), F	NHL, APL, AML	different function
RARA	RARB, RARG	66,93	84,12	T (PML, ZNF145, TIF1, NUMA1, NPM1)	APL	different function
KDM5C	KDM5D	78,08	91,75	N, F, S	clear cell renal carcinoma	dosage duplicates
CLTC	CLTCL1	77,79	85,95	T (ALK, TFE3)	ALCL, renal	functional duplicates
GNAQ	GNA11	89,97	89,94	M	uveal melanoma	functional duplicates
PBX1	PBX3	60,79	92,31	T (TCF3, EWSR1)	ALL, myoepithelioma	functional duplicates
SMARCA4	SMARCA2	61,41	87,14	F, N, M	NSCLC	functional duplicates
VHL	VHLL	61,05	76,71	D, M, N, F, S	renal, hemangioma, pheochromocytoma	functional duplicates
DUX4	LOC653543	82,27	99,01	T (CIC)	soft tissue sarcoma	unknown
EIF4A1, EIF4A2	EIF4A1, EIF4A2	86,49	92,13	T (BCL6)	NHL	unknown
HSP90AA1, HSP90AB1	HSP90AA1, HSP90AB1	71,55	87,48	T (BCL6)	NHL	unknown
LMO1	LMO3	69,7	95,52	T (TRDA)	ALL	unknown
SSX1, SSX2	SSX1, SSX2, SSX3, SSX4, SSX4B	61,11	81,48	T (SS18)	synovial sarcoma	unknown

The 35 cancer genes from the cancer gene census (CGC) that have paralogs at 60% conservation are shown. Orange represents dominant genes, while blue represents recessive genes. Coverage corresponds to the percentage of the gene length that is conserved in its paralog. The mutation type is derived from the cancer gene census: A, amplification; D, large deletion; F, frameshift; N, nonsense; S, splice site; T translocation. In case of translocations, the translocation partner is indicated. The cancer type is also derived from the cancer gene census: NHL, non-Hodgkin lymphoma; ALCL, anaplastic large-cell lymphoma; AML, acute myeloid leukemia; AL, acute leukemia; NSCLC, non-small cell lung cancer; APL, acute promyelocytic leukemia. The paralog type may be functional, dosage, diverging (different function or expression in different tissues) or unknown. These information retrieved from a literature-based search.

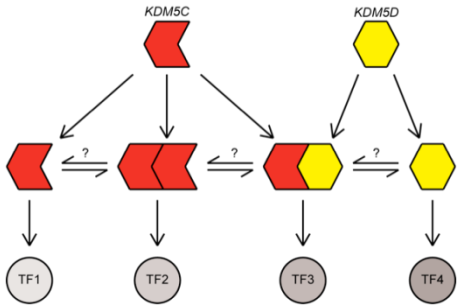


Figure 65: KDM5C and KDM5D

KDM5C and KDM5D may act as monomers, homodimers (only KDM5C) or heterodimers, each activating different transcription factors (Kim et al., 2008).

CLTC is a ubiquitous protein involved in receptor-mediated endocytosis, intracellular trafficking and recycling of receptors (Hood and Royle, 2009). Its active form is a trimer, with each subunit bound to a clathrin light chain (Hood and Royle, 2009). Its paralog has an equivalent function and is also expressed ubiquitously (Hood and Royle, 2009). CLTC was found translocated in cancer with TFE3 and ALK. TFE3 is a transcription factor involved in cell growth and proliferation. The fusion protein acts as aberrant transcription factor, with the promoters of CLTC and the DNA-binding activity of TFE3 (Argani et al., 2003) (Figure 66). ALK is a receptor tyrosine kinase involved in nervous system development. The fusion protein includes almost all the CLTC sequence, which most likely does not lose its function, while ALK is activated in a ligand-independent way (Gascoyne et al., 2003; Patel et al., 2007) (Figure 66).

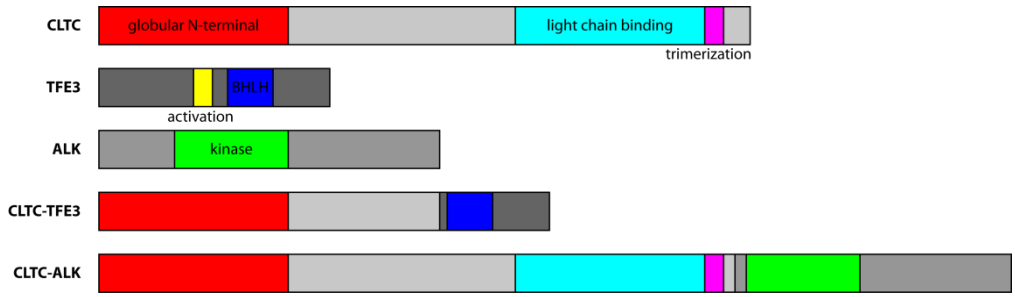


Figure 66: CLTC and its translocations in cancer

CLTC possesses a globular N-terminus and two C-terminal domains: the light chain binding and the trimerization domain. TFE3 has an activation domain and a basic helix-loop-helix (BHLH) domain that binds DNA. ALK has a kinase domain. The CLTC-TFE3 chimeric protein loses the CLTC C-terminus and the TFE3 N-terminus, but keeps the DNA-binding properties of TFE3 (Argani et al., 2003). The CLTC-ALK translocation maintains almost all the CLTC structure and the kinase activity of ALK, which remains constitutively active {Gascoyne, 2003 #216; Patel et al., 2007).

PBX1 is a ubiquitous transcription regulator associated to HOX proteins. It is frequently translocated with TCF3, a widely expressed transcription factor, in acute lymphoid leukemia (Hunger, 1996), and with EWSR1, a ubiquitous RNA-binding protein involved in splicing regulation, in myoepithelioma (Antonescu et al.).

GNAQ is part of heterotrimeric G-proteins involved in NF-κB activation. Its mutations in uveal melanoma impair the GTP binding site, keeping the G protein always active (Van Raamsdonk et al., 2009). Its paralog GNA11 was recently found mutated with high frequency in the same tumors, with a mechanism highly similar to GNAQ mutations (Van Raamsdonk et al.). This implies that not only the function of the genes is maintained between the paralogs, but also the mechanisms that promote tumorigenesis are conserved.

Both SMARCA4 (also known as BRG1) and its paralog SMARCA2 (BRM) are part of the SWI/SNF chromatin-remodeling complex, which regulates the expression of 5-7% of all genes in yeast (Sudarsanam et al., 2000; Zraly et al., 2006) and between 1-2% and the whole genome in *D. melanogaster* (Armstrong et al., 2002). SMARCA4 and SMARCA2 have 61% conserved sequence and share the same domains. They differ for one expansion repeat of 33 glutamines (polyQ) present only in SMARCA4 (Figure 67)

(Reisman et al., 2009). Their functions are redundant *in vivo* (Reisman et al., 2009; Reisman et al., 2002), although it was demonstrated that only the knockout of SMARCA4 is embryonic lethal in mouse (Reisman et al., 2002), while only SMARCA2 is epigenetically silenced *in vitro* (Glaros et al., 2007; Mizutani et al., 2002; Yamamichi et al., 2005). Although the two paralogs seem very similar in terms of both domain composition and function, only SMARCA4 presents strong evidence of involvement in cancer and is included in the Cancer Gene Census.

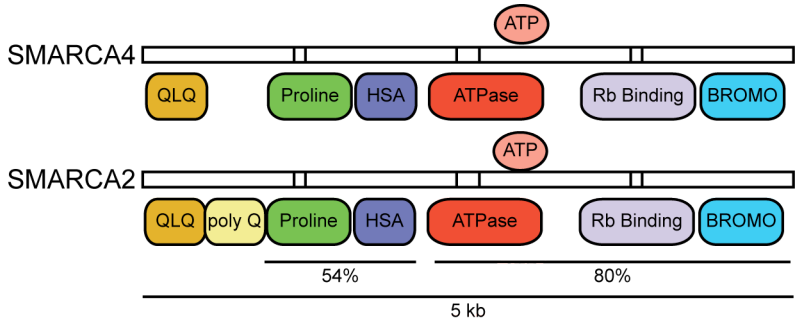


Figure 67: domain composition of SMARCA4 and SMARCA2

Adapted from (Reisman et al., 2009). The figure shows that SMARCA2 and SMARCA4 have a very similar domain composition. QLQ, glutamine-leucine/glutamine domain; poly Q, polyglutamine, HSA, helicase/*SA*NT-associated domain.

VHL and its paralog von Hippel-Lindau like (VHLL) are involved in the response to hypoxia by regulating hypoxia-inducible factor α (HIFα). VHL is a component of the E3 ubiquitin ligase complex and contains two domains: the α domain is required for the binding to the rest of the E3 ubiquitin ligase complex, while the β domain binds directly HIFα (Stebbins et al., 1999). Under normal oxygen concentrations, HIFα has a hydroxylated proline residue that is required for the binding with VHL, which promotes HIFα ubiquitination and degradation by the E3 ubiquitin ligase complex (Figure 68). VHLL shares 60% sequence conservation with VHL, but does not have the α domain. Indeed it is able to bind HIFα and protects it from ubiquitination by the E3 complex (Qi et al., 2004). Mutations in VHL have been associated with several tumor types (Leung and

Ohh, 2002; Maher and Kaelin, 1997). Mutated VHL is unable to recognize either the E3 complex or HIFα, which remains always active and induces the transcription of several target genes, such as VEGF, EPO, TF, TFRC, and GLUT1, keeping the cell in an inappropriate response to hypoxia (Leung and Ohh, 2002).

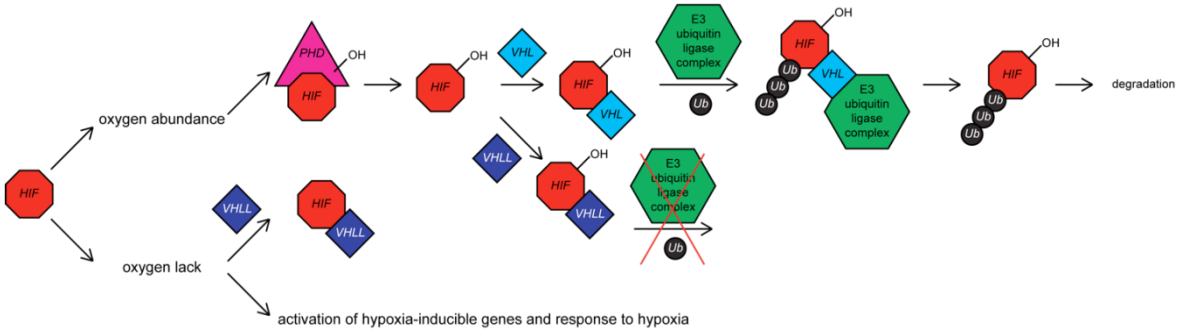


Figure 68: VHL and VHLL pathway

The mechanism of action of VHL and VHLL is shown. VHL binds to hydroxylated HIF α and allows its degradation by binding to E3 ubiquitin ligase complex (Stebbins et al., 1999). VHLL, instead, binds to hydroxylated HIFα and sequester it from degradation because it cannot bind to E3 ubiquitin ligase complex. In case of lack of oxygen, VHL cannot bind HIFα, which remains free or bound to VHLL and can activate the hypoxia-inducible genes (Qi et al., 2004).

The detection of cancer genes that have functional paralogs has two major implications:

1. The discovery of new pharmaceutical targets of tumors with one of these genes mutated, and
2. The identification of new candidate cancer genes.

The functional redundancy between two paralogs may imply a similar involvement in tumorigenesis of both paralogs, as was demonstrated for GNAQ and GNA11 (Van Raamsdonk et al.). It was also shown that mutations of these two paralogs are mutually exclusive (Van Raamsdonk et al.), indirectly indicating the presence of negative genetic interactions between functional paralogs in human.

In *S. cerevisiae*, functional paralogs have negative genetic interactions between them, because they are able to buffer one another's function (VanderSluis et al.). When

one paralog is deleted, the other is able to rescue its original function, but, if also the other paralog is deleted, the cell has lower fitness than the expected product of the two single deletions. In human, homozygous mutations in tumor suppressor genes impair their function, promoting tumorigenesis. This roughly corresponds to a single gene deletion in yeast. If one manages to block the function of the tumor suppressor's paralog, the fitness of the cancer cell will be impaired. Among all cancer genes, we identified candidates that may be used to demonstrate this hypothesis. We analyzed several cancer cell lines to discover those that have homozygous mutations in duplicated cancer genes. We plan to infect these cells with short hairpin RNAs (shRNAs) that target the functional paralog of the duplicated cancer gene in order to prove that a synthetic lethal (or, at least, a negative genetic) interaction exists between the couple of paralogs. This experimental part will be performed by Francesco Nicassio in Pier Paolo Di Fiore's group.

2.1. Dataset of cancer cell lines

We retrieved mutation data in cancer cell lines from two sources inside the Cancer Genome Project: NCI-60 (http://www.sanger.ac.uk/genetics/CGP/NCI60/) (Ikediobi et al., 2006) and the Cancer Cell Line Project (http://www.sanger.ac.uk/genetics/CGP/CellLines/). NCI-60 includes a set of 59 cancer cell lines derived from several tissues that have been extensively characterized: they were treated with more than 100,000 chemical compounds, their expression profile was characterized using several microarray platforms and their karyotype was determined. Furthermore, 24 cancer genes (13 dominant, 11 recessive) were sequenced in all 24 cell lines (Ikediobi et al., 2006). The Cancer Cell Line Project tries a systematic characterization of the genetics and genomics of a large number of cancer cell lines. To date, 67 cancer genes were sequenced in 683 cell lines.

In total, we were able to gather mutational data for 68 cancer genes in 694 cancer cell lines.

2.2. Detection of candidates for synthetic lethal screenings

We applied several filters to the mutation data in the 694 cell lines in order to detect candidate cell lines for screenings of synthetic lethal interactions between a cancer gene and its functional paralog:

1. The cancer gene must have at least one paralogous gene with 10% sequence conservation;
2. The cancer gene must have homozygous mutations in at least one cell line, which must have only one gene with homozygous mutations;
3. The mutations must be missense or nonsense.

The 68 cancer genes included 40 duplicated genes, of which 11 had a highly conserved paralog. Seven genes were further eliminated because they did not bear homozygous mutations in any of the 694 cell lines or they had homozygous mutations in cell lines with more than one homozygous mutation (Table 21). From the 144 cell lines with one homozygous mutation in one duplicated cancer genes, we retrieved 15 that bore frameshift or nonsense mutations. We focused only on these mutations, because they disrupt whole gene function by impairing the whole protein structure. We found eight genes with missense or nonsense mutations in these cell lines (Table 22). Three genes (PTEN, SMARCA4 and VHL) have a highly conserved paralog (PTENP1, SMARCA2 and VHLL, respectively). We decided to try to validate synthetic lethal interactions for VHL and SMARCA4, because their functional paralog is highly conserved and expressed in the same tissues (Qi et al., 2004; Reisman et al., 2009). PTEN was discarded because its paralog does not produce a functional protein, since it lacks the initiator methionine codon (Fujii et al., 1999).

Table 21: filters on cell lines

filter	cell lines	duplicated cancer genes	with highly conserved	with lowly conserved

filter		cell lines	duplicated cancer genes (N=68)	paralogs (≥60%)	paralogs (<60%)
Initial data		**694**	**40**	11	29
homozygous mutations		**144**	**33**	9	24
mutation type	frameshift	9	6	3	3
	nonsense	3	3	1	2
	deletion	2	2	1	1
	splice site	1	1	-	1
	total	**15**	**8**	3	5
ploidy	2n	**4**	**3**	PTEN, SMARCA4	TP53
	3n	3	3	VHL	ERBB2, TP53
	4n	3	2	PTEN	TP53
	NA	5	4	SMARCA4	JAK2, KDR, NOTCH1

From the initial data, three filters are applied to identify the best candidate cell lines: first, only cell lines with homozygous mutations in one gene are considered. Second, missense mutations are eliminated. Third, the filter on ploidy is applied.

Table 22: candidate cell lines

cell line	tissue	ploidy	cancer gene	mutation effect	highly conserved paralogs	lowly conserved paralogs
HUTU-80	small intestine	2n	SMARCA4	Helicase and chromatin-binding domains are lost	SMARCA2	
CCRF-CEM	haematopoietic	2n+/-	PTEN	PTP domain is lost	PTENP1 (*pseudogene*)	
HL-60	haematopoietic	2n+/-	TP53	Deletion (?)		TP73
NCI-H522	lung	2n+/-	TP53	DNA-binding motif is lost		TP73
K-562	haematopoietic	3n-	TP53	DNA-binding motif is lost		TP73
MDA-MB-468	breast	3n-	ERBB2	Almost all protein is disrupted		EGFR, ERBB4
A498	kidney	3n	VHL	C-terminus is lost (E2A binding?)	VHLL	
MOLT-4	haematopoietic	4n	PTEN	C2 domain is lost	PTENP1 (*pseudogene*)	
SK-OV-3	ovary	4n+/-	TP53	DNA-binding motif is lost		TP73
HOP-62	lung	4n+	TP53	DNA-binding motif is lost		TP73
NCI-H1573	lung	na	JAK2	Tyrosine kinase domain is lost		JAK3
NCI-H838	lung	na	KDR	Tyrosine kinase domain is lost		FLT1, FLT4
TE-1	oesophagus	na	NOTCH1	ANK repeats and C-terminus are lost		NOTCH2, NOTCH4
TE-11	oesophagus	na	NOTCH1	Transmembrane receptor domain is lost		NOTCH2, NOTCH4
GAMG	CNS	na	SMARCA4	chromatin-binding domain is lost	SMARCA2	

The 15 cell lines with homozygous mutations in one cancer gene are shown. Highly conserved paralogs refer to 60% conservation, while lowly conserved corresponds to 10% conservation.

We identified one cell line (A498) that has a frameshift mutation in VHL that disrupts its α domain, thereby inhibiting its role in the degradation of HIFα, and two cell lines (GAMG and HUTU-80) that have frameshift and nonsense mutations in SMARCA4, which both disrupt the chromatin-binding domain. Without this domain, the SWI/SNF complex cannot bind to chromatin, therefore it cannot regulate gene expression. For each of these cell lines we investigated the presence of negative genetic (or synthetic lethal) interactions between the mutated gene and its paralog. We also identified two controls for each cell line: these are cell lines from the same cancer tissue that did not present mutations in the analyzed genes (Table 23).

Table 23: candidate controls

cell line	Histology	Tissue	Ploidy	Mutations			Mutated Genes	
				Homozygous	Heterozygous	Total	Homozygous	Heterozygous
HUTU-80	**carcinoma**	**small intestine**	**NA**	**1**	**4**	**5**	**SMARCA4**	**ERBB2, RB1, TSC1, WT1**
HT-29	carcinoma	large intestine	3n+/-, Near-triploid 69+/- (58-80)	2	3	5	TP53, SMAD4	APC, BRAF, PIK3CA
SK-CO-1	carcinoma	large intestine	3n+, hypertriploid (73)	0	3	3		FLCN, PDGFRA, RET
GAMG	**glioma**	**central nervous system**	**NA**	**1**	**3**	**4**	**SMARCA4**	**FBXW7, NOTCH1, TET2**
SW1088	glioma	central nervous system	NA	1	1	2	SUFU	TET2
LN-405	glioma	central nervous system	NA	0	5	5		ALK, FBXW7, MLH1, SMO, TET2
A498	**carcinoma**	**kidney**	**3n, Triploid (69)**	**1**	**1**	**2**	**VHL**	**SMARCA4**
CAKI-1	carcinoma	kidney	3n, Triploid (69)	1	3	4	CDKN2A	MET, PTCH, RB1
ACHN	carcinoma	kidney	2n+/-, Near-diploid 46+/- (35-57)	1	1	2	CDKN2A	SMO

Two controls for each cell line are identified. Each control is the most similar to the cell line it refers to, in terms of histology, tissue, ploidy and mutated genes.

3. The human genetic interaction network

Since no data of genetic interactions is available in human, we constructed a map of genetic interactions involving cancer genes, by integrating the data from several large-scale studies of RNA interference (RNAi) in human cancer cell lines. The purpose of these studies was to identify genes that are essential for the survival of a cancer cell but not of a normal cell, in order to discover new drugs that may act on one of these essential genes in order to kill only tumor cells.

3.1. Construction of a cancer gene-centered network of genetic interactions

We identified 13 experiments that described essential genes in 64 human cancer cell lines (Table 24). 29 of these (45%) have at least one mutated gene reported in COSMIC (Forbes et al.) (Table 24). The other 36 were discarded from the analysis. We eliminated an additional cell line (HCC1954) because it was hypermutated. Indeed it had 89 mutated genes, while for the others the median value of mutated genes was 2. The 28 cell lines had 25 mutated genes in total, of which 24 were included in the Cancer Gene Census (11 dominant and 13 recessive).

Table 24: experiments used to identify genetic interactions of cancer genes

Study	Cancer	Normal_cells	Cancer_cells	Screened_genes
(Arora et al., 2010)	Ewing sarcoma	normal fibroblast	TC-32, TC-71, SK-ES-1, RD-ES	572 kinases
(Baldwin et al., 2010)	cervical intraepithelial neoplasia (CIN)	human foreskin keratinocytes (HFK)	Ca-Ski, SiHa, HeLa	88 kinases
(Barbie et al., 2009)	KRAS-driven cancers		19 cell lines	1028 (Moffat)
Bommi-Reddy2008 (Bommi-Reddy et al., 2008)	clear cell renal carcinoma	786-0, RCC-4 with reconstituted VHL	786-0, RCC-4	88 kinases
Firestein2008 (Firestein et al., 2008)	colorectal cancer		DLD-1, HCT116	1000
Gruenberg2008a (Grueneberg et al., 2008a)	non-small cell lung adenocarcinoma, kidney, cervical cancer	human foreskin keratinocytes (HFK), human foreskin fibroblasts (HFF)	HeLa, 293T, NCI-H1299, NCI-H358, NCI-H1975, NCI-H23, 786-0, A498, ACHN, Calu-1, A549, MCF7, 293T, WI38, BJ, MCF10A, Ca-Ski, SiHa, RKO	88 kinases
(Grueneberg et al., 2008b)	cervical, renal cancer		786-0, HeLa, Ca-Ski, siHA, C-33-A, ACHN, A498, RCC-4	88 kinases + 25 in other cell lines
Luo2008 (Luo et al., 2008)	small-cell lung cancer, non-small-cell lung cancer, glioblastoma, CML, lymphocytic leukemia		H82, H87-772, A549, H1650, H1975, HCC827, LN-229, U251, K-562, Jurkat, SUP-T1, REH	9500
Moffat2006 (Moffat et al., 2006)	colon cancer		HT29	1028
Schlabach2008 (Schlabach et al., 2008)	colon, breast cancer	HMEC	DLD-1, HCT116, HCC1954	2924
Silva2008 (Silva et al., 2008)	breast cancer		MCF-10A, MDA-MB-435, MDA-MB-231, ZR-65.1, T47D	?
Thaker2009 (Thaker et al., 2009)	glioblastoma	HA (control)	T98G, U87, U373-MG, A172, A549, LN-308, LN-428, HUVEC, SG388	5520
Tyner2008 (Tyner et al., 2008)	acute myeloid leukemia		K-562, CMK, HEL, HMC-1_1	tyrosine kinases
Yang2010 (Yang et al.)	osteosarcoma	HOB-c	KHOS, TC-71, U2-OS, MES-SA, SK-OV-3, OSA344, CS1, SS-1,	673 kinases

The number of screened genes is derived from the Methods sections of the respective experiments. Barbie *et al.* (Barbie et al., 2009) identified all the genes that are essential for KRAS-driven cancer. They do not provide information about the cell line used to identify these genes, therefore we can only assume a genetic interactions between each of these genes and KRAS.

We defined the interaction between a mutated gene and an essential gene in a particular cancer cell line as negative genetic interaction, because the sum of the mutation of one gene and the silencing of the second gene impairs the viability of the cell, while the mutation only is not sufficient. Essential genes for each cell line were defined in different ways in every experiment, although the experimental procedures to determine essential genes were highly similar. Briefly, a cell culture is infected with lentiviruses able to encode shRNAs that suppress specific genes. The cells are infected with a high number of different lentiviral RNAs: the experiments that we analyzed used libraries that targeted from less than 100 to almost 10,000 genes. After a period of time that ranges between 10 days and four weeks, microarray analysis is preformed to measure the abundance of the shRNAs in the cells. The levels of shRNAs are compared with those of a control, which may be of two types: the same cells right after infection or a normal cell line. If the abundance of a shRNA that targets a particular gene is higher in the control, then this gene is likely to be essential in the cell line and, if it is silenced, the cell cannot survive (Berns et al., 2004).

In order to identify the essential genes in each cell line, we used the definitions given by the authors. These varied from simple differences between the expression values in the cancer cell line and the control to complex statistical methods (Luo et al., 2008). In total, we identified 1,735 essential genes (94 cancer genes from the Cancer Gene Census and 111 candidate cancer genes) in the 28 cell lines. Each cell line had between three and 312 essential genes (median 94). We identified also 525 genes (30 cancer genes) that were found essential in a normal cell line (HMEC).

We identified 8,525 genetic interactions between the 25 mutated genes and the 1,735 essential genes.

3.2. Genetic interactors of cancer genes are ancient duplicated hubs

Of all the 1,735 essential genes, 1,045 (60.2%) are specific for one cell line, while 74 (4.3%) are essential in at least five cell lines (Figure 69). We analyzed origin, duplicability and network properties of essential genes and found that they significantly differ from the rest of the human genes (Table 25). All essential genes, independently from the number of cell lines where they were found essential, are ancient, having 49% of the genes that originated with the last universal common ancestor (*p*-value 5e-35, Fisher's exact test, Figure 70), highly connected and central (*p*-value 1e-74 and 2e-44, respectively, Wilcoxon test). Duplicability, instead, depends on the number of cell lines. Considering low coverage (10% sequence conservation of the additional hit on the genome), genes that are essential in 4 cell lines at most are highly duplicated, while at high coverage only genes that are essential in more than 4 cell lines are highly duplicated (37.0%, *p*-value 5.8e-3 from Fisher's exact test, Figure 71).

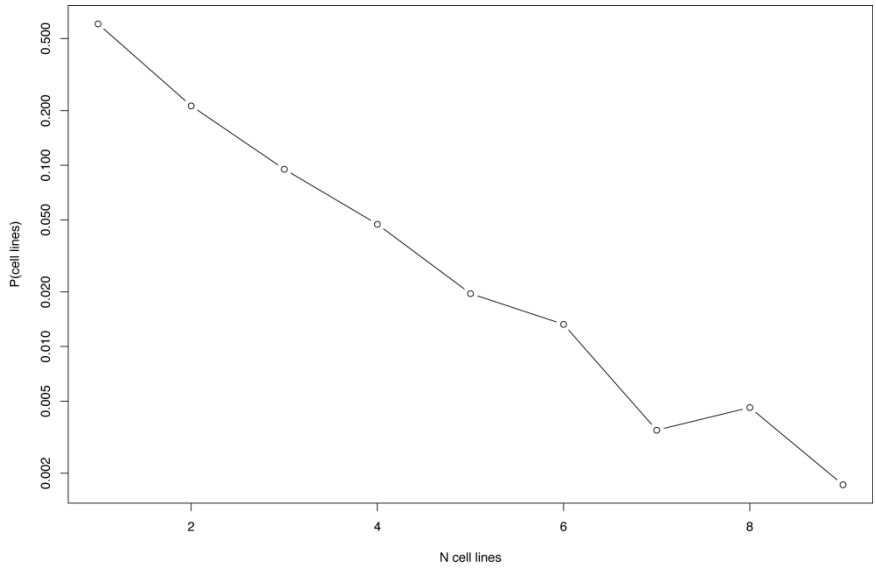


Figure 69: essential genes and cancer cell lines

The relationship between essential genes and the number of cell lines they are found essential is shown.

Table 25: properties of essential genes

Property		essential in cancer		essential in 1 cell line		essential in 2-4 cell lines		essential in >4 cell lines		rest of human genes
		N	p-value	N	p-value	N	p-value	N	p-value	
total		1,735	NA	1,045	NA	616	NA	74	NA	20,392
with origin		1,683	NA	1,006	NA	603	NA	74	NA	18,427
LUCA		49.0	5.48E-35	46.4	3.06E-18	50.1	1.35E-15	74.3	8.40E-08	27.5
eukaryotes		23.4	5.07E-02	22.3	3.28E-02	25.5	8.20E-01	20.3	4.32E-01	26.2
opisthokonts		0.8	5.25E-01	1.1	7.50E-01	0.5	2.96E-01	0.0	1.00E+00	1.0
metazoans		13.3	7.84E-05	14.2	1.91E-02	12.8	7.29E-03	4.1	3.48E-03	17.6
vertebrates		10.2	7.97E-11	11.7	1.63E-04	8.6	6.96E-07	1.4	2.01E-04	16.9
mammals		3.3	3.94E-20	4.0	5.52E-10	2.5	1.54E-10	0.0	1.67E-03	9.8
primates		0.2	1.53E-04	0.3	1.97E-02	0.0	4.82E-03	0.0	1.00E+00	1.0
with duplicability		1658	NA	988	NA	597	NA	73	NA	16,797
60%		21.7	4.27E-02	20.1	5.16E-01	22.4	1.13E-01	37.0	5.82E-03	19.2
10%		62.7	8.40E-04	61.6	2.05E-02	63.5	2.44E-02	69.9	1.87E-01	54.5
60% exonic		17.0	2.51E-02	16.3	2.03E-01	16.4	2.98E-01	31.5	3.01E-03	14.6
10% exonic		56.9	9.28E-05	56.6	2.79E-03	56.8	1.57E-02	61.6	2.00E-01	48.0
with protein interaction network		1441	NA	845	NA	526	NA	70	NA	12,078
degree	mean	27.5	1.36E-74	23.0	1.94E-34	34.4	3.23E-42	30.2	2.01E-08	12.9
	median	12		11		14		20		5
betweenness	mean	49,461	2.10E-44	36,262	2.33E-22	71,138	7.31E-24	40,656	5.40E-05	16,976
	median	6,058		5,676		6,683		9,592		1,574

Essential genes are divided into three categories, on the basis of the number of cell lines where they are found essential. From the rest of the human genes, all genes that are essential in the normal cell line are excluded. *P*-values are calculated with Fisher's exact test for origin and duplicability, with Wilcoxon test for the network properties. Red represents enrichment in comparison with the rest of the human genes, while green represents depletion.

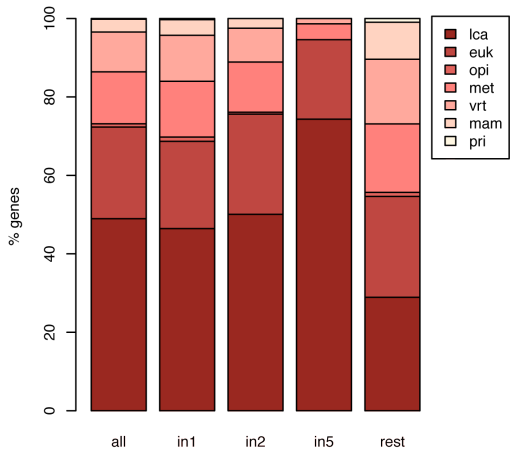


Figure 70: origin of essential genes

Essential genes are divided into three categories, on the basis of the number of cell lines where they were found essential: 1 cell line, 2-4 cell lines and at least 4 cell lines. The first column represents all essential genes in cancer cell lines, while the last column corresponds to the rest of the human genes. Origin is calculated as in Figure 26.

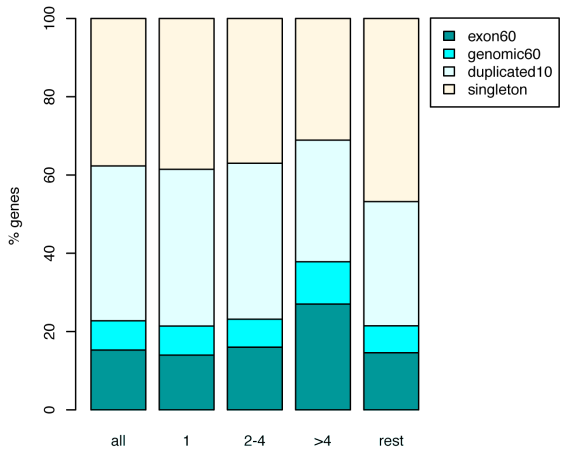


Figure 71: duplicability of essential genes and cancer cell lines

Duplicated genes are divided into three categories: duplicates at high coverage (>60%) that overlap known genes, genomic duplicates at high coverage, duplicates at low coverage (10%). Duplicability is calculated as in Figure 28.

Among the essential genes, we identified 93 cancer genes from the cancer gene census (5.4% of all essential genes), while, among the non-essential genes, cancer genes represent less than 2% of all human genes (437, 1.7%). Therefore essential genes are enriched in known cancer genes (*p*-value 1.4e-25 from chi-squared test). These findings are independent from the number of cell lines where a gene was found essential, having 49 cancer genes essential in one cell line (4.7%), 41 in two, three or four cell lines (6.7%) and 3 in more than four cell lines (4.1%).

3.3. The integration of protein-protein interactions and genetic interactions allows the identification of putative cancer genes

The identification of genetic interactions that involve cancer genes allows the detection of direct or indirect interactors that may be important in the development of cancer or essential for the survival of the tumor cells. These interactors are enriched in known cancer genes, although their properties do not fully correspond to those of cancer genes. Essential genes are highly connected and central in the human protein interaction network and they are ancient like recessive genes, but their duplicability does not reflect that of recessive genes. In particular, genes that are essential in many cell lines are enriched in highly conserved duplicates.

The enrichment of cancer genes among essential genes, in addition to the fact that cancer genes are near in the human protein interaction network (Figure 48), suggests that interactors of already known cancer genes are likely to have an active role in tumor development, when mutated. Given this premise and knowing the systems-level properties of cancer genes, we are developing a method to identify putative new cancer genes.

A first screening on the basis of origin, duplicability and network properties allowed us to narrow down the search for new putative cancer genes to two classes of genes:

- Singleton hubs that originated with the last universal common ancestor or eukaryotes;
- Duplicated hubs that originated with metazoans or vertebrates.

The identification of these two categories allowed us to set a first filter to identify putative recessive and dominant genes, respectively. We detected 844 recessive genes and 78 dominant genes that were not included in either the cancer genes census or in the collection of candidate cancer genes. Among these putative cancer genes, 406 (48%) recessive and 20 (25.6%) dominant had known non-synonymous mutations in high-

throughput mutational screenings of cancer tissues, but were not considered as candidates because they were not found mutated in the validations screenings. Both putative dominant and recessive genes are enriched in genes that are essential in cancer cell lines, having 14 (18.0%, *p*-value 0.006 from Fisher's exact test) and 203 (24.0%, *p*-value 3.25e-46 from Fisher's exact test), respectively. These findings support our hypothesis that, by analyzing the systems-level properties of cancer genes and studying their interactors, new putative cancer genes may be identified. However, these initial results need to be further investigated, in order to narrow down the pool of possible new cancer genes. The filter on genes that have genetic interactions with known cancer genes is strict and may cause the loss of real new candidates, because only less than 1,700 genes have this type of information. A further way to proceed is to identify, among the 922 putative cancer genes, genes that are enriched in protein interaction network neighbors that are known cancer genes. However, in order to identify more accurately the genes whose mutations drive tumorigenesis, we need to consider other properties that, in this Thesis, we did not take into account.

A further filter would be the identification of functional interactions. To identify them, the association of genes with GO terms may be used. An interaction is present between two genes if they are associated with at least one common GO term. The functional network would be weighted and the weight of each interaction would be calculated on the basis of the number of common GO terms. However, a drawback of this method is the fact that it is based on manual annotations of genes to GO terms. This is not an issue when considering global functional classifications of gene lists or when comparing the functional characteristics of different gene lists, but the analysis of functional features of single genes may biased due to misclassifications or lack of knowledge in the functional classification.

In order to have a more complete view of the interactions between proteins and genes, in addition to protein-protein and genetic interactions, we are planning to identify

the protein-DNA interaction networks. This will add information about the genes that are regulated by the numerous DNA-binding proteins that are involved in cancer and will allow the identification of new cancer candidates defined on the basis of their common binding sites on the genome. Mutations in many transcription factors or DNA-binding proteins have often been associated with cancer. We will be able to identify putative new cancer genes by studying the similarities with the DNA-binding sites of known cancer genes and with the genes that they regulate at the transcriptional level.

References

Albert, R. (2005). Scale-free networks in cell biology. *J Cell Sci* **118**, 4947-57.

Albert, R. and Barabasi, A. L. (2002). Statistical mechanics of complex networks. *Rev. Modern Phys.* **74**, 47-97.

Alpert, J. and Hajaj, N. (2008). We knew the Web was big. *http://googleblog.blogspot.com/2008/07/we-knew-web-was-big.html*.

Antonescu, C. R., Zhang, L., Chang, N. E., Pawel, B. R., Travis, W., Katabi, N., Edelman, M., Rosenberg, A. E., Nielsen, G. P., Dal Cin, P. et al. (2010). EWSR1-POU5F1 fusion in soft tissue myoepithelial tumors. A molecular analysis of sixty-six cases, including soft tissue, bone, and visceral lesions, showing common involvement of the EWSR1 gene. *Genes Chromosomes Cancer* **49**, 1114-24.

Argani, P., Lui, M. Y., Couturier, J., Bouvier, R., Fournet, J. C. and Ladanyi, M. (2003). A novel CLTC-TFE3 gene fusion in pediatric renal adenocarcinoma with t(X;17)(p11.2;q23). *Oncogene* **22**, 5374-8.

Armstrong, J. A., Papoulas, O., Daubresse, G., Sperling, A. S., Lis, J. T., Scott, M. P. and Tamkun, J. W. (2002). The Drosophila BRM complex facilitates global transcription by RNA polymerase II. *EMBO J* **21**, 5245-54.

Arora, S., Gonzales, I. M., Hagelstrom, R. T., Beaudry, C., Choudhary, A., Sima, C., Tibes, R., Mousses, S. and Azorsa, D. O. (2010). RNAi phenotype profiling of kinases identifies potential therapeutic targets in Ewing's sarcoma. *Mol Cancer* **9**, 218.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T. et al. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25**, 25-9.

Avery, L. and Wasserman, S. (1992). Ordering gene function: the interpretation of epistasis in regulatory hierarchies. *Trends Genet* **8**, 312-6.

Bader, J. S., Chaudhuri, A., Rothberg, J. M. and Chant, J. (2004). Gaining confidence in high-throughput protein interaction networks. *Nat Biotechnol* **22**, 78-85.

Bailey, J. A., Gu, Z., Clark, R. A., Reinert, K., Samonte, R. V., Schwartz, S., Adams, M. D., Myers, E. W., Li, P. W. and Eichler, E. E. (2002). Recent segmental duplications in the human genome. *Science* **297**, 1003-7.

Bailey, J. A., Yavor, A. M., Massa, H. F., Trask, B. J. and Eichler, E. E. (2001). Segmental duplications: organization and impact within the current human genome project assembly. *Genome Res* **11**, 1005-17.

Baldwin, A., Grueneberg, D. A., Hellner, K., Sawyer, J., Grace, M., Li, W., Harlow, E. and Munger, K. (2010). Kinase requirements in human cells: V. Synthetic lethal interactions between p53 and the protein kinases SGK2 and PAK3. *Proc Natl Acad Sci U S A* **107**, 12463-8.

Baldwin, A., Li, W., Grace, M., Pearlberg, J., Harlow, E., Munger, K. and Grueneberg, D. A. (2008). Kinase requirements in human cells: II. Genetic interaction screens identify kinase requirements following HPV16 E7 expression in cancer cells. *Proc Natl Acad Sci U S A* **105**, 16478-83.

Barabasi, A. L. and Albert, R. (1999). Emergence of scaling in random networks. *Science* **286**, 509-12.

Barabasi, A. L. and Oltvai, Z. N. (2004). Network biology: understanding the cell's functional organization. *Nat Rev Genet* **5**, 101-13.

Barbie, D. A., Tamayo, P., Boehm, J. S., Kim, S. Y., Moody, S. E., Dunn, I. F., Schinzel, A. C., Sandy, P., Meylan, E., Scholl, C. et al. (2009). Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. *Nature* **462**, 108-12.

Barretina, J., Taylor, B. S., Banerji, S., Ramos, A. H., Lagos-Quintana, M., Decarolis, P. L., Shah, K., Socci, N. D., Weir, B. A., Ho, A. et al. (2010). Subtype-specific genomic alterations define new targets for soft-tissue sarcoma therapy. *Nat Genet* **42**, 715-21.

Barrett, T., Troup, D. B., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., Marshall, K. A., Phillippy, K. H., Sherman, P. M. et al. (2011). NCBI GEO: archive for functional genomics data sets--10 years on. *Nucleic Acids Res* **39**, D1005-10.

Bartel, D. P. (2004). MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* **116**, 281-97.

Beltrao, P. and Serrano, L. (2007). Specificity and evolvability in eukaryotic protein interaction networks. *PLoS Comput Biol* **3**, e25.

Berns, K., Hijmans, E. M., Mullenders, J., Brummelkamp, T. R., Velds, A., Heimerikx, M., Kerkhoven, R. M., Madiredjo, M., Nijkamp, W., Weigelt, B. et al. (2004). A large-scale RNAi screen in human cells identifies new components of the p53 pathway. *Nature* **428**, 431-7.

Bommi-Reddy, A., Almeciga, I., Sawyer, J., Geisen, C., Li, W., Harlow, E., Kaelin, W. G., Jr. and Grueneberg, D. A. (2008). Kinase requirements in human cells: III. Altered kinase requirements in VHL-/- cancer cells detected in a pilot synthetic lethal screen. *Proc Natl Acad Sci U S A* **105**, 16484-9.

Boveri, T. H. (1914). The Origin of malignant tumors. *The Williams and Wilkins Co., Baltimore*.

Breitkreutz, B. J., Stark, C., Reguly, T., Boucher, L., Breitkreutz, A., Livstone, M., Oughtred, R., Lackner, D. H., Bahler, J., Wood, V. et al. (2008). The BioGRID Interaction Database: 2008 update. *Nucleic Acids Res* **36**, D637-40.

Bridge, A. J., Pebernard, S., Ducraux, A., Nicoulaz, A. L. and Iggo, R. (2003). Induction of an interferon response by RNAi vectors in mammalian cells. *Nat Genet* **34**, 263-4.

Burki, F. and Kaessmann, H. (2004). Birth and adaptive evolution of a hominoid gene that supports high neurotransmitter flux. *Nat Genet* **36**, 1061-3.

Cai, J., Zhao, R., Jiang, H. and Wang, W. (2008). De novo origination of a new protein-coding gene in Saccharomyces cerevisiae. *Genetics* **179**, 487-96.

Cesareni, G., Chatr-aryamontri, A., Licata, L. and Ceol, A. (2008). Searching the MINT database for protein interaction information. *Curr Protoc Bioinformatics* **Chapter 8**, Unit 8 5.

Chien, C. T., Bartel, P. L., Sternglanz, R. and Fields, S. (1991). The two-hybrid system: a method to identify and clone genes for proteins that interact with a protein of interest. *Proc Natl Acad Sci U S A* **88**, 9578-82.

Christoffels, A., Koh, E. G., Chia, J. M., Brenner, S., Aparicio, S. and Venkatesh, B. (2004). Fugu genome analysis provides evidence for a whole-genome duplication early during the evolution of ray-finned fishes. *Mol Biol Evol* **21**, 1146-51.

Ciccarelli, F. D. (2010). The (r)evolution of cancer genetics. *BMC Biol* **8**, 74.

Clark, M. J., Homer, N., O'Connor, B. D., Chen, Z., Eskin, A., Lee, H., Merriman, B. and Nelson, S. F. (2010). U87MG decoded: the genomic sequence of a cytogenetically aberrant human cancer cell line. *PLoS Genet* **6**, e1000832.

Coffey, A. J., Kokocinski, F., Calafato, M. S., Scott, C. E., Palta, P., Drury, E., Joyce, C. J., Leproust, E. M., Harrow, J., Hunt, S. et al. (2011). The GENCODE exome: sequencing the complete human exome. *Eur J Hum Genet*.

Cole, G. M., Stone, D. E. and Reed, S. I. (1990). Stoichiometry of G protein subunits affects the Saccharomyces cerevisiae mating pheromone signal transduction pathway. *Mol Cell Biol* **10**, 510-7.

Conant, G. C. and Wagner, A. (2004). Duplicate genes and robustness to transient gene knock-downs in Caenorhabditis elegans. *Proc Biol Sci* **271**, 89-96.

Conant, G. C. and Wolfe, K. H. (2008). Turning a hobby into a job: how duplicated genes find new functions. *Nat Rev Genet* **9**, 938-50.

Cote, R. G., Jones, P., Martens, L., Kerrien, S., Reisinger, F., Lin, Q., Leinonen, R., Apweiler, R. and Hermjakob, H. (2007). The Protein Identifier Cross-Referencing (PICR) service: reconciling protein identifiers across multiple source databases. *BMC Bioinformatics* **8**, 401.

Cusick, M. E., Klitgord, N., Vidal, M. and Hill, D. E. (2005). Interactome: gateway into systems biology. *Hum Mol Genet* **14 Spec No. 2**, R171-81.

Cusick, M. E., Yu, H., Smolyar, A., Venkatesan, K., Carvunis, A. R., Simonis, N., Rual, J. F., Borick, H., Braun, P., Dreze, M. et al. (2009). Literature-curated protein interaction datasets. *Nat Methods* **6**, 39-46.

Dalgliesh, G. L., Furge, K., Greenman, C., Chen, L., Bignell, G., Butler, A., Davies, H., Edkins, S., Hardy, C., Latimer, C. et al. (2010). Systematic sequencing of renal carcinoma reveals inactivation of histone modifying genes. *Nature* **463**, 360-3.

Davies, H., Bignell, G. R., Cox, C., Stephens, P., Edkins, S., Clegg, S., Teague, J., Woffendin, H., Garnett, M. J., Bottomley, W. et al. (2002). Mutations of the BRAF gene in human cancer. *Nature* **417**, 949-54.

Davies, H., Hunter, C., Smith, R., Stephens, P., Greenman, C., Bignell, G., Teague, J., Butler, A., Edkins, S., Stevens, C. et al. (2005). Somatic mutations of the protein kinase gene family in human lung cancer. *Cancer Res* **65**, 7591-5.

Dehal, P. and Boore, J. L. (2005). Two rounds of whole genome duplication in the ancestral vertebrate. *PLoS Biol* **3**, e314.

Des Marais, D. L. and Rausher, M. D. (2008). Escape from adaptive conflict after duplication in an anthocyanin pathway gene. *Nature* **454**, 762-5.

Ding, L., Ellis, M. J., Li, S., Larson, D. E., Chen, K., Wallis, J. W., Harris, C. C., McLellan, M. D., Fulton, R. S., Fulton, L. L. et al. (2010). Genome remodelling in a basal-like breast cancer metastasis and xenograft. *Nature* **464**, 999-1005.

Ding, L., Getz, G., Wheeler, D. A., Mardis, E. R., McLellan, M. D., Cibulskis, K., Sougnez, C., Greulich, H., Muzny, D. M., Morgan, M. B. et al. (2008). Somatic mutations affect key pathways in lung adenocarcinoma. *Nature* **455**, 1069-75.

Dixon, S. J., Costanzo, M., Baryshnikova, A., Andrews, B. and Boone, C. (2009). Systematic mapping of genetic interaction networks. *Annu Rev Genet* **43**, 601-25.

DomainTools. (2011). Domain Counts and Internet Statistics. *http://www.domaintools.com/internet-statistics/*.

Domazet-Loso, T. and Tautz, D. (2008). An ancient evolutionary origin of genes associated with human genetic diseases. *Mol Biol Evol* **25**, 2699-707.

Domazet-Loso, T. and Tautz, D. (2010). Phylostratigraphic tracking of cancer genes suggests a link to the emergence of multicellularity in metazoa. *BMC Biol* **8**, 66.

Drysdale, R. (2008). FlyBase : a database for the Drosophila research community. *Methods Mol Biol* **420**, 45-59.

Echeverri, C. J., Beachy, P. A., Baum, B., Boutros, M., Buchholz, F., Chanda, S. K., Downward, J., Ellenberg, J., Fraser, A. G., Hacohen, N. et al. (2006). Minimizing the risk of reporting false positives in large-scale RNAi screens. *Nat Methods* **3**, 777-9.

Engel, S. R., Balakrishnan, R., Binkley, G., Christie, K. R., Costanzo, M. C., Dwight, S. S., Fisk, D. G., Hirschman, J. E., Hitz, B. C., Hong, E. L. et al. (2010). Saccharomyces Genome Database provides mutant phenotype data. *Nucleic Acids Res* **38**, D433-6.

Evlampiev, K. and Isambert, H. (2008). Conservation and topology of protein interaction networks under duplication-divergence evolution. *Proc Natl Acad Sci U S A* **105**, 9863-8.

Fernandez, A. and Chen, J. (2009). Human capacitance to dosage imbalance: coping with inefficient selection. *Genome Res* **19**, 2185-92.

Fields, S. and Song, O. (1989). A novel genetic system to detect protein-protein interactions. *Nature* **340**, 245-6.

Firestein, R., Bass, A. J., Kim, S. Y., Dunn, I. F., Silver, S. J., Guney, I., Freed, E., Ligon, A. H., Vena, N., Ogino, S. et al. (2008). CDK8 is a colorectal cancer oncogene that regulates beta-catenin activity. *Nature* **455**, 547-51.

Flicek, P., Amode, M. R., Barrell, D., Beal, K., Brent, S., Chen, Y., Clapham, P., Coates, G., Fairley, S., Fitzgerald, S. et al. (2011). Ensembl 2011. *Nucleic Acids Res* **39**, D800-6.

Forbes, S. A., Bindal, N., Bamford, S., Cole, C., Kok, C. Y., Beare, D., Jia, M., Shepherd, R., Leung, K., Menzies, A. et al. (2011). COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. *Nucleic Acids Res* **39**, D945-50.

Friedberg, E. C. (2003). DNA damage and repair. *Nature* **421**, 436-40.

Fujii, G. H., Morimoto, A. M., Berson, A. E. and Bolen, J. B. (1999). Transcriptional analysis of the PTEN/MMAC1 pseudogene, psiPTEN. *Oncogene* **18**, 1765-9.

Futreal, P. A., Coin, L., Marshall, M., Down, T., Hubbard, T., Wooster, R., Rahman, N. and Stratton, M. R. (2004). A census of human cancer genes. *Nat Rev Cancer* **4**, 177-83.

Gandhi, T. K., Zhong, J., Mathivanan, S., Karthick, L., Chandrika, K. N., Mohan, S. S., Sharma, S., Pinkert, S., Nagaraju, S., Periaswamy, B. et al. (2006). Analysis of the human protein interactome and comparison with yeast, worm and fly interaction datasets. *Nat Genet* **38**, 285-93.

Gascoyne, R. D., Lamant, L., Martin-Subero, J. I., Lestou, V. S., Harris, N. L., Muller-Hermelink, H. K., Seymour, J. F., Campbell, L. J., Horsman, D. E., Auvigne, I. et al. (2003). ALK-positive diffuse large B-cell lymphoma is associated with Clathrin-ALK rearrangements: report of 6 cases. *Blood* **102**, 2568-73.

Ge, X., Yamamoto, S., Tsutsumi, S., Midorikawa, Y., Ihara, S., Wang, S. M. and Aburatani, H. (2005). Interpreting expression profiles of cancers by genome-wide survey of breadth of expression in normal tissues. *Genomics* **86**, 127-41.

Gerstein, M. B., Bruce, C., Rozowsky, J. S., Zheng, D., Du, J., Korbel, J. O., Emanuelsson, O., Zhang, Z. D., Weissman, S. and Snyder, M. (2007). What is a gene, post-ENCODE? History and updated definition. *Genome Res* **17**, 669-81.

Giot, L., Bader, J. S., Brouwer, C., Chaudhuri, A., Kuang, B., Li, Y., Hao, Y. L., Ooi, C. E., Godwin, B., Vitols, E. et al. (2003). A protein interaction map of Drosophila melanogaster. *Science* **302**, 1727-36.

Glaros, S., Cirrincione, G. M., Muchardt, C., Kleer, C. G., Michael, C. W. and Reisman, D. (2007). The reversible epigenetic silencing of BRM: implications for clinical targeted therapy. *Oncogene* **26**, 7058-66.

Goh, K. I., Cusick, M. E., Valle, D., Childs, B., Vidal, M. and Barabasi, A. L. (2007). The human disease network. *Proc Natl Acad Sci U S A* **104**, 8685-90.

Goh, K. I., Kahng, B. and Kim, D. (2001). Universal behavior of load distribution in scale-free networks. *Phys Rev Lett* **87**, 278701.

Gonzalez, J., Lenkov, K., Lipatov, M., Macpherson, J. M. and Petrov, D. A. (2008). High rate of recent transposable element-induced adaptation in Drosophila melanogaster. *PLoS Biol* **6**, e251.

Greenman, C., Stephens, P., Smith, R., Dalgliesh, G. L., Hunter, C., Bignell, G., Davies, H., Teague, J., Butler, A., Stevens, C. et al. (2007). Patterns of somatic mutation in human cancer genomes. *Nature* **446**, 153-8.

Greif, P. A., Eck, S. H., Konstandin, N. P., Benet-Pages, A., Ksienzyk, B., Dufour, A., Vetter, A. T., Popp, H. D., Lorenz-Depiereux, B., Meitinger, T. et al. (2011). Identification of recurring tumor-specific somatic mutations in acute myeloid leukemia by transcriptome sequencing. *Leukemia* **25**, 821-7.

Grueneberg, D. A., Degot, S., Pearlberg, J., Li, W., Davies, J. E., Baldwin, A., Endege, W., Doench, J., Sawyer, J., Hu, Y. et al. (2008a). Kinase requirements in human cells: I. Comparing kinase requirements across various cell types. *Proc Natl Acad Sci U S A* **105**, 16472-7.

Grueneberg, D. A., Li, W., Davies, J. E., Sawyer, J., Pearlberg, J. and Harlow, E. (2008b). Kinase requirements in human cells: IV. Differential kinase requirements in cervical and renal human tumor cell lines. *Proc Natl Acad Sci U S A* **105**, 16490-5.

Gu, Z., Steinmetz, L. M., Gu, X., Scharfe, C., Davis, R. W. and Li, W. H. (2003). Role of duplicate genes in genetic robustness against null mutations. *Nature* **421**, 63-6.

Hahn, M. W. and Kern, A. D. (2005). Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks. *Mol Biol Evol* **22**, 803-6.

Haider, S., Ballester, B., Smedley, D., Zhang, J., Rice, P. and Kasprzyk, A. (2009). BioMart Central Portal--unified access to biological data. *Nucleic Acids Res* **37**, W23-7.

Han, J. D., Bertin, N., Hao, T., Goldberg, D. S., Berriz, G. F., Zhang, L. V., Dupuy, D., Walhout, A. J., Cusick, M. E., Roth, F. P. et al. (2004). Evidence for dynamically organized modularity in the yeast protein-protein interaction network. *Nature* **430**, 88-93.

Han, M. V., Demuth, J. P., McGrath, C. L., Casola, C. and Hahn, M. W. (2009). Adaptive evolution of young gene duplicates in mammals. *Genome Res* **19**, 859-67.

Hanahan, D. and Weinberg, R. A. (2000). The hallmarks of cancer. *Cell* **100**, 57-70.

Hanahan, D. and Weinberg, R. A. (2010). Hallmarks of cancer: the next generation. *Cell* **144**, 646-74.

Harrow, J., Denoeud, F., Frankish, A., Reymond, A., Chen, C. K., Chrast, J., Lagarde, J., Gilbert, J. G., Storey, R., Swarbreck, D. et al. (2006). GENCODE: producing a reference annotation for ENCODE. *Genome Biol* **7 Suppl 1**, S4 1-9.

Hart, G. T., Ramani, A. K. and Marcotte, E. M. (2006). How complete are current yeast and human protein-interaction networks? *Genome Biol* **7**, 120.

Hood, F. E. and Royle, S. J. (2009). Functional equivalence of the clathrin heavy chains CHC17 and CHC22 in endocytosis and mitosis. *J Cell Sci* **122**, 2185-90.

Hughes, A. L. (1994). The evolution of functionally novel proteins after gene duplication. *Proc Biol Sci* **256**, 119-24.

Hughes, A. L. and Friedman, R. (2005). Gene duplication and the properties of biological networks. *J Mol Evol* **61**, 758-64.

Hunger, S. P. (1996). Chromosomal translocations involving the E2A gene in acute lymphoblastic leukemia: clinical features and molecular pathogenesis. *Blood* **87**, 1211-24.

Ikediobi, O. N., Davies, H., Bignell, G., Edkins, S., Stevens, C., O'Meara, S., Santarius, T., Avis, T., Barthorpe, S., Brackenbury, L. et al. (2006). Mutation analysis of 24 known cancer genes in the NCI-60 cell line set. *Mol Cancer Ther* **5**, 2606-12.

Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M. and Sakaki, Y. (2001). A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc Natl Acad Sci U S A* **98**, 4569-74.

Jaillon, O., Aury, J. M., Brunet, F., Petit, J. L., Stange-Thomann, N., Mauceli, E., Bouneau, L., Fischer, C., Ozouf-Costaz, C., Bernot, A. et al. (2004). Genome duplication in the teleost fish Tetraodon nigroviridis reveals the early vertebrate proto-karyotype. *Nature* **431**, 946-57.

Jensen, L. J., Julien, P., Kuhn, M., von Mering, C., Muller, J., Doerks, T. and Bork, P. (2008). eggNOG: automated construction and annotation of orthologous groups of genes. *Nucleic Acids Res* **36**, D250-4.

Jeong, H., Mason, S. P., Barabasi, A. L. and Oltvai, Z. N. (2001). Lethality and centrality in protein networks. *Nature* **411**, 41-2.

Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N. and Barabasi, A. L. (2000). The large-scale organization of metabolic networks. *Nature* **407**, 651-4.

Jiao, Y., Shi, C., Edil, B. H., de Wilde, R. F., Klimstra, D. S., Maitra, A., Schulick, R. D., Tang, L. H., Wolfgang, C. L., Choti, M. A. et al. (2011). DAXX/ATRX, MEN1, and mTOR pathway genes are frequently altered in pancreatic neuroendocrine tumors. *Science* **331**, 1199-203.

Johnston, M. (1987). A model fungal gene regulatory mechanism: the GAL genes of Saccharomyces cerevisiae. *Microbiol Rev* **51**, 458-76.

Jones, S., Zhang, X., Parsons, D. W., Lin, J. C., Leary, R. J., Angenendt, P., Mankoo, P., Carter, H., Kamiyama, H., Jimeno, A. et al. (2008). Core signaling pathways in human pancreatic cancers revealed by global genomic analyses. *Science* **321**, 1801-6.

Jonsson, P. F. and Bates, P. A. (2006). Global topological features of cancer proteins in the human interactome. *Bioinformatics* **22**, 2291-7.

Kan, Z., Jaiswal, B. S., Stinson, J., Janakiraman, V., Bhatt, D., Stern, H. M., Yue, P., Haverty, P. M., Bourgon, R., Zheng, J. et al. (2010). Diverse somatic mutation patterns and pathway alterations in human cancers. *Nature* **466**, 869-73.

Kellis, M., Birren, B. W. and Lander, E. S. (2004). Proof and evolutionary analysis of ancient genome duplication in the yeast Saccharomyces cerevisiae. *Nature* **428**, 617-24.

Kent, W. J. (2002). BLAT--the BLAST-like alignment tool. *Genome Res* **12**, 656-64.

Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M. and Haussler, D. (2002). The human genome browser at UCSC. *Genome Res* **12**, 996-1006.

Kerrien, S., Alam-Faruque, Y., Aranda, B., Bancarz, I., Bridge, A., Derow, C., Dimmer, E., Feuermann, M., Friedrichsen, A., Huntley, R. et al. (2007). IntAct--open source resource for molecular interaction data. *Nucleic Acids Res* **35**, D561-5.

Keseler, I. M., Bonavides-Martinez, C., Collado-Vides, J., Gama-Castro, S., Gunsalus, R. P., Johnson, D. A., Krummenacker, M., Nolan, L. M., Paley, S., Paulsen, I. T. et al. (2009). EcoCyc: a comprehensive view of Escherichia coli biology. *Nucleic Acids Res* **37**, D464-70.

Keshava Prasad, T. S., Goel, R., Kandasamy, K., Keerthikumar, S., Kumar, S., Mathivanan, S., Telikicherla, D., Raju, R., Shafreen, B., Venugopal, A. et al. (2009). Human Protein Reference Database--2009 update. *Nucleic Acids Res* **37**, D767-72.

Kim, P. M., Korbel, J. O. and Gerstein, M. B. (2007). Positive selection at the protein network periphery: evaluation in terms of structural constraints and cellular context. *Proc Natl Acad Sci U S A* **104**, 20274-9.

Kim, T. D., Shin, S. and Janknecht, R. (2008). Repression of Smad3 activity by histone demethylase SMCX/JARID1C. *Biochem Biophys Res Commun* **366**, 563-7.

Kinzler, K. W. and Vogelstein, B. (1997). Cancer-susceptibility genes. Gatekeepers and caretakers. *Nature* **386**, 761, 763.

Knowles, D. G. and McLysaght, A. (2009). Recent de novo origin of human protein-coding genes. *Genome Res* **19**, 1752-9.

Kondrashov, F. A. and Kondrashov, A. S. (2006). Role of selection in fixation of gene duplications. *J Theor Biol* **239**, 141-51.

Koonin, E. V. (2005). Orthologs, paralogs, and evolutionary genomics. *Annu Rev Genet* **39**, 309-38.

Krause, R., von Mering, C., Bork, P. and Dandekar, T. (2004). Shared components of protein complexes--versatile building blocks or biochemical artefacts? *Bioessays* **26**, 1333-43.

Kuhner, S., van Noort, V., Betts, M. J., Leo-Macias, A., Batisse, C., Rode, M., Yamada, T., Maier, T., Bader, S., Beltran-Alvarez, P. et al. (2009). Proteome organization in a genome-reduced bacterium. *Science* **326**, 1235-40.

Kunin, V., Pereira-Leal, J. B. and Ouzounis, C. A. (2004). Functional evolution of the yeast protein interaction network. *Mol Biol Evol* **21**, 1171-6.

Lee, T. I., Rinaldi, N. J., Robert, F., Odom, D. T., Bar-Joseph, Z., Gerber, G. K., Hannett, N. M., Harbison, C. T., Thompson, C. M., Simon, I. et al. (2002). Transcriptional regulatory networks in Saccharomyces cerevisiae. *Science* **298**, 799-804.

Lemke, N., Heredia, F., Barcellos, C. K., Dos Reis, A. N. and Mombach, J. C. (2004). Essentiality and damage in metabolic networks. *Bioinformatics* **20**, 115-9.

Leung, S. K. and Ohh, M. (2002). Playing Tag with HIF: The VHL Story. *J Biomed Biotechnol* **2**, 131-135.

Ley, T. J., Mardis, E. R., Ding, L., Fulton, B., McLellan, M. D., Chen, K., Dooling, D., Dunford-Shore, B. H., McGrath, S., Hickenbotham, M. et al. (2008). DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. *Nature* **456**, 66-72.

Li, S., Armstrong, C. M., Bertin, N., Ge, H., Milstein, S., Boxem, M., Vidalain, P. O., Han, J. D., Chesneau, A., Hao, T. et al. (2004). A map of the interactome network of the metazoan C. elegans. *Science* **303**, 540-3.

Liang, H. and Li, W. H. (2007). Gene essentiality, gene duplicability and protein connectivity in human and mouse. *Trends Genet* **23**, 375-8.

Liao, B. Y. and Zhang, J. (2007). Mouse duplicate genes are as essential as singletons. *Trends Genet* **23**, 378-81.

Long, M., Betran, E., Thornton, K. and Wang, W. (2003). The origin of new genes: glimpses from the young and old. *Nat Rev Genet* **4**, 865-75.

Lopes, C. T., Franz, M., Kazi, F., Donaldson, S. L., Morris, Q. and Bader, G. D. (2010). Cytoscape Web: an interactive web-based network browser. *Bioinformatics* **26**, 2347-8.

Luo, B., Cheung, H. W., Subramanian, A., Sharifnia, T., Okamoto, M., Yang, X., Hinkle, G., Boehm, J. S., Beroukhim, R., Weir, B. A. et al. (2008). Highly parallel identification of essential genes in cancer cells. *Proc Natl Acad Sci U S A* **105**, 20380-5.

Luscombe, N. M., Babu, M. M., Yu, H., Snyder, M., Teichmann, S. A. and Gerstein, M. (2004). Genomic analysis of regulatory network dynamics reveals large topological changes. *Nature* **431**, 308-12.

Lynn, J. (2010). Internet users to exceed 2 billion this year. *http://www.reuters.com/article/2010/10/19/us-telecoms-internet-idUSTRE69I24720101019*.

Ma'ayan, A., Jenkins, S. L., Neves, S., Hasseldine, A., Grace, E., Dubin-Thaler, B., Eungdamrong, N. J., Weng, G., Ram, P. T., Rice, J. J. et al. (2005). Formation of regulatory patterns during signal propagation in a Mammalian cellular network. *Science* **309**, 1078-83.

Maeda, I., Kohara, Y., Yamamoto, M. and Sugimoto, A. (2001). Large-scale analysis of gene function in Caenorhabditis elegans by high-throughput RNAi. *Curr Biol* **11**, 171-6.

Maere, S., De Bodt, S., Raes, J., Casneuf, T., Van Montagu, M., Kuiper, M. and Van de Peer, Y. (2005). Modeling gene and genome duplications in eukaryotes. *Proc Natl Acad Sci U S A* **102**, 5454-9.

Maglott, D., Ostell, J., Pruitt, K. D. and Tatusova, T. (2011). Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res* **39**, D52-7.

Maher, E. R. and Kaelin, W. G., Jr. (1997). von Hippel-Lindau disease. *Medicine (Baltimore)* **76**, 381-91.

Makino, T., Hokamp, K. and McLysaght, A. (2009). The complex relationship of gene duplication and essentiality. *Trends Genet* **25**, 152-5.

Makino, T. and McLysaght, A. (2010). Ohnologs in the human genome are dosage balanced and frequently associated with disease. *Proc Natl Acad Sci U S A* **107**, 9270-4.

Makova, K. D. and Li, W. H. (2003). Divergence in the spatial pattern of gene expression between human duplicate genes. *Genome Res* **13**, 1638-45.

Mani, R., St Onge, R. P., Hartman, J. L. t., Giaever, G. and Roth, F. P. (2008). Defining genetic interaction. *Proc Natl Acad Sci U S A* **105**, 3461-6.

Mardis, E. R., Ding, L., Dooling, D. J., Larson, D. E., McLellan, M. D., Chen, K., Koboldt, D. C., Fulton, R. S., Delehaunty, K. D., McGrath, S. D. et al. (2009). Recurring mutations found by sequencing an acute myeloid leukemia genome. *N Engl J Med* **361**, 1058-66.

Marques, A. C., Dupanloup, I., Vinckenbosch, N., Reymond, A. and Kaessmann, H. (2005). Emergence of young human genes after a burst of retroposition in primates. *PLoS Biol* **3**, e357.

McLendon, R. Friedman, A. Bigner, D. Van, M., EG., Brat, D. Mastrogianakis, G. Olson, J. Mikkelsen, T. Lehman, N. Aldape, K. et al. (2008). Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **455**, 1061-8.

Merlo, L. M., Pepper, J. W., Reid, B. J. and Maley, C. C. (2006). Cancer as an evolutionary and ecological process. *Nat Rev Cancer* **6**, 924-35.

Meyer, A. and Van de Peer, Y. (2005). From 2R to 3R: evidence for a fish-specific genome duplication (FSGD). *Bioessays* **27**, 937-45.

Mirkin, B., Muchnik, I. and Smith, T. F. (1995). A biologically consistent model for comparing molecular phylogenies. *J Comput Biol* **2**, 493-507.

Mizutani, T., Ito, T., Nishina, M., Yamamichi, N., Watanabe, A. and Iba, H. (2002). Maintenance of integrated proviral gene expression requires Brm, a catalytic subunit of SWI/SNF complex. *J Biol Chem* **277**, 15859-64.

Moffat, J., Grueneberg, D. A., Yang, X., Kim, S. Y., Kloepfer, A. M., Hinkle, G., Piqani, B., Eisenhaure, T. M., Luo, B., Grenier, J. K. et al. (2006). A lentiviral RNAi library for human and mouse genes applied to an arrayed viral high-content screen. *Cell* **124**, 1283-98.

Muller, J., Szklarczyk, D., Julien, P., Letunic, I., Roth, A., Kuhn, M., Powe, ll, S., von Mering, C., Doerks, T. et al. (2010). eggNOG v2.0: extending the evolutionary genealogy of genes with enhanced non-supervised orthologous groups, species and functional annotations. *Nucleic Acids Res* **38**, D190-5.

Nakatani, Y., Takeda, H., Kohara, Y. and Morishita, S. (2007). Reconstruction of the vertebrate ancestral genome reveals dynamic genome reorganization in early vertebrates. *Genome Res* **17**, 1254-65.

Negrini, S., Gorgoulis, V. G. and Halazonetis, T. D. (2010). Genomic instability--an evolving hallmark of cancer. *Nat Rev Mol Cell Biol* **11**, 220-8.

Ohno, S. (1970). Evolution by gene duplication. Berlin-Heidelberg-New York: Springer-Verlang.

Ostlund, G., Schmitt, T., Forslund, K., Kostler, T., Messina, D. N., Roopra, S., Frings, O. and Sonnhammer, E. L. (2010). InParanoid 7: new algorithms and tools for eukaryotic orthology analysis. *Nucleic Acids Res* **38**, D196-203.

Papadopoulos, G. L., Reczko, M., Simossis, V. A., Sethupathy, P. and Hatzigeorgiou, A. G. (2009). The database of experimentally supported targets: a functional update of TarBase. *Nucleic Acids Res* **37**, D155-8.

Papp, B., Pal, C. and Hurst, L. D. (2003). Dosage sensitivity and the evolution of gene families in yeast. *Nature* **424**, 194-7.

Patel, A. S., Murphy, K. M., Hawkins, A. L., Cohen, J. S., Long, P. P., Perlman, E. J. and Griffin, C. A. (2007). RANBP2 and CLTC are involved in ALK

rearrangements in inflammatory myofibroblastic tumors. *Cancer Genet Cytogenet* **176**, 107-14.

Petrov, D. A. (2002). DNA loss and evolution of genome size in Drosophila. *Genetica* **115**, 81-91.

Petrov, D. A. and Hartl, D. L. (1998). High rate of DNA loss in the Drosophila melanogaster and Drosophila virilis species groups. *Mol Biol Evol* **15**, 293-302.

Plaitakis, A., Spanaki, C., Mastorodemos, V. and Zaganas, I. (2003). Study of structure-function relationships in human glutamate dehydrogenases reveals novel molecular mechanisms for the regulation of the nerve tissue-specific (GLUD2) isoenzyme. *Neurochem Int* **43**, 401-10.

Pleasance, E. D., Stephens, P. J., O'Meara, S., McBride, D. J., Meynert, A., Jones, D., Lin, M. L., Beare, D., Lau, K. W., Greenman, C. et al. (2010). A small-cell lung cancer genome with complex signatures of tobacco exposure. *Nature* **463**, 184-90.

Poliseno, L., Salmena, L., Zhang, J., Carver, B., Haveman, W. J. and Pandolfi, P. P. (2010). A coding-independent function of gene and pseudogene mRNAs regulates tumour biology. *Nature* **465**, 1033-8.

Prachumwat, A. and Li, W. H. (2006). Protein function, connectivity, and duplicability in yeast. *Mol Biol Evol* **23**, 30-9.

Proost, S., Pattyn, P., Gerats, T. and Van de Peer, Y. (2011). Journey through the past: 150 million years of plant genome evolution. *Plant J* **66**, 58-65.

Pruitt, K. D., Tatusova, T. and Maglott, D. R. (2007). NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res* **35**, D61-5.

Puig, O., Caspary, F., Rigaut, G., Rutz, B., Bouveret, E., Bragado-Nilsson, E., Wilm, M. and Seraphin, B. (2001). The tandem affinity purification (TAP) method: a general procedure of protein complex purification. *Methods* **24**, 218-29.

Qi, H., Gervais, M. L., Li, W., DeCaprio, J. A., Challis, J. R. and Ohh, M. (2004). Molecular cloning and characterization of the von Hippel-Lindau-like protein. *Mol Cancer Res* **2**, 43-52.

Qian, W. and Zhang, J. (2008). Gene dosage and gene duplicability. *Genetics* **179**, 2319-24.

Rain, J. C., Selig, L., De Reuse, H., Battaglia, V., Reverdy, C., Simon, S., Lenzen, G., Petel, F., Wojcik, J., Schachter, V. et al. (2001). The protein-protein interaction map of Helicobacter pylori. *Nature* **409**, 211-5.

Rambaldi, D., Giorgi, F. M., Capuani, F., Ciliberto, A. and Ciccarelli, F. D. (2008). Low duplicability and network fragility of cancer genes. *Trends Genet* **24**, 427-30.

Ravasz, E., Somera, A. L., Mongru, D. A., Oltvai, Z. N. and Barabasi, A. L. (2002). Hierarchical organization of modularity in metabolic networks. *Science* **297**, 1551-5.

Reddy, E. P., Reynolds, R. K., Santos, E. and Barbacid, M. (1982). A point mutation is responsible for the acquisition of transforming properties by the T24 human bladder carcinoma oncogene. *Nature* **300**, 149-52.

Reisman, D., Glaros, S. and Thompson, E. A. (2009). The SWI/SNF complex and cancer. *Oncogene* **28**, 1653-68.

Reisman, D. N., Strobeck, M. W., Betz, B. L., Sciariotta, J., Funkhouser, W., Jr., Murchardt, C., Yaniv, M., Sherman, L. S., Knudsen, E. S. and Weissman, B. E. (2002). Concomitant down-regulation of BRM and BRG1 in human tumor cell lines: differential effects on RB-mediated growth arrest vs CD44 expression. *Oncogene* **21**, 1196-207.

Rual, J. F., Venkatesan, K., Hao, T., Hirozane-Kishikawa, T., Dricot, A., Li, N., Berriz, G. F., Gibbons, F. D., Dreze, M., Ayivi-Guedehoussou, N. et al. (2005). Towards a proteome-scale map of the human protein-protein interaction network. *Nature* **437**, 1173-8.

Ruan, J., Li, H., Chen, Z., Coghlan, A., Coin, L. J., Guo, Y., Heriche, J. K., Hu, Y., Kristiansen, K., Li, R. et al. (2008). TreeFam: 2008 Update. *Nucleic Acids Res* **36**, D735-40.

Salmena, L., Poliseno, L., Tay, Y., Kats, L. and Pandolfi, P. P. (2011). A ceRNA Hypothesis: The Rosetta Stone of a Hidden RNA Language? *Cell* **146**, 353-8.

Salwinski, L., Miller, C. S., Smith, A. J., Pettit, F. K., Bowie, J. U. and Eisenberg, D. (2004). The Database of Interacting Proteins: 2004 update. *Nucleic Acids Res* **32**, D449-51.

Santarius, T., Shipley, J., Brewer, D., Stratton, M. R. and Cooper, C. S. (2010). A census of amplified and overexpressed human cancer genes. *Nat Rev Cancer* **10**, 59-64.

Schlabach, M. R., Luo, J., Solimini, N. L., Hu, G., Xu, Q., Li, M. Z., Zhao, Z., Smogorzewska, A., Sowa, M. E., Ang, X. L. et al. (2008). Cancer proliferation gene discovery through functional genomics. *Science* **319**, 620-4.

Semon, M. and Wolfe, K. H. (2007). Consequences of genome duplication. *Curr Opin Genet Dev* **17**, 505-12.

Sempere, L. F., Cole, C. N., McPeek, M. A. and Peterson, K. J. (2006). The phylogenetic distribution of metazoan microRNAs: insights into evolutionary complexity and constraint. *J Exp Zool B Mol Dev Evol* **306**, 575-88.

Shah, S. P., Morin, R. D., Khattra, J., Prentice, L., Pugh, T., Burleigh, A., Delaney, A., Gelmon, K., Guliany, R., Senz, J. et al. (2009). Mutational evolution in a lobular breast tumour profiled at single nucleotide resolution. *Nature* **461**, 809-13.

Silva, J. M., Marran, K., Parker, J. S., Silva, J., Golding, M., Schlabach, M. R., Elledge, S. J., Hannon, G. J. and Chang, K. (2008). Profiling essential genes in human mammary cells by multiplex RNAi screening. *Science* **319**, 617-20.

Sjoblom, T., Jones, S., Wood, L. D., Parsons, D. W., Lin, J., Barber, T. D., Mandelker, D., Leary, R. J., Ptak, J., Silliman, N. et al. (2006). The consensus coding sequences of human breast and colorectal cancers. *Science* **314**, 268-74.

Stankiewicz, P. and Lupski, J. R. (2002). Genome architecture, rearrangements and genomic disorders. *Trends Genet* **18**, 74-82.

Stark, C., Breitkreutz, B. J., Chatr-Aryamontri, A., Boucher, L., Oughtred, R., Livstone, M. S., Nixon, J., Van Auken, K., Wang, X., Shi, X. et al. (2011). The BioGRID Interaction Database: 2011 update. *Nucleic Acids Res* **39**, D698-704.

Stebbins, C. E., Kaelin, W. G., Jr. and Pavletich, N. P. (1999). Structure of the VHL-ElonginC-ElonginB complex: implications for VHL tumor suppressor function. *Science* **284**, 455-61.

Stransky, N., Egloff, A. M., Tward, A. D., Kostic, A. D., Cibulskis, K., Sivachenko, A., Kryukov, G. V., Lawrence, M., Sougnez, C., McKenna, A. et al. (2011). The Mutational Landscape of Head and Neck Squamous Cell Carcinoma. *Science*.

Stratton, M. R., Campbell, P. J. and Futreal, P. A. (2009). The cancer genome. *Nature* **458**, 719-24.

Su, A. I., Wiltshire, T., Batalov, S., Lapp, H., Ching, K. A., Block, D., Zhang, J., Soden, R., Hayakawa, M., Kreiman, G. et al. (2004). A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc Natl Acad Sci U S A* **101**, 6062-7.

Sudarsanam, P., Iyer, V. R., Brown, P. O. and Winston, F. (2000). Whole-genome expression analysis of snf/swi mutants of Saccharomyces cerevisiae. *Proc Natl Acad Sci U S A* **97**, 3364-9.

Syed, A. S., D'Antonio, M. and Ciccarelli, F. D. (2010). Network of Cancer Genes: a web resource to analyze duplicability, orthology and network properties of cancer genes. *Nucleic Acids Res* **38**, D670-5.

Tatusov, R. L., Fedorova, N. D., Jackson, J. D., Jacobs, A. R., Kiryutin, B., Koonin, E. V., Krylov, D. M., Mazumder, R., Mekhedov, S. L., Nikolskaya, A. N. et al. (2003). The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* **4**, 41.

Tatusov, R. L., Koonin, E. V. and Lipman, D. J. (1997). A genomic perspective on protein families. *Science* **278**, 631-7.

Taylor, J. S. and Raes, J. (2004). Duplication and divergence: the evolution of new genes and old ideas. *Annu Rev Genet* **38**, 615-43.

Taylor, J. S., Van de Peer, Y., Braasch, I. and Meyer, A. (2001). Comparative genomics provides evidence for an ancient genome duplication event in fish. *Philos Trans R Soc Lond B Biol Sci* **356**, 1661-79.

Thaker, N. G., Zhang, F., McDonald, P. R., Shun, T. Y., Lewen, M. D., Pollack, I. F. and Lazo, J. S. (2009). Identification of survival genes in human glioblastoma cells by small interfering RNA screening. *Mol Pharmacol* **76**, 1246-55.

Timmons, L., Court, D. L. and Fire, A. (2001). Ingestion of bacterially expressed dsRNAs can produce specific and potent genetic interference in Caenorhabditis elegans. *Gene* **263**, 103-12.

Toll-Riera, M., Bosch, N., Bellora, N., Castelo, R., Armengol, L., Estivill, X. and Alba, M. M. (2009). Origin of primate orphan genes: a comparative genomics approach. *Mol Biol Evol* **26**, 603-12.

Truong, K. and Ikura, M. (2001). The use of FRET imaging microscopy to detect protein-protein interactions and protein conformational changes in vivo. *Curr Opin Struct Biol* **11**, 573-8.

Tyner, J. W., Walters, D. K., Willis, S. G., Luttropp, M., Oost, J., Loriaux, M., Erickson, H., Corbin, A. S., O'Hare, T., Heinrich, M. C. et al. (2008). RNAi screening of the tyrosine kinome identifies therapeutic targets in acute myeloid leukemia. *Blood* **111**, 2238-45.

Uetz, P., Giot, L., Cagney, G., Mansfield, T. A., Judson, R. S., Knight, J. R., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P. et al. (2000). A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. *Nature* **403**, 623-7.

Uetz, P. and Pankratz, M. J. (2004). Protein interaction maps on the fly. *Nat Biotechnol* **22**, 43-4.

Vallabhajosyula, R. R., Chakravarti, D., Lutfeali, S., Ray, A. and Raval, A. (2009). Identifying hubs in protein interaction networks. *PLoS One* **4**, e5344.

Van Raamsdonk, C. D., Bezrookove, V., Green, G., Bauer, J., Gaugler, L., O'Brien, J. M., Simpson, E. M., Barsh, G. S. and Bastian, B. C. (2009). Frequent somatic mutations of GNAQ in uveal melanoma and blue naevi. *Nature* **457**, 599-602.

Van Raamsdonk, C. D., Griewank, K. G., Crosby, M. B., Garrido, M. C., Vemula, S., Wiesner, T., Obenauf, A. C., Wackernagel, W., Green, G., Bouvier, N. et al. (2010). Mutations in GNA11 in uveal melanoma. *N Engl J Med* **363**, 2191-9.

VanderSluis, B., Bellay, J., Musso, G., Costanzo, M., Papp, B., Vizeacoumar, F. J., Baryshnikova, A., Andrews, B., Boone, C. and Myers, C. L. (2010). Genetic interactions reveal the evolutionary trajectories of duplicate genes. *Mol Syst Biol* **6**, 429.

Veitia, R. A. (2002). Exploring the etiology of haploinsufficiency. *Bioessays* **24**, 175-84.

Veitia, R. A. (2004). Gene dosage balance in cellular pathways: implications for dominance and gene duplicability. *Genetics* **168**, 569-74.

Veitia, R. A., Bottani, S. and Birchler, J. A. (2008). Cellular reactions to gene dosage imbalance: genomic, transcriptomic and proteomic effects. *Trends Genet* **24**, 390-7.

Venkatesan, K., Rual, J. F., Vazquez, A., Stelzl, U., Lemmens, I., Hirozane-Kishikawa, T., Hao, T., Zenkner, M., Xin, X., Goh, K. I. et al. (2009). An empirical framework for binary interactome mapping. *Nat Methods* **6**, 83-90.

Vogel, C. and Chothia, C. (2006). Protein family expansions and biological complexity. *PLoS Comput Biol* **2**, e48.

Vogelstein, B. and Kinzler, K. W. (2004). Cancer genes and the pathways they control. *Nat Med* **10**, 789-99.

Vogelstein, B., Lane, D. and Levine, A. J. (2000). Surfing the p53 network. *Nature* **408**, 307-10.

von Hansemann, D. P. (1890). On primary cancer of the liver. *Berl. klin. Wschr.* **27** 353-356.

von Mering, C., Krause, R., Snel, B., Cornell, M., Oliver, S. G., Fields, S. and Bork, P. (2002). Comparative assessment of large-scale data sets of protein-protein interactions. *Nature* **417**, 399-403.

Walhout, A. J. and Vidal, M. (1999). A genetic strategy to eliminate self-activator baits prior to high-throughput yeast two-hybrid screens. *Genome Res* **9**, 1128-34.

Wan, P. T., Garnett, M. J., Roe, S. M., Lee, S., Niculescu-Duvaz, D., Good, V. M., Jones, C. M., Marshall, C. J., Springer, C. J., Barford, D. et al. (2004). Mechanism of activation of the RAF-ERK signaling pathway by oncogenic mutations of B-RAF. *Cell* **116**, 855-67.

Watts, D. J. and Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature* **393**, 440-2.

Wheeler, B. M., Heimberg, A. M., Moy, V. N., Sperling, E. A., Holstein, T. W., Heber, S. and Peterson, K. J. (2009). The deep evolution of metazoan microRNAs. *Evol Dev* **11**, 50-68.

Wolfe, K. (2000). Robustness--it's not where you think it is. *Nat Genet* **25**, 3-4.

Wolfe, K. H. (2001). Yesterday's polyploids and the mystery of diploidization. *Nat Rev Genet* **2**, 333-41.

Xiao, F., Zuo, Z., Cai, G., Kang, S., Gao, X. and Li, T. (2009). miRecords: an integrated resource for microRNA-target interactions. *Nucleic Acids Res* **37**, D105-10.

Yachida, S., Jones, S., Bozic, I., Antal, T., Leary, R., Fu, B., Kamiyama, M., Hruban, R. H., Eshleman, J. R., Nowak, M. A. et al. (2010). Distant metastasis occurs late during the genetic evolution of pancreatic cancer. *Nature* **467**, 1114-7.

Yamamichi, N., Yamamichi-Nishina, M., Mizutani, T., Watanabe, H., Minoguchi, S., Kobayashi, N., Kimura, S., Ito, T., Yahagi, N., Ichinose, M. et al. (2005). The Brm gene suppressed at the post-transcriptional level in various human cell lines is inducible by transient HDAC inhibitor treatment, which exhibits antioncogenic potential. *Oncogene* **24**, 5471-81.

Yang, C., Ji, D., Weinstein, E. J., Choy, E., Hornicek, F. J., Wood, K. B., Liu, X., Mankin, H. and Duan, Z. (2010). The kinase Mirk is a potential therapeutic target in osteosarcoma. *Carcinogenesis* **31**, 552-8.

Yang, J., Lusk, R. and Li, W. H. (2003). Organismal complexity, protein complexity, and gene duplicability. *Proc Natl Acad Sci U S A* **100**, 15661-5.

Yu, H., Braun, P., Yildirim, M. A., Lemmens, I., Venkatesan, K., Sahalie, J., Hirozane-Kishikawa, T., Gebreab, F., Li, N., Simonis, N. et al. (2008a). High-quality binary protein interaction map of the yeast interactome network. *Science* **322**, 104-10.

Yu, J., Pacifico, S., Liu, G. and Finley, R. L., Jr. (2008b). DroID: the Drosophila Interactions Database, a comprehensive resource for annotated gene and protein interactions. *BMC Genomics* **9**, 461.

Zhou, Q., Zhang, G., Zhang, Y., Xu, S., Zhao, R., Zhan, Z., Li, X., Ding, Y., Yang, S. and Wang, W. (2008). On the origin of new genes in Drosophila. *Genome Res* **18**, 1446-55.

Zraly, C. B., Middleton, F. A. and Dingwall, A. K. (2006). Hormone-response genes are direct in vivo regulatory targets of Brahma (SWI/SNF) complex function. *J Biol Chem* **281**, 35305-15.