European School of Molecular Medicine

Thesis submitted by

ÖMER AN

For the PhD in

MOLECULAR MEDICINE
(Curriculum in Computational Biology)

Thesis Title

Role of somatic copy number variations in cancer

Supervising team

Supervisor	**Dr. Francesca Ciccarelli**
Internal advisor	**Prof. Giuseppe Testa**
External advisor	**Prof. Alexandre Reymond**

Thesis approved by the supervisor

supervisor's signature

This thesis is dedicated to my nephew Efekan and my niece Ebrar,

who will follow a better path than mine so that they can achieve further.

"If you mix the mashed potatoes and sauce, you can't separate them later. It's forever. The smoke comes out of Daddy's cigarette, but it never goes back in. We cannot go back. That's why it's hard to choose. You have to make the right choice. As long as you don't choose, everything remains possible."

the movie **"Mr. Nobody"**, 2009

I hope that I have made the right choice.

November 2015

List of Abbreviations

aCGH	:	array Comparative Genomic Hybridisation
BLAT	:	BLAST-Like Alignment Tool
CCLE	:	Cancer Cell Line Encyclopedia
CGC	:	Cancer Gene Census
CGH	:	Comparative Genomic Hybridisation
CNV	:	Copy Number Variation
CNVD	:	Copy Number Variation in Disease
CNVDB	:	Copy Number Variation Control Database
COSMIC	:	Catalogue of Somatic Mutations in Cancer
CRISPR	:	Clustered Regularly Interspaced Short Palindromic Repeats
DAISY	:	DAta mIning SYnthetic lethality identification pipeline
DECIPHER	:	DatabasE of genomiC varIation and Phenotype in Humans using Ensembl Resources
DGV	:	Database of Genomic Variants
DNA	:	Deoxyribonucleic acid
E-MAP	:	Epistatic Miniarray Profiles
ECARUCA	:	European Cytogeneticists Association Register of Unbalanced Chromosome Aberrations
FISH	:	Fluorescence *in situ* Hybridisation
GISTIC	:	Genomic Identification of Significant Targets in Cancer
GTEx	:	Genotype-Tissue Expression project
GWAS	:	Genome-Wide Association Studies
HGP	:	Human Genome Project
ICGC	:	International Cancer Genome Consortium
LINE	:	Long Interspersed Nuclear Elements
LTR	:	Long terminal repeats
MuSiC	:	Mutational Significance in Cancer
MutSig	:	Mutation Significance
NCG	:	Network of Cancer Genes
PARP	:	Poly ADP Ribose Polymerase
PolyPhen-2	:	Polymorphism Phenotyping
PPIN	:	Protein Protein Interaction Network
RNA-Seq	:	RNA Sequencing
RNAi	:	RNA interference
RT-qPCR	:	Reverse Transcription quantitative Polymerase Chain Reaction

SFE	:	Selected Functional Elements
SGA	:	Synthetic Genetic Array
shRNA	:	small/short hairpin RNA
SIFT	:	Sorting Intolerant from Tolerant
SINE	:	Short Interspersed Nuclear Elements
siRNA	:	small/short interfering RNA
SLAM	:	Synthetic Lethality Analysis by Microarray
SNP	:	Single Nucleotide Polymorphism
SV	:	Structural Variation
TCGA	:	The Cancer Genome Atlas

TCGA Cancer Types

BLCA	:	Bladder urothelial carcinoma
BRCA	:	Breast invasive carcinoma
CESC	:	Cervical squamous cell carcinoma
COAD	:	Colon adenocarcinoma
DLBC	:	Lymphoid neoplasm diffuse large B-cell lymphoma
GBM	:	Glioblastoma multiforme
HNSC	:	Head and neck squamous cell carcinoma
KICH	:	Kidney chromophobe
KIRC	:	Kidney renal clear cell carcinoma
KIRP	:	Kidney renal papillary cell carcinoma
LGG	:	Brain lower grade glioma
LIHC	:	Liver hepatocellular carcinoma
LUAD	:	Lung adenocarcinoma
LUSC	:	Lung squamous cell carcinoma
OV	:	Ovarian serous cystadenocarcinoma
PAAD	:	Pancreatic adenocarcinoma
PRAD	:	Prostate adenocarcinoma
READ	:	Rectum adenocarcinoma
SARC	:	Sarcoma
SKCM	:	Skin cutaneous melanoma
STAD	:	Stomach adenocarcinoma
THCA	:	Thyroid carcinoma
UCEC	:	Uterine corpus endometrial carcinoma

List of Figures

List of Tables

Abstract

Genetic variation is the main reason of the phenotypic differences among individuals, as well as of many human genetic diseases. Recent advances in the methods to study the human genetic variation allow better identification of its different forms, in particular of copy number variations (CNVs). The causative role of germline CNVs in Mendelian diseases and in cancer predisposition is well established. Moreover, the driver role of cancer somatic CNVs is recently emerging, and large-scale quantitative analyses elucidating their functional role in cancer genomes are needed. To achieve this, we have analysed the genomic landscape of somatic CNVs in cancer genomes in comparison to germline CNVs in the genomes of healthy individuals. We observed that somatic CNVs substantially affect the genic portion of the genome, preferentially targeting cancer genes. Moreover, this is independent of genomic features, such as DNA repeating elements and recombination rate. In particular, we confirmed that oncogenes are preferentially amplified and tumour suppressors are preferentially deleted. To investigate their functional impact, we measured the gene expression changes upon copy number variation. We observed that amplification of a gene leads to its higher expression whereas deletion results in decreased gene expression, which suggests that amplifications activate dominant genes and deletions inactivate recessive genes. The two classes of cancer genes are vastly modified consistent with their functional roles as oncogenes and tumour suppressors, with the few exceptions of frequently amplified recessive genes underlying complex epigenetic regulation.

The mutational spectrum of the human genes in cancer, together with their systems-level properties, can be exploited to identify novel targets for anti-cancer therapy, in which synthetic lethality emerges as a promising approach. Based on the working hypothesis that paralogous genes may engage in synthetic lethal interactions due to the functional redundancy between them, we combined several gene properties to predict synthetic lethality between paralogous gene pairs. Out of 37 candidate gene pairs, we experimentally

validated the synthetic lethal interaction between two components of the cohesin complex, *STAG1* and *STAG2*.

Finally, we present the latest release of Network of Cancer Genes (NCG 5.0), a manually curated database of cancer genes and their systems-level properties. NCG 5.0 collects a list of 1,571 cancer genes mutated in 13,315 cancer samples and 24 primary sites from 175 published papers. NCG has been increasingly appreciated as a central resource for cancer genomics research, facilitating candidate prioritization for hypothesis testing and experimental planning in a wide range of studies.

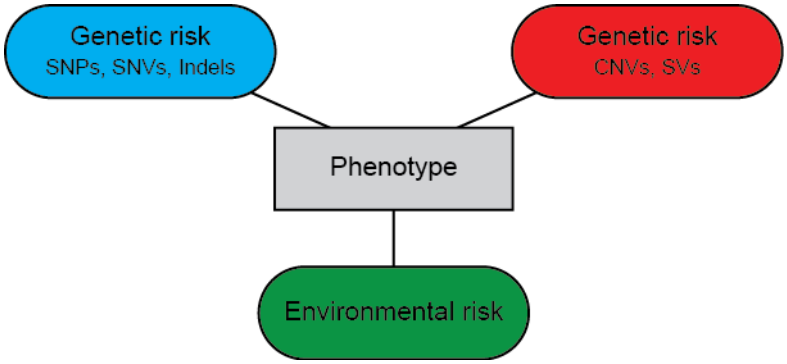
1 Introduction

1.1 Genetic variability in the human genome

Not two individuals have the same genetic make-up. Even monozygous twins meet variability soon after the fertilisation, becoming genetically different (1). Why do not twins look identical? Why do the patients respond to the same drug differently? Why do some individuals in a population develop certain genetic disorders but not others? These types of questions can be explained by the presence of genetic variability, underlying the uniqueness of each individual's genetic make-up. Although the concept of genetic variability is well established (2-5), it took decades for scientists to identify the elements that contribute to it, yet, elucidating its role on phenotype is still ongoing. This is challenging because most phenotypes have a complex origin, due to the involvement of a combination of genetic and environmental risk factors (Figure). Assessing the individual contribution of all possible factors is not only important to understand the aetiology of a disease, but also to develop treatment strategies. The ultimate goal of such studies is to decipher a genotype-phenotype map of the entire genome, and variation is the starting material for this difficult task.

The completion of the Human Genome Project (HGP) made it clear that we are far from understanding human genomic variability (2,5). Initially, single nucleotide polymorphisms (SNPs) have been thought to be the major source of genetic variability between individuals, simply because SNPs were widely studied and better defined owing to the availability of traditional detection platforms (6). Emergence of a variety of genotyping platforms led to a better characterisation of such variants in the human genome. For example, the International HapMap Project catalogued several millions of common SNPs (allele frequency >1%) in 270 individuals from several nations (7). Currently, dbSNP (v145) stores more than 85 million human SNPs derived from a large range of genotyping and other studies, and this number is likely to increase in the new releases (8).

Figure : Combined impact of genetic and environmental risks in a complex phenotype



Legend: In a complex phenotype, both genetic and environmental risks are involved. Genetic risks can be in different forms, which are grouped here according to their genomic size. Adapted from Beckmann *et. al.*, Nature Reviews, 2007.

Soon after the deciphering of the human genome sequence, the thought that SNPs are the major source of the genetic variability was challenged by the introduction of Next Generation Sequencing (NGS) technology, which revolutionised human genetic studies. NGS platforms were quickly adapted in genomic studies because they offer a number of advantages over previous techniques, such as unbiased detection of novel variants, increased specificity and sensitivity, and easier detection of rare variants (9). A huge progress towards deciphering a complete genetic variation map has been made by large-scale projects based on NGS, such as 1000 Genomes Project (3). Consequently, the spectrum of known genomic variation enlarged from common SNPs to structural variations (SVs), and the detection power improved to include also low-frequency variants (allele frequency < 1%). Together, collective efforts with an unbiased perspective and an innovative methodology advanced our knowledge and understanding of the genomic variation. With all these data and opportunities available, now the challenge has been changing from uncovering missing variability to interpreting this variability.

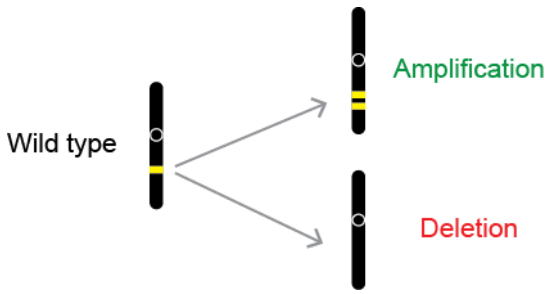
1.2 Copy number variations and disease

1.2.1 Definition of copy number variations

Discovery of structural variations revealed that human genome is much more dynamic than previously thought. An obvious reason is that structural variations account for a larger

portion of the genomic variation than do SNPs (10). Among SVs, the most common type is copy number variations (CNVs), which can be in the form of insertions, deletions and duplications. CNVs are initially defined as DNA segments between 1Kb and 3Mb present at a variable copy numbers in comparison with a reference genome (11). These thresholds are, however, arbitrarily chosen to distinguish CNVs from SNPs and large microscopic alterations, which are not strictly used today. For example, the Database of Genomic Variants (DGV) (12) collects germline CNVs between 50bp and 3Mb. In cancer, somatic CNVs are often defined as arm-level that covers an entire chromosome arm or focal that affects shorter genomic regions than an arm (13-15). Despite the flexibility in the definition of the CNV length, increase in the copy number of a DNA segment is generally referred as amplification, whereas copy number decrease as compared to the reference genome is regarded as deletion (Figure). Today, DGV (12) reports about 500,000 CNV regions and stands as the major resource for germline CNVs, among few others such as CNV DB (https://gwas.biosciencedbc.jp/cgi-bin/cnvdb/cnv_top.cgi) and The Copy Number Variation Project (http://www.sanger.ac.uk/research/areas/humangenetics/cnv/).

Figure : Types of copy number variations



Legend: Copy number variations can be broadly classified into two groups: Amplification results in gain of a DNA segment (highlighted in yellow) and deletion causes loss of a genomic region compared to the wild type.

1.2.2 Significance of copy number variations in diseases

In addition to their contribution to genetic variability (10), CNVs are often involved in disease onset. Indeed, CNVs were first recognized due to their causative role in rare genetic disorders (16). Since then, several rare human genetic syndromes as well as common disorders (such as Parkinson's disease, autism, schizophrenia, epilepsy) have been associated with CNVs (6,10,17) (Table). Today, a large collection of pathogenic CNVs associated with human genetic disorders is stored in public databases, (DECIPHER (18), ECARUCA (19), CNVD (20)). These public data are useful for researchers to study CNVs in a systematic way, as well as for clinicians referring to diagnostic purposes. For example, a deletion of 1.40Mb on chromosome 7 containing the elastin gene (*ELN*) is causative for Williams-Beuren Syndrome (DECIPHER, https://decipher.sanger.ac.uk).

table continued on the next page…

Table : Mendelian and complex diseases associated with CNVs

Phenotype	Locus/Gene	CNV	References
Mendelian (autosomal dominant)			
Williams-Beuren syndrome	7q11.23	del	(21)
7q11.23 duplication syndrome	7q11.23	dup	(22)
Spinocerebellar ataxia type 20	11q12	dup	(23)
Smith-Magenis syndrome	17p11.2/*RAI1*	del	(24)
Potocki-Lupski syndrome	17p11.2	dup	(25)
HNPP	17p12/*PMP22*	del	(26)
CMT1A	17p12/*PMP22*	dup	(27)
Miller-Dieker lissencephaly syndrome	17p13.3/*LIS1*	del	(28,29)
Mental retardation	17p13.3/*LIS1*	dup	(30)
DGS/VCFS	22q11.2/*TBX1*	del	(31,32)
Microduplication 22q11.2	22q11.2	dup	(33-35)
Adult-onset leukodystrophy	*LMNB1*	dup	(36)
Mendelian (autosomal recessive)			
Familial juvenile nephronophthisis	2q13/*NPHP1*	del	(37,38)
Gaucher disease	1q21/*GBA*	del	(39)
Pituitary dwarfism	17q24/*GH1*	del	(40,41)
Spinal muscular atrophy	5q13/*SMN1*	del	(42,43)
beta-thalassemia	11p15/*beta-globin*	del	(44)
alpha-thalassemia	16p13.3/*HBA*	del	(45)
Mendelian (X-linked)			
Hemophilia A	*F8*	inv/del	(46)
Hunter syndrome	*IDS*	del/inv	(47-49)
Ichthyosis	*STS*	del	(50)
Mental retardation	*HUWE1*	dup	(51)
Pelizaeus-Merzbacher disease	*PLP1*	del/dup/tri	(52-56)
Progressive neurological symptoms (MR+SZ)	*MECP2*	dup	(57-59)
Red-green colour blindness	opsin genes	del	(60)
Complex traits			
Alzheimer disease	*APP*	dup	(61)
Autism	3q24	inherited homozygous del	(62)

	16p11.2	del/dup	(63-66)
Crohn disease	*HBD-2*	copy number loss	(67)
	IRGM	del	(68)
HIV susceptibility	*CCL3L1*	copy number loss	(69,70)
Mental retardation	15q13.3	del	(71)
	17q21.31	del	(72-74)
	Xp11.22	dup	(51)
Pancreatitis	*PRSS1*	tri	(75)
Parkinson disease	*SNCA*	dup/tri	(76-80)
Psoriasis	*DEFB*	copy number gain	(81)
Schizophrenia	1q21.1	del	(82-84)
	15q11.2	del	(82)
	15q13.3	del	(82,83)
Systemic lupus erythematosus	*FCGR3B*	copy number loss	(85-87)
	C4	copy number loss	(88)

Legend: A literature-derived collection of genomic disorders associated with CNVs. del: deletion, dup: duplication, inv: inversion, tri: triplication. Adapted from Zhang *et. al.*, Annu. Rev. Genomics Hum. Genet., 2009.

For years, genome wide association studies (GWAS) used SNPs to identify biomarkers for susceptibility to complex diseases. Due to the improved detectability and the increasing significance of CNVs in disease, it did not take long for researchers to notice that CNVs may also be incorporated into GWAS (89,90). As a result, many *de novo* CNVs increasing disease risk have been discovered (91-93). Moreover, together with SNPs, CNVs also contribute to differential response to drugs. At present, certain forms of genetic disorders (developmental delay/intellectual disability (94), multiple congenital anomalies (95) and neuropsychiatric disorders (96)) can be routinely diagnosed via the presence of abnormal copy numbers of specified genes or DNA segments.

1.2.3 Copy number variations and cancer

In addition to their pathogenic role in human genetic diseases, CNVs also play a major role in cancer. Owing to the advance in the CNV detection techniques, a large amount of CNV data from tumour samples have become available, leading to the systematic analysis of such variants in cancer. Initially, the role of CNVs in cancer predisposition gained supporting evidence (97-100) (Table). For example, individuals with genomic rearrangements involving *BRCA1* and *BRCA2* genes carry a higher risk for hereditary breast and ovarian cancers (101-103).

Subsequent studies revealed that cancer genomes might be heavily affected by somatic CNVs, which substantially changed the interpretation of cancer genomics from SNP-only analysis to an integrated approach including structural variations. In cancer, CNVs may arise due to the increased genomic instability that favours copy number gain or loss, resulting in an increase (amplification) or decrease (deletion) of the genomic regions that often contains genes. For example, *RB1* and *APC* deletions are associated with familial retinoblastoma and colorectal cancer, respectively (104,105). The detrimental effects of CNVs mainly arise from the unbalance in gene copy numbers that lead to altered gene expression (106). This may bring to the activation of oncogenes or to the inactivation of tumour suppressor genes (107). Many oncogenes (such as *MYC* (108)) and tumour

suppressor genes (such as *PTEN* (109)) are found amplified or deleted, respectively, in several cancer types. The recent advent of technology such as microarray-based hybridisation and next-generation sequencing has made it possible to precisely identify CNVs in large number of samples with higher resolution and at lower cost (3,13). Currently, genomic databases such as TCGA, Tumorscape (13) and COSMIC (110) store huge data on somatic CNVs from large cohorts of cancer samples.

Somatic CNVs may substantially vary in quantity, length and genomic position across and within cancer types (13,15,111). Therefore it is important to quantify and characterize somatic CNVs to better understand their impact on cancer. Despite such a prominent role, it is challenging to identify the genomic regions that bear the actual effect on phenotype (driver events). Because a single CNV can contain several to hundreds of genes and there may be several such variant regions within a single genome. A common approach to find driver events is to consider recurrent CNV regions across samples (13,112). For the same purpose, increasing number of studies use an integrative approach by combining CNV data with other types of information, such as somatic mutations, DNA methylation and gene expression (14,113).

table continued on the next page…

Table : Known cancer predisposition genes associated with CNVs

Gene	Cancer type	Locus	References
APC	Colorectal, pancreatic, desmoid, hepatoblastoma, glioma, other CNS cancers	5q22.2	(105,114)
BMPR1A	Gastrointestinal polyps	10q22.3	(115)
BRCA1	Breast, ovarian	17q21	(101,116)
BRCA2	Breast, ovarian, pancreatic, leukaemia (*FANCB*, *FANCD1*)	13q12.3	(103)
CDH1	Gastric, breast	16q22.1	(117)
CDKN1B	Pituitary tumour, testicular tumour	12p13.1	(118)
CDKN2A	Melanoma, pancreatic	9p21	(119)
CHEK2	Breast, prostate	22q12.1	(120,121)
CREBBP	Nervous system, brain, leukaemia	16p13.3	(122)
CYLD	Multiple skin appendage tumours	16q12.1	(123)
EPCAM	Colorectal, endometrial	2p21	(124)
EXT1	Exostoses, osteosarcoma	8q24.11	(125)
EXT2	Exostoses, osteosarcoma	11p11.2	(125)
FANCA	Acute myeloid leukaemia	16q24.3	(126)
FH	Lieomyomatosis, renal	1q42.1	(127)
FLCN	Renal cell carcinoma	17p11.2	(118)
GPC3	Wilms' tumours	Xq26	(118)
HRPT2	Parathyroid carcinoma, renal cell carcinoma	1q31.2	(128)
JAG1	Hepatocellular carcinoma, papillary thyroid carcinoma	20p12	(129,130)
MADH4	Gastrointestinal polyps	18q21.1	(131)
MEN1	Parathyroid adenoma, pituitary adenoma, pancreatic islet cell, carcinoid	11q13	(132)
MSH2	Colorectal, endometrial, ovarian	2p21	(114)
MSH6	Colorectal, endometrial, ovarian	2p16	(133)
NF1	Neurofibroma, glioma	17q11.2	(134)
NF2	Meningioma, acoustic neuroma	22q12.2	(135)
NSD1	Increased risk of benign or malignant tumours, including neuroblastoma and gastric carcinoma	5q35.3	(118)
PMS2	Colorectal, endometrial, ovarian, medulloblastoma, glioma	7p22	(136)
PRKAR1A	Myxoma, endocrine, papillary thyroid	17q24.2	(137)
PTCH1	Skin basal cell, medulloblastoma	9q22.3	(138)
PTEN	Breast cancer, leukaemia, renal cell adenocarcinoma, neuroendocrine carcinoma, Merkel cell carcinoma	10q23.31	(139)
RB1	Retinoblastoma, sarcoma, breast, small cell lung	13q14.2	(140)

RUNX1	Acute myeloid leukaemia	21q22.12	(141)
SDHB	Paraganglioma, pheochromocytoma	1p36.13	(142)
SDHC	Paraganglioma, pheochromocytoma	1q21	(143)
SDHD	Paraganglioma, pheochromocytoma	11q23	(143)
SMAD4	Colon, stomach, small bowel and pancreas	18q21.2	(131)
SMARCB1	Schwannomas, malignant rhabdoid	22q11	(144)
STK11	Jejunal harmartoma, ovarian, testicular, pancreatic	19p13.3	(145)
TP53	Breast, sarcoma, adrenocortical carcinoma, glioma, multiple other tumour types	17p13.1	(146)
TSC1	Hamartoma, renal cell	9q34	(147)
TSC2	Hamartoma, renal cell	16p13.3	(147)
VHL	Renal, hemangioma, pheochromocytoma	3p25.3	(148)
WT1	Wilms' tumour	11p13	(149)

Legend: A literature-derived collection of cancer predisposing genes associated with CNVs. Adapted from Krepischi *et. al.*, Future Oncol., 2012.

1.3 Detection of copy number variations

CNVs can be detected using a variety of platforms, which can be broadly classified into two groups: hybridisation-based and sequencing-based platforms. Between the two groups, there are substantial differences in terms of the approach used, therefore the resolution of the output greatly varies.

Hybridisation-based platforms span a wide range of methods, which are essentially evolving forms of the same biological principal throughout time. For a long time, cytogenetics has been in play, which includes karyotyping, fluorescence *in situ* hybridisation (FISH), comparative genomic hybridisation (CGH), array comparative genomic hybridisation (aCGH) and SNP arrays (150,151). The earliest studies used karyotyping, where unique banding patterns of the chromosomes are observed by using a microscope, allowing of the detection of chromosomal abnormalities (152). In FISH, fluorescently labelled probes are hybridized to specific DNA segments, which are then analysed by fluorescence microscopy (153). CGH was used to reveal the ploidy of cells by labelling the DNA of a test and a reference sample hybridized to metaphase chromosomes (154). All of these techniques, however, are limited to detect copy number changes at chromosome level, or down to 5 Mbp in ideal conditions. More recent techniques, therefore, focused on detecting smaller CNVs. Use of microarrays together with CGH (aCGH) led to a locus-specific measure of CNVs, which increased the resolution to detect variants as short as 100Kb (155). Moreover, compared to CGH, aCGH does not require the use of metaphase chromosomes (156). However, both techniques are unable to detect aberrations that do not result in copy number changes (balanced structural variations) (155). SNP arrays have overcome this limitation: they are capable of detecting loss of heterozygosity events, which are commonly observed in tumorigenesis (157). However, the resolution of SNP arrays is still confined to the designed probes and prior knowledge of the variants to be analysed is required.

Sequencing-based platforms provide a completely different approach to detect CNVs, overcoming the aforementioned limitations. Ideally, NGS-based technology allows detection of much smaller SVs compared to the hybridisation-based platforms. This improvement led to the refinement of the definition of CNV length i.e. currently CNVs are widely accepted as segments longer than 50bp (158). Moreover, the boundaries of the CNVs can be identified at single-base resolution, fostering studies on CNV breakpoint analyses (13,159,160). Another powerful feature of sequencing-based approach is the capability of detecting novel variants (158).

Currently SNP arrays and sequencing-based platforms are widely used in CNV analyses as they provide a quick and cost-effective solution. However, an important aspect of such platforms is the accuracy of the downstream analysis, i.e. identification of the actual regions altered in the genome, known as CNV calling. A number of efficient tools were developed to call copy number variations from both SNP arrays and sequencing data. The design of SNP arrays is rather straightforward, whereas the sequencing strategies can be diverse, reflecting on the diversity of the tools developed. Each tool has its own advantages and limitations, and there is not always a high concordance among them. Comparison of different platforms and CNV calling tools has been extensively reviewed in the literature (161,162).

1.4 Systems-level properties of cancer genes

To understand the impact of diseases on the cell fitness, it is crucial to elucidate how they alter the gene function. Copy number variations that we have explained so far represent a class of mechanical changes to the genome, with an immediate potential to alter the dosage of gene function (163). In a broader context, human genes have also adapted some intrinsic properties throughout evolution that can help characterise their functional role in diseases. For example, disease genes tend to be non-essential and their products occupy functionally peripheral positions in protein networks (164).

The better understanding of the complexity of biological systems urged to analyse biological data differently than the classical methods. In many fields, scientific research has substantially changed perspective from single-case analysis to a broader outlook. New studies started to integrate all potential contributors of a given condition to better assess the actual causes and mechanisms leading to that condition. As a result, also triggered by the availability of advanced computational tools to explore genomics data, a new research field has emerged: systems biology.

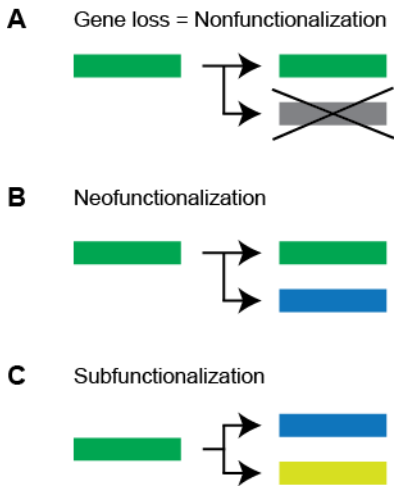
Biological systems are dynamic and complex, and their behaviour may be challenging to predict from the properties of individual parts (165). Systems biology is thus the study of interacting biological components as a whole. Emergence of systems biology not only expanded the vision of the approaches to the biological problems, but also helped better identify the components of the whole picture (166). In this context, a wide range of gene properties is studied to address different biological questions. Here we overview several systems-level gene properties which we use to distinguish the cancer genes from the rest of human genes and to identify novel targets for anti-cancer therapy:

1.4.1 Gene duplicability

Gene duplicability is defined as the tendency of a gene to preserve its duplicates in the genome (167). Gene duplications can arise from different mechanisms such as retroposition (168), segmental duplication (169), chromosomal duplication or whole genome duplication (170). Following the duplication event, most of the cases the new gene copies are lost due to the deleterious mutations, which render them non-functional (Figure A). Rarely, and more importantly in evolutionary terms, duplication may result in creation of new genes via two possible processes: neofunctionalization or subfunctionalization (171). In neofunctionalization, one of the copies gains a new function by acquiring a beneficial mutation, which is positively selected and fixed in the population, while the other copy retains its original function (Figure B) (172). In subfunctionalization, both copies go through functional divergence, specializing to perform different parts of the

original gene function, which gives rise to paralogs with complementary functions (Figure C) (172). A variety of methods have been developed to measure gene duplicability and identify paralogous genes. In our group, we use our pipeline which relies on the sequence conservation (173) and utilises the sequence alignment tool BLAT (174). Based on our pipeline, for example, *TP53* has only one duplicate, *TP73*, covering 13% of its sequence (NCG 5.0, http://ncg.kcl.ac.uk/). In general, cancer genes tend to be less duplicated than the rest of human genes (173,175).

Figure : Evolutionary processes following gene duplication



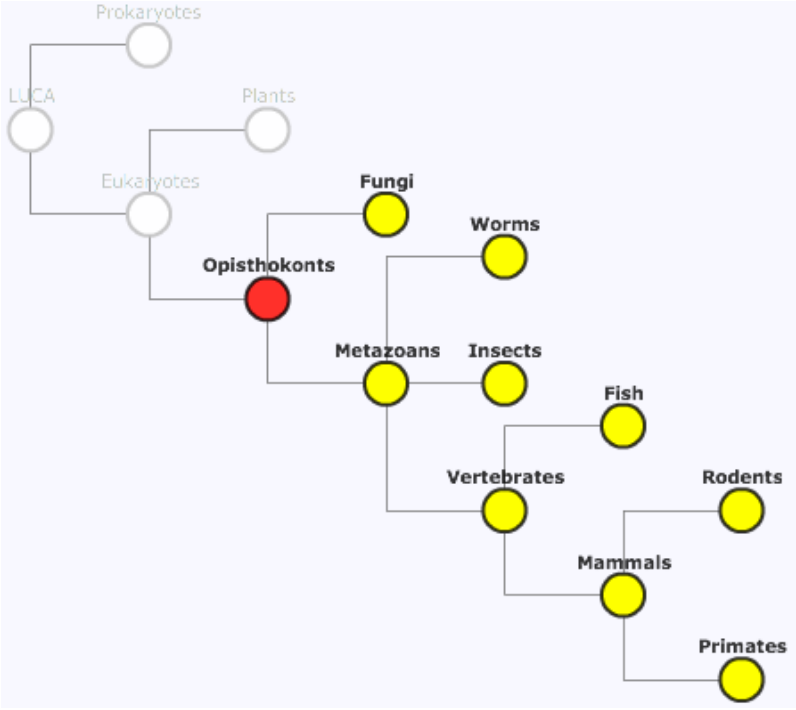
Legend: Three possible fates after gene duplication. Each rectangle represents a gene and each colour represents a different function. **A)** Most common outcome is a non-functional gene copy which is lost in time. **B)** Duplicated copy may gain a new function which is fixed if beneficial while the original gene retains its function. **C)** Both original gene and duplicated copy specialize on complementary functions. The functional divergence in each process is driven by a mutational event. Adapted from B. Conrad and S.E. Antonarakis, Annu. Rev. Genomics Hum. Genet., 2007.

1.4.2 Orthology and evolutionary appearance

Unlike paralogs, which originate via a duplication event within the same genome, orthologs are genes in different species which diverged from a common ancestral gene by a speciation event. Orthologs often have similar functions (176), but not always (177). The identification of orthologs is not an easy task, which led to the development of many different tools with a variety of approaches. Briefly, some approaches use graphs cluster of

genes from different species into groups based on all-against-all sequence comparison. Reciprocal best hits between genes of two different species are then regarded as orthologs (178-183). Other methods start from multiple sequence alignments of gene families to build phylogenetic trees. These trees are reconciled based on the species tree, and used to detect speciation and duplication events within each gene family. Then speciation and duplication events are associated with orthology and paralogy relationships, respectively (184-187). In addition to detect functional counterparts, orthology assignment is also useful to trace back to the evolutionary origin of a gene, defined as the most ancient internal node where an ortholog can be found (188). For example, *TP53* originated in opisthokonts according to the eggNOG database (v4.0) (189) (Figure).

Figure : Evolutionary origin of *TP53*



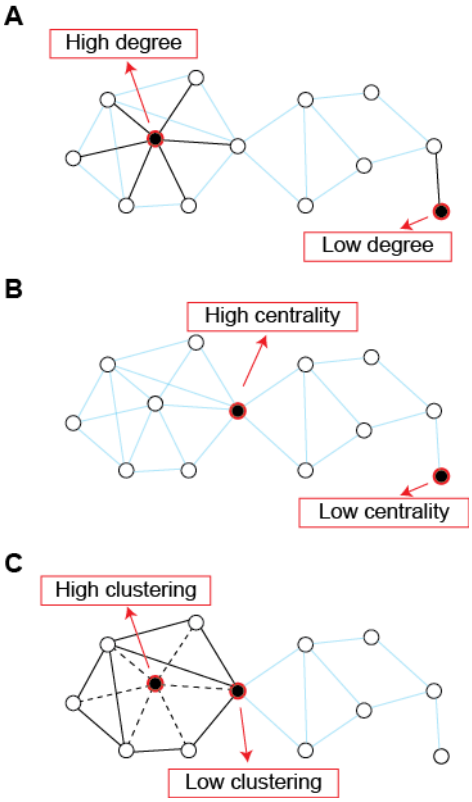
Legend: Tree of life representing the evolutionary origin and presence of orthologs of *TP53*. Shown are seven branches in evolution and representative groups of species at each branch. Yellow denotes the presence of orthologs and red stands for the most ancient node in which an ortholog is found, i.e. the origin. Adapted from (NCG 5.0, http://ncg.kcl.ac.uk/).

1.4.3 Network properties

A network is a graphical representation of entities (nodes) and connections between them (edges). In a biological network, for example, nodes may represent proteins and edges may represent physical interactions between them (190). Each network has its own structure, which can be assessed through many properties including size, density and diameter. Within a network, nodes can also be characterized by their individual properties. The most widely analysed node properties are degree, betweenness and clustering coefficient (191).

In a protein-protein interaction network (PPIN), degree is the number of direct interactions of a protein, which is a measure of the connectedness of that protein (Figure A). Highly connected proteins are regarded as hubs, defined as proteins with a degree above a certain threshold in the degree distribution. For example, p53 and Myc are two important hubs with a degree of 859 and 560 in human PPIN, respectively (NCG 5.0). Betweenness is a measure of centrality, defined as the number of shortest paths from all nodes to all others that pass through that node (Figure B). Central nodes, thus, represent key proteins that are potentially involved in many cellular processes. For example, FBXO6 and MDM2 are two extreme proteins with high centrality, with a betweenness score of 1,458,093 and 1,400,662, respectively (NCG 5.0). Clustering coefficient is the ratio of existing links between the neighbours of a node to the maximum possible number of such links (Figure C). In other words, it shows how much the neighbours of a protein is connected or clustered together. For example, INO80D has a degree of 14, and all of them are connected to each other, hence the clustering coefficient is the maximum, i.e. 1 (NCG 5.0). In general, proteins encoded by cancer genes are enriched in hubs and in central positions in human PPIN (167,173).

Figure : Basic network properties



Legend: A simple undirected graph with 13 nodes and 23 edges representing basic network metrics. Highlighted on the graph are **A)** a highly connected node with a degree of 6 and a peripheral node with a degree of 1, **B)** a highly central node located on most of the paths between any two nodes and a peripheral node rarely visited across the paths, **C)** a highly clustered node whose neighbours have 8 out of 15 possible edges and a lowly clustered node with 1 out of 10 possible edges. Note that the same node can be highly central **(B)** but lowly clustered **(C)**, showing that each measure indicates a distinct property. Adapted from Sporns, Front. Comput. Neurosci., 2011 and Sporns, Dialogues Clin. Neurosci., 2013.

1.4.4 Gene expression

The central dogma of molecular biology explains the information flow from DNA to RNA (transcription), and RNA to protein (translation) as a unidirectional process (192). In this flow, gene expression refers to the synthesis of a functional product (i.e. RNA or protein) from the genetic code, and they are synthesized only when and as much as they are needed (gene regulation). In functional studies, assessing the expression of the genes of interest is a crucial step because it gives information on the cellular processes or pathways operating in the cell. This might be particularly important when comparing a disease condition to the normal state. For example, expression of *TNFSF10*, a member of the

tumor necrosis factor (TNF) ligand family, induces apoptosis (193) and overexpression of the oncogene *EGFR* triggers cell proliferation in tumour cells (194).

Gene expression can be measured by quantifying the mRNA or protein levels. mRNA quantification is more widely used as it is easier to detect and can be performed by a variety of lab techniques such as nuclease protection assay, northern blotting and RT-qPCR. For high-throughput expression profiling, microarrays and RNA sequencing are commonly used. Microarrays are based on the hybridisation of the pre-designed probe sequences to the mRNA of the target genes. Instead, RNA sequencing gives information on the abundance of RNA transcripts that are present at the cell. One of the many applications of gene expression measurement is to assess the functional impact and the underlying mechanisms of gene copy number changes on the genome (195,196).

1.5 Synthetic lethality

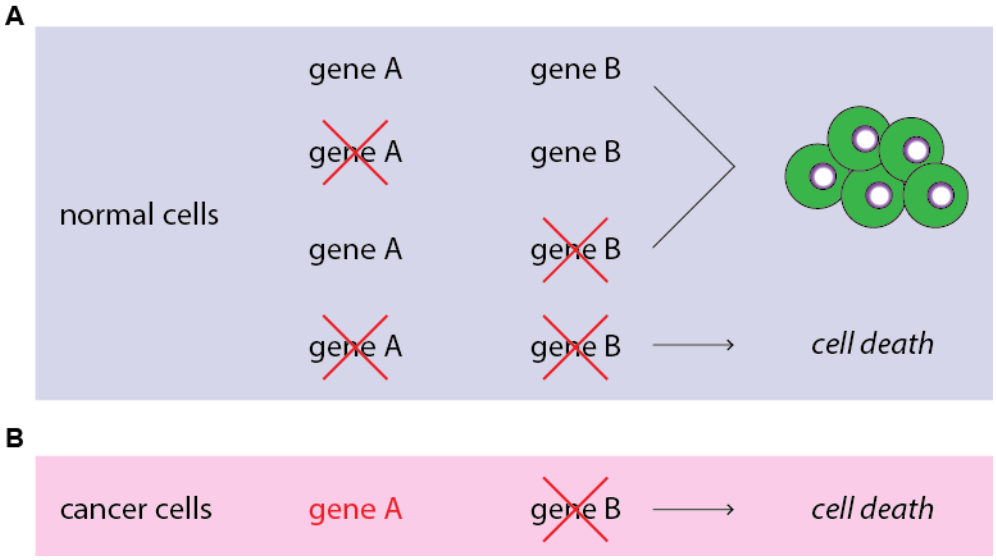
We have explained some of the systems-level properties of genes and illustrated with some examples. In principle, these and other gene properties can be helpful in addressing many different research questions, for example to distinguish cancer genes from the rest of human genes, as well as in practical implications in developing treatment strategies. Next, we demonstrate how such gene properties can be used to identify potential targets in anti-cancer therapy.

Increasing amount and better accessibility of the genomic data led to the rapid development of novel tools and strategies fostering our understanding of the human genetic diseases. However, there is still a gap between genomic research and patient treatment, which urges the need to translate biological knowledge into practical applications in therapy. In this respect, one of the strategies proved to be promising in anti-cancer therapy is known as synthetic lethality. The concept of synthetic lethality has been very popular and widely exploited in the recent literature as it represents an alternative and effective approach to identify novel therapeutic targets. This is much appreciated when

considering the complexity of targeting cancer, such as low rate of druggability of cancer drivers (197) and “non-oncogenic addiction” of cancer state (i.e. dependency on genes that are not oncogenes themselves but act in oncogenic pathways) (198).

Two genes are regarded as synthetic lethal if mutation in one of them does not interfere with cell viability while mutation in both leads to cell death (Figure A). Such a scenario implies the existence of a specific function required for the cell survival that is dependent on either of these two genes. In addition, both genes must be non-essential because the cell would not survive as well in case of mutation in one of them. In principle, synthetic lethality provides a powerful framework for killing cancer cells in a specific manner. Because only cancer cells carry the mutated gene whose function is impaired, therefore knockout of its synthetic lethal partner will result in the death of the cancer cells only (Figure B). In addition to loss-of-function mutations, synthetic lethality with gain of function mutations, gene overexpression (199,200), epigenetic changes (201) and cell extrinsic differences (202,203) have also been reported. Detecting such alterations in the genome is straightforward, but how can their synthetic lethal partners be identified?

Figure : Synthetic lethality



Legend: A representative scheme of synthetic lethality between two genes. In the simplest terms, a red cross represent mutations or perturbations that lead to the functional impairment of the gene. **A)** In the presence of synthetic lethality, co-mutation of the genes leads to cell death, while in any other combination the cells are viable. **B)** In cancer cells, one of the synthetic lethal genes is already mutated (highlighted in red), and mutagenesis of its partner selectively kills cancer cells. Adapted from Hühn *et. al.*, Swiss Medical Weekly, 2013.

A variety of approaches have been used to discover synthetic lethality. Early studies mainly focused on the high throughput screenings of the yeast genome (*S. cerevisiae*) by using RNA interference (RNAi) or drug libraries to identify synthetic lethal interactions (204). Yeast has been a popular model organism for this purpose because of abundance of genetic interactions reported in functional analyses, also known as yeast knockouts (205). For example, a genome-wide strategy through synthetic genetic array (SGA) analysis has been developed for the systematic construction of double mutants (206,207). Another technique introduced the use of microarrays to probe genome-wide gene-chemical and gene-gene interactions, called synthetic lethality analysis by microarray (SLAM) (208,209). Moreover, a modified version of the SGA method, called epistatic miniarray profiles (E-MAPs), was developed for quantitative pairwise measurements of the genetic interactions in a selected subset of the genome (210-212). In addition to these experimental designs, a number of computational algorithms utilizing *in silico* predictions were

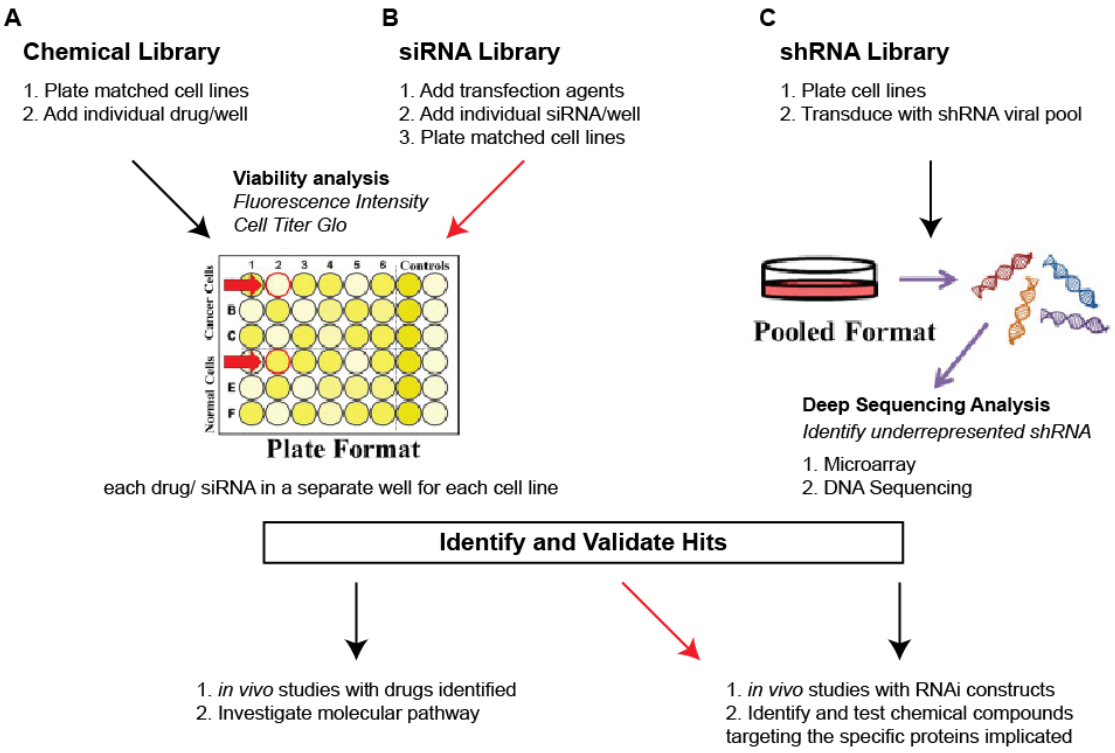
developed aiming to understand the patterns at genome scale resulting in synthetic lethality (213-220). These approaches collectively identified numerous synthetic lethal interactions in yeast and other model organisms (221-224). However, such genome-wide screening studies have a number of limitations including high cost, high rate of false positives, variation among different strains, different growth conditions and lack of mechanistic explanation. Most importantly, synthetic lethal pairs found in yeast are not readily applicable to human genes, which require orthologous mapping and investigation of additional factors intrinsic to the human genome (225).

More recently, synthetic lethality has gained great importance in human genome studies. Several reasons can be listed for this: better understanding of the human interactome owing to the functional analyses and RNA sequencing, development of efficient molecular tools for cell engineering, and urging need for alternative therapeutic strategies in cancer (226). As a result, a large number of synthetic lethal interactions have been identified and experimentally validated, yet many more were predicted *in silico* (227). Approaches to identify synthetic lethal interactions can be divided into 3 groups: screening-based, hypothesis-driven and computational approach.

Screening-based approaches rely on the genome-wide comparison of the effect of chemical compounds or RNAi interference between a test and genetically matched cell line. In this design, the only difference between the two cell lines is the expression/activation status of the gene of interest. Different libraries can be prepared for such a screening. For example, chemical libraries (Figure A) were used to identify the synthetic dependencies of VHL-deficient renal cell carcinomas (228,229) and Fanconi anaemia pathway-deficient ovarian cancer (230). Small interfering RNAs (siRNA) libraries, which are prepared on a plate wherein each siRNA is transfected separately in its own well (Figure B), led to the identification of *FAT1* as an antagonist of caspase-8 in extrinsic apoptosis in patient-derived glioblastoma cell lines (231), and of *ATR* sensitized by the topoisomerase I inhibition in a breast cancer cell line (232). Use of small hairpin RNA (shRNA) libraries,

which consist of shRNA virus pools (Figure C), led to the identification of synthetic lethality between *CCNE1* amplification and loss of *BRCA1* (233), and *KRAS* and *STK33* suppression (234). In addition, an emerging method of genome editing called clustered regularly interspaced short palindromic repeats (CRISPR) (235) is likely to be used in synthetic lethality screens more frequently in the near future.

Figure : Screening-based approaches to detecting synthetic lethal interactions



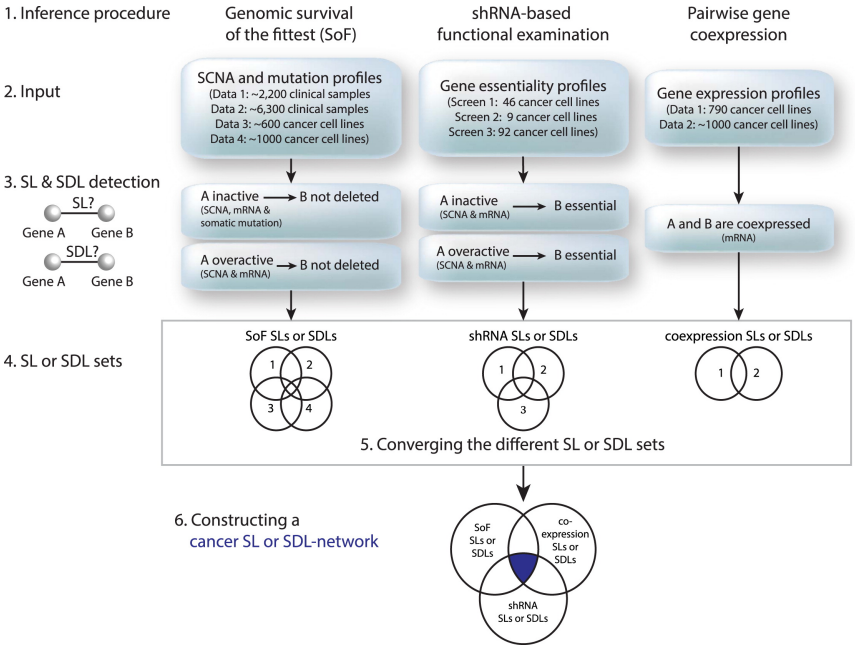
Legend: Steps of library preparation and target identification for synthetic lethality screens of **A)** chemical, **B)** siRNA, and **C)** shRNA library approaches. Adapted from Thompson *et. al.*, Yale J. Biol. Med., 2015.

Hypothesis-driven approaches rely on established knowledge of well-characterized patterns of genomic alterations or other events that may give hint of specific vulnerabilities in cancer. For example, frequent mutations in the *BRCA1* and *BRCA2* genes in breast and ovarian cancers and the functional role of these genes in homologous recombination and DNA double-strand break repair are well known. This led to the hypothesis that targeting the DNA repair defect would be synthetically lethal with loss of *BRCA1/2*, which led to discovery of PARP inhibitors as a therapeutic target in these tumours (236,237). Likewise,

paralogous genes may have functional redundancy due to their common ancestral origin and certain level of sequence conservation, particularly in the early course of evolution after a duplication event. Based on this, hypothesizing that functionally paralogous cancer genes may act as negative genetic interactors, our group previously demonstrated synthetic lethal interactions between three novel gene combinations, *SMARCA4-SMARCA2*, *CDH1-CDH3* and *DNMT3A-DNMT3B-DNMT1* (188). Another study utilised the mutual exclusivity of *KRAS* and *EGFR* mutations in human lung adenocarcinomas, and observed that co-expression of both genes are toxic to cells, implying synthetic lethality (238). The same concept of mutual exclusivity was also used at the genome scale in a study of breast, prostate, ovarian and uterine cancers, revealing 718 genes that are likely to be synthetic lethal with six key DNA-damage response genes (239).

Computational approaches integrate multiple types of data to predict synthetic lethal interactions based on a model. These approaches take advantage of high calculation capability to exploit large genomic profiles. A comprehensive pipeline for this purpose (data mining synthetic lethality identification pipeline, DAISY) statistically infers synthetic lethal interactions from cancer genomic data of both cell lines and clinical samples (240). DAISY builds a cancer synthetic lethality network collecting mutational and functional profiles of the human genes in cancer samples (Figure). Another tool uses a hybridised method, which combines a data-driven model with knowledge of signalling pathways to simulate the influence of single and double gene knock-down to cell death, and assigns a probability score to the synthetic lethal candidates based on the cell viability (241).

Figure : The workflow of DAISY



Legend: DAISY is a computational approach to predict synthetic lethal interactions in cancer. Shown are the steps of three different inference procedures leading to the construction of a synthetic lethality network. Adapted from Jerby-Arnon *et. al.*, Cell, 2014.

1.6 Aim and rationale of the thesis

The aim of this thesis is to understand the role of somatic CNVs in cancer. In extension, systems-level properties of cancer genes are utilised to identify novel therapeutic targets that confer synthetic lethality.

Variability in the human genome is the ultimate source of genetic diseases. Although each genome is unique, certain chromosomal abnormalities or gene aberrations result in the same phenotype in different individuals. For example, extra copy of chromosome 21 (trisomy 21) leads to Down syndrome which is essentially characterized by intellectual disability (242). At a smaller scale, the functional impairment of a single gene, *PRF1*, leads to an autosomal recessive disorder known as familial hemophagocytic lymphohistiocytosis type 2 (243). These examples indicate the presence of a certain level of robustness in the human genome, which allows a systematic analysis of disease-phenotype associations. However, in complex diseases such as cancer, linking a phenotype to its causal genotype is often much more complicated. This is mainly due to the multiple

genomic aberrations that simultaneously occur in cancer cells, which include only a few driver events and many passenger events. For example, *TP53*, a well-known tumour suppressor involved in cell cycle, DNA repair and other key regulatory mechanisms, is frequently mutated in more than 20 different cancer types (International Cancer Genome Consortium, ICGC, https://dcc.icgc.org/). However, *TP53*-mutated samples also include several to hundreds of other mutated genes, yet some samples lack mutation in *TP53*. In such a scenario, it is not possible to link the same cancer phenotype to a single driver in each tumour; it is actually a challenging task to identify the drivers in each sample. On the other hand, some drivers mutate at a higher frequency in certain cancer types. For example, mutations in *APC* are observed in more than 30% of the colorectal cancers, similarly, *PIK3CA* is mutated in above 30% of the breast cancer samples (ICGC). Recurrence of a genomic alteration, thus, may be informative of the driver role of the altered gene in certain contexts. Taken together, these observations imply that most of the cancer genes are tumour and even sample specific (244,245).

Nevertheless, answers to the challenging questions raised above ultimately lie in the sequence of the human genome. The genomic sequence is the raw material that must be processed to understand the mechanistic causes of diseases. The sequence alterations can be in many different forms, including single nucleotide variations, copy number variations and genomic rearrangements. Among these, this thesis particularly deals with copy number variations. Briefly, the properties of somatic CNVs that are acquired in the cancer genomes are characterised and compared to those of germline CNVs that are inherited. Then the impact of somatic CNVs on gene expression is investigated to explain how this can lead to the activation or inactivation of cancer genes. The genomic features, which may potentially impact on CNV formation, are further inquired. Finally, the genes that are frequently amplified and deleted in cancer are analysed to see the potential contributors of the disease.

2 Methods

2.1 Dataset of human genes and cancer genes

2.1.1 Human genes

To derive the human gene set, we used the pipeline previously developed in our group (173). First, we downloaded the protein sequences of all human genes from NCBI RefSeq database (version 51) and aligned them to the reference human genome (Hg19) using a sequence similarity search tool (BLAST-like Alignment Tool (BLAT) (174)). For each protein sequence, the tool gives an alignment hit in the genome with the highest score (best hit), and any possible additional hits at lower coverages (duplicates). The best hit of each protein is considered as the gene locus. We retained only the hits that map to the autosomal or sex chromosomes on the reference genome, and discarded those mapping to alternate haplotype sequences or random chromosomes. In cases of overlapping best hits, we retained only the longest one as the representative of the locus, to have only one isoform for each gene corresponding to a unique genomic region. Finally, we removed all the genes with their best hit shorter than 60% of the original protein length. This was done to eliminate spurious hits, which do not correspond to the original locus of the gene. This resulted in 19,045 unique human genes.

2.1.2 Cancer genes

Cancer genes are the mutated genes that are causally implicated in oncogenesis (246). Based on this definition, we collected a union of 501 cancer genes from the Cancer Gene Census (CGC (246), 448 genes (January 2012) and the literature ((247), 77 genes). Genes from the CGC are divided into dominant (349) and recessive (103) genes, and 4 of them can act as both dominant and recessive. The dominant and recessive cancer genes can be defined as the following:

1) Dominant cancer genes acquire mutation on single allele and this is sufficient to promote cancer
2) Recessive cancer genes require both alleles to be mutated to promote cancer

Usually, oncogenes correspond to dominant genes, since these require a gain-of-function mutation. Tumour suppressors are instead recessive genes, since loss-of-function mutations usually require complete gene inactivation. In our analyses, we considered dominant and recessive cancer genes separately for comparison with the rest of human genes. For more recent analyses, we used the updated list of cancer genes from the Cancer Gene Census, of which the latest one included 518 cancer genes (February 2014).

2.2 Dataset of CNVs

2.2.1 Germline CNVs

In order to understand the genomic landscape of CNVs in normal population, we initially studied germline CNVs in healthy individuals. We obtained germline CNVs from the Database of Genomic Variants (DGV) (12), which stands as the largest curated repository of human genomic structural variations identified in apparently healthy control or normal samples. Before inclusion in the database, DGV applies a set of filters to ensure high quality data:

1) CNVs from patients are excluded, only those from controls are included,
2) CNVs that map to alternate haplotype sequences, random chromosomes or mitochondrial chromosome are excluded, only those that map to the autosomal or sex chromosomes are kept,
3) CNVs from the same study that overlap by at least 70% in length and position are merged (i.e. these are likely to be the same variant),
4) CNVs smaller than 50bp, larger than 3Mb, found in gaps and associated with known disorders are removed.

We downloaded 133,693 CNVs from DGV (March 2012), which contained 120,883 deletions, 2,054 amplifications, 9,843 insertions and 913 complex CNVs. The data was derived from 35 studies and represented a non-homogeneous collection due to the various methods used to detect the CNVs (such as fluorescence *in situ* hybridisation (FISH), comparative genomic hybridisation (CGH), PCR-based copy number arrays, microarray and sequencing). Deletions constituted the vast majority of germline CNVs, owing to the large contribution from the pilot phase of 1000 Genomes Project (3). To reduce the bias towards deletions, we integrated 48,931 segmental duplications into amplifications, which are blocks of highly conserved repeated DNA segment longer than 1 kb, as derived from UCSC Table Browser (248). To obtain unique regions, we merged individual CNVs of the same type that overlap by genomic coordinates. This resulted in a final dataset of 67,782 unique regions (Table). DGV is regularly updated and current version (July 2015) includes 491,894 CNVs derived from 67 studies.

Table : Dataset of germline and somatic CNVs

CNV Dataset	CNV Type	Number	Genome Coverage	Samples	Cancer Types
Germline	Amplifications	8,635	7%	NA	NA
	Deletions	50,943	22%		
	Total	67,782	25%		
Tumorscape All	Amplifications	75,700	97%	3,056	54
	Deletions	55,101	97%		
	Total	130,801	97%		
Tumorscape Peaks	Amplifications	76	6%		
	Deletions	82	13%		
	Total	158	18%		
TCGA All	Amplifications	242,745	96%	6,213	23
	Deletions	342,501	96%		
	Total	585,246	96%		
TCGA Recurrent Regions	Amplifications	1,425	72%		
	Deletions	2,214	76%		
	Total	3,639	90%		
TCGA Core Regions	Amplifications	61	22%		
	Deletions	234	36%		
	Total	295	55%		

Legend: For each CNV dataset, reported are the numbers of amplifications, deletions and all CNVs, the percentages of their genome coverage, and the corresponding number of samples and cancer types. Germline CNVs are merged by overlapping coordinates and segmental duplications from UCSC Table Browser (248) are integrated as amplifications. Genome coverage shows the cumulative percentage of genome that undergoes CNV. For germline CNVs, only amplifications and deletions are shown for comparison with other datasets (Insertions and Complex CNVs are included in the total).

2.2.2 Somatic CNVs

2.2.2.1 Tumorscape

As the first dataset of somatic CNVs, we downloaded data from the Tumorscape database (13) (http://www.broadinstitute.org/tumorscape/pages/portalHome.jsf). This study represents one of the first large-scale analyses of high-resolution somatic copy number variations, which collects 130,801 somatic CNVs from 3056 samples corresponding to 54 cancer types (Table). All the copy-number measurements are obtained on the same platform (Affymetrix 250K Sty array), representing a homogeneous set of CNVs across samples.

The authors further identified focal regions of somatic copy number alterations by using GISTIC 2.0 algorithm (249), defined as "peak regions". Briefly, GISTIC 2.0 is a gene-centric approach that identifies regions that occur significantly more frequently than the expected background rate. For our analysis, we used 158 peak regions from the original study (Table).

2.2.2.2 TCGA

As a second dataset of somatic CNVs, we used data from The Cancer Genome Atlas (TCGA, https://tcga-data.nci.nih.gov/tcga/), which is a comprehensive and on-going project to profile the genomes of more than 10,000 patients from above 30 cancer types. We downloaded somatic CNVs available in 6,213 tumour samples from 23 cancer types (January 2013) (Table).

To reduce possible false positives, we applied filters on the amplitude of the alterations. Amplitude is the log2 ratio of the copy numbers in a given genomic region between the tumour sample and the reference sample, and calculated as the mean of all amplitudes of the individual probes for that region (defined as segment). We considered the segments with amplitude higher than 0.3 as amplified, and those with amplitude lower than -0.3 as deleted. These thresholds are widely used in the literature (250-254).

2.3 Dataset of somatic mutations

We used a modified version of Level 2 somatic mutation data on TCGA samples in the mutation annotation format (MAF) downloaded from the Synapse repository (255) (https://www.synapse.org/#!Synapse:syn1729383/wiki/, March 2013). The curators applied data cleaning steps to ensure accurate annotation and high quality of variants, i.e. mapping genomic loci of all variants to Hg19, filtering out variants that are unlikely to be somatic, classifying variants in pseudogenes as silent, excluding variants from metastases or recurrences if variants from the primary tumour are already present. The clean dataset consisted of 19 MAF files corresponding to 20 cancer types (COAD and READ are combined as colorectal). For our purpose, we used a subset of this dataset including 427,995 non-silent mutations identified in 11 cancer types.

2.4 Dataset of gene expression

We downloaded gene expression data from TCGA for the cancer types with also CNV data, which was available for 11 cancer types out of 23. Each expression file included expression values for 17,814 genes processed by the same microarray platform (Agilent 244K Custom Gene Expression G4502A-07). All the tumour samples in the array were hybridized against a common reference, Stratagene's Universal Human Reference RNA (http://www.chem-agilent.com/pdf/strata/740000.pdf). This reference sample is composed of equal quantities of total RNA from 10 human cell lines and designed to be used as a control for spot intensity rather than a biological reference. Then the expression value obtained for each sample was (lowess) normalized and the log2 ratio between the sample and the reference was given as the output. We directly used these normalized expression values for our analysis.

2.5 Combined dataset

In order to track the genetic and regulatory modifications of genes together, we created a dataset by combining the 3 types of data from TCGA at sample level (Table). This resulted in 1,245 samples from 11 cancer types in which somatic CNV, mutation and gene expression data were available for each sample. In the combined dataset, we labelled each gene in each sample for different data types as the following:

1) CNV → Amplified, Deleted, Wild-type

2) Mutation → Mutated, Wild-type

3) Gene expression → Highly expressed, Medium expressed, Lowly expressed

For CNV categories, we considered 25% of the gene length for the intersection. For mutation categories, we accounted for the presence or absence of a non-silent mutation based on the annotation in the original file. For the gene expression, we determined the categories based on the gene expression distribution in each sample; labelling the gene as highly expressed if it had an expression value falling in the top 10% of the distribution, as lowly expressed if in the bottom 10% and as medium expressed if anywhere else.

Table : Combined dataset of somatic CNVs, mutation and expression

Cancer Type	Samples	Human Genes	Cancer Genes	Dominant genes	Recessive genes
BRCA	259	5,438	204	145	62
COAD	110	9,668	318	242	80
GBM	253	6,213	224	168	59
KIRC	46	1,900	85	52	34
KIRP	15	790	37	26	12
LGG	25	661	30	17	13
LUAD	31	4,812	182	128	57
LUSC	114	9,766	319	232	90
OV	304	6,745	245	173	73
READ	46	3,031	121	82	39
UCEC	42	9,936	325	240	86
Unique total	**1,245**	**14,288**	**427**	**328**	**103**

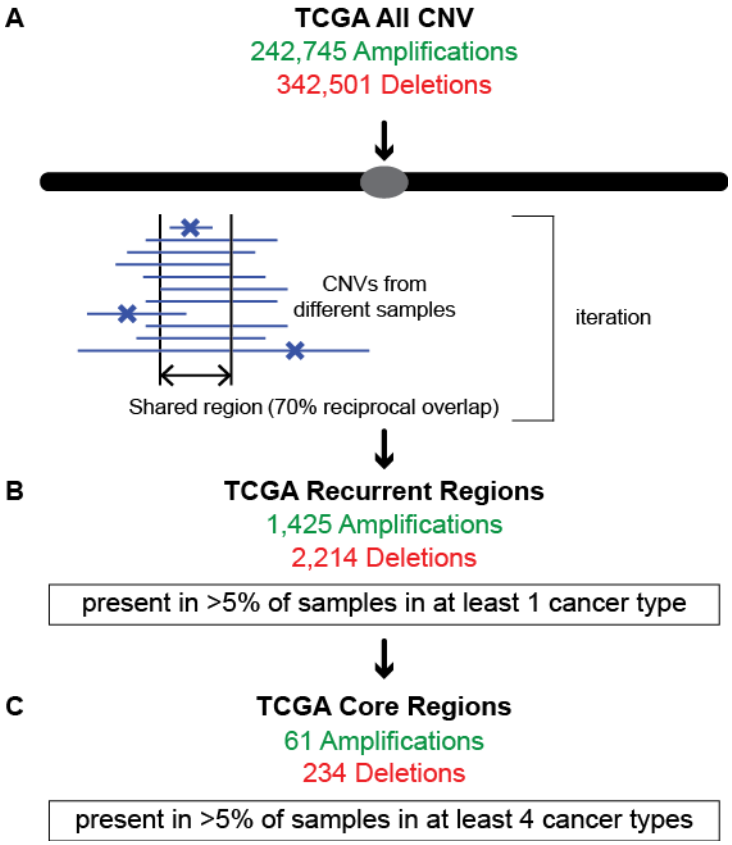
Legend: For each of 11 cancer types, reported are the number of samples with somatic CNVs, mutation and expression data from TCGA, and the corresponding number of genes with all 3 types of data.

2.6 Identification of recurrent regions of copy number alteration

We observed that somatic copy number variations, unlike the germline copy number variations, are pervasive in the genome, covering more than 95% of the genome when merged. To analyze them in the same way as for germline CNVs, we applied a different methodology to the somatic CNVs than that we applied to the germline CNVs. Instead of merging all the individual CNVs from different samples (which was the case for germline CNVs), we first defined recurrent CNV regions that occur frequently across the samples. We identified recurrent regions in each cancer type based on an iterative procedure (Figure A). Among CNVs that overlap at least 70% reciprocally by coordinates, the shared region (minimal common region) is identified, whereas CNVs that overlap less than 70% reciprocally are discarded. This step is repeated iteratively until no two regions with >70% overlap remains. Among the surviving regions, only those present in >5% of the samples of the corresponding cancer type are retained. This led to 1,245 amplifications and 2,214 deletions, defined as "recurrent regions" (Figure B). This approach resulted in a reduction of genome coverage of somatic CNVs from ~95% to ~70%, which was still high compared to that of the germline CNVs.

Above approach is limited to define cancer-specific recurrent regions. To obtain the regions that are common in cancer, we further refined the recurrent regions retaining those present in at least 4 cancer types and merging if the distance between the regions is minimal (<1 Mbp). This resulted in 61 amplifications and 234 deletions, defined as "core regions" (Figure C). Biologically, these regions should possess higher propensity to undergo copy number variations than the rest of the genome as they follow a pan-cancer pattern. The core regions we identified were comparable to the germline CNVs in terms of genome coverage, and to the peak regions from Tumorscape dataset in terms of both coverage and number.

Figure : Identification of recurrent somatic CNV datasets



Legend: Methodology to identify recurrent regions of somatic CNVs and core regions in TCGA dataset. **A)** Starting from the original data from TCGA, the shared region (minimal common region) is identified among CNVs that overlap at least 70% reciprocally by coordinates, whereas those that overlap less than 70% are discarded. This step is repeated iteratively until no two regions with >70% overlap remains. **B)** Among the surviving regions, only those present in >5% of samples of the corresponding cancer type are retained, resulting in "Recurrent regions". **C)** Recurrent regions are further refined by retaining those present in at least 4 cancer types and such regions are merged if the distance between them is minimal (<1 Mbp), resulting in "Core regions".

Our approach to define recurrent regions was a simple alternative to the widely used GISTIC algorithm (249), which is essentially based on the frequency (recurrence across samples) and the amplitude (copy number) to identify significantly recurrently altered CNV regions in tumour samples.

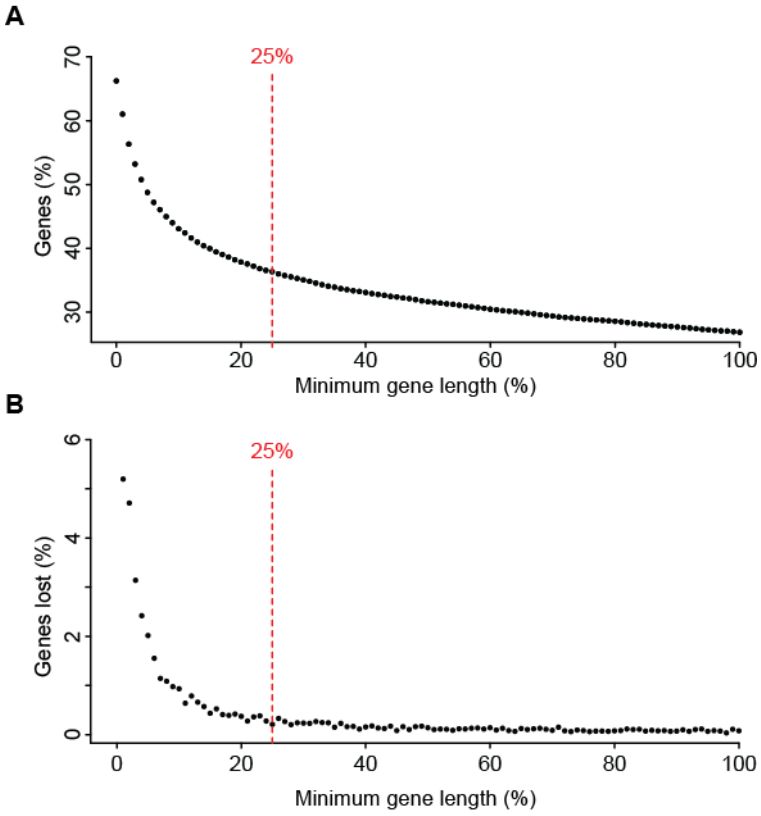
2.7 Intersection of CNVs with genes

To assess the gene content of CNVs, we intersected genes with CNVs by the genomic coordinates using BEDTools (256). For the intersection, we used the merged regions from all the individuals for the dataset of germline CNVs, instead the recurrent and the core regions that we identified for the datasets of somatic CNVs. At least 25% of the gene length was required to overlap with a copy number variation to mark the gene as altered. This threshold was the best compromise between the gene portion spanned by a CNV to consider the gene as affected and the number of CNVs that still overlap with a gene (at 1% threshold the gene is not likely to be affected and at 100% threshold there are too few CNVs that overlap with genes) (Figure). Each gene can be amplified or deleted in the same sample, as amplifications and deletions do not overlap. However, the same region can be amplified and deleted in different samples, indicative of cancer type-specific alteration or involvement of different mechanisms leading to copy number variation in different set of individuals.

2.8 CNV coverage and distribution along the chromosomes

To see how the somatic CNVs are distributed along the human genome, we first measured the CNV coverage of each chromosome and chromosome arm. For each chromosome/arm, we calculated the non-redundant number of bases that undergo copy number variation and divided by the total number of bases in that chromosome/arm. This led us to rank the chromosomes/arms according to their overall CNV coverage. We further assessed the chromosome arms that are most frequently altered across the samples and the cancer types and explored the cancer genes within these regions.

Figure : Determining threshold for gene length overlapping with CNVs



Legend: A) For each percentile of gene length as the minimum portion to be contained within a CNV, shown is the percentage of genes overlapping with CNVs. **B)** For each percentile increment in gene length, shown is the decrease in the percentage of genes overlapping with CNVs. For values higher than 25%, the decrease is constant.

2.9 Intersection of CNVs with genomic features

In order to distinguish whether somatic copy number variations in cancer occur in some genomic regions in a recurrent fashion due to their involvement in cancer-specific processes or simply due to the genomic features underlying their fragility, we investigated the distribution of genomic features along the genome and their overlap with CNVs. We first checked the fragile sites of the genome, which are thought to have an increased local rate of DNA breakage (257), to see if CNVs have a propensity to lye within them. Then we obtained other relevant genomic features from the UCSC Table Browser (258) that might potentially lead to CNV formation. We divided the genomic features into two as discrete and continuous features. The former consisted of those which have a definitive quantity and length (such as genes, SNPs) that can be attributed as present or absent within CNVs.

Instead the latter are continuous throughout the genome which show varying local rates per unit (such as recombination rate, level of expression). For discrete features, we compared the proportion of each feature within CNVs to the proportion within the rest of the genome by Chi-square test. For continuous features, the original data is given in a format that the genome is divided into bins of equal length and each bin is assigned to a score reflecting the abundance of the feature in that bin. For each CNV, we calculated the average score of a given feature over the spanned bins. Similarly, we calculated the average scores of non-CNV regions, where the bins between CNV regions were considered. Having the scores for each CNV and non-CNV regions, we compared the distribution between the two by Wilcoxon rank-sum test. We realized that comparing distributions is more accurate than comparing proportions in case of such large quantity of data, therefore we applied this method also to the discrete features.

We intersected each set of genomic features with recurrent somatic CNV datasets (Tumorscape peaks, TCGA recurrent and TCGA core regions). We compared the CNV regions with the rest of human genome for their overlap with a given feature. We excluded the regions in the reference genome with poor mappability (telomeric ends, centromeres and the short arms of the acrocentric chromosomes from UCSC Table Browser Hg19 gap file) to eliminate possible bias, as CNVs from TCGA SNP array did not map to these regions. We merged the overlapping discrete features to count each base once.

2.10 Gene expression change upon copy number alteration

In order to measure the impact of somatic CNVs on gene expression, we used a method previously used in similar studies (14). This method is simple yet useful where expression data from the normal samples is not available, which was the case in our analysis.

Starting from the normalized gene expression levels in the array, we calculated the average expression levels of genes in 3 conditions by using the formula:

$$average\ expression\ level_g = \frac{\sum_{1}^{n} expression\ level_{g,n}}{n}$$

where n is equal to the number of samples where gene g:

1) is amplified,

2) is deleted,

3) does not undergo copy number variation.

Then we compared the distribution of average expression levels of genes in the amplified samples (1) and the deleted samples (2) to those in the samples in which genes do not undergo copy number variation (3) by Wilcoxon rank-sum test. This allowed us to see the general trend on gene expression change upon copy number alteration.

2.11 Preferential modification of cancer genes

We used somatic CNV and mutation data to determine the primary modification that genes have incurred. For each gene, the fraction of samples in which the gene is amplified, deleted and mutated over the total number of modified samples for that gene were calculated by using the formulae:

$$f_{g,\,amplified} = \frac{n_{g,\,amplified}}{n_{g,\,modified}}, \quad f_{g,\,deleted} = \frac{n_{g,\,deleted}}{n_{g,\,modified}}, \quad f_{g,\,mutated} = \frac{n_{g,\,mutated}}{n_{g,\,modified}}$$

$$n_{g,\,modified} = n_{g,\,amplified} + n_{g,\,deleted} + n_{g,\,mutated}$$

where

- f denotes fraction of samples,
- g denotes gene,
- n denotes number of samples.

However, due to the few number of mutations per sample, the fractions of mutated samples were too low to compare to the fractions of amplified and deleted samples. Therefore only amplifications and deletions were considered to determine the primary modification between the two that affect the cancer genes. In general, to determine the primary modification of a gene, we considered more than 90% of the samples to have the same type of alteration.

3 Results

In this section, we first describe the features of germline copy number variations in healthy population. This provides an understanding of how changes inherited in the human genome may remain compatible with a normal phenotype. Next we compare these features to those of somatic copy number variations that occur in cancer samples. Comparison between germline and somatic CNVs reveals substantial differences, suggesting a role for the latter in cancer. We then quantify the link between somatic CNVs and genes, in particular with dominant and recessive cancer genes, leading to interesting findings. Enrichment of amplifications in dominant cancer genes and of deletions in recessive cancer genes suggest two opposite patterns of tumourigenic activation. This result is supported by the high correspondence of dominant and recessive cancer genes to oncogenes and tumour suppressors, respectively. As a further support, we found a positive correlation between amplification and overexpression and between deletion and decreased expression across cancer samples. More generally, such correlations were also shown between germline CNVs and contained genes across healthy individuals (89,196,259). Taken together, we propose that dominant cancer genes, which are mostly oncogenes, are activated through their genomic amplification, instead recessive genes, which are mostly tumour suppressors, are inactivated via deletion.

Next we look into gene copy number variation, in particular gene amplification, in a broader context in the evolution extending to the gene duplicability, defined as the propensity of multiple copies of a gene retained in the genome (167). Gene duplicability has important implications in genome evolution, such as emergence of new genes, which give rise to paralogous genes. We use gene duplicability as well as other gene properties to define paralogous gene pairs for a practical purpose, that is the identification of novel synthetic lethal gene pairs in cancer, which has been previously demonstrated to be a working hypothesis by our group (188). By integrating numerous gene properties (duplicability, network properties, protein domain composition, functional redundancy,

pathway information) with the genomics data derived from cancer samples (somatic mutation and gene expression from The Cancer Genome Atlas (TCGA), https://tcga-data.nci.nih.gov/tcga/), we predict putative gene pairs with potential synthetic lethal interaction.

Finally, we present the last release of the Network of Cancer Genes (NCG), a database curated in our group (175,260-262). NCG provides literature information and system-levels properties on a manually curated set of cancer genes. In addition to the general description of the database, we demonstrate how NCG can be useful for hypothesis testing, as was the case for prioritizing our candidates with putative synthetic lethal interaction.

3.1 Somatic CNVs are pervasive in the genome compared to germline CNVs

To investigate the genomic landscape of CNVs in normal individuals and cancer samples, we derived a dataset of germline CNVs from the Database of Genomic Variants (DGV) (12) and two datasets of somatic CNVs from Tumorscape (13) and TCGA (https://tcga-data.nci.nih.gov/tcga/). Within these datasets, we characterized CNVs in terms of their length, genome coverage and gene correspondence (number of genes contained within a CNV).

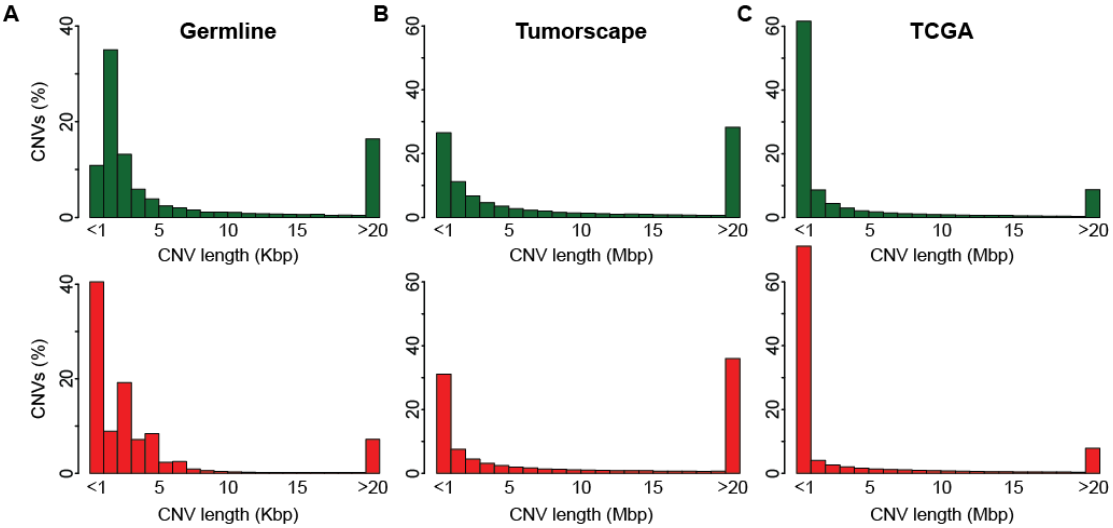
We found that the vast majority of germline CNVs is less than 5kb long (Figure A), and they cover 25% of the genome when merged by coordinates (amplifications cover 7% and deletions cover 22%, separately) (Table). In general, genes tend to fall within long CNVs while CNVs tend to overlap with short genes (Table). In addition, there is a trend for one-to-one correspondence between germline CNVs and genes (i.e. one gene undergoes only one CNV and one CNV affects only one gene) (Figure A, B). This trend is more prominent for cancer genes (Figure C, D).

Somatic CNVs, however, are significantly longer than germline CNVs (in the order of megabase pairs) and in some cases they may span the entire arm of a chromosome (Figure B, C). This is not surprising considering that cancer genomes are often associated with large-scale genomic alterations (263). In the last years, somatic CNVs affecting an entire

chromosome arm (arm-level CNVs) were extensively identified and distinguished from smaller ones (focal CNVs) (13,15,264). As a result, the original sets of somatic CNVs cover the entire genome when merged. On the other hand, recurrent somatic CNV regions that we identified have more comparable genome coverage to that of germline dataset. For this reason, we used only recurrent regions of somatic CNVs for further analyses.

In addition, we observed that the average number of somatic CNVs per sample varies among cancer types, which might suggest a distinguishing characteristic for each cancer type (Table). For example, ovarian cancer and sarcoma have the highest number CNVs per sample in average (104 amplifications and 125 deletions per sample for ovarian cancer, and 170 amplifications and 104 deletions per sample for sarcoma). In a pan-cancer analysis, ovarian cancer was shown to be primarily characterized by CNVs rather than mutations (14). On the other extreme, thyroid cancer has the lowest number of CNVs per sample in average (2 amplifications and 13 deletions per sample).

Figure : Length distribution of CNVs



Legend: Shown are length distributions of amplifications and deletions in **A)** germline, **B)** Tumorscape (13) and **C)** TCGA original datasets. Distributions of germline and somatic CNVs are given in kilobase pairs (Kbp) and megabase pairs (Mbp), respectively. Wilcoxon rank-sum test is applied to compare the length distributions of both somatic CNV datasets to that of the germline CNVs (p-value < 2.2e-16).

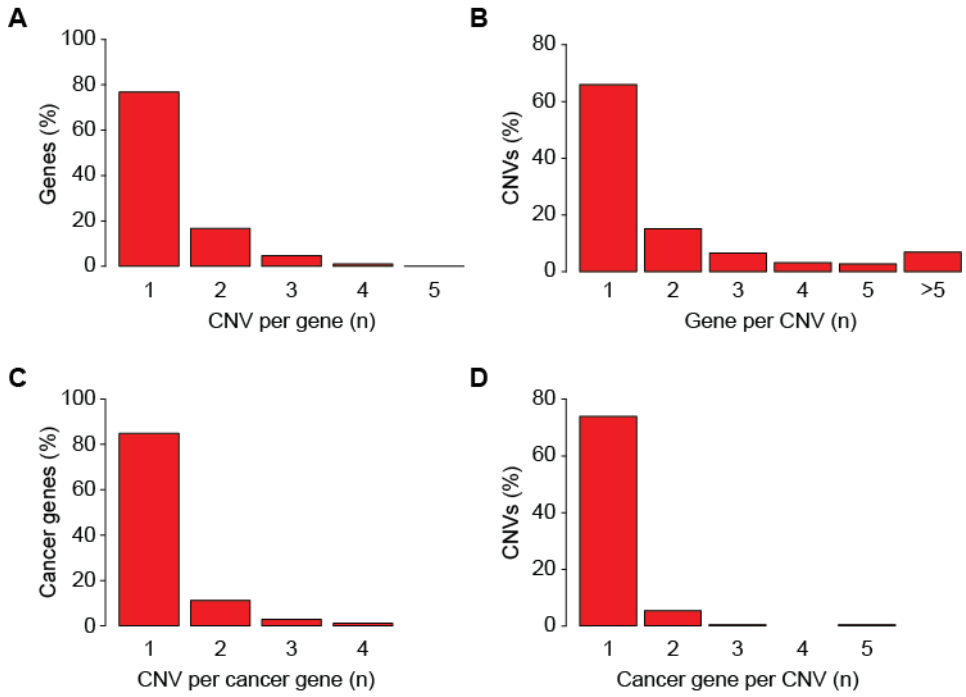
These results imply that germline CNVs tend to minimally affect the protein-coding part of the genome. This is expected considering the fact that abnormal number of gene copies may cause certain genetic diseases (6,17) and germline CNVs used here derive from apparently healthy individuals. On the other hand, somatic CNVs alter the genetic makeup substantially, complying with the cancer phenotype. For example, somatic copy number alteration of certain genes were shown to play a driver role in several cancer types, such as amplification of *MYC* (265), *ERBB2* (266), and deletion of *CDKN2A* (267), *CHD1* (268).

Table : Length distribution of germline CNVs and genes

	Number	Median length (kb)	Average length (kb)	p-value
All CNVs	67,782	2.000	15.460	
Amplifications	8,635	2.242	25.463	
Deletions	50,943	2.050	13.025	NA
Insertions	7,430	0.342	13.203	
Complex CNVs	774	10.155	85.830	
All Genes	19,045	15.558	48.421	
Cancer Genes	501	35.194	75.061	< 2.2e-16
Rest of Human Genes	17,336	14.268	42.634	
Genes That Overlap with CNVs	6,919	9.202	34.518	< 2.2e-16
Genes That Do not Overlap with CNVs	12,126	19.811	56.353	
Cancer Genes That Overlap with CNVs	178	18.080	62.041	1.96E-06
Cancer Genes That Do not Overlap with CNVs	323	44.666	82.235	
CNVs That Overlap with Genes	4,278	46.069	148.832	< 2.2e-16
CNVs That Do not Overlap with Genes	63,504	1.850	6.476	
CNVs That Overlap with Cancer Genes	195	157.070	384.315	< 2.2e-16
CNVs That Do not Overlap with Cancer Genes	67,587	2.000	14.400	

Legend: Reported are the numbers, median and average length of germline CNVs and genes. The length distribution of cancer genes is compared to that of the rest of human genes, where candidate cancer genes are excluded. Wilcoxon rank-sum test is applied to compare the distributions between the two groups. P-value is highlighted in red in case there is enrichment and in green in case there is depletion for the first observation.

Figure : Correspondence between genes and germline CNVs



Legend: Shown are distributions of numbers of A) CNVs that overlap with genes, B) genes that overlap with CNVs, C) CNVs that overlap with cancer genes, D) cancer genes that overlap with CNVs.

Table : Dataset of TCGA somatic copy number variations

Cancer Type	Samples (n)	Amplifications (n)	Average Amplifications per Sample	Deletions (n)	Average Deletions per Sample (n)
BLCA	114	6,047	53	7,461	65
BRCA	841	49,415	60	46,196	55
CESC	103	2,223	22	4,044	39
COAD	403	7,913	20	20,177	50
DLBC	14	252	21	628	45
GBM	515	13,820	27	37,754	73
HNSC	297	8,524	29	13,073	44
KICH	65	1,338	22	2,693	41
KIRC	519	5,008	10	15,605	30
KIRP	111	1,635	15	4,013	36
LGG	176	2,148	13	5,150	29
LIHC	96	5,125	55	4,528	47
LUAD	359	13,057	37	15,752	44
LUSC	347	17,262	50	24,020	69
OV	567	58,891	104	70,952	125
PAAD	45	427	10	965	21
PRAD	156	1,624	12	6,856	44
READ	144	3,780	26	8,207	57
SARC	28	4,746	170	2,907	104
SKCM	271	10,685	40	13,794	51
STAD	227	9,093	42	11,802	52
THCA	341	576	2	4,536	13
UCEC	474	19,156	43	21,388	45
Total	**6,213**	**242,745**	**41**	**342,501**	**55**

Legend: For each of 23 cancer types from TCGA with available somatic CNV data, reported are the total numbers of samples and of the alterations, and average numbers of alterations per sample. Only those samples that have at least one alteration are considered to calculate the average numbers.

3.2 Somatic CNV coverage varies throughout the genome

To assess to which extent somatic CNVs affect the individual chromosomes, we measured the cumulative CNV coverage of each chromosome and chromosome arm (Table , Table). Cumulative coverage was defined as the percentage of bases covered by all CNVs divided by the total number of bases on that chromosome/arm. Unlike the germline CNVs, which never span more than half of a chromosome arm (Table , Table), somatic CNVs often cover almost an entire chromosome (>90%) (Table , Table). For example, chromosomes 7,5,3,18,17 have more than 90% of their length as recurrently (in more than 5% of the samples) amplified in at least one cancer type, of which chromosome 7 is recurrently amplified by 83% in at least 5 cancer types (Table). There are also chromosome arms that are highly (>90%) amplified (12p, 8q, 11q, 1q, 6p, 10p). Similarly, chromosomes 2,4,10,6,5,11,9 have more than 90% of their length as recurrently deleted in at least one cancer type, of which chromosome 10 is recurrently deleted by 87% in at least 4 cancer types, and 18q, 1p, 16q, 3p, 8p, Xq are deleted by >90% (Table). Although most of these chromosomal regions were already known to undergo recurrent somatic copy number alterations (13,15,111), a few of them are novel to our study (18, 11q, 10p as amplified by >90%; 2, 1p as deleted by >90%).

With this analysis, we confirm our previous observation at the chromosome arm level that germline CNVs have much lower coverages than somatic CNVs (Table). Moreover, we found that germline CNVs have a homogeneous distribution along the genome (ranging from 3% to 12% per chromosome), while somatic CNVs accumulate on some chromosomes more than the others (ranging from 0% to 98% per chromosome, depending on the dataset) (Table , Table). We also discovered novel chromosome arms with frequent alterations that are not previously reported.

Table : Coverage of amplifications per chromosome

		Germline CNVs				Tumorscape Peaks				TCGA Recurrent Regions				TCGA Core Regions			
Chr	Chr Length (Mbp)	No of Amplified Regions	Chr Coverage (%)	p-arm Coverage (%)	q-arm Coverage (%)	No of Amplified Regions	Chr Coverage (%)	p-arm Coverage (%)	q-arm Coverage (%)	No of Amplified Regions	Chr Coverage (%)	p-arm Coverage (%)	q-arm Coverage (%)	No of Amplified Regions	Chr Coverage (%)	p-arm Coverage (%)	q-arm Coverage (%)
chr1	249	709	7	5	9	9	7	6	8	131	68	39	96	6	47	2	**90**
chr2	243	540	6	7	6	1	0	0	0	38	22	56	1	0	0	0	0
chr3	198	490	4	4	3	3	16	6	25	138	92	86	**98**	4	56	35	75
chr4	191	372	5	5	5	1	1	0	1	23	87	83	88	0	0	0	0
chr5	181	441	4	4	4	2	4	0	5	94	95	**96**	94	7	25	67	10
chr6	171	461	4	4	4	4	15	**40**	2	62	53	95	31	3	17	47	0
chr7	159	581	11	8	12	3	6	1	10	161	**97**	**96**	**98**	8	**83**	**80**	85
chr8	146	320	5	13	2	7	10	1	14	153	88	77	93	12	57	0	83
chr9	141	351	11	18	7	1	2	0	3	32	86	79	90	0	0	0	0
chr10	136	437	8	9	7	1	5	0	8	24	52	93	35	1	2	7	0
chr11	135	364	5	6	5	4	6	11	3	42	85	77	91	0	0	0	0
chr12	134	354	6	11	5	14	7	8	6	181	88	**96**	85	2	14	50	0
chr13	115	285	4	0	4	3	**25**	0	**29**	50	76	0	89	3	32	0	38
chr14	107	217	3	0	4	2	1	0	1	21	76	0	91	0	0	0	0
chr15	103	293	**12**	0	**14**	1	1	0	1	11	60	0	73	1	3	0	4
chr16	90	246	**12**	**25**	3	0	0	0	0	30	56	65	50	1	16	14	18
chr17	81	391	10	14	8	8	6	5	6	39	90	85	91	3	28	0	40
chr18	78	160	3	12	1	1	6	0	8	43	91	89	91	1	3	13	0
chr19	59	376	11	11	11	3	5	0	8	44	86	81	89	1	15	0	28
chr20	63	162	5	7	3	3	3	0	5	68	87	88	86	7	57	33	77
chr21	48	108	6	10	5	0	0	0	0	5	29	0	39	0	0	0	0
chr22	51	202	10	0	**14**	1	1	0	2	9	43	0	60	1	20	0	28
chrX	155	621	10	11	10	2	2	0	3	26	40	51	33	0	0	0	0
chrY	59	154	10	63	17	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Total	**3,036**	**8,635**	**7%**	**9%**	**7%**	**74**	**6%**	**5%**	**7%**	**1,425**	**72%**	**69%**	**73%**	**61**	**22%**	**18%**	**25 %**

Legend: For each chromosome, reported are the chromosome length, number of amplifications, and cumulative coverage of amplified regions in germline and somatic CNV datasets. Coverages are given in percentage and shown separately for p and q arms. Only germline dataset included CNVs mapping to chrY. Total coverage is calculated on the entire genome, which may differ than the average of individual chromosomes, as chromosome lengths vary. The highest percentage of coverage in each column is highlighted in yellow; green denotes amplifications.

Table : Coverage of deletions per chromosome

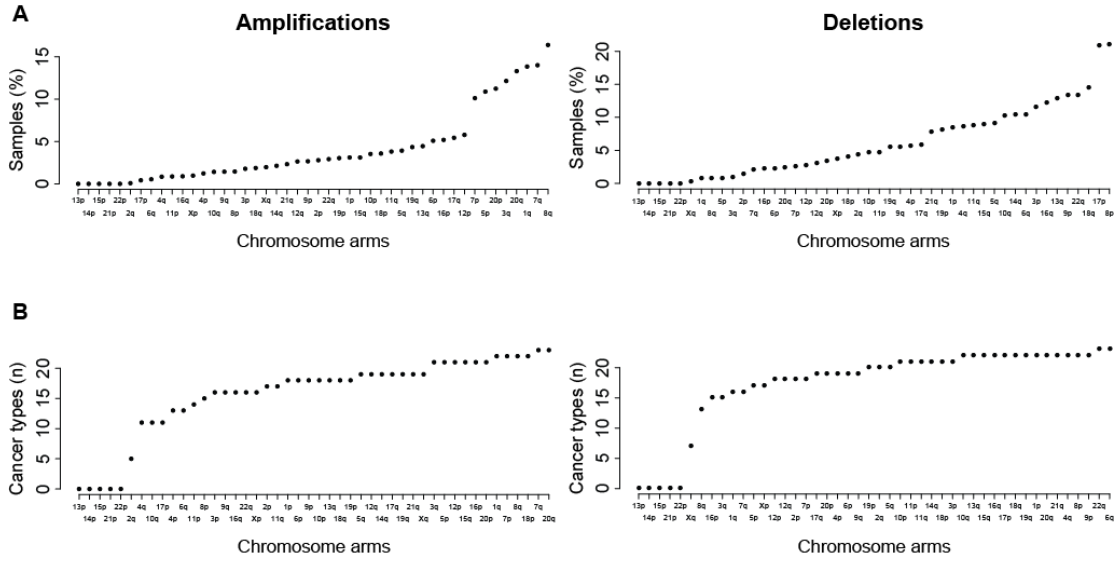
		Germline CNVs				Tumorscape Peaks				TCGA Recurrent Regions				TCGA Core Regions			
Chr	Chr Length (Mbp)	No of Deleted Regions	Chr Coverage (%)	p-arm Coverage (%)	q-arm Coverage (%)	No of Deleted Regions	Chr Coverage (%)	p-arm Coverage (%)	q-arm Coverage (%)	No of Deleted Regions	Chr Coverage (%)	p-arm Coverage (%)	q-arm Coverage (%)	No of Deleted Regions	Chr Coverage (%)	p-arm Coverage (%)	q-arm Coverage (%)
chr1	249	3937	20	22	18	5	16	13	19	111	84	91	78	12	37	**75**	0
chr2	243	4003	20	21	19	4	10	16	5	88	**97**	95	**98**	3	6	0	9
chr3	198	3253	21	20	22	4	6	5	6	126	81	94	70	2	32	68	0
chr4	191	3445	19	21	18	3	4	1	5	154	95	93	96	19	63	52	67
chr5	181	3124	18	22	16	3	29	0	39	133	91	79	96	11	52	0	71
chr6	171	3059	19	19	19	3	28	2	**42**	131	93	84	**98**	14	45	0	70
chr7	159	2861	23	21	25	4	13	2	20	46	52	73	40	0	0	0	0
chr8	146	2589	19	25	17	5	31	**38**	27	128	65	**97**	50	18	20	66	0
chr9	141	2269	21	30	17	4	6	16	0	195	90	83	94	19	65	67	64
chr10	136	2272	21	19	22	4	7	2	8	151	95	90	97	29	**87**	66	**96**
chr11	135	2410	22	21	22	4	13	3	19	121	91	95	88	11	45	54	39
chr12	134	2420	17	18	17	3	2	2	2	50	44	49	42	1	4	13	0
chr13	115	1938	18	0	21	3	5	0	6	123	76	0	89	14	64	0	76
chr14	107	1421	19	0	23	2	26	0	31	65	68	0	81	10	48	0	57
chr15	103	1267	20	0	25	1	8	0	10	65	67	0	81	11	45	0	55
chr16	90	1605	25	37	16	6	**33**	21	41	146	83	69	93	16	44	2	73
chr17	81	1526	27	29	26	5	6	9	4	80	84	80	86	9	24	74	3
chr18	78	1605	18	22	17	3	6	3	7	77	89	83	91	13	72	53	77
chr19	59	1239	**41**	**47**	**36**	3	4	2	5	72	74	76	72	11	20	25	15
chr20	63	1289	21	22	20	2	3	8	0	15	22	51	0	1	4	10	0
chr21	48	783	18	4	24	1	7	0	10	27	44	0	60	2	28	0	38
chr22	51	798	23	0	32	2	9	0	13	69	51	0	71	8	46	0	64
chrX	155	1725	34	38	32	2	4	11	0	41	16	40	100	0	0	0	0
chrY	59	105	28	43	25	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Total	**3,036**	**50,943**	**22%**	**23%**	**21%**	**76**	**13%**	**9%**	**15%**	**2,214**	**76%**	**75%**	**76%**	**234**	**36%**	**33%**	**38%**

Legend: For each chromosome, reported are the chromosome length, number of deletions, and cumulative coverage of deleted regions in germline and somatic CNV datasets. Coverages are given in percentage and shown separately for p and q arms. Only germline dataset included CNVs mapping to chrY. Total coverage is calculated on the entire genome, which may differ than the average of individual chromosomes, as chromosome lengths vary. The highest percentage of coverage in each column is highlighted in yellow; red denotes deletions.

3.3 Somatic CNV coverage associates with somatic CNV frequency

The cumulative coverage does not explain the actual frequency of CNVs. In other words, it cannot be assessed only by cumulative coverage whether the high coverage of a chromosome arm is due to a few long CNVs or to many short ones. A few long CNVs from the samples of a single cancer type spanning the entire arm will result in high coverage. Similarly, many short CNVs from many samples of several cancer types distributed along the entire arm will also result in high coverage. To distinguish between the two cases, we calculated the CNV frequency across samples and cancer types. Low CNV frequency with a high coverage would imply the former case, where a few CNVs span the arm. However, high CNV frequency with a high coverage would imply the latter case, where high coverage accounts for many individual CNVs. For this analysis, we selected the dataset of TCGA recurrent regions that we defined (Figure B) as they represent the highest genome coverages. As each CNV has different frequency, we picked only the one with the highest frequency on each arm as the representative of that arm. We observed that certain chromosome arms contain amplifications (8q, 7q, 1q, 20q, 3q, 20p, 5p, 7p) and deletions (8p, 17p) at a higher frequency than others, in agreement with the arms with high CNV coverage, but not necessarily having the same order (Figure A). Moreover, the high CNV frequency in these arms is not due to samples from few cancer types (Figure B). Overall, there is a close relationship between cumulative coverage and frequency of somatic CNVs across chromosome arms.

Figure : Somatic CNV frequency on chromosome arms



Legend: For each chromosome arm, shown are highest observed CNV frequencies in terms of **A)** samples and **B)** cancer types. For each CNV region, frequency is calculated by the number of affected samples divided by the total number of samples in **(A)**, and the number of affected cancer types out of 23 in **(B)**. Then the region with the highest frequency on each arm is selected as the representative of that arm.

3.4 CNV datasets have poor overlap

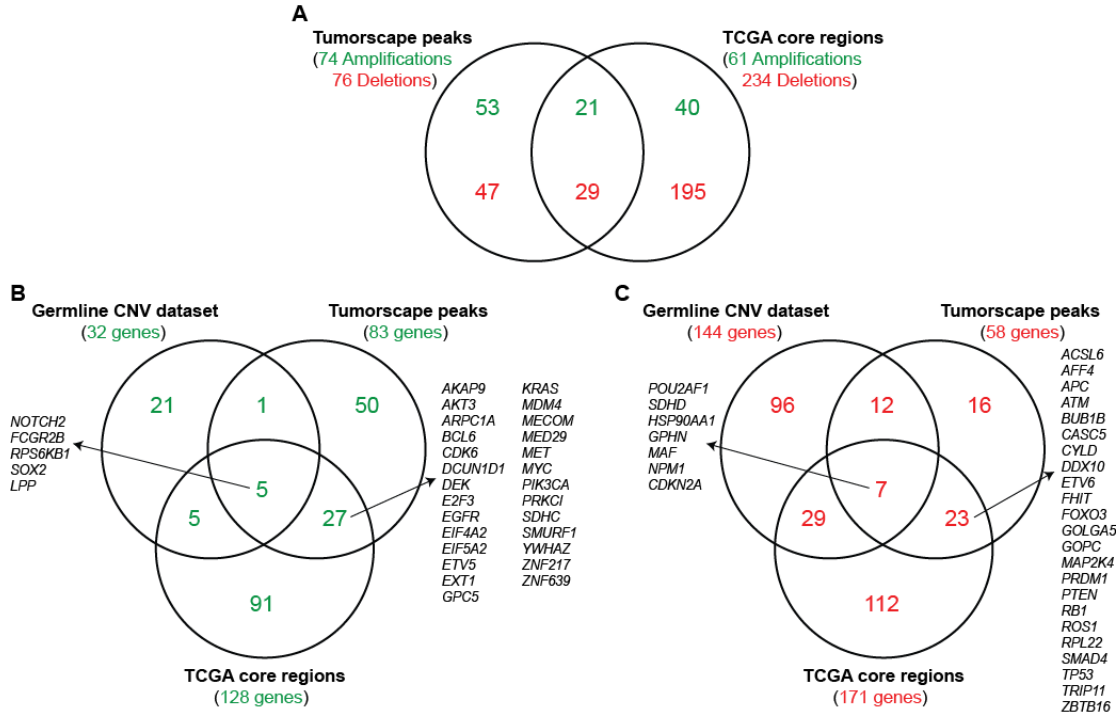
For our CNV analyses, we used three different datasets from two different sources. Before interpreting any result, we wondered to see the overlap between the datasets in terms of CNVs and genes. To assess this, we first identified the CNV regions that overlap between the two datasets (Figure A). Out of 74 amplifications in Tumorscape peaks, 21 of them (28%) overlapped with the amplifications from TCGA core regions. Similarly, 29 deletions out of 76 (38%) overlapped with the deletions from TCGA core regions. Next we identified the altered cancer genes that are shared between the two datasets. At this step we also included genes altered in the germline CNVs. Overall, 32 cancer genes are amplified in both somatic CNV datasets, of which 5 are also amplified in germline CNVs (Figure B). Similarly, 30 cancer genes are deleted in both somatic CNV datasets, of which 7 are also deleted in germline CNVs (Figure C).

We further analysed the genes that are shared between the datasets. Among the 32 amplified cancer genes, 17 are dominant, 13 are known to be amplified and overexpressed

in cancer (247), and 2 are recessive (*EXT1*, *SBDC*). These contain functionally validated oncogenes reported to be activated by amplification such as *MYC*, *EGFR*, *KRAS* and *PIK3CA* (13). Among the 30 deleted cancer genes, 12 are recessive and 18 are dominant. These contain functionally validated tumour suppressor genes reported to be inactivated by deletion such as *TP53*, *PTEN*, *CDKN2A* and *RB1* (13).

The poor overlap between the somatic CNV datasets suggests that our conclusions are independent of the CNV dataset used. However, there are some regions and well-known cancer genes that frequently undergo CNVs in different cohorts.

Figure : Overlap between CNV datasets



Legend: A) Number of amplifications and deletions overlapping between somatic CNVs from two different sources. **B)** Number of cancer genes amplified in germline and somatic CNVs. **C)** Number of cancer genes deleted in germline and somatic CNVs.

3.5 Germline CNVs are intergenic while somatic CNVs are genic

Intersection of CNVs with all human genes shows remarkable differences between the germline and the somatic CNV datasets. Before the intersection, we noticed that deletions constitute the majority (75%) of CNVs in the germline CNV dataset (Table). This is due to the study-specific bias that deletions mostly derive from the pilot study of 1000 Genomes Project, which reported many more deletions than other types of variants (4). This bias was later resolved by the reanalysis of the whole data in the phase 1 of the project (3). After the intersection, despite the bias towards deletions, we observed that only 5% of germline deletions overlaps with human genes along the genome, showing that germline CNVs, particularly deletions, are mostly (95%) intergenic. On the contrary, 14% of amplifications overlap with genes.

In the somatic CNV datasets, instead, we observed that more than 70% of somatic CNVs overlap with genes (Table). In particular, more than 95% of Tumorscape peaks that we derived from the literature and TCGA core regions that we identified overlap with genes. Moreover, the proportions of amplifications and deletions that overlap with genes are similar. In overall, these results show that somatic CNVs are mostly genic, which seems to be expected given the high genome coverages of somatic CNV datasets. However, we observed the same pattern also in Tumorscape peaks, which have comparable genome coverage to that of germline CNVs.

Overall, the low overlap between germline CNVs and genes suggest that germline CNVs minimally affect genes, as a result, no disease phenotype is observed. On the other hand, somatic CNVs are preferentially located in coding part of the genome, leading to the cancer phenotype.

Table : Intersection of CNVs with all human genes

Dataset	CNV Type	CNVs that overlap with genes			Genes that overlap with CNVs		
		Overlapping	Total	Percentage	Overlapping	Total	Percentage
Germline	Amplifications	1,180	8,635	14%	1,824	19,045	9.5%
	Deletions	2,442	50,943	5%	5,315	19,045	28%
	Total	4,278	67,782	6.5%	6,919	19,045	36.5%
Tumorscape Peaks	Amplifications	70	74	95%	1,509	19,045	8
	Deletions	76	76	100%	1,876	19,045	10
	Total	146	150	97%	3,263	19,045	17
TCGA Recurrent Regions	Amplifications	1,194	1,425	84%	13,758	19,045	72%
	Deletions	1,617	2,214	73%	14,372	19,045	75.5%
	Total	2,811	3,639	77.5%	17,237	19,045	90.5%
TCGA Core Regions	Amplifications	60	61	98%	3,991	19,045	21%
	Deletions	229	234	98%	6,456	19,045	34%
	Total	289	295	98%	9,670	19,045	51%

Legend: For each CNV dataset, reported are the numbers and the percentages of CNVs that overlap with genes and of genes that overlap with CNVs. Genes are counted as overlapping if at least 25% of their length intersect with a CNV.

3.6 Somatic CNVs are enriched in cancer genes

Observing that somatic CNVs are vastly genic unlike germline CNVs, we next sought how this impacts on cancer genes in particular. In overall dataset, we observed no difference between the proportions of cancer genes and of the rest of human genes that occur in germline CNVs (35.5% and 37%, respectively, Table). In particular, amplifications are depleted in cancer genes (Table). In case of somatic CNVs, on the other hand, amplifications are enriched in cancer genes regardless of the somatic CNV dataset used (Table). For example, 16.5% of amplifications in Tumorscape peaks overlap with cancer genes, whereas only 8% of them contain non-cancer genes. Instead, there is no significant difference between the proportions of amplifications that overlap with cancer genes and non-cancer genes (11% compared to 10%, respectively). We first thought that the overall enrichment of CNVs in cancer genes in Tumorscape dataset could be due to the GISTIC algorithm that attributes a higher score to gene-containing regions. To eliminate this possible bias, we repeated our analysis with TCGA recurrent and core regions, and observed a similar trend. Remarkably, in all three datasets, the signal came from the enrichment of amplifications in cancer genes.

The tendency of cancer genes to increase their copies in cancer genomes led us to question if this is valid for all the cancer genes. To answer this, we divided the cancer genes into two groups as dominant and recessive cancer genes based on the annotation from the Cancer Gene Census (CGC) (246) (see Methods), and repeated the same analysis on each group separately.

We observed an opposite signal between dominant and recessive cancer genes when we intersected them with Tumorscape peaks (Table). Amplifications show enrichment in dominant genes and depletion in recessive genes when compared to the rest of cancer genes (Table). We did not observe the same pattern when we analyzed somatic CNVs from TCGA datasets. However, by using a different approach (see Methods), we confirmed the same pattern also in this dataset. We found that dominant genes are

amplified in a significantly higher number of samples than the rest of the human genes (Figure A). Similarly, recessive genes are deleted in a significantly higher number of samples than the rest of the human genes (Figure B). This suggests that dominant genes are preferentially amplified and recessive genes are preferentially deleted compared to the rest of human genes.

Table : Intersection of CNVs with cancer genes

Dataset	CNV Type	Cancer Genes			Rest of Human Genes			p-value
		Overlapping	Total	Percentage	Overlapping	Total	Percentage	
Germline	Amplifications	30	501	6%	1,716	17,336	10%	2.83E-03
	Deletions	140	501	28%	4,887	17,336	28%	9.60E-01
	Total	178	501	35.5%	6,376	17,336	37%	6.05E-01
Tumorscape Peaks	Amplifications	83	501	16.5%	1,426	18,544	8%	1.30E-10
	Deletions	56	501	11%	1,820	18,544	10%	3.23E-01
	Total	138	501	27.5%	3,125	18,544	17%	8.98E-14
TCGA Recurrent Regions	Amplifications	402	523	77%	13,356	18,522	72%	1.74E-02
	Deletions	403	523	77%	13,969	18,522	75.5%	4.20E-01
	Total	489	523	93.5%	16,748	18,522	90.5%	1.55E-02
TCGA Core Regions	Amplifications	128	523	24.5%	3,863	18,522	21%	4.97E-02
	Deletions	171	523	33%	6,285	18,522	34%	5.74E-01
	Total	280	523	53.5%	9,390	18,522	51%	2.14E-01

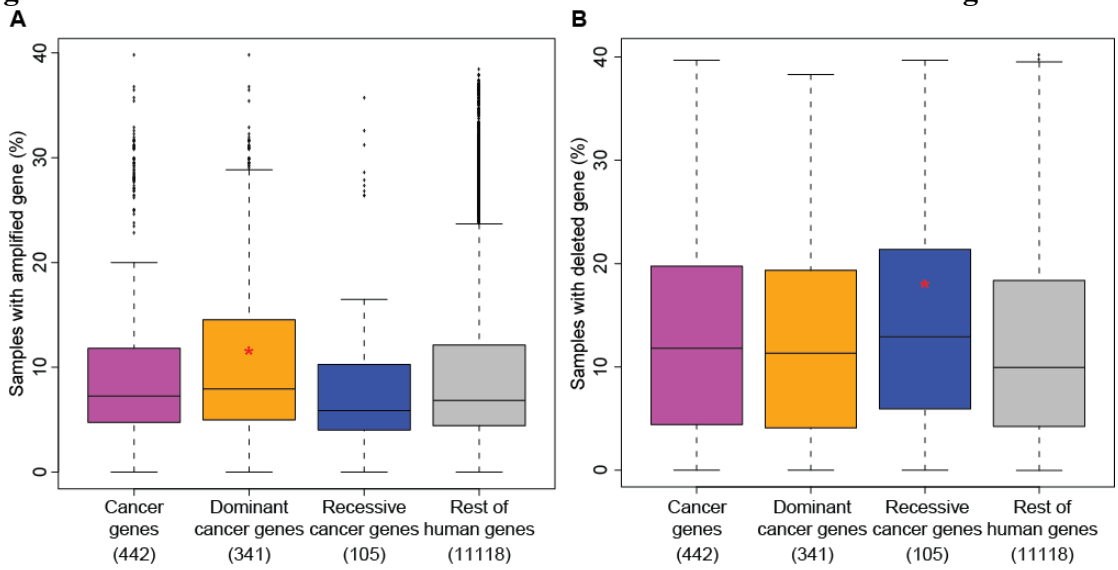
Legend: For each CNV dataset, reported are the numbers and the percentages of cancer genes and other genes that overlap with CNVs. The numbers of cancer genes show differences as the updated list of cancer genes from the Cancer Gene Census is used at the time of each analysis. Fisher's exact test is applied to compare the proportions between the two groups. P-value is highlighted in red in case there is enrichment and in green in case there is depletion for cancer genes (p-value < 0.05).

Table : Intersection of CNVs with dominant and recessive cancer genes

Dataset	CNV Type	Cancer Genes			Rest of Cancer Genes			Fisher's Exact Test	Cancer Genes			Rest of Cancer Genes			p-value
		Dominant	Total	%	Dominant	Total	%	p-value	Recessive	Total	%	Recessive	Total	%	
Germline	Amps	23	26	88	326	422	77	0.2282	3	26	12	100	422	24	0.228
	Dels	98	123	80	251	325	77	0.6121	25	123	20	78	325	24	0.4518
Tumorscape Peaks	Amps	55	61	90	294	387	76	1.23E-02	6	61	10	97	387	24	8.22E-03
	Dels	37	53	70	312	395	79	1.57E-01	16	53	30	87	395	21	2.23E-01
TCGA Recurrent Regions	Amps	274	353	78	89	117	76	7.99E-01	83	353	24	28	117	24	1.00E+00
	Dels	288	372	77	75	98	77	8.92E-01	87	372	23	24	98	24	7.91E-01
TCGA Core Regions	Amps	84	106	79	279	364	77	6.93E-01	24	106	23	87	364	24	8.97E-01
	Dels	125	165	76	238	305	78	5.67E-01	42	165	25	69	305	23	4.97E-01

Legend: For each CNV dataset, reported are the numbers and the percentages of dominant and recessive cancer genes that overlap with CNVs compared to those of cancer genes that do not overlap with CNVs. Fisher's exact test is applied to compare the proportions between the two groups. P-value is highlighted in red in case there is enrichment and in green in case there is depletion for cancer genes (p-value < 0.05). Amps: Amplifications, Dels: Deletions.

Figure : Enrichment of somatic CNVs in dominant and recessive cancer genes



Legend: Distributions of proportions of samples in which genes are **A)** amplified and **B)** deleted. Wilcoxon rank-sum test is applied to compare the distributions of dominant and recessive cancer genes to that of the rest of human genes and statistical significance is shown with an asterisk (p-value < 0.05). Numbers of genes overlapping with CNVs are shown in parenthesis.

3.7 Somatic CNVs show poor correlation with genomic features

To investigate whether the enrichment of somatic CNVs in cancer genes are confounded by other genomic factors, we collected 12 groups of genomic features from UCSC Table Browser and checked for their overlap with somatic CNVs (Table , Table). Overall, we did not observe any consistent enrichment/depletion of CNVs within the genomic features across the different datasets. However, the enrichment of Tumorscape peaks in the genes, exons and transcribed regions from UCSC confirmed our finding that somatic CNVs are highly genic (Table). On the other hand, we did not notice any remarkable feature within the DNA repeating elements that consistently correlates with the CNVs, and those that correlate with CNVs such as short interspersed nuclear elements (SINEs) and long interspersed nuclear elements (LINEs) show differences between amplifications and deletions, unlike a previous observation (159). Additionally, this study reported two classes of breakpoint hotspots for somatic CNVs, cancer-type specific and common hotspots. Cancer-type specific hotspots, are found enriched in cancer genes but

poorly correlated with genomic features, similar to our results, although the common hotspots showed the opposite patterns.

table continued on the next page…

Table : Intersection of somatic amplifications with genomic features

Genomic Feature	Tumorscape Peaks	Rest of the Genome	p-value	TCGA Recurrent Regions	Rest of the Genome	p-value	TCGA Core Regions	Rest of the Genome	p-value
Fragile Sites	21.55%	20.67%	3.09E-02	17.37%	18.24%	2.08E-01	25.20%	27.15%	9.66E-01
Gene Content									
Our dataset (19,045 genes)	37.05%	28.95%	1.58E-03	29.51%	34.10%	1.04E-01	33.05%	30.87%	2.56E-01
UCSC RefSeq Genes	51.41%	38.37%	1.62E-05	41.41%	45.89%	6.29E-02	42.38%	42.95%	5.26E-01
UCSC Known Genes	55.25%	41.68%	2.19E-05	45.86%	50.74%	2.64E-02	46.07%	46.21%	3.59E-01
Exon Density									
Our dataset (181,690 exons)	2.03%	1.20%	9.52E-04	1.50%	1.60%	9.51E-01	1.07%	1.46%	1.13E-01
UCSC RefSeq Exons	4.74%	2.59%	3.78E-05	3.17%	3.16%	9.89E-01	2.32%	3.15%	1.17E-01
UCSC Known Exons	5.52%	3.11%	2.56E-05	3.71%	4.29%	7.10E-01	2.76%	3.77%	1.11E-01
Association with Variants									
COSMIC variants	0.08%	0.24%	3.36E-01	0.15%	0.35%	1.72E-01	0.22%	0.21%	8.69E-02
Uniprot variants	0.83%	0.09%	1.04E-02	0.06%	0.16%	2.85E-01	0.13%	0.09%	9.74E-02
GWAS variants	0.00%	0.00%	6.61E-02	0.00%	0.00%	7.51E-01	0.00%	0.00%	3.40E-01
SNP Density									
Common SNPs	0.47%	0.44%	8.92E-01	0.53%	0.53%	2.35E-01	0.49%	0.52%	6.58E-01
Flagged SNPs	0.01%	0.00%	7.34E-01	0.00%	0.01%	7.91E-01	0.00%	0.01%	2.09E-01
Multi SNPs	0.06%	0.15%	3.83E-04	0.08%	0.10%	1.94E-01	0.07%	0.10%	1.29E-02
All SNPs	0.71%	0.65%	8.98E-01	0.91%	0.96%	9.43E-01	1.01%	0.96%	6.69E-01
Repeating Elements									
All repeats	48.13%	44.40%	4.67E-01	50.05%	49.66%	6.42E-01	46.88%	46.61%	8.69E-02
SINE	18.84%	12.81%	4.00E-05	15.22%	15.99%	7.49E-01	13.13%	14.30%	4.85E-01
LINE	17.17%	18.21%	2.90E-02	21.68%	20.18%	5.04E-01	19.76%	18.67%	8.04E-02
LTR	7.04%	8.16%	3.03E-03	8.04%	8.55%	9.44E-01	8.77%	8.19%	1.33E-01
DNA repeat elements	3.04%	2.91%	9.55E-01	3.15%	2.89%	5.70E-02	3.40%	3.08%	7.78E-03
Simple repeats (micro-satellites)	0.92%	0.92%	5.82E-01	0.96%	1.12%	3.98E-01	0.83%	0.99%	8.20E-02
Low complexity repeats	0.64%	0.51%	2.07E-01	0.56%	0.59%	2.84E-01	0.56%	0.57%	6.12E-01
Satellite repeats	0.22%	0.67%	6.39E-08	0.23%	0.12%	5.03E-01	0.22%	0.61%	3.66E-02
Microsatellite	0.05%	0.05%	8.60E-01	0.06%	0.05%	4.47E-02	0.06%	0.05%	1.65E-02
CpG Island	1.64%	0.82%	1.10E-03	1.09%	1.42%	3.43E-01	0.57%	1.19%	7.67E-04
Level of Transcription	60.79%	46.81%	4.00E-05	51.86%	56.78%	5.20E-02	51.53%	53.48%	9.53E-01

Regulatory Elements	0.79%	0.41%	2.20E-02	0.48%	0.45%	7.32E-01	0.38%	0.48%	8.36E-02
Recombination Rate									
decodeAvg	1.38	1.31	8.27E-01	1.52	1.50	6.76E-01	1.42	1.69	1.31E-01
decodeFemale	1.71	1.53	8.71E-01	1.64	1.48	2.67E-02	1.85	1.76	6.00E-01
decodeMale	1.03	1.07	8.71E-02	1.36	1.48	5.32E-01	0.99	1.60	2.62E-02
marshfieldAvg	1.31	1.30	6.73E-01	1.56	1.56	8.46E-01	1.38	1.63	5.13E-02
marshfieldFemale	1.56	1.46	4.90E-01	1.73	1.57	3.71E-02	1.77	1.69	5.87E-01
marshfieldMale	1.08	1.11	4.57E-01	1.40	1.52	6.32E-01	0.97	1.57	1.11E-02
genethonAvg	1.24	1.26	4.76E-01	1.52	1.46	6.45E-01	1.37	1.65	4.27E-02
genethonFemale	1.49	1.45	6.11E-01	1.65	1.47	1.96E-02	1.77	1.68	8.87E-01
genethonMale	0.96	1.00	1.60E-01	1.33	1.38	9.15E-01	0.96	1.60	1.04E-02
Level of Expression	995.59	994.27	2.31E-03	994.17	993.80	7.20E-01	994.81	994.08	1.38E-01

Legend: Reported are 12 groups of genomic features and their overlap with the somatic amplifications and the rest of the genome. The average values for the overlap are shown in percentage for discrete features and in number for continuous features. Wilcoxon rank-sum test is applied to compare the distributions between the two groups along the genome. P-value is highlighted in red in case there is enrichment and in green in case there is depletion for the somatic amplifications. SINE: Short interspersed nuclear elements, LINE: Long interspersed nuclear elements, LTR: Long terminal repeat elements.

table continued on the next page…

Table : Intersection of somatic deletions with genomic features

Genomic Feature	Tumorscape Peaks	Rest of the Genome	p-value	TCGA Recurrent Regions	Rest of the Genome	p-value	TCGA Core Regions	Rest of the Genome	p-value
Fragile Sites	29.12%	17.97%	7.70E-01	23.10%	22.54%	6.00E-02	22.83%	20.25%	3.47E-01
Gene Content									
Our dataset (19,045 genes)	41.75%	26.75%	4.98E-05	29.49%	31.96%	5.61E-02	33.65%	34.26%	9.45E-01
UCSC RefSeq Genes	53.57%	36.16%	4.30E-06	39.70%	44.32%	3.24E-02	44.87%	47.42%	6.89E-01
UCSC Known Genes	58.80%	38.96%	2.36E-07	43.31%	48.68%	2.93E-02	48.08%	51.24%	5.25E-01
Exon Density									
Our dataset (181,690 exons)	1.40%	1.17%	7.22E-01	0.94%	1.34%	7.33E-05	1.52%	1.63%	5.30E-01
UCSC RefSeq Exons	2.92%	2.52%	9.42E-01	1.99%	2.82%	6.74E-06	3.12%	3.56%	2.55E-01
UCSC Known Exons	3.50%	3.00%	8.89E-01	2.35%	3.38%	4.67E-06	3.69%	4.26%	1.59E-01
Association with Variants									
COSMIC variants	2.24%	0.21%	9.04E-01	0.48%	0.20%	5.30E-04	0.20%	0.24%	8.31E-01
Uniprot variants	0.35%	0.08%	1.19E-02	0.06%	0.14%	1.89E-01	0.08%	0.24%	2.59E-01
GWAS variants	0.00%	0.00%	1.59E-01	0.00%	0.00%	8.24E-03	0.00%	0.00%	7.22E-01
SNP Density									
Common SNPs	0.56%	0.42%	3.01E-04	0.85%	0.51%	2.86E-04	0.56%	0.56%	5.27E-01
Flagged SNPs	0.00%	0.01%	5.93E-03	0.00%	0.01%	3.51E-02	0.00%	0.01%	4.54E-01
Multi SNPs	0.13%	0.18%	4.67E-01	0.06%	0.09%	8.45E-05	0.11%	0.10%	9.87E-02
All SNPs	0.95%	0.74%	4.67E-01	0.63%	0.66%	8.06E-01	0.62%	0.69%	7.92E-01
Repeating Elements									
All repeats	47.54%	40.83%	1.33E-01	47.29%	47.43%	8.09E-01	46.58%	45.78%	5.67E-02
SINE	14.45%	12.02%	3.54E-01	12.21%	13.16%	4.42E-02	15.66%	15.57%	9.80E-01
LINE	19.18%	16.01%	7.16E-02	20.46%	20.14%	3.03E-02	17.96%	17.79%	5.41E-01
LTR	8.56%	7.69%	6.43E-02	9.15%	8.86%	9.11E-03	7.97%	7.45%	1.39E-01
DNA repeat elements	3.20%	2.60%	3.34E-02	3.55%	3.18%	4.97E-03	3.00%	2.79%	6.21E-02
Simple repeats (micro-satellites)	1.14%	1.00%	1.95E-02	1.08%	1.02%	5.92E-04	0.99%	1.09%	1.89E-01
Low complexity repeats	0.64%	0.47%	1.00E-03	0.56%	0.60%	7.57E-04	0.58%	0.57%	7.39E-01
Satellite repeats	0.18%	0.92%	5.07E-04	0.14%	0.20%	9.74E-01	0.22%	0.29%	1.22E-02
Microsatellite	0.05%	0.04%	7.43E-03	0.13%	0.04%	1.17E-02	0.05%	0.05%	7.98E-02
CpG Island	1.42%	0.94%	2.97E-01	0.95%	1.50%	2.89E-06	1.10%	2.09%	1.23E-02
Level of Transcription	62.33%	43.57%	5.85E-06	47.88%	53.19%	4.60E-02	54.65%	58.36%	3.57E-01

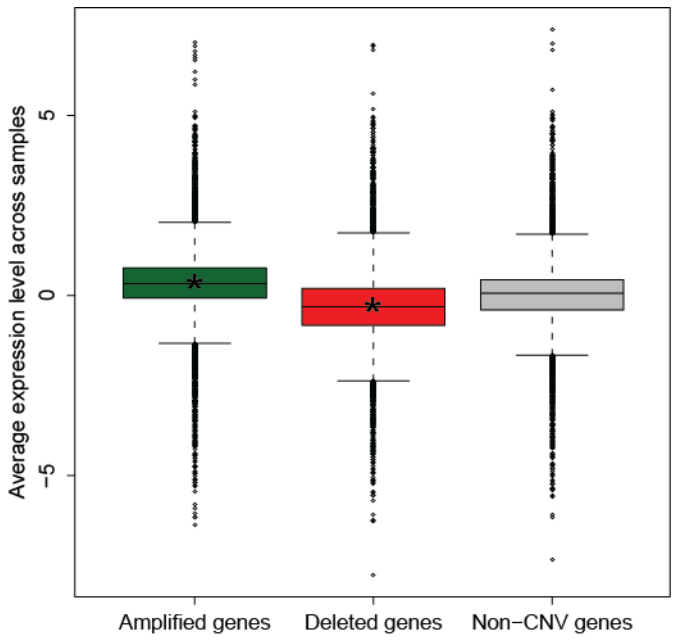
Regulatory Elements	0.48%	0.37%	8.37E-01	0.38%	0.52%	1.06E-05	0.52%	0.64%	3.61E-01
Recombination Rate									
decodeAvg	1.55	1.60	3.76E-01	1.67	1.61	3.78E-01	1.88	1.82	8.48E-01
decodeFemale	1.45	1.67	7.99E-02	1.80	1.67	4.28E-02	2.07	1.76	3.71E-02
decodeMale	1.58	1.46	9.60E-01	1.36	1.38	7.89E-01	1.70	1.87	5.45E-01
marshfieldAvg	1.65	1.58	9.26E-01	1.58	1.49	2.80E-01	1.85	1.74	7.07E-01
marshfieldFemale	1.59	1.61	3.99E-01	1.78	1.51	2.20E-02	2.07	1.61	1.91E-02
marshfieldMale	1.67	1.55	7.28E-01	1.37	1.41	5.63E-01	1.65	1.82	4.96E-01
genethonAvg	1.62	1.46	7.62E-01	1.65	1.50	1.55E-01	1.87	1.82	9.72E-01
genethonFemale	1.51	1.47	4.17E-01	1.73	1.57	7.42E-02	2.05	1.73	3.59E-02
genethonMale	1.61	1.42	9.02E-01	1.40	1.35	5.45E-01	1.64	1.91	4.23E-01
Level of Expression	993.78	994.92	3.60E-01	994.53	995.37	8.20E-02	993.71	994.51	9.44E-01

Legend: Reported are 12 groups of genomic features and their overlap with the somatic deletions and the rest of the genome. The average values for the overlap are shown in percentage for discrete features and in number for continuous features. Wilcoxon rank-sum test is applied to compare the distributions between the two groups along the genome. P-value is highlighted in red in case there is enrichment and in green in case there is depletion for the somatic deletions. SINE: Short interspersed nuclear elements, LINE: Long interspersed nuclear elements, LTR: Long terminal repeat elements.

3.8 Amplifications activate oncogenes and deletions inactivate tumour suppressors

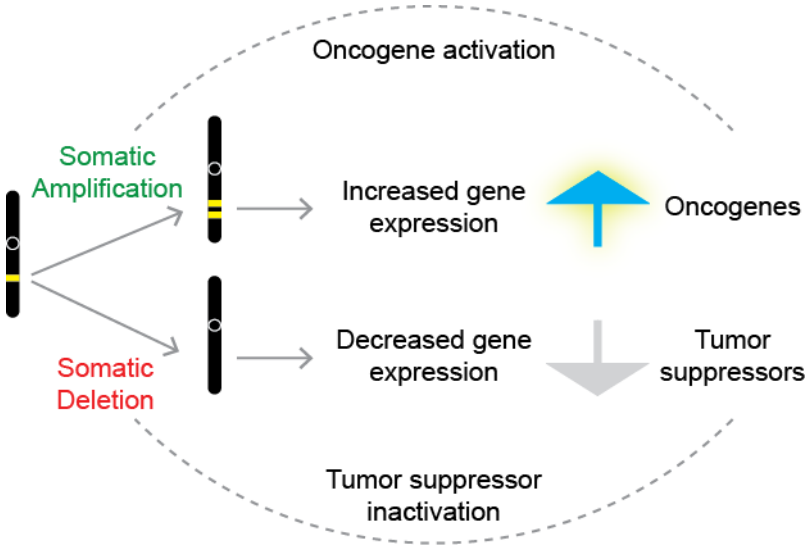
We used gene expression to measure the functional consequences of somatic copy number variations in the cancer genomes. We observed that amplification of a gene usually leads to its higher expression whereas deletion results in decreased gene expression (Figure). We previously showed that amplifications are enriched in dominant genes, while deletions are enriched in recessive genes (Figure). Taken together, these results suggest that dominant genes may be activated through their genomic amplification and recessive genes may be inactivated via deletion. This is interesting because usually dominant genes are oncogenes that are activated by a gain-of-function mutation. Most recessive genes, instead, encode tumour suppressors whose loss-of-function mutations require complete gene inactivation. Therefore, amplifications may activate dominant genes whereas deletion may cause loss-of-function in recessive genes (Figure).

Figure : Gene expression change upon copy number alteration



Legend: Distribution of average expression levels of the genes that are amplified, deleted and that do not undergo copy number variation across samples. Only samples with at least one amplified or deleted gene are considered. Wilcoxon rank-sum test is applied to compare the distributions of amplified genes and deleted genes to that of non-CNV genes and statistical significance is shown with an asterisk (p-value < 0.05).

Figure : Activation and inactivation of cancer genes via gene expression

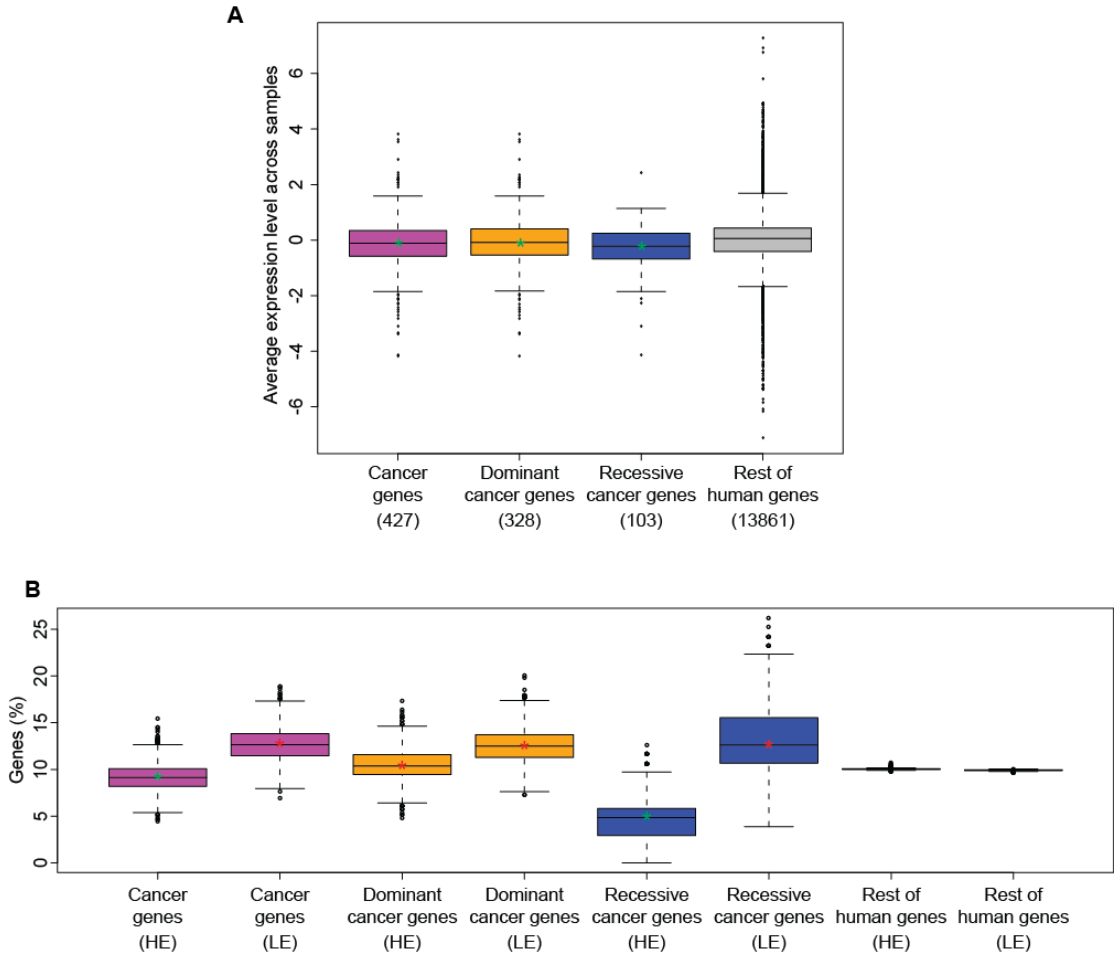


Legend: Amplification of a gene on average increases its expression; on the other hand, dominant genes in particular are enriched in somatic amplifications. Since dominant genes are mostly oncogenes, this may provide a link between somatic amplifications and oncogenes towards the activation. Likewise, deletion of a gene on average decreases its expression; on the other hand, recessive genes in particular are enriched in somatic deletions. Since recessive genes are mostly tumour suppressors, this may provide a link between somatic deletions and tumour suppressors towards inactivation.

3.9 Cancer genes are less expressed than the rest of human genes in cancer samples

In cancer samples, we observed a different expression pattern for cancer genes compared to the rest of human genes. For each gene, we first measured the average expression levels across all 1,245 cancer samples. Although a small fraction of genes are consistently overexpressed or underexpressed, the distribution of average expression levels for all genes are expected to be centered around 0. This is due to the fact that the original expression data is normalized around 0 at the sample level. Likewise, any randomly selected subset of genes is expected to have such a distribution. Cancer genes, however, show significantly lower expression than the rest of human genes, particularly owing to recessive cancer genes (Figure A). To better assess the enrichment of cancer genes in lowly expressed genes, we next divided the genes into highly expressed, medium expressed and lowly expressed based on the gene expression distribution. In the bottom 10% of the distribution, the fraction of cancer genes is significantly higher than expected (Figure B). With these results, we confirm that cancer genes, in particular recessive cancer genes, are less expressed than the rest of the human genes in cancer samples.

Figure : Gene expression in cancer samples



Legend: A) Distribution of average expression levels of genes in the combined dataset of 1,245 cancer samples grouped by the gene type. Numbers in parenthesis denote the number of genes with expression information in the combined dataset. **B)** Percentages of genes that are highly expressed (HE) and lowly expressed (LE) in the cancer samples. The upper and bottom 10% of the gene expression distribution in each sample is used to define HE and LE genes, respectively. Wilcoxon rank-sum test is applied to compare the distributions of all, dominant and recessive cancer genes to that of the rest of human genes, and statistical significance is shown with an asterisk (red: enrichment, green: depletion, p-value < 0.05).

3.10 Frequently amplified recessive cancer genes are involved in epigenetic regulation

Despite the general trend that dominant cancer genes are preferentially amplified and recessive cancer genes are preferentially deleted, however, we found interesting exceptions. Around 3% of the dominant cancer genes are mainly (>90%) deleted in the samples in which they have been modified, whereas 4% of the recessive cancer genes are mainly amplified (Table). This may reflect the fact that dominant cancer genes are not

always oncogenes and not all recessive cancer genes are tumour suppressors. However, there are also cases of real oncogenes that are mostly deleted and of tumour suppressors that are mostly amplified. For example, *ASXL1* is a tumour suppressor gene that is amplified in the overwhelming majority of cancers where it is altered (95%).

We searched for such cases in the literature on an extended list (also including genes from individual cancer types) and found supporting evidence for some of these genes for their modifications opposite to the expected. For example, *PER1*, an essential component of the circadian clock, is reported to have tumour suppressor properties in breast (269,270) and lung (271) cancers, which may explain its frequent deletion in our samples (94% of the modified samples). On the other hand, *EZH2*, catalytic subunit of the PRC2 complex involved in the transcriptional repression of genes, is reported to have a dual role as oncogene and tumour suppressor in haematological malignancies (272). (*EZH2* is deleted in 86% of the modified samples in the overall dataset, and in >90% in 5 individual cancer types.)

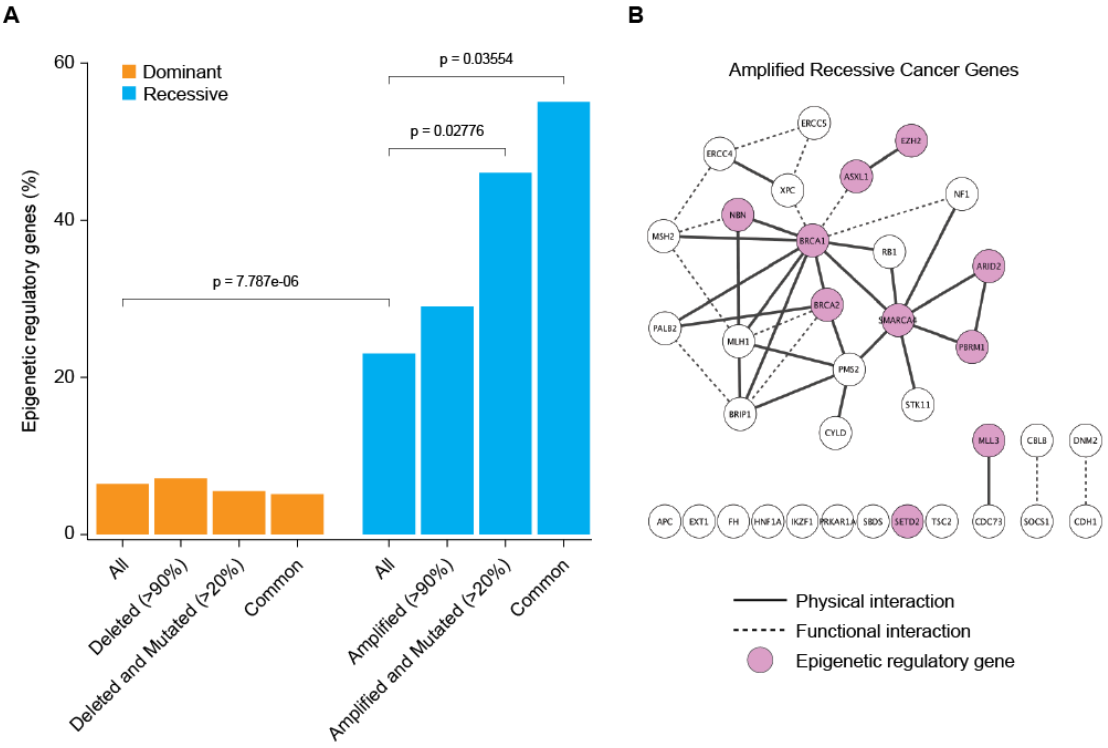
We further investigated other cases that we could not explain through their oncogenic role in cancer, if their modifications highlight certain molecular processes related to their unexpected behaviour. We found that such recessive cancer genes are overrepresented in epigenetic regulation (Figure A). Interestingly, recessive cancer genes in overall are enriched in epigenetic regulators compared to the dominant cancer genes (Figure A). Moreover, we observed that recessive genes that are mainly amplified (>90% of the samples) are highly interconnected via epigenetic regulatory genes (Figure B). Overall, these findings suggest that genes involved in epigenetic mechanisms may undergo genomic modifications in a complex fashion, underlying the importance of the net functional impact of their alteration in tumourigenesis.

Table : List of cancer genes with unexpected genetic modifications

Gene	Gene type	Amplified (n)	Amplified (%)	Deleted (n)	Deleted (%)	Mutated (n)	Mutated (%)	Total Modified (n)
GAS7	Dominant	18	3.58	477	94.83	9	1.79	503
PER1	Dominant	23	4.62	466	93.57	12	2.41	498
RABEP1	Dominant	32	6.56	450	92.21	6	1.23	488
RAP1GDS1	Dominant	19	6.29	278	92.05	6	1.99	302
MAF	Dominant	39	8.92	397	90.85	1	0.23	437
USP6	Dominant	29	5.85	449	90.52	24	4.84	496
TLX1	Dominant	37	9	372	90.51	2	0.49	411
CBFB	Dominant	37	8.85	378	90.43	6	1.44	418
IL2	Dominant	25	8.45	267	90.2	4	1.35	296
ASXL1	Recessive	406	94.64	8	1.86	21	4.90	429
CDC73	Recessive	334	91.76	21	5.77	12	3.30	364
SBDS	Recessive	356	91.75	29	7.47	3	0.77	388
EXT1	Recessive	445	91.56	30	6.17	13	2.67	486

Legend: Reported are the list of 9 dominant cancer genes that are deleted in more than 90% of the samples and 4 recessive cancer genes that are amplified in more than 90% of the samples in which they are modified. Somatic amplifications, deletions and mutations are considered for detecting the modified samples. Given numbers and percentages denote of samples.

Figure : Enrichment of recessive cancer genes within epigenetic regulatory genes



Legend: A) Proportions of dominant and recessive cancer genes within epigenetic regulatory genes grouped according to their unexpected genomic modifications. Modification refers to any of the three types of genomic alterations: amplification, deletion and mutation. Deleted (>90%): genes deleted in >90% of the samples in which they are modified. Deleted and Mutated (20%): genes both deleted and mutated in >20% of the samples in which they are mutated. Amplified (>90%): genes amplified in >90% of the samples in which they are modified. Amplified and Mutated (20%): genes both amplified and mutated in >20% of the samples in which they are mutated. Common: genes present in both groups with unexpected modifications. The thresholds of percentages are selected to define the most extreme cases of the unexpected modification. Fisher's exact test is applied to compare proportions between indicated groups. **B)** Out of 35 recessive genes that are mainly amplified, 20 of them are interconnected in PPI network via a physical or functional interaction, of which 8 are involved in epigenetic regulation. The network representation is generated in Cytoscape (http://www.cytoscape.org/) by using physical PPI data from NCG (http://ncg.kcl.ac.uk/) and functional interactome data from Reactome (http://wiki.reactome.org/index.php/Reactome_FI_Cytoscape_Plugin_4). A list of 633 epigenetic regulatory genes is obtained from the literature (273).

3.11 Identification of novel synthetic lethal interactors in cancer

So far we have mainly shown how somatic CNVs can impact on cancer genomes, which helps extend our understanding of the mutational causes of the cancer. It is equally important to make use of this knowledge in translational research. For this purpose, we used the mutational landscape and the systems-level properties of genes to identify targetable cancer-specific vulnerabilities, in particular, synthetic lethality. We hypothesized that paralogous genes, i.e. genes duplicated in the genome, are good candidates for conferring synthetic lethality due to the functional redundancy between them, as both originate from the same ancestral gene. This hypothesis was proved to be working for recessive cancer genes and their functional paralogs (188). In the presence of synthetic lethal interaction between the paralogous pair, this approach can be a powerful strategy to kill cancer cells specifically, where one of the paralogs is mutated and the cells are dependent on the remaining paralog. To identify such paralogous pairs in cancer samples, I first predicted the mutated genes with loss-of-function mutations in cancer samples, and then used several gene properties and additional information to assess their functional paralogs. My work included all the steps of the computational work in the pipeline described next.

3.11.1 Prediction of putative candidates

We exploited the data curated in this work to predict putative synthetic lethal interactors in cancer. Starting from the dataset of 1,245 cancer samples from 11 cancer types derived from TCGA for our somatic CNV analysis (Table), we first annotated the genes with loss-of-function mutations. Our hypothesis is that if synthetic lethality occurs between two genes that exert a specific function that is required for cell survival, impairing both of them will lead to cell death. With this aim, we assessed the impact of somatic mutations on the protein function based on a combined score of six prediction tools (SIFT (274), PolyPhen-2 HumDiv (275), PolyPhen-2 HumVar (275), Likelihood Ratio Test (276), MutationTaster2 (277) and MutationAssessor (278)). We considered only those

mutations predicted to be damaging by at least 4 out of the 6 tools (>66%). We identified 11,285 genes that acquired loss-of-function mutations in at least one of the cancer samples. We also identified paralogs of these mutated genes by using the method we described before (173). We preferentially focused on 2,028 paralogous pairs, *i.e.* cases where the mutated gene has only one paralog. The presence of only one paralog reduces the chances of off-target effects in the experimental validation.

Out of these 2,028 paralogous pairs, we prioritized the best candidates for experimental validation considering the following conditions:

1) Sequence identity and domain composition between the paralogs

Function of a protein is closely related to its structure, and structure is dictated by its primary sequence. Therefore sequence of a protein can be informative of its function. Although not always true, two proteins with highly identical sequences can exert similar functions. Using this information, we prioritized our candidates based on their sequence identity, expecting higher levels of functional redundancy between highly identical paralogs. In addition, we focused on the paralogous pairs with shared domains, particularly if they are present only in these two genes rather than being a common domain in many proteins. In this case, impairment of the shared domain in the gene will sensitize the paralog for the specific function, i.e. reducing the possibility of functional compensation by another gene with the same domain.

2) Relative expression levels of the paralogs

In the presence of functional compensation between paralogous pairs, we hypothesized reduced expression of the impaired gene and/or overexpression of its paralog. We therefore analysed the expression profile of the paralogous pairs across all 1,245 cancer samples. For each pair, we compared expression distribution of the mutated gene to that of the wild type paralog in the same samples. We prioritized those cases

supporting our hypothesis, i.e. down-regulated genes with loss-of-function mutations and up-regulated paralogs with wild type phenotype.

3) Literature and network support of functional redundancy

Through the literature we searched for evidence of functional redundancy between our paralogous pairs, considering that proteins part of the same complex or pathway are more likely to be involved in similar biological function. In complement, we assessed the proximity of the paralogs in the protein-protein interaction network by measuring the shortest path (i.e. the minimum number of nodes) between the two encoded proteins. We prioritized cases where the distance between the paralogs are minimal, which may indicate the involvement of the two proteins in the same biological pathway.

4) Availability of cell lines with the ideal genetic make up

For the experimental validation of synthetic dependence, we used cell-based in vitro assays. To decide on the most suited experimental set up for our purpose, we analysed the mutational landscape of 1,417 cell lines from two public sources (1,036 cell lines from Cancer Cell Line Encyclopedia (279) and 1,030 cell lines from COSMIC Cell Lines Project (280)). We prioritized cell lines with homozygous loss-of-function mutations in the same genes that were found mutated in the cancer samples but with wild-type paralogs. If possible, we used the cell lines from the same cancer type as that of the samples, in which the genes were mutated.

At the end of this selection procedure, we extracted 37 paralogous pairs (Table). Among these we focused on *STAG1*/*STAG2* pair for further experimental validation for the following reasons:

1) *STAG1* and *STAG2* has a relatively high sequence coverage (49%) and high sequence identity (83%) within the covered region (NCG 5.0, http://ncg.kcl.ac.uk/)

2) Both proteins contain STAG domain (Pfam, (281))

3) The encoded proteins of the two genes are part of the same complex, i.e. cohesin complex (NCG 5.0, http://ncg.kcl.ac.uk/)

table continued on the next page…

Table : Predicted putative synthetic lethal gene pairs

Gene	Gene type	Paralog	Gene type	Sequence coverage	Samples (n)	Expression of mutated gene	Expression of wild-type paralog
CYFIP2	Non-cancer	*CYFIP1*	Candidate	83%	9	down	-
VHL	Recessive	*VHLL*	Non-cancer	60%	23	down	-
TLN2	Non-cancer	*TLN1*	Non-cancer	56%	9	down	-
TRIM3	Non-cancer	*TRIM2*	Non-cancer	52%	2	down	-
SOS2	Non-cancer	*SOS1*	Non-cancer	51%	7	down	-
CDKN2A	Recessive	*CDKN2B*	Candidate	49%	12	-	up
STAG2	Recessive	*STAG1*	Non-cancer	49%	17	down	-
STXBP5	Non-cancer	*STXBP5L*	Non-cancer	38%	3	down	up
KCTD3	Non-cancer	*SHKBP1*	Candidate	34%	11	down	-
OLFML2A	Non-cancer	*OLFML2B*	Non-cancer	34%	7	down	-
MEGF10	Non-cancer	*MEGF11*	Non-cancer	32%	9	down	-
ADAMTS6	Non-cancer	*ADAMTS10*	Non-cancer	29%	7	down	-
ARID1A	Recessive	*ARID1B*	Candidate	29%	31	down	up
HAND2	Non-cancer	*HAND1*	Non-cancer	28%	3	down	-
FRY	Non-cancer	*FRYL*	Non-cancer	28%	13	down	-
RING1	Non-cancer	*RNF2*	Non-cancer	27%	2	-	up
RBM10	Candidate	*RBM5*	Non-cancer	25%	11	down	-
SDK1	Non-cancer	*SDK2*	Non-cancer	22%	18	down	-
PTPRG	Non-cancer	*PTPRZ1*	Non-cancer	20%	7	down	-
GYG2	Non-cancer	*GYG1*	Non-cancer	19%	4	down	-
ARNTL	Non-cancer	*ARNTL2*	Non-cancer	19%	2	-	up
GPRASP1	Non-cancer	*GPRASP2*	Non-cancer	18%	6	down	-
E2F7	Non-cancer	*E2F8*	Non-cancer	18%	5	-	up
E2F8	Non-cancer	*E2F7*	Non-cancer	15%	2	-	up
MPDZ	Non-cancer	*INADL*	Non-cancer	15%	5	down	-
ZDHHC17	Non-cancer	*ZDHHC13*	Non-cancer	15%	4	-	up
MAP1B	Non-cancer	*MAP1A*	Non-cancer	14%	7	down	-
PLEKHH2	Non-cancer	*PLEKHH1*	Non-cancer	14%	9	down	-
COL4A5	Non-cancer	*COL4A1*	Non-cancer	14%	15	down	-
ADAMTS19	Non-cancer	*ADAMTS17*	Non-cancer	14%	9	down	-

TP53	Recessive	*TP73*	Non-cancer	13%	531	down	up
NFE2L2	Dominant	*NFE2L1*	Non-cancer	12%	16	-	up
ASTN2	Non-cancer	*ASTN1*	Candidate	12%	15	down	-
PRDM1	Recessive	*ZNF683*	Non-cancer	12%	3	-	up
PTEN	Recessive	*ANKFN1*	Non-cancer	11%	119	down	up
PPM1E	Non-cancer	*PPM1F*	Non-cancer	11%	9	down	-
CIC	Recessive	*SEPT14*	Non-cancer	10%	8	down	-

Legend: Reported is the list of 37 predicted synthetic lethal pairs and associated properties used for the selection (not all properties are shown). Gene types refer to the classification in NCG 5.0 database (http://ncg.kcl.ac.uk/). Sequence coverage is the portion of the gene also covered in the paralog. Samples refer to the cancer samples from TCGA carrying the mutated gene and the wild type paralog. Gene expression distributions of the mutated gene and of the wild type paralog in the given samples were compared to those in the rest of the 1,245 samples. down: significantly decreased expression, up: significantly increased expression.

3.11.2 *STAG1* and *STAG2* is a novel synthetic lethal gene pair

We performed in our lab *in vitro* experiments to test the predicted synthetic lethal interaction between *STAG1* and *STAG2*. My colleague Lorena Benedetti performed all the wet lab experiments described below, and I produced the plots based on these experimental data. Then we both discussed on the results and interpreted them, leading to the design of the next steps. Briefly, we used transient RNA interference (RNAi) to block the expression of *STAG1* and *STAG2* with small interfering RNA (siRNA) and then measured the cell proliferation in Cal-51 breast cell lines on the following conditions:

1) Wild type *STAG2* and wild type *STAG1*

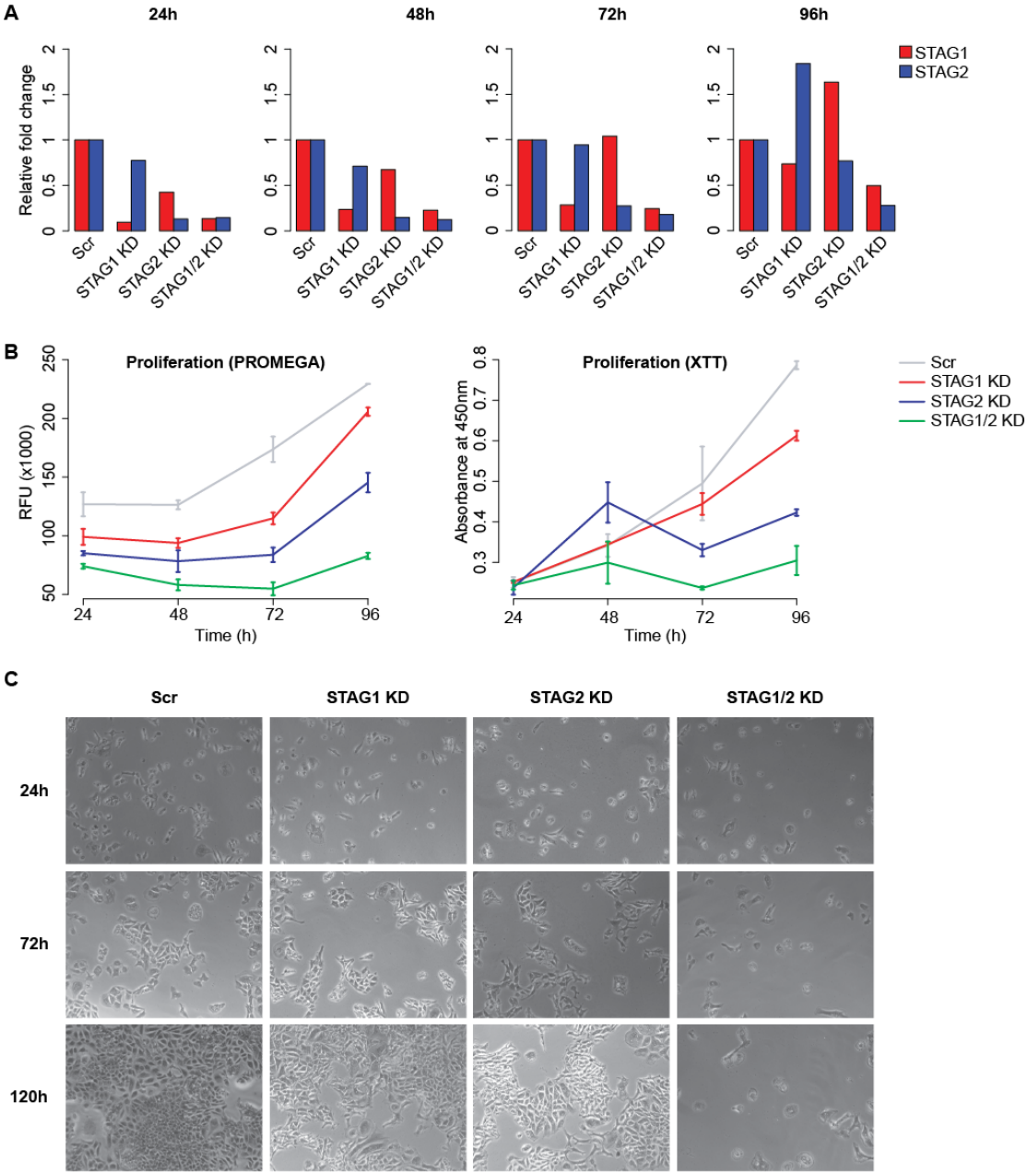
2) *STAG2* treated with siRNA and wild type *STAG1*

3) Wild type *STAG2* and *STAG1* treated with siRNA

4) *STAG2* and *STAG1* treated with siRNA

First, we measured the expression levels of both genes in four time points (24h, 48h, 72h, 96h). We verified decreased expression levels of both genes when they were knocked down in Cal-51 breast cell lines (Figure A). We also observed an increased expression of the paralog (after 96 hours) when we blocked the expression of only one gene. Next we tested for the cell proliferation in the same conditions by measuring the proliferation rate of the cells at the same time points by using two different proliferation assays (Promega ApoLive-Glo Multiplex Assay and XTT Cell Proliferation Assay) and colony formation assay. We found that when the expression of only one paralog is knocked down, the cells continue to proliferate, although with a reduced rate compared to scrambled. On the other hand, cells cease to proliferate when the expression of both genes are knocked down, confirmed by using two different proliferation assays (Figure B) and colony formation assay (Figure C). Overall, our findings suggest that impairment of both *STAG1* and *STAG2* is incompatible with cell viability.

Figure : Experimental validation of *STAG1* and *STAG2* synthetic dependence



Legend: A) Expression of STAG1 and STAG2 upon treatment with siRNA oligos assessed by qPCR. The ratios are relativized to scrambled. **B)** Proliferation of cells treated with siRNA oligos compared to scrambled assessed by using two different assays. **C)** Colony formation of cells treated with siRNA oligos compared to scrambled. Scr: scrambled, KD: transient knockdown, RFU: Relative fluorescence units.

3.12 Network of Cancer Genes (NCG 5.0)

In order to facilitate the research based on cancer genes and their properties, we maintain a public database developed in our group: Network of Cancer Genes (NCG). NCG is a manually curated repository of cancer genes and associated systems-level properties (Figure). The website was first published in 2010 (261), and is updated every two years. The latest version (NCG 5.0) includes 1,571 cancer genes from 175 cancer genomics studies and from the CGC (262). These cancer genes were analysed in 13,315 cancer samples from 49 cancer types corresponding to 24 primary sites. As the main developer of the last two versions of NCG, I manually curated all the 175 papers, maintained a database to store all the data and implemented the updated content and new features on the website. My work included the extraction of cancer genes along with complementary cancer information, analysis of their systems-level properties by integrating data from internal pipelines and external sources, creating web content to display the updated and new features, and improving the visualisation and the performance of the website. My colleagues Giovanni Dall'Olio and Thanos Mourikis contributed to the development of the latest version as the second curator of the papers, and helping me in part to update the website.

The pipeline for the NCG development consists of two main steps: data curation and implementation.

3.12.1 Data curation

The main purpose of NCG is to provide a reliable collection of cancer genes. Cancer genes in NCG derive from the published cancer genomics studies. We continuously review the literature for new publications to include and manually extract cancer genes satisfying the following conditions (Figure A):

1) **Genes detected by DNA sequencing in cancer samples:** High-throughput screenings of cancer samples provide an unbiased approach to detect cancer

genes. Studies that analyse only cell lines or that use other approaches than DNA sequencing to identify cancer genes are discarded.

2) **Genes altered with point mutations or indels:** Genetic alterations are the driving force of cancer. Currently NCG collects only mutated genes with non-silent point mutations or indels. Genes undergoing other types of alterations will be added to the database in the future.
3) **Genes identified as drivers based on a method:** Studies must apply a method to prove the driver role of the mutated genes in cancer. A method can be the application of any known tools or software, or other approaches as described in the original study. This step is crucial to confine the curation to reliable cancer drivers and to avoid any possible false positives detected without a proper method.

We divide curated cancer genes into two groups (Figure B):

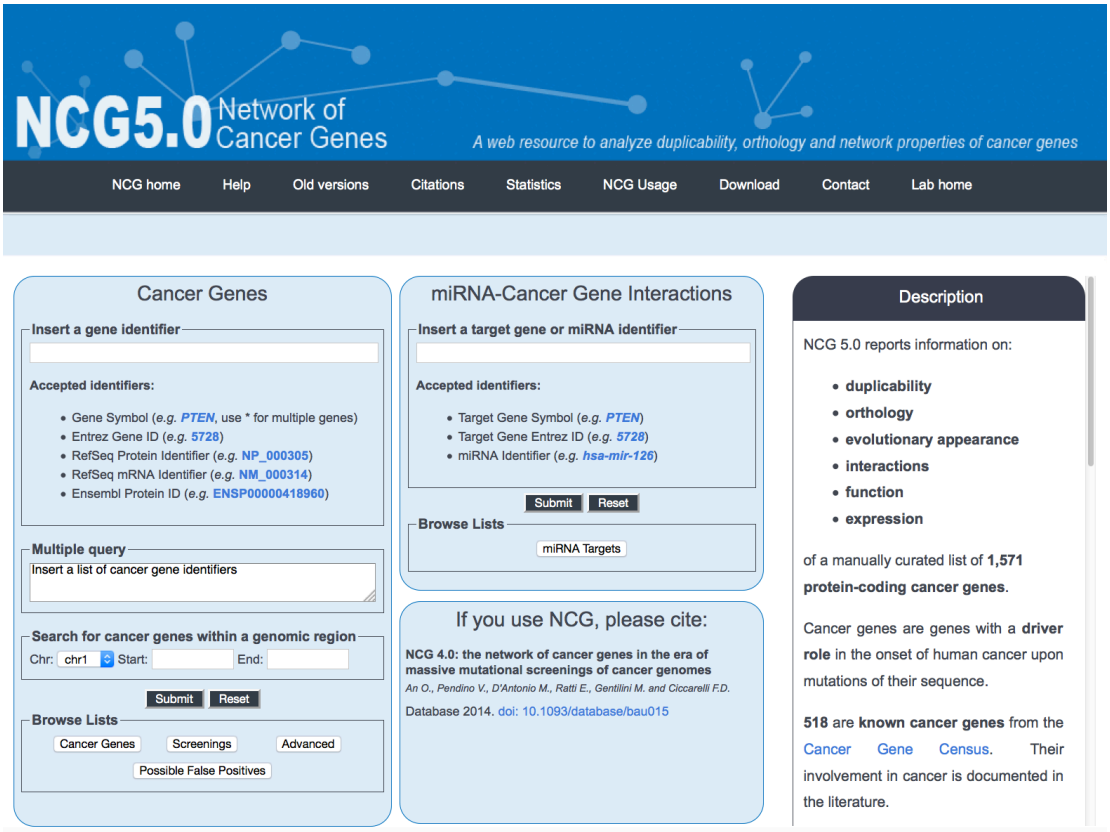
1) **Known cancer genes:** Genes also reported in CGC and defined as causally implicated in oncogenesis (246). Together with other genes from this source, we report 518 known cancer genes in NCG.
2) **Candidate cancer genes:** The remaining genes derived from the literature. We report 1053 candidate cancer genes whose driver role in cancer was described in the original studies.

In addition to cancer genes, we also collect complementary data from each study: screening type, tumour description, primary site, cancer type, number of patients, method description, experimental validation, and reference (Figure C). We further provide several properties for each cancer gene either calculated in-house (duplicability, evolutionary origin, network properties) or integrated from external sources (orthologs, protein-protein interactions, expression, function, miRNA interactions) (Figure D).

3.12.2 Implementation

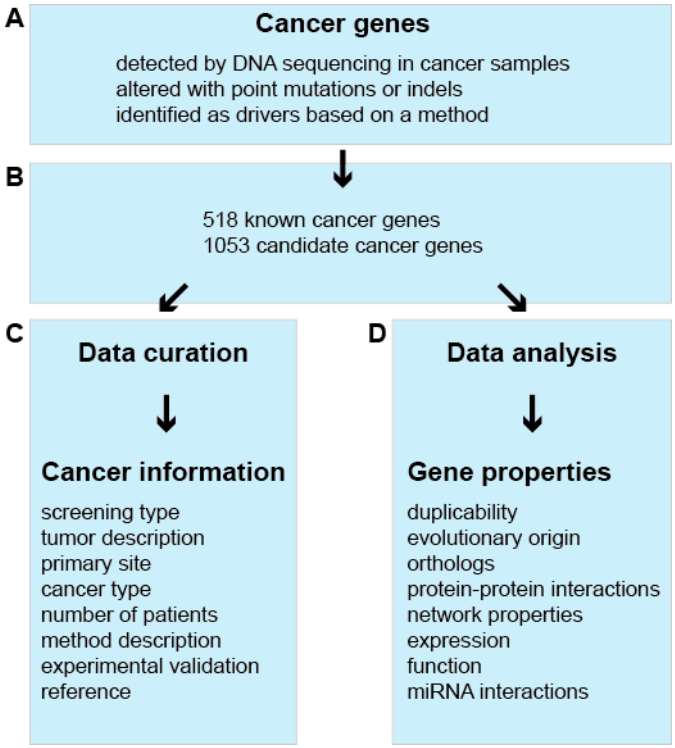
We developed the web interface of NCG in PHP and implemented network visualisation of protein-protein and miRNA-target interactions in Cytoscape Web (http://cytoscapeweb.cytoscape.org/, (282)). NCG 5.0 runs on an Apache web server and data are stored in a MySQL database on a virtual Windows machine. The data content can be downloaded as a text file in batch, or individually for the specific queries.

Figure : NCG 5.0 home page



Legend: Screenshot of the home page of NCG 5.0, the latest version published in October 2015 (262).

Figure : Overview of cancer genes curation and data content in NCG 5.0



Legend: A) Criteria for inclusion of cancer genes in the database. **B)** Number of cancer genes in NCG 5.0. **C)** Curated complementary data from the literature along with the cancer genes. **D)** Analyzed and integrated gene properties.

3.12.3 Other features

In addition to the manual curation, another powerful feature of NCG is the annotation of experimentally validated candidate cancer genes. Focused on providing a reliable set of cancer genes, we searched in the literature for any experimental validation which demonstrates the oncogenic role of the curated candidate cancer genes. Out of 1053 candidate cancer genes, we found experimental evidence for 120 genes spanning a variety of methods such as gene overexpression, gene silencing, immunohistochemistry and protein activity assay. Thus experimental validation adds a complementary measure to the methods used to identify the cancer genes in the original studies.

On the other hand, NCG also reports a list of 48 possible false positives among the candidate cancer genes. These genes are regarded as possible false positive due to the following reasons:

1) Literature evidence ((283), 14 genes)
2) Functional irrelevance (i.e. olfactory receptors, 30 genes)
3) Gene length (i.e. long exons and/or introns, 9 genes)

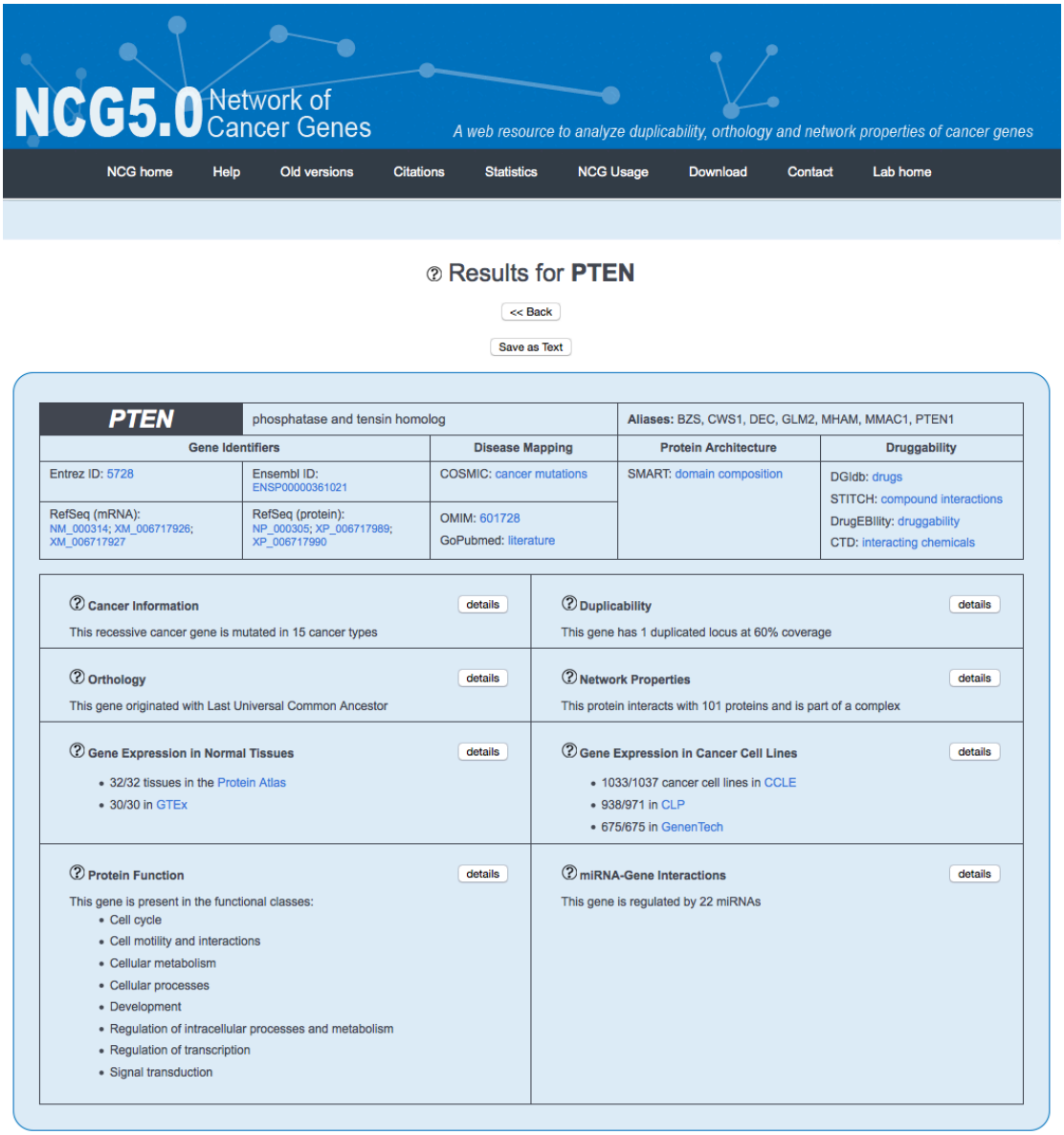
Interestingly, we noticed that a gene, *CSMD3*, is in common between the list of 120 experimentally validated and 48 possible false positive genes. *CSMD3* is proposed as a possible false positive due to its length, sequence composition and proximity to fragile sites (283), however in another study it is experimentally shown that its stable knockout leads to increased cell proliferation (284). In this case, NCG reports information from the both sources and leaves the final decision to the user on the gene annotation.

3.12.4 User interface

NCG interface allows the users to search for a single gene or a list of genes by inserting one of the accepted identifiers (i.e. gene symbol, Entrez id, RefSeq protein id, RefSeq mRNA id or Ensembl protein id). An option to retrieve all the genes within a genomic region is also provided. Alternatively, the complete list of cancer genes and pre-compiled list of cancer genes from individual screenings can be browsed. The advanced search allows for specific queries based on the shared gene properties. A successful query returns information on the annotation and properties of each gene (Figure):

1) **Gene summary:** Description, aliases and cross-links to external databases.
2) **Cancer information:** List of all the screenings and complementary data demonstrating the involvement in cancer.
3) **Duplicability:** Additional copies at different thresholds of coverages.
4) **Orthology:** Evolutionary origin and orthologs in the tree of life.
5) **Network properties:** Interactions and network properties in human PPI network.
6) **Gene expression:** Expression levels in 38 normal tissues and in 1,543 cell lines.
7) **Protein function:** List of all functional classes of the encoded protein.
8) **miRNA-gene interactions:** miRNAs that regulate the gene and their targets.

Figure : Annotation and properties of cancer genes in NCG 5.0



Legend: Screenshot of results page of NCG 5.0, showing a summary of annotation and systems-level properties of an example cancer gene, *PTEN*.

3.13 NCG can be used to prioritize candidates for experimental validation: A case study of *STAG1* and *STAG2*

NCG database can be used as a central repository for cancer genes and their properties to help experimental planning. By using NCG, we explored the properties of our candidate gene pairs predicted to have synthetic dependency and decided to start with *STAG1/STAG2* pair for experimental validation. In NCG, *STAG2* is reported as a recessive cancer gene based on CGC and also as mutated in 4 different cancer types in 11 studies (Figure A). In our mutation dataset, we found *STAG2* mutated in 17 cancer samples from TCGA, which overlapped by one study curated in NCG (285). *STAG1*, on the other hand, cannot be queried in NCG since it is not a cancer gene, however, some of its properties can be still derived via *STAG2* (see below). For more details on *STAG1*, we used the external sources cross-linked from NCG. We used the following information in NCG for the selection of this pair:

1) *STAG2* has only one paralog, *STAG1*, which has 49% coverage (Figure B). Literature mining showed that a second paralog exists, *STAG3*; however, we found that this gene shares very low sequence coverage with *STAG2* (<10%) and expressed in a tissue-specific manner (only in testis). Therefore we excluded *STAG3* as a potential gene which can functionally compensate *STAG2*.

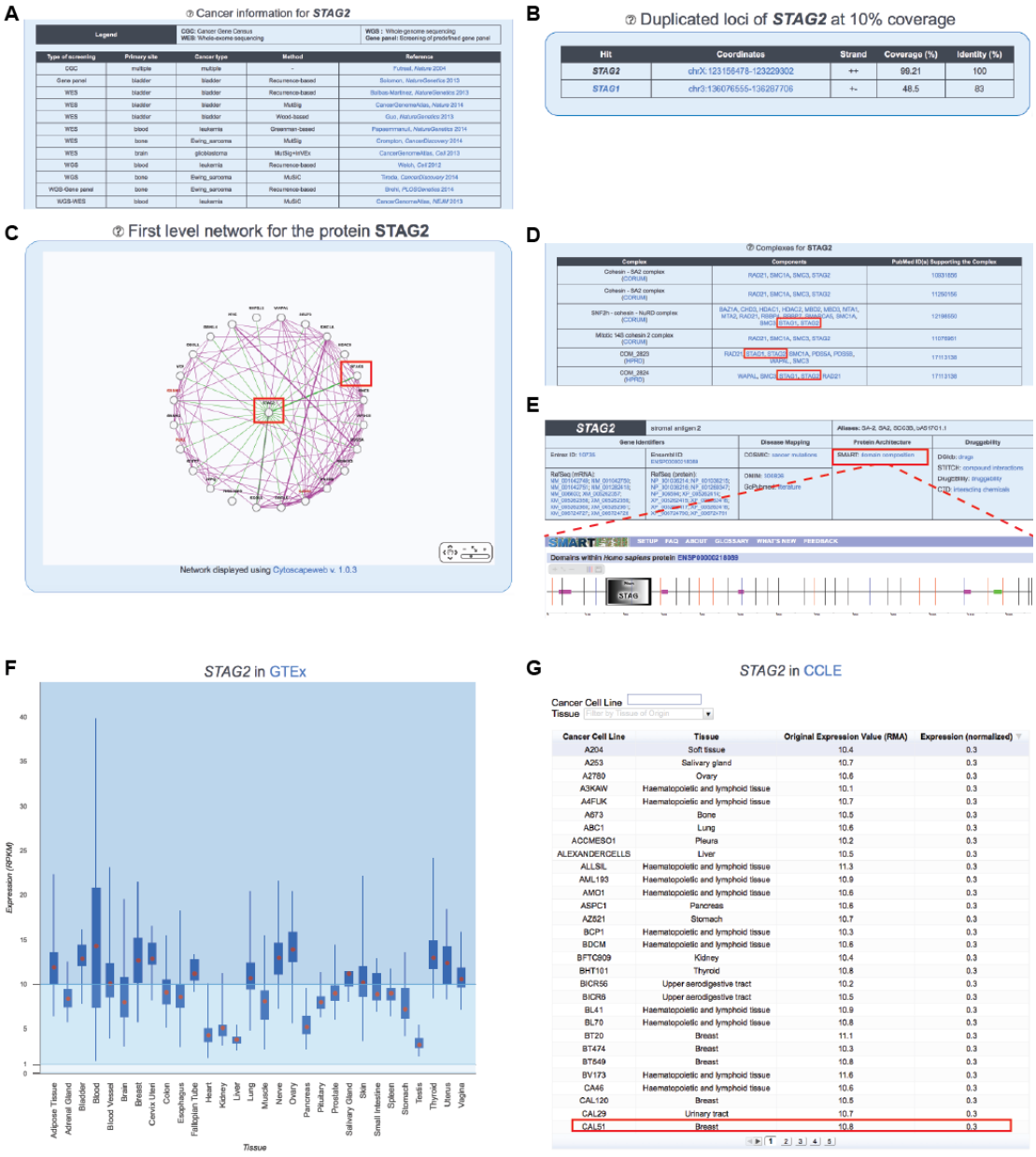
2) Both genes are physical interactors in binary human PPI network (Figure C), and components of the same complex ("cohesin" complex) (Figure D).

3) Both genes have the same protein domain ("STAG" domain) (Figure E).

4) Both genes are widely expressed across normal human tissues (Figure F).

5) Both genes are expressed in many cancer cell lines from different tissues (Figure G). We selected Cal-51 breast cancer cell line for our experiment, because we found *STAG2* mutated also in breast cancer samples from TCGA, among 6 other cancer types, and Cal-51 was readily available.

Figure : Annotation and properties of *STAG2* in NCG 5.0



Legend: A) List of studies in which *STAG2* is identified as a cancer gene. **B)** Duplicates of *STAG2* above 10% coverage of its sequence. **C)** Primary interactions of *STAG2* in human PPI network. **D)** Examples of complexes in which *STAG1* and *STAG2* are involved. **E)** Cross-linking from NCG to SMART database (286) depicting its domain composition. **F)** Expression levels of *STAG2* in 30 normal human tissues from the Genotype-Tissue Expression (GTEx) project (287). **G)** List of cancer cell lines from the Cancer Cell Line Encyclopedia (CCLE) (279) in which *STAG2* is expressed.

4 Discussion

To study the role of somatic copy number variations (CNVs) in cancer, it is important first to understand the landscape of germline CNVs in normal population. This is not only required for a grounded comparison, but also provides a basis on how the somatic CNVs should be interpreted. In order to map the landscape of CNVs in normal population, we have explored germline CNVs from a large collection of healthy individuals (Table). Comparing them to somatic CNVs from cancer samples allowed us to distinguish the features of cancer-specific CNVs from inherited CNVs. Then we focused on cancer-specific CNVs with the aim of understanding their driver role in cancer. With this aim, we performed two lines of analyses: 1) quantification of genes contained within CNVs 2) assessment of the functional impact of altered genes via gene expression.

CNVs are not always biomarkers for a disease state. They are also common in normal genomes with no apparent phenotype (12), which makes it challenging to distinguish disease-causing variants from neutral variants. In cancer, it is more complicated to identify the driver variants contributing to cancer onset and progression owing to the passenger events. Several tools were developed for this purpose (112,249,288-294) and were extensively used in later studies. Here we have used a simple approach to identify recurrent regions of CNVs in cancer without attempting to annotate them as driver or passenger. Our approach relied on the previous observations that these regions tend to recur because they contain potential events favouring the cancer progression (positive selection) (13,15,257). Among these events, we explored the importance of genes, in particular cancer genes, and several genomic features.

We observed that somatic CNVs load a much heavier burden on the genome compared to the germline CNVs, which is expected due to the genomic instability that arise in cancer (295). They may involve focal regions containing driver genes for cancer progression (13,264) or span large portions of the genome with an oncogenic role (296). On the other hand, germline CNVs are mainly distributed along the intergenic regions, or

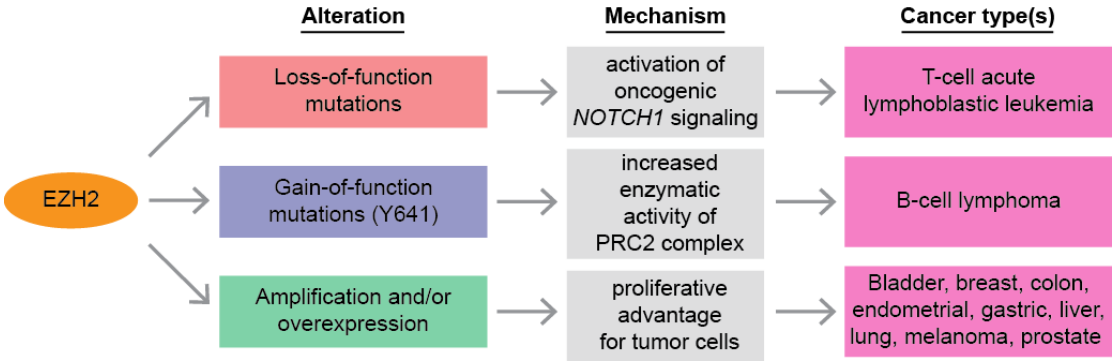
involve one or few genes with no apparent disease phenotype. This is also compatible with one of the suggested roles for germline CNVs, that they are substrates for human evolution and adaptation (297).

A large number of studies reported amplification and deletion of cancer genes in individual cancer types, and a few others analysed somatic CNVs in cancer on a large scale (13,15,111). Here we also collect a large cohort of samples, and further investigate cancer genes in two classes as dominant and recessive genes. We show that there is an enrichment of somatic amplifications within dominant genes and an enrichment of somatic deletions within recessive genes (Figure). This result agrees with the functional relevance of these two classes of cancer genes to oncogenes and tumour suppressors; moreover, carries the discussion further to the genetic basis of cancer genes.

Primary modification of cancer genes support our proposal that dominant cancer genes can be activated by somatic amplifications and recessive cancer genes can be inactivated by somatic deletions (Figure). However, we also found a few cases in which the primary modification of the genes seems to be contradictory to this, that dominant genes that are mainly deleted (such as *GAS7*, *PER1*, *RABEP1* deleted in >90% of the tumour samples) and recessive genes that are mainly amplified (such as *ASXL1*, *CDC73*, *SBDS* amplified in >90% of the tumour samples) (Table). Investigation of these genes in the literature revealed that dominant cancer genes can act as tumour suppressors (*EBF1* (298), *PER1* (269-271), *ZNF331* (299), *CAMTA1* (300-302)) and recessive cancer genes can act as oncogenes (*EZH2* (272), *CBLB* (303)) (former is more common). Furthermore, a portion of cancer genes can act as both oncogene and tumour suppressor depending on the context, such as mutation type or cancer type (such as *EZH2* (304), *MN1* (305), *NSD1* (306)). These cases indeed validate our findings and strengthen our proposal (Figure). Classifying cancer genes as oncogenes and tumour suppressors rather than as dominant and recessive thereby could give a stronger signal in our analysis. However, such an annotation for all known cancer genes is not yet available.

We relied on Cancer Gene Census (CGC) (246) for the classification of cancer genes as dominant and recessive. CGC is an ongoing effort to catalogue genes for which mutations have been causally implicated in cancer that is evidenced in the literature (246). The cancer genes in the census are annotated on the basis of molecular genetics of their driver mutations, i.e. dominant and recessive. As we discussed before, dominant and recessive cancer genes usually correspond to oncogenes and tumour suppressors, respectively, however, this should not be taken as a reference. During literature mining, we encountered cases in which dominant cancer genes are reported to have tumour suppressor roles (*EBF1* (298), *ZNF331* (299), *CAMTA1* (300-302)), and recessive cancer genes are regarded as oncogenes (*EZH2* (272), *CBLB* (303)). Among these, we found the case of *EZH2* very interesting. *EZH2* is a histone-lysine N-methyltransferase and the catalytic component of Polycomb Repressive Complex 2 (PRC2), acting mainly as a transcriptional repressor of a wide range of genes. In T-cell acute lymphoblastic leukaemia (T-ALL), *EZH2* was found to have loss of function mutations and deletions in 25% of the samples together with *SUZ12*, another important component of the PRC2 complex. Disrupting the function of the complex due to the impairment of these two genes in *NOTCH1*-mutated samples triggered oncogenic NOTCH1 signalling, leading to T-ALL (307). In diffuse large B-cell lymphoma, *EZH2* was found to have a gain of function mutation (Y641), which increased the enzymatic activity of the PRC2 complex (308). Moreover, *EZH2* was found amplified and/or overexpressed that confers proliferative advantage to the cells in a variety of cancer types including bladder (309-311), breast (312), colon (313), endometrial (314), gastric (315), liver (316), lung (317), melanoma (314) and prostate (318). Overall these data suggest multiple roles for *EZH2* in cancer in a context dependent manner (i.e. mutation type, cancer type) (Figure), a phenomenon commonly observed among epigenetic regulators (319) and signalling proteins (320).

Figure : Role of *EZH2* in cancer



Legend: *EZH2* is a key gene in epigenetic regulation of transcriptional repression, which has been widely studied owing to its important function and increasing significance in cancer. A variety of oncogenic alterations in *EZH2* have been identified so far, leading to different outcomes triggering various mechanisms based on the alteration type and the cancer context. The multi-faceted role of *EZH2* in cancer has been well established, a phenomenon commonly observed among epigenetic regulators (319) and signalling proteins (320).

Copy number variations are mechanical changes in chromosome structure, thereby they are prone to be affected by a variety of genomic features. Considering that the human genome is not homogeneous along its sequence and CNVs occur more frequently on certain regions of the genome than others, we wondered if there is an association between the genomic features and CNV formation. To investigate this, we compared the overlap between the genomic features and somatic CNVs. We would expect a similar distribution between a particular feature and CNVs along the genome in the case of positive correlation. Although previous studies found positive correlations between CNVs and several genomic features (such as indel rate, exon density and substitution rate), this may depend on several factors, such as cancer types included and CNV datasets used (159). Similarly, we observed varying patterns depending on the CNV dataset (i.e. Tumorscape peaks, TCGA recurrent regions, TCGA core regions) and type of alteration (i.e. amplification, deletion) (Table , Table). There was no feature consistently correlated with CNVs in the same direction across all three datasets and two alteration types. On the other

hand, there was no inconsistency between amplifications and deletions from the same dataset for the same feature, i.e. we did not find enrichment in amplifications and depletion in deletions for the same feature. Here we propose two possible explanations for these observations: 1) CNV formation is influenced by a combination of genomic features, possibly including other features that we did not consider here. 2) Amplifications and deletions are selected in the same direction by the genomic features, but not necessarily at the same strength. For example, both amplifications and deletions tend to occur in the SINEs, but only amplifications show statistically significant enrichment (Table , Table).

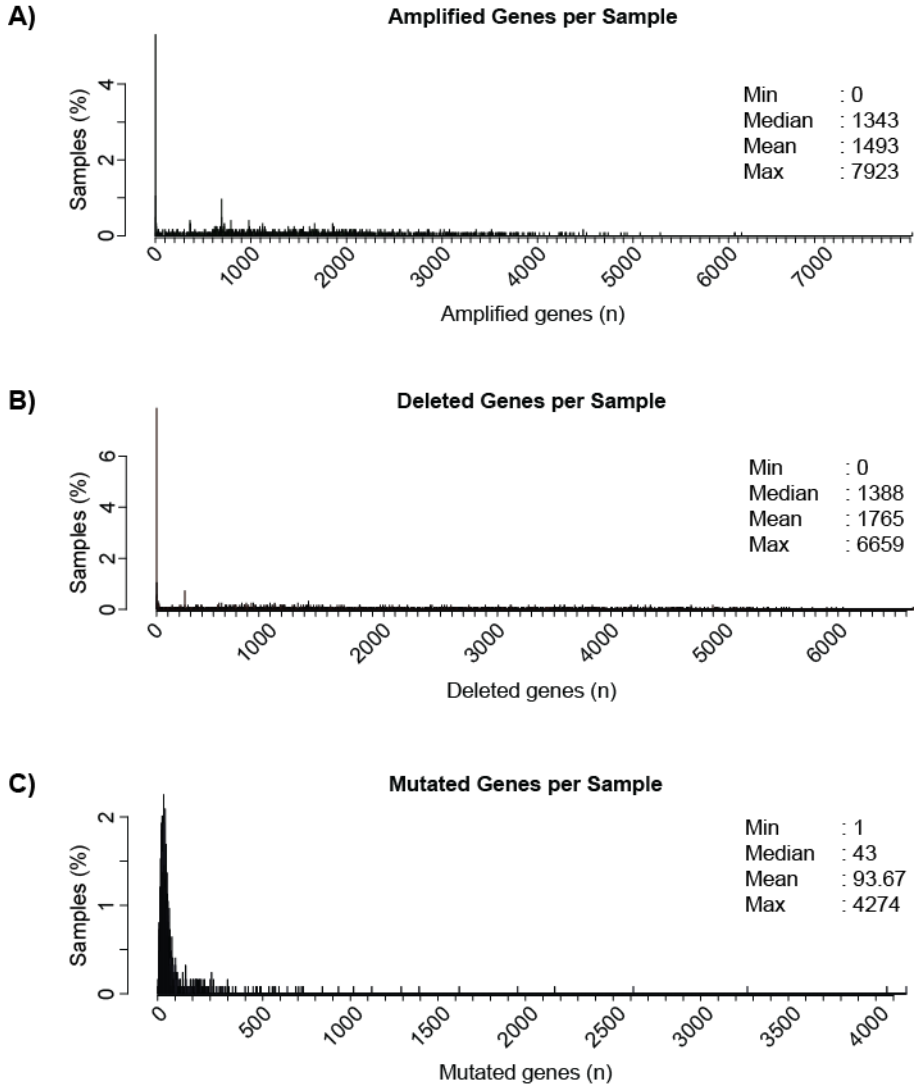
In our CNV analyses, we did not take the zygosity of the somatic copy number variations into account. For our purpose, we used CNV data as a categorical variable, i.e. present or absent, to define each gene as amplified, deleted or non-altered. Such an assignment led to a clear signal between CNVs and gene expression: a positive correlation between amplifications and gene expression, and a negative correlation between deletions and gene expression. However, absolute copy numbers can be useful for more detailed quantitative analysis, and it is possible to infer this information from the amplitude of each segment (defined as the log2 ratio of the signal intensity in a given genomic region between the test and the reference sample) given in the original datasets. Illustratively, a previous study showed a positive correlation between the copy number and the expression of a set of genes involved in colorectal cancer (321), adding an extra layer of information to what we have shown here (Figure).

Our effort to determine the primary genetic modification that affects cancer genes was limited to CNVs, and did not include mutations. This was due to the sparsity of the mutation data. In TCGA dataset, a typical cancer sample contained an average of 1493 amplified and 1765 deleted genes, but only 94 mutated genes (Figure). The fractions of samples with mutated genes, therefore, were too low compared to the fractions of samples with CNVs to see any correlation. A previous study, however, reported anticorrelation between somatic copy number variations and mutations by using a similar cohort of

samples from TCGA (14). The approach used in this study considered only the recurrent elements (defined as selected functional elements, SFEs) for both copy number variations and mutations, detected by GISTIC (322) and MuSiC/MutSig (323,324), respectively. Using these SFEs, the authors could obtain a comparable number of samples with CNVs and mutations, and observe an anticorrelation between the two (14).

We used gene expression changes as a measure of the functional impact of somatic CNVs. For such an analysis, gene expression values from tumour samples should be ideally compared to the original values in the matched normal samples. This will minimize the confounding factors on gene expression such as tissue-specificity, and allow for the measurement of changes only due to somatic CNVs, as they are present only in the tumour. However, due to the lack of expression data for the matched normal samples in TCGA samples, we made the comparison by using only the tumour samples, i.e. between tumour samples in which a gene is altered versus non-altered. This allowed us to observe a clear difference between the two sets of samples, which we concluded as due to the presence of copy number variations (Figure).

Figure : Distribution of modified genes across cancer samples



Legend: Distribution of **A)** amplified, **B)** deleted, and **C)** mutated genes across 1,245 cancer samples from TCGA. The summary statistics of each distribution is given for comparison.

Microarray-based platforms have been widely used for decades for gene expression detection, as they offered a rapid and low cost solution for functional studies. On the other hand, new technologies based on RNA sequencing (RNA-Seq) has several advantages over the microarrays, such as unbiased detection of novel transcripts, easier detection of rare or weakly expressed genes, and increased specificity and sensitivity. In our analyses, we used only microarray-based gene expression data from TCGA, because it was available for a wider cohort of samples at the time of data retrieval. At present, public resources on cancer genomics (such as TCGA and ICGC) provide RNA-Seq data more widely than

microarray-based data for gene expression. Although microarray-based data served well for our purpose that is to measure gene expression change upon copy number variation, we encountered a limitation: identification of not expressed genes. We classified genes as highly, medium or lowly expressed based on the gene expression distribution, however, we could not identify the genes that are not expressed, as we used processed data from TCGA where each gene had an expression value. Identification of not expressed genes could be used to better assess whether the lack of expression of a gene is due to its deletion.

Large-scale analysis of genomics data not only provides a high statistical power, but also depicts a more complete profile of the investigated features. Therefore, we started with all available somatic CNVs data from TCGA at the time of analysis, which consisted of 6,213 samples from 23 cancer types. For the later analyses, we integrated mutation and gene expression data for the same samples wherever available, which led to a smaller dataset of 1,245 samples from 11 cancer types. Also the number of human genes with corresponding information decreased from 19,045 to 14,288. Although this data reduction did not effect our initial observations regarding the characterisation of somatic CNVs on the overall dataset, it is possible that we have missed some cancer-type specific patterns. For example, previous studies reported cancer genes recurrently altered in cancer types which were not part of our dataset (264). It is also possible that in a larger cohort of samples, our predicted set of paralogous pairs with functional redundancy would include other potential candidates. However, these limitations are likely to disappear in the future, because TCGA as well as other public resources on genomics data are continuously growing, providing a more complete data profiling on the cancer samples.

Novel methods and their applications on cancer genomics data hold the promise revolutionizing personalised anticancer therapy in the near future. Among these, the concept of synthetic lethality has recently gained great importance as it provides an opportunity of selective targeting of tumour cells. Utilizing this concept already led to the identification of numerous synthetic lethal pairs in cancer (325). Previous studies used

different approaches to identify such pairs, including genome-wide screening (199) and mutual exclusivity analysis (238). Here we exploited the curated data for copy number variation analysis also to identify novel synthetic lethal pairs. In addition to the genomics data, we also gathered a variety of information at the gene level (such as gene duplicability, sequence identity, protein domain composition, functional redundancy, network properties, pathway and complex information) to predict and prioritize the best candidate pairs with a possible synthetic lethal interaction. We extended the work previously published in our group, which demonstrated that recessive cancer genes sharing certain properties are potential candidates for having synthetic lethality with their functional paralogs (188). Experimental validation of the pair *STAG1/STAG2* presented here is a further confirmation of this idea. Efficient molecular tools recently developed for cellular engineering such as CRISPR (326,327) will ease and enlarge the applicability of similar ideas in the field.

Our approach to identify putative synthetic lethal pairs has several advantages as well as limitations compared to the previous methods (199,216,238,239). First, our analysis is biased to the genes that are mutated in at least one cancer sample in our dataset and have only one paralog on the genome, which represent about 11% of the annotated genes (2,028 out of 19,014). Although we might miss many potential candidates among the remaining genes, we principally focus on the most relevant selection of genes for our purpose, supported by the working hypothesis that paralogous genes are engaged in synthetic lethal interaction due to their functional redundancy (188). Second, we curated as many features as possible that can support the functional redundancy of the paralogous pairs, by using both computational and manual approach. We did not apply strict criteria on the features for the prediction, rather we used the data to prioritize the candidate pairs for experimental validation. Ideally, all of 2,028 pairs should be experimentally tested to assess our prediction accuracy. Such an assessment will also quantify the applicability of our initial hypothesis. However, this is not practically feasible in terms of laboratory work, time and

cost that it requires. Nevertheless, considering the virtually infinite combinations of gene pairs in the genome, our approach is a good approximation of potentially relevant candidates, reducing the data to a manageable size. In the future, it will be useful to develop a computational tool based on the methodology that we used here to readily predict and prioritize candidate pairs on the updated datasets.

The prediction of putative gene pairs with synthetic lethality relied on several assumptions. First, we assumed paralogous genes with certain properties would have functional redundancy, such as high sequence identity, shared functional domains and being part of the same biological pathway or complex. This is a widely accepted assumption based on a number of real examples from the literature (328), although there are cases in which two proteins with high sequence identity have different functions (329,330). Second, we assumed that the decreased expression of a gene is due to its loss-of-function mutation. For each gene, we coupled loss-of-function mutations with gene expression changes in the cancer samples in which it is mutated, however, there is not necessarily a cause-and-effect relationship between the two. Third, we assumed that the increased expression of wild-type paralog is due to the functional compensation. We took increased expression as an indicative of increased gene activity that replaces the specific function disrupted due to the impaired gene. All these assumptions, however, were made only to predict and prioritize our candidates for the experimental validation, and we were able to validate our prediction for the pair of *STAG1/STAG2*.

Appendix: Published papers

The two published papers are attached at the end of the thesis:

1) An, O., Dall'Olio, G.M., Mourikis, T.P. and Ciccarelli, F.D. (2015) **NCG 5.0: updates of a manually curated repository of cancer genes and associated properties from cancer mutational screenings.** *Nucleic Acids Res*.
2) An, O., Pendino, V., D'Antonio, M., Ratti, E., Gentilini, M. and Ciccarelli, F.D. (2014) **NCG 4.0: the network of cancer genes in the era of massive mutational screenings of cancer genomes.** *Database (Oxford)*, **2014**, bau015.

References

1. Li, R., Montpetit, A., Rousseau, M., Wu, S.Y., Greenwood, C.M., Spector, T.D., Pollak, M., Polychronakos, C. and Richards, J.B. (2014) Somatic point mutations occurring early in development: a monozygotic twin study. *J Med Genet*, **51**, 28-34.
2. Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W. *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860-921.
3. 1000 Genomes Project Consortium, Abecasis, G.R., Auton, A., Brooks, L.D., DePristo, M.A., Durbin, R.M., Handsaker, R.E., Kang, H.M., Marth, G.T. and McVean, G.A. (2012) An integrated map of genetic variation from 1,092 human genomes. *Nature*, **491**, 56-65.
4. 1000 Genomes Project Consortium, Abecasis, G.R., Altshuler, D., Auton, A., Brooks, L.D., Durbin, R.M., Gibbs, R.A., Hurles, M.E. and McVean, G.A. (2010) A map of human genome variation from population-scale sequencing. *Nature*, **467**, 1061-1073.
5. Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A. *et al.* (2001) The sequence of the human genome. *Science*, **291**, 1304-1351.
6. Almal, S.H. and Padh, H. (2012) Implications of gene copy-number variation in health and diseases. *J Hum Genet*, **57**, 6-13.
7. International HapMap Consortium. (2003) The International HapMap Project. *Nature*, **426**, 789-796.
8. Sherry, S.T., Ward, M.H., Kholodov, M., Baker, J., Phan, L., Smigielski, E.M. and Sirotkin, K. (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res*, **29**, 308-311.
9. Zhang, J., Chiodini, R., Badr, A. and Zhang, G. (2011) The impact of next-generation sequencing on genomics. *J Genet Genomics*, **38**, 95-109.
10. Stankiewicz, P. and Lupski, J.R. (2010) Structural variation in the human genome and its role in disease. *Annu Rev Med*, **61**, 437-455.
11. Feuk, L., Carson, A.R. and Scherer, S.W. (2006) Structural variation in the human genome. *Nat Rev Genet*, **7**, 85-97.
12. MacDonald, J.R., Ziman, R., Yuen, R.K., Feuk, L. and Scherer, S.W. (2014) The Database of Genomic Variants: a curated collection of structural variation in the human genome. *Nucleic Acids Res*, **42**, D986-992.
13. Beroukhim, R., Mermel, C.H., Porter, D., Wei, G., Raychaudhuri, S., Donovan, J., Barretina, J., Boehm, J.S., Dobson, J., Urashima, M. *et al.* (2010) The landscape of somatic copy-number alteration across human cancers. *Nature*, **463**, 899-905.
14. Ciriello, G., Miller, M.L., Aksoy, B.A., Senbabaoglu, Y., Schultz, N. and Sander, C. (2013) Emerging landscape of oncogenic signatures across human cancers. *Nat Genet*, **45**, 1127-1133.
15. Kim, T.M., Xi, R., Luquette, L.J., Park, R.W., Johnson, M.D. and Park, P.J. (2013) Functional genomic analysis of chromosomal aberrations in a compendium of 8000 cancer genomes. *Genome Res*, **23**, 217-227.
16. Zhang, F., Gu, W., Hurles, M.E. and Lupski, J.R. (2009) Copy number variation in human health, disease, and evolution. *Annu Rev Genomics Hum Genet*, **10**, 451-481.
17. Fanciulli, M., Petretto, E. and Aitman, T.J. (2010) Gene copy number variation and common human disease. *Clin Genet*, **77**, 201-213.
18. Firth, H.V., Richards, S.M., Bevan, A.P., Clayton, S., Corpas, M., Rajan, D., Van Vooren, S., Moreau, Y., Pettett, R.M. and Carter, N.P. (2009) DECIPHER:

Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources. *Am J Hum Genet*, **84**, 524-533.
19. Vulto-van Silfhout, A.T., van Ravenswaaij, C.M., Hehir-Kwa, J.Y., Verwiel, E.T., Dirks, R., van Vooren, S., Schinzel, A., de Vries, B.B. and de Leeuw, N. (2013) An update on ECARUCA, the European Cytogeneticists Association Register of Unbalanced Chromosome Aberrations. *Eur J Med Genet*, **56**, 471-474.
20. Qiu, F., Xu, Y., Li, K., Li, Z., Liu, Y., DuanMu, H., Zhang, S., Li, Z., Chang, Z., Zhou, Y. *et al.* (2012) CNVD: text mining-based copy number variation in disease database. *Hum Mutat*, **33**, E2375-2381.
21. Bayes, M., Magano, L.F., Rivera, N., Flores, R. and Perez Jurado, L.A. (2003) Mutational mechanisms of Williams-Beuren syndrome deletions. *Am J Hum Genet*, **73**, 131-151.
22. Somerville, M.J., Mervis, C.B., Young, E.J., Seo, E.J., del Campo, M., Bamforth, S., Peregrine, E., Loo, W., Lilley, M., Perez-Jurado, L.A. *et al.* (2005) Severe expressive-language delay related to duplication of the Williams-Beuren locus. *N Engl J Med*, **353**, 1694-1701.
23. Knight, M.A., Hernandez, D., Diede, S.J., Dauwerse, H.G., Rafferty, I., van de Leemput, J., Forrest, S.M., Gardner, R.J., Storey, E., van Ommen, G.J. *et al.* (2008) A duplication at chromosome 11q12.2-11q12.3 is associated with spinocerebellar ataxia type 20. *Hum Mol Genet*, **17**, 3847-3853.
24. Chen, K.S., Manian, P., Koeuth, T., Potocki, L., Zhao, Q., Chinault, A.C., Lee, C.C. and Lupski, J.R. (1997) Homologous recombination of a flanking repeat gene cluster is a mechanism for a common contiguous gene deletion syndrome. *Nat Genet*, **17**, 154-163.
25. Potocki, L., Bi, W., Treadwell-Deering, D., Carvalho, C.M., Eifert, A., Friedman, E.M., Glaze, D., Krull, K., Lee, J.A., Lewis, R.A. *et al.* (2007) Characterization of Potocki-Lupski syndrome (dup(17)(p11.2p11.2)) and delineation of a dosage-sensitive critical interval that can convey an autism phenotype. *Am J Hum Genet*, **80**, 633-649.
26. Chance, P.F., Alderson, M.K., Leppig, K.A., Lensch, M.W., Matsunami, N., Smith, B., Swanson, P.D., Odelberg, S.J., Disteche, C.M. and Bird, T.D. (1993) DNA deletion associated with hereditary neuropathy with liability to pressure palsies. *Cell*, **72**, 143-151.
27. Lupski, J.R., de Oca-Luna, R.M., Slaugenhaupt, S., Pentao, L., Guzzetta, V., Trask, B.J., Saucedo-Cardenas, O., Barker, D.F., Killian, J.M., Garcia, C.A. *et al.* (1991) DNA duplication associated with Charcot-Marie-Tooth disease type 1A. *Cell*, **66**, 219-232.
28. Reiner, O., Carrozzo, R., Shen, Y., Wehnert, M., Faustinella, F., Dobyns, W.B., Caskey, C.T. and Ledbetter, D.H. (1993) Isolation of a Miller-Dieker lissencephaly gene containing G protein beta-subunit-like repeats. *Nature*, **364**, 717-721.
29. Cardoso, C., Leventer, R.J., Ward, H.L., Toyo-Oka, K., Chung, J., Gross, A., Martin, C.L., Allanson, J., Pilz, D.T., Olney, A.H. *et al.* (2003) Refinement of a 400-kb critical region allows genotypic differentiation between isolated lissencephaly, Miller-Dieker syndrome, and other phenotypes secondary to deletions of 17p13.3. *Am J Hum Genet*, **72**, 918-930.
30. Bi, W., Sapir, T., Shchelochkov, O.A., Zhang, F., Withers, M.A., Hunter, J.V., Levy, T., Shinder, V., Peiffer, D.A., Gunderson, K.L. *et al.* (2009) Increased LIS1 expression affects human and mouse brain development. *Nat Genet*, **41**, 168-177.
31. Shaikh, T.H., Kurahashi, H., Saitta, S.C., O'Hare, A.M., Hu, P., Roe, B.A., Driscoll, D.A., McDonald-McGinn, D.M., Zackai, E.H., Budarf, M.L. *et al.* (2000) Chromosome 22-specific low copy repeats and the 22q11.2 deletion syndrome: genomic organization and deletion endpoint analysis. *Hum Mol Genet*, **9**, 489-501.

32. Edelmann, L., Pandita, R.K., Spiteri, E., Funke, B., Goldberg, R., Palanisamy, N., Chaganti, R.S., Magenis, E., Shprintzen, R.J. and Morrow, B.E. (1999) A common molecular basis for rearrangement disorders on chromosome 22q11. *Hum Mol Genet*, **8**, 1157-1167.
33. Yu, S., Cox, K., Friend, K., Smith, S., Buchheim, R., Bain, S., Liebelt, J., Thompson, E. and Bratkovic, D. (2008) Familial 22q11.2 duplication: a three-generation family with a 3-Mb duplication and a familial 1.5-Mb duplication. *Clin Genet*, **73**, 160-164.
34. Ensenauer, R.E., Adeyinka, A., Flynn, H.C., Michels, V.V., Lindor, N.M., Dawson, D.B., Thorland, E.C., Lorentz, C.P., Goldstein, J.L., McDonald, M.T. *et al.* (2003) Microduplication 22q11.2, an emerging syndrome: clinical, cytogenetic, and molecular analysis of thirteen patients. *Am J Hum Genet*, **73**, 1027-1040.
35. Ou, Z., Berg, J.S., Yonath, H., Enciso, V.B., Miller, D.T., Picker, J., Lenzi, T., Keegan, C.E., Sutton, V.R., Belmont, J. *et al.* (2008) Microduplications of 22q11.2 are frequently inherited and are associated with variable phenotypes. *Genet Med*, **10**, 267-277.
36. Padiath, Q.S., Saigoh, K., Schiffmann, R., Asahara, H., Yamada, T., Koeppen, A., Hogan, K., Ptacek, L.J. and Fu, Y.H. (2006) Lamin B1 duplications cause autosomal dominant leukodystrophy. *Nat Genet*, **38**, 1114-1123.
37. Saunier, S., Calado, J., Benessy, F., Silbermann, F., Heilig, R., Weissenbach, J. and Antignac, C. (2000) Characterization of the NPHP1 locus: mutational mechanism involved in deletions in familial juvenile nephronophthisis. *Am J Hum Genet*, **66**, 778-789.
38. Konrad, M., Saunier, S., Heidet, L., Silbermann, F., Benessy, F., Calado, J., Le Paslier, D., Broyer, M., Gubler, M.C. and Antignac, C. (1996) Large homozygous deletions of the 2q13 region are a major cause of juvenile nephronophthisis. *Hum Mol Genet*, **5**, 367-371.
39. Beutler, E. and Gelbart, T. (1994) Erroneous assignment of Gaucher disease genotype as a consequence of a complete gene deletion. *Hum Mutat*, **4**, 212-216.
40. Braga, S., Phillips, J.A., 3rd, Joss, E., Schwarz, H. and Zuppinger, K. (1986) Familial growth hormone deficiency resulting from a 7.6 kb deletion within the growth hormone gene cluster. *Am J Med Genet*, **25**, 443-452.
41. Goossens, M., Brauner, R., Czernichow, P., Duquesnoy, P. and Rappaport, R. (1986) Isolated growth hormone (GH) deficiency type 1A associated with a double deletion in the human GH gene cluster. *J Clin Endocrinol Metab*, **62**, 712-716.
42. Rodrigues, N.R., Owen, N., Talbot, K., Ignatius, J., Dubowitz, V. and Davies, K.E. (1995) Deletions in the survival motor neuron gene on 5q13 in autosomal recessive spinal muscular atrophy. *Hum Mol Genet*, **4**, 631-634.
43. Matthijs, G., Schollen, E., Legius, E., Devriendt, K., Goemans, N., Kayserili, H., Apak, M.Y. and Cassiman, J.J. (1996) Unusual molecular findings in autosomal recessive spinal muscular atrophy. *J Med Genet*, **33**, 469-474.
44. Kan, Y.W., Golbus, M.S. and Trecartin, R. (1975) Prenatal diagnosis of homozygous beta-thalassaemia. *Lancet*, **2**, 790-791.
45. Higgs, D.R., Pressley, L., Old, J.M., Hunt, D.M., Clegg, J.B., Weatherall, D.J. and Serjeant, G.R. (1979) Negro alpha-thalassaemia is caused by deletion of a single alpha-globin gene. *Lancet*, **2**, 272-276.
46. Antonarakis, S.E., Kazazian, H.H. and Tuddenham, E.G. (1995) Molecular etiology of factor VIII deficiency in hemophilia A. *Hum Mutat*, **5**, 1-22.
47. Wraith, J.E., Cooper, A., Thornley, M., Wilson, P.J., Nelson, P.V., Morris, C.P. and Hopwood, J.J. (1991) The clinical phenotype of two patients with a complete deletion of the iduronate-2-sulphatase gene (mucopolysaccharidosis II--Hunter syndrome). *Hum Genet*, **87**, 205-206.

48. Wilson, P.J., Suthers, G.K., Callen, D.F., Baker, E., Nelson, P.V., Cooper, A., Wraith, J.E., Sutherland, G.R., Morris, C.P. and Hopwood, J.J. (1991) Frequent deletions at Xq28 indicate genetic heterogeneity in Hunter syndrome. *Hum Genet*, **86**, 505-508.
49. Bondeson, M.L., Dahl, N., Malmgren, H., Kleijer, W.J., Tonnesen, T., Carlberg, B.M. and Pettersson, U. (1995) Inversion of the IDS gene resulting from recombination with IDS-related sequences is a common cause of the Hunter syndrome. *Hum Mol Genet*, **4**, 615-621.
50. Shapiro, L.J., Yen, P., Pomerantz, D., Martin, E., Rolewic, L. and Mohandas, T. (1989) Molecular studies of deletions at the human steroid sulfatase locus. *Proc Natl Acad Sci U S A*, **86**, 8477-8481.
51. Froyen, G., Corbett, M., Vandewalle, J., Jarvela, I., Lawrence, O., Meldrum, C., Bauters, M., Govaerts, K., Vandeleur, L., Van Esch, H. *et al.* (2008) Submicroscopic duplications of the hydroxysteroid dehydrogenase HSD17B10 and the E3 ubiquitin ligase HUWE1 are associated with mental retardation. *Am J Hum Genet*, **82**, 432-443.
52. Lee, J.A., Carvalho, C.M. and Lupski, J.R. (2007) A DNA replication mechanism for generating nonrecurrent rearrangements associated with genomic disorders. *Cell*, **131**, 1235-1247.
53. Inoue, K., Osaka, H., Thurston, V.C., Clarke, J.T., Yoneyama, A., Rosenbarker, L., Bird, T.D., Hodes, M.E., Shaffer, L.G. and Lupski, J.R. (2002) Genomic rearrangements resulting in PLP1 deletion occur by nonhomologous end joining and cause different dysmyelinating phenotypes in males and females. *Am J Hum Genet*, **71**, 838-853.
54. Combes, P., Bonnet-Dupeyron, M.N., Gauthier-Barichard, F., Schiffmann, R., Bertini, E., Rodriguez, D., Armour, J.A., Boespflug-Tanguy, O. and Vaurs-Barriere, C. (2006) PLP1 and GPM6B intragenic copy number analysis by MAPH in 262 patients with hypomyelinating leukodystrophies: Identification of one partial triplication and two partial deletions of PLP1. *Neurogenetics*, **7**, 31-37.
55. Lee, J.A., Inoue, K., Cheung, S.W., Shaw, C.A., Stankiewicz, P. and Lupski, J.R. (2006) Role of genomic architecture in PLP1 duplication causing Pelizaeus-Merzbacher disease. *Hum Mol Genet*, **15**, 2250-2265.
56. Wolf, N.I., Sistermans, E.A., Cundall, M., Hobson, G.M., Davis-Williams, A.P., Palmer, R., Stubbs, P., Davies, S., Endziniene, M., Wu, Y. *et al.* (2005) Three or more copies of the proteolipid protein gene PLP1 cause severe Pelizaeus-Merzbacher disease. *Brain*, **128**, 743-751.
57. Van Esch, H., Bauters, M., Ignatius, J., Jansen, M., Raynaud, M., Hollanders, K., Lugtenberg, D., Bienvenu, T., Jensen, L.R., Gecz, J. *et al.* (2005) Duplication of the MECP2 region is a frequent cause of severe mental retardation and progressive neurological symptoms in males. *Am J Hum Genet*, **77**, 442-453.
58. del Gaudio, D., Fang, P., Scaglia, F., Ward, P.A., Craigen, W.J., Glaze, D.G., Neul, J.L., Patel, A., Lee, J.A., Irons, M. *et al.* (2006) Increased MECP2 gene copy number as the result of genomic duplication in neurodevelopmentally delayed males. *Genet Med*, **8**, 784-792.
59. Bauters, M., Van Esch, H., Friez, M.J., Boespflug-Tanguy, O., Zenker, M., Vianna-Morgante, A.M., Rosenberg, C., Ignatius, J., Raynaud, M., Hollanders, K. *et al.* (2008) Nonrecurrent MECP2 duplications mediated by genomic architecture-driven DNA breaks and break-induced replication repair. *Genome Res*, **18**, 847-858.
60. Nathans, J., Piantanida, T.P., Eddy, R.L., Shows, T.B. and Hogness, D.S. (1986) Molecular genetics of inherited variation in human color vision. *Science*, **232**, 203-210.
61. Rovelet-Lecrux, A., Hannequin, D., Raux, G., Le Meur, N., Laquerriere, A., Vital, A., Dumanchin, C., Feuillette, S., Brice, A., Vercelletto, M. *et al.* (2006) APP locus

duplication causes autosomal dominant early-onset Alzheimer disease with cerebral amyloid angiopathy. *Nat Genet*, **38**, 24-26.
62. Morrow, E.M., Yoo, S.Y., Flavell, S.W., Kim, T.K., Lin, Y., Hill, R.S., Mukaddes, N.M., Balkhy, S., Gascon, G., Hashmi, A. *et al.* (2008) Identifying autism loci and genes by tracing recent shared ancestry. *Science*, **321**, 218-223.
63. Weiss, L.A., Shen, Y., Korn, J.M., Arking, D.E., Miller, D.T., Fossdal, R., Saemundsen, E., Stefansson, H., Ferreira, M.A., Green, T. *et al.* (2008) Association between microdeletion and microduplication at 16p11.2 and autism. *N Engl J Med*, **358**, 667-675.
64. Kumar, R.A., KaraMohamed, S., Sudi, J., Conrad, D.F., Brune, C., Badner, J.A., Gilliam, T.C., Nowak, N.J., Cook, E.H., Jr., Dobyns, W.B. *et al.* (2008) Recurrent 16p11.2 microdeletions in autism. *Hum Mol Genet*, **17**, 628-638.
65. Sebat, J., Lakshmi, B., Malhotra, D., Troge, J., Lese-Martin, C., Walsh, T., Yamrom, B., Yoon, S., Krasnitz, A., Kendall, J. *et al.* (2007) Strong association of de novo copy number mutations with autism. *Science*, **316**, 445-449.
66. Marshall, C.R., Noor, A., Vincent, J.B., Lionel, A.C., Feuk, L., Skaug, J., Shago, M., Moessner, R., Pinto, D., Ren, Y. *et al.* (2008) Structural variation of chromosomes in autism spectrum disorder. *Am J Hum Genet*, **82**, 477-488.
67. Fellermann, K., Stange, D.E., Schaeffeler, E., Schmalzl, H., Wehkamp, J., Bevins, C.L., Reinisch, W., Teml, A., Schwab, M., Lichter, P. *et al.* (2006) A chromosome 8 gene-cluster polymorphism with low human beta-defensin 2 gene copy number predisposes to Crohn disease of the colon. *Am J Hum Genet*, **79**, 439-448.
68. McCarroll, S.A., Huett, A., Kuballa, P., Chilewski, S.D., Landry, A., Goyette, P., Zody, M.C., Hall, J.L., Brant, S.R., Cho, J.H. *et al.* (2008) Deletion polymorphism upstream of IRGM associated with altered IRGM expression and Crohn's disease. *Nat Genet*, **40**, 1107-1112.
69. Kuhn, L., Schramm, D.B., Donninger, S., Meddows-Taylor, S., Coovadia, A.H., Sherman, G.G., Gray, G.E. and Tiemessen, C.T. (2007) African infants' CCL3 gene copies influence perinatal HIV transmission in the absence of maternal nevirapine. *AIDS*, **21**, 1753-1761.
70. Gonzalez, E., Kulkarni, H., Bolivar, H., Mangano, A., Sanchez, R., Catano, G., Nibbs, R.J., Freedman, B.I., Quinones, M.P., Bamshad, M.J. *et al.* (2005) The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility. *Science*, **307**, 1434-1440.
71. Sharp, A.J., Mefford, H.C., Li, K., Baker, C., Skinner, C., Stevenson, R.E., Schroer, R.J., Novara, F., De Gregori, M., Ciccone, R. *et al.* (2008) A recurrent 15q13.3 microdeletion syndrome associated with mental retardation and seizures. *Nat Genet*, **40**, 322-328.
72. Sharp, A.J., Hansen, S., Selzer, R.R., Cheng, Z., Regan, R., Hurst, J.A., Stewart, H., Price, S.M., Blair, E., Hennekam, R.C. *et al.* (2006) Discovery of previously unidentified genomic disorders from the duplication architecture of the human genome. *Nat Genet*, **38**, 1038-1042.
73. Shaw-Smith, C., Pittman, A.M., Willatt, L., Martin, H., Rickman, L., Gribble, S., Curley, R., Cumming, S., Dunn, C., Kalaitzopoulos, D. *et al.* (2006) Microdeletion encompassing MAPT at chromosome 17q21.3 is associated with developmental delay and learning disability. *Nat Genet*, **38**, 1032-1037.
74. Koolen, D.A., Vissers, L.E., Pfundt, R., de Leeuw, N., Knight, S.J., Regan, R., Kooy, R.F., Reyniers, E., Romano, C., Fichera, M. *et al.* (2006) A new chromosome 17q21.31 microdeletion syndrome associated with a common inversion polymorphism. *Nat Genet*, **38**, 999-1001.
75. Le Marechal, C., Masson, E., Chen, J.M., Morel, F., Ruszniewski, P., Levy, P. and Ferec, C. (2006) Hereditary pancreatitis caused by triplication of the trypsinogen locus. *Nat Genet*, **38**, 1372-1374.

76. Chartier-Harlin, M.C., Kachergus, J., Roumier, C., Mouroux, V., Douay, X., Lincoln, S., Levecque, C., Larvor, L., Andrieux, J., Hulihan, M. *et al.* (2004) Alpha-synuclein locus duplication as a cause of familial Parkinson's disease. *Lancet*, **364**, 1167-1169.
77. Singleton, A.B., Farrer, M., Johnson, J., Singleton, A., Hague, S., Kachergus, J., Hulihan, M., Peuralinna, T., Dutra, A., Nussbaum, R. *et al.* (2003) alpha-Synuclein locus triplication causes Parkinson's disease. *Science*, **302**, 841.
78. Ibanez, P., Bonnet, A.M., Debarges, B., Lohmann, E., Tison, F., Pollak, P., Agid, Y., Durr, A. and Brice, A. (2004) Causal relation between alpha-synuclein gene duplication and familial Parkinson's disease. *Lancet*, **364**, 1169-1171.
79. Farrer, M., Kachergus, J., Forno, L., Lincoln, S., Wang, D.S., Hulihan, M., Maraganore, D., Gwinn-Hardy, K., Wszolek, Z., Dickson, D. *et al.* (2004) Comparison of kindreds with parkinsonism and alpha-synuclein genomic multiplications. *Ann Neurol*, **55**, 174-179.
80. Fuchs, J., Nilsson, C., Kachergus, J., Munz, M., Larsson, E.M., Schule, B., Langston, J.W., Middleton, F.A., Ross, O.A., Hulihan, M. *et al.* (2007) Phenotypic variation in a large Swedish pedigree due to SNCA duplication and triplication. *Neurology*, **68**, 916-922.
81. Hollox, E.J., Huffmeier, U., Zeeuwen, P.L., Palla, R., Lascorz, J., Rodijk-Olthuis, D., van de Kerkhof, P.C., Traupe, H., de Jongh, G., den Heijer, M. *et al.* (2008) Psoriasis is associated with increased beta-defensin genomic copy number. *Nat Genet*, **40**, 23-25.
82. Stefansson, H., Rujescu, D., Cichon, S., Pietilainen, O.P., Ingason, A., Steinberg, S., Fossdal, R., Sigurdsson, E., Sigmundsson, T., Buizer-Voskamp, J.E. *et al.* (2008) Large recurrent microdeletions associated with schizophrenia. *Nature*, **455**, 232-236.
83. International Schizophrenia Consortium. (2008) Rare chromosomal deletions and duplications increase risk of schizophrenia. *Nature*, **455**, 237-241.
84. Walsh, T., McClellan, J.M., McCarthy, S.E., Addington, A.M., Pierce, S.B., Cooper, G.M., Nord, A.S., Kusenda, M., Malhotra, D., Bhandari, A. *et al.* (2008) Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia. *Science*, **320**, 539-543.
85. Willcocks, L.C., Lyons, P.A., Clatworthy, M.R., Robinson, J.I., Yang, W., Newland, S.A., Plagnol, V., McGovern, N.N., Condliffe, A.M., Chilvers, E.R. *et al.* (2008) Copy number of FCGR3B, which is associated with systemic lupus erythematosus, correlates with protein expression and immune complex uptake. *J Exp Med*, **205**, 1573-1582.
86. Aitman, T.J., Dong, R., Vyse, T.J., Norsworthy, P.J., Johnson, M.D., Smith, J., Mangion, J., Roberton-Lowe, C., Marshall, A.J., Petretto, E. *et al.* (2006) Copy number polymorphism in Fcgr3 predisposes to glomerulonephritis in rats and humans. *Nature*, **439**, 851-855.
87. Molokhia, M., Fanciulli, M., Petretto, E., Patrick, A.L., McKeigue, P., Roberts, A.L., Vyse, T.J. and Aitman, T.J. (2011) FCGR3B copy number variation is associated with systemic lupus erythematosus risk in Afro-Caribbeans. *Rheumatology (Oxford)*, **50**, 1206-1210.
88. Yang, Y., Chung, E.K., Wu, Y.L., Savelli, S.L., Nagaraja, H.N., Zhou, B., Hebert, M., Jones, K.N., Shu, Y., Kitzmiller, K. *et al.* (2007) Gene copy-number variation and associated polymorphisms of complement component C4 in human systemic lupus erythematosus (SLE): low copy number is a risk factor for and high copy number is a protective factor against SLE susceptibility in European Americans. *Am J Hum Genet*, **80**, 1037-1054.
89. McCarroll, S.A. (2008) Extending genome-wide association studies to copy-number variation. *Hum Mol Genet*, **17**, R135-142.

90. Casey, J.P., Magalhaes, T., Conroy, J.M., Regan, R., Shah, N., Anney, R., Shields, D.C., Abrahams, B.S., Almeida, J., Bacchelli, E. *et al.* (2012) A novel approach of homozygous haplotype sharing identifies candidate genes in autism spectrum disorder. *Hum Genet*, **131**, 565-579.
91. Lee, K.W., Woon, P.S., Teo, Y.Y. and Sim, K. (2012) Genome wide association studies (GWAS) and copy number variation (CNV) studies of the major psychoses: what have we learnt? *Neurosci Biobehav Rev*, **36**, 556-571.
92. Dauber, A., Yu, Y., Turchin, M.C., Chiang, C.W., Meng, Y.A., Demerath, E.W., Patel, S.R., Rich, S.S., Rotter, J.I., Schreiner, P.J. *et al.* (2011) Genome-wide association of copy-number variation reveals an association between short stature and the presence of low-frequency genomic deletions. *Am J Hum Genet*, **89**, 751-759.
93. Moon, S., Keam, B., Hwang, M.Y., Lee, Y., Park, S., Oh, J.H., Kim, Y.J., Lee, H.S., Kim, N.H., Kim, Y.J. *et al.* (2015) A genome-wide association study of copy-number variation identifies putative loci associated with osteoarthritis in Koreans. *BMC Musculoskelet Disord*, **16**, 76.
94. Cooper, G.M., Coe, B.P., Girirajan, S., Rosenfeld, J.A., Vu, T.H., Baker, C., Williams, C., Stalker, H., Hamid, R., Hannig, V. *et al.* (2011) A copy number variation morbidity map of developmental delay. *Nat Genet*, **43**, 838-846.
95. Stankiewicz, P. and Beaudet, A.L. (2007) Use of array CGH in the evaluation of dysmorphology, malformations, developmental delay, and idiopathic mental retardation. *Curr Opin Genet Dev*, **17**, 182-192.
96. Lupski, J.R. (2012) Brain copy number variants and neuropsychiatric traits. *Biol Psychiatry*, **72**, 617-619.
97. Shlien, A. and Malkin, D. (2010) Copy number variations and cancer susceptibility. *Curr Opin Oncol*, **22**, 55-63.
98. Kuiper, R.P., Ligtenberg, M.J., Hoogerbrugge, N. and Geurts van Kessel, A. (2010) Germline copy number variation and cancer risk. *Curr Opin Genet Dev*, **20**, 282-289.
99. Krepischi, A.C., Pearson, P.L. and Rosenberg, C. (2012) Germline copy number variations and cancer predisposition. *Future Oncol*, **8**, 441-450.
100. Stadler, Z.K., Vijai, J., Thom, P., Kirchhoff, T., Hansen, N.A., Kauff, N.D., Robson, M. and Offit, K. (2010) Genome-wide association studies of cancer predisposition. *Hematol Oncol Clin North Am*, **24**, 973-996.
101. Montagna, M., Dalla Palma, M., Menin, C., Agata, S., De Nicolo, A., Chieco-Bianchi, L. and D'Andrea, E. (2003) Genomic rearrangements account for more than one-third of the BRCA1 mutations in northern Italian breast/ovarian cancer families. *Hum Mol Genet*, **12**, 1055-1061.
102. Petrij-Bosch, A., Peelen, T., van Vliet, M., van Eijk, R., Olmer, R., Drusedau, M., Hogervorst, F.B., Hageman, S., Arts, P.J., Ligtenberg, M.J. *et al.* (1997) BRCA1 genomic deletions are major founder mutations in Dutch breast cancer patients. *Nat Genet*, **17**, 341-345.
103. Casilli, F., Tournier, I., Sinilnikova, O.M., Coulet, F., Soubrier, F., Houdayer, C., Hardouin, A., Berthet, P., Sobol, H., Bourdon, V. *et al.* (2006) The contribution of germline rearrangements to the spectrum of BRCA2 mutations. *J Med Genet*, **43**, e49.
104. Bremner, R., Du, D.C., Connolly-Wilson, M.J., Bridge, P., Ahmad, K.F., Mostachfi, H., Rushlow, D., Dunn, J.M. and Gallie, B.L. (1997) Deletion of RB exons 24 and 25 causes low-penetrance retinoblastoma. *Am J Hum Genet*, **61**, 556-570.
105. Michils, G., Tejpar, S., Thoelen, R., van Cutsem, E., Vermeesch, J.R., Fryns, J.P., Legius, E. and Matthijs, G. (2005) Large deletions of the APC gene in 15% of

mutation-negative patients with classical polyposis (FAP): a Belgian study. *Hum Mutat*, **25**, 125-134.
106. Tang, Y.C. and Amon, A. (2013) Gene copy-number alterations: a cost-benefit analysis. *Cell*, **152**, 394-405.
107. Zhao, X., Li, C., Paez, J.G., Chin, K., Janne, P.A., Chen, T.H., Girard, L., Minna, J., Christiani, D., Leo, C. *et al.* (2004) An integrated view of copy number and allelic alterations in the cancer genome using single nucleotide polymorphism arrays. *Cancer Res*, **64**, 3060-3071.
108. Little, C.D., Nau, M.M., Carney, D.N., Gazdar, A.F. and Minna, J.D. (1983) Amplification and expression of the c-myc oncogene in human lung cancer cell lines. *Nature*, **306**, 194-196.
109. Li, J., Yen, C., Liaw, D., Podsypanina, K., Bose, S., Wang, S.I., Puc, J., Miliaresis, C., Rodgers, L., McCombie, R. *et al.* (1997) PTEN, a putative protein tyrosine phosphatase gene mutated in human brain, breast, and prostate cancer. *Science*, **275**, 1943-1947.
110. Forbes, S.A., Beare, D., Gunasekaran, P., Leung, K., Bindal, N., Boutselakis, H., Ding, M., Bamford, S., Cole, C., Ward, S. *et al.* (2015) COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Res*, **43**, D805-811.
111. Baudis, M. (2007) Genomic imbalances in 5918 malignant epithelial tumors: an explorative meta-analysis of chromosomal CGH data. *BMC Cancer*, **7**, 226.
112. Diskin, S.J., Eck, T., Greshock, J., Mosse, Y.P., Naylor, T., Stoeckert, C.J., Jr., Weber, B.L., Maris, J.M. and Grant, G.R. (2006) STAC: A method for testing the significance of DNA copy number aberrations across multiple array-CGH experiments. *Genome Res*, **16**, 1149-1158.
113. Zhang, J., Zhang, S., Wang, Y. and Zhang, X.S. (2013) Identification of mutated core cancer modules by integrating somatic mutation, copy number variation, and gene expression data. *BMC Syst Biol*, **7 Suppl 2**, S4.
114. Bunyan, D.J., Eccles, D.M., Sillibourne, J., Wilkins, E., Thomas, N.S., Shea-Simonds, J., Duncan, P.J., Curtis, C.E., Robinson, D.O., Harvey, J.F. *et al.* (2004) Dosage analysis of cancer predisposition genes by multiplex ligation-dependent probe amplification. *Br J Cancer*, **91**, 1155-1159.
115. Aretz, S., Stienen, D., Uhlhaas, S., Stolte, M., Entius, M.M., Loff, S., Back, W., Kaufmann, A., Keller, K.M., Blaas, S.H. *et al.* (2007) High proportion of large genomic deletions and a genotype phenotype update in 80 unrelated families with juvenile polyposis syndrome. *J Med Genet*, **44**, 702-709.
116. Casilli, F., Di Rocco, Z.C., Gad, S., Tournier, I., Stoppa-Lyonnet, D., Frebourg, T. and Tosi, M. (2002) Rapid detection of novel BRCA1 rearrangements in high-risk breast-ovarian cancer families using multiplex PCR of short fluorescent fragments. *Hum Mutat*, **20**, 218-226.
117. Oliveira, C., Senz, J., Kaurah, P., Pinheiro, H., Sanges, R., Haegert, A., Corso, G., Schouten, J., Fitzgerald, R., Vogelsang, H. *et al.* (2009) Germline CDH1 deletions in hereditary diffuse gastric cancer families. *Hum Mol Genet*, **18**, 1545-1555.
118. Pichert, G., Mohammed, S.N., Ahn, J.W., Ogilvie, C.M. and Izatt, L. (2011) Unexpected findings in cancer predisposition genes detected by array comparative genomic hybridisation: what are the issues? *J Med Genet*, **48**, 535-539.
119. Lesueur, F., de Lichy, M., Barrois, M., Durand, G., Bombled, J., Avril, M.F., Chompret, A., Boitier, F., Lenoir, G.M., French Familial Melanoma Study, G. *et al.* (2008) The contribution of large genomic deletions at the CDKN2A locus to the burden of familial melanoma. *Br J Cancer*, **99**, 364-370.
120. Cybulski, C., Wokolorczyk, D., Huzarski, T., Byrski, T., Gronwald, J., Gorski, B., Debniak, T., Masojc, B., Jakubowska, A., Gliniewicz, B. *et al.* (2006) A large

germline deletion in the Chek2 kinase gene is associated with an increased risk of prostate cancer. *J Med Genet*, **43**, 863-866.
121. Cybulski, C., Wokolorczyk, D., Huzarski, T., Byrski, T., Gronwald, J., Gorski, B., Debniak, T., Masojc, B., Jakubowska, A., van de Wetering, T. *et al.* (2007) A deletion in CHEK2 of 5,395 bp predisposes to breast cancer in Poland. *Breast Cancer Res Treat*, **102**, 119-122.
122. Breuning, M.H., Dauwerse, H.G., Fugazza, G., Saris, J.J., Spruit, L., Wijnen, H., Tommerup, N., van der Hagen, C.B., Imaizumi, K., Kuroki, Y. *et al.* (1993) Rubinstein-Taybi syndrome caused by submicroscopic deletions within 16p13.3. *Am J Hum Genet*, **52**, 249-254.
123. Jenner, M.W., Leone, P.E., Walker, B.A., Ross, F.M., Johnson, D.C., Gonzalez, D., Chiecchio, L., Dachs Cabanas, E., Dagrada, G.P., Nightingale, M. *et al.* (2007) Gene mapping and expression analysis of 16q loss of heterozygosity identifies WWOX and CYLD as being important in determining clinical outcome in multiple myeloma. *Blood*, **110**, 3291-3300.
124. Ligtenberg, M.J., Kuiper, R.P., Chan, T.L., Goossens, M., Hebeda, K.M., Voorendt, M., Lee, T.Y., Bodmer, D., Hoenselaar, E., Hendriks-Cornelissen, S.J. *et al.* (2009) Heritable somatic methylation and inactivation of MSH2 in families with Lynch syndrome due to deletion of the 3' exons of TACSTD1. *Nat Genet*, **41**, 112-117.
125. Vink, G.R., White, S.J., Gabelic, S., Hogendoorn, P.C., Breuning, M.H. and Bakker, E. (2005) Mutation screening of EXT1 and EXT2 by direct sequence analysis and MLPA in patients with multiple osteochondromas: splice site mutations and exonic deletions account for more than half of the mutations. *Eur J Hum Genet*, **13**, 470-474.
126. Levran, O., Diotti, R., Pujara, K., Batish, S.D., Hanenberg, H. and Auerbach, A.D. (2005) Spectrum of sequence variations in the FANCA gene: an International Fanconi Anemia Registry (IFAR) study. *Hum Mutat*, **25**, 142-149.
127. Ahvenainen, T., Lehtonen, H.J., Lehtonen, R., Vahteristo, P., Aittomaki, K., Baynam, G., Dommering, C., Eng, C., Gruber, S.B., Gronberg, H. *et al.* (2008) Mutation screening of fumarate hydratase by multiplex ligation-dependent probe amplification: detection of exonic deletion in a patient with leiomyomatosis and renal cell cancer. *Cancer Genet Cytogenet*, **183**, 83-88.
128. Caron, P., Simonds, W.F., Maiza, J.C., Rubin, M., Cantor, T., Rousseau, L., Bilezikian, J.P., Souberbielle, J.C. and D'Amour, P. (2011) Nontruncated amino-terminal parathyroid hormone overproduction in two patients with parathyroid carcinoma: a possible link to HRPT2 gene inactivation. *Clin Endocrinol (Oxf)*, **74**, 694-698.
129. Laufer-Cahana, A., Krantz, I.D., Bason, L.D., Lu, F.M., Piccoli, D.A. and Spinner, N.B. (2002) Alagille syndrome inherited from a phenotypically normal mother with a mosaic 20p microdeletion. *Am J Med Genet*, **112**, 190-193.
130. Krantz, I.D., Piccoli, D.A. and Spinner, N.B. (1999) Clinical and molecular genetics of Alagille syndrome. *Curr Opin Pediatr*, **11**, 558-564.
131. van Hattem, W.A., Brosens, L.A., de Leng, W.W., Morsink, F.H., Lens, S., Carvalho, R., Giardiello, F.M. and Offerhaus, G.J. (2008) Large genomic deletions of SMAD4, BMPR1A and PTEN in juvenile polyposis. *Gut*, **57**, 623-627.
132. Kikuchi, M., Ohkura, N., Yamaguchi, K., Obara, T. and Tsukada, T. (2004) Gene dose mapping delineated boundaries of a large germline deletion responsible for multiple endocrine neoplasia type 1. *Cancer Lett*, **208**, 81-88.
133. Plaschke, J., Ruschoff, J. and Schackert, H.K. (2003) Genomic rearrangements of hMSH6 contribute to the genetic predisposition in suspected hereditary non-polyposis colorectal cancer syndrome. *J Med Genet*, **40**, 597-600.

134. Bentivegna, A., Venturin, M., Gervasini, C., Corrado, L., Larizza, L. and Riva, P. (2001) FISH with locus-specific probes on stretched chromosomes: a useful tool for genome organization studies. *Chromosome Res*, **9**, 167-170.
135. Tsilchorozidou, T., Menko, F.H., Lalloo, F., Kidd, A., De Silva, R., Thomas, H., Smith, P., Malcolmson, A., Dore, J., Madan, K. *et al.* (2004) Constitutional rearrangements of chromosome 22 as a cause of neurofibromatosis 2. *J Med Genet*, **41**, 529-534.
136. Overbeek, L.I., Kets, C.M., Hebeda, K.M., Bodmer, D., van der Looij, E., Willems, R., Goossens, M., Arts, N., Brunner, H.G., van Krieken, J.H. *et al.* (2007) Patients with an unexplained microsatellite instable tumour have a low risk of familial cancer. *Br J Cancer*, **96**, 1605-1612.
137. Horvath, A., Bossis, I., Giatzakis, C., Levine, E., Weinberg, F., Meoli, E., Robinson-White, A., Siegel, J., Soni, P., Groussin, L. *et al.* (2008) Large deletions of the PRKAR1A gene in Carney complex. *Clin Cancer Res*, **14**, 388-395.
138. Shimkets, R., Gailani, M.R., Siu, V.M., Yang-Feng, T., Pressman, C.L., Levanat, S., Goldstein, A., Dean, M. and Bale, A.E. (1996) Molecular analysis of chromosome 9q deletions in two Gorlin syndrome patients. *Am J Hum Genet*, **59**, 417-422.
139. Arch, E.M., Goodman, B.K., Van Wesep, R.A., Liaw, D., Clarke, K., Parsons, R., McKusick, V.A. and Geraghty, M.T. (1997) Deletion of PTEN in a patient with Bannayan-Riley-Ruvalcaba syndrome suggests allelism with Cowden disease. *Am J Med Genet*, **71**, 489-493.
140. Francke, U. (1976) Retinoblastoma and chromosome 13. *Birth Defects Orig Artic Ser*, **12**, 131-134.
141. Preudhomme, C., Renneville, A., Bourdon, V., Philippe, N., Roche-Lestienne, C., Boissel, N., Dhedin, N., Andre, J.M., Cornillet-Lefebvre, P., Baruchel, A. *et al.* (2009) High frequency of RUNX1 biallelic alteration in acute myeloid leukemia secondary to familial platelet disorder. *Blood*, **113**, 5583-5587.
142. Cascon, A., Montero-Conde, C., Ruiz-Llorente, S., Mercadillo, F., Leton, R., Rodriguez-Antona, C., Martinez-Delgado, B., Delgado, M., Diez, A., Rovira, A. *et al.* (2006) Gross SDHB deletions in patients with paraganglioma detected by multiplex PCR: a possible hot spot? *Genes Chromosomes Cancer*, **45**, 213-219.
143. Bayley, J.P., Weiss, M.M., Grimbergen, A., van Brussel, B.T., Hes, F.J., Jansen, J.C., Verhoef, S., Devilee, P., Corssmit, E.P. and Vriends, A.H. (2009) Molecular characterization of novel germline deletions affecting SDHD and SDHC in pheochromocytoma and paraganglioma patients. *Endocr Relat Cancer*, **16**, 929-937.
144. Swensen, J.J., Keyser, J., Coffin, C.M., Biegel, J.A., Viskochil, D.H. and Williams, M.S. (2009) Familial occurrence of schwannomas and malignant rhabdoid tumour associated with a duplication in SMARCB1. *J Med Genet*, **46**, 68-72.
145. Aretz, S., Stienen, D., Uhlhaas, S., Loff, S., Back, W., Pagenstecher, C., McLeod, D.R., Graham, G.E., Mangold, E., Santer, R. *et al.* (2005) High proportion of large genomic STK11 deletions in Peutz-Jeghers syndrome. *Hum Mutat*, **26**, 513-519.
146. Bougeard, G., Brugieres, L., Chompret, A., Gesta, P., Charbonnier, F., Valent, A., Martin, C., Raux, G., Feunteun, J., Bressac-de Paillerets, B. *et al.* (2003) Screening for TP53 rearrangements in families with the Li-Fraumeni syndrome reveals a complete deletion of the TP53 gene. *Oncogene*, **22**, 840-846.
147. Kozlowski, P., Roberts, P., Dabora, S., Franz, D., Bissler, J., Northrup, H., Au, K.S., Lazarus, R., Domanska-Pakiela, D., Kotulska, K. *et al.* (2007) Identification of 54 large deletions/duplications in TSC1 and TSC2 using MLPA, and genotype-phenotype correlations. *Hum Genet*, **121**, 389-400.
148. Richards, F.M., Phipps, M.E., Latif, F., Yao, M., Crossey, P.A., Foster, K., Linehan, W.M., Affara, N.A., Lerman, M.I., Zbar, B. *et al.* (1993) Mapping the

Von Hippel-Lindau disease tumour suppressor gene: identification of germline deletions by pulsed field gel electrophoresis. *Hum Mol Genet*, **2**, 879-882.
149. Hittner, H.M., Riccardi, V.M. and Francke, U. (1979) Aniridia caused by a heritable chromosome 11 deletion. *Ophthalmology*, **86**, 1173-1183.
150. Wan, T.S. (2014) Cancer cytogenetics: methodology revisited. *Ann Lab Med*, **34**, 413-425.
151. Wang, N. (2002) Methodologies in cancer cytogenetics and molecular cytogenetics. *Am J Med Genet*, **115**, 118-124.
152. O'Connor, C. (2008) Karyotyping for chromosomal abnormalities. *Nature Education*, **1**, 27.
153. O'connor, C. (2008) Fluorescence in situ hybridization (FISH). *Nature Education*, **1**, 171.
154. Weiss, M.M., Hermsen, M.A., Meijer, G.A., van Grieken, N.C., Baak, J.P., Kuipers, E.J. and van Diest, P.J. (1999) Comparative genomic hybridisation. *Mol Pathol*, **52**, 243-251.
155. Oostlander, A.E., Meijer, G.A. and Ylstra, B. (2004) Microarray-based comparative genomic hybridization and its applications in human genetics. *Clin Genet*, **66**, 488-495.
156. Pinkel, D., Segraves, R., Sudar, D., Clark, S., Poole, I., Kowbel, D., Collins, C., Kuo, W.L., Chen, C., Zhai, Y. *et al.* (1998) High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nat Genet*, **20**, 207-211.
157. Ryland, G.L., Doyle, M.A., Goode, D., Boyle, S.E., Choong, D.Y., Rowley, S.M., Li, J., Australian Ovarian Cancer Study, G., Bowtell, D.D., Tothill, R.W. *et al.* (2015) Loss of heterozygosity: what is it good for? *BMC Med Genomics*, **8**, 45.
158. Alkan, C., Coe, B.P. and Eichler, E.E. (2011) Genome structural variation discovery and genotyping. *Nat Rev Genet*, **12**, 363-376.
159. Li, Y., Zhang, L., Ball, R.L., Liang, X., Li, J., Lin, Z. and Liang, H. (2012) Comparative analysis of somatic copy-number alterations across different human cancer types reveals two distinct classes of breakpoint hotspots. *Hum Mol Genet*, **21**, 4957-4965.
160. De, S. and Michor, F. (2011) DNA secondary structures and epigenetic determinants of cancer genome evolution. *Nat Struct Mol Biol*, **18**, 950-955.
161. Pinto, D., Darvishi, K., Shi, X., Rajan, D., Rigler, D., Fitzgerald, T., Lionel, A.C., Thiruvahindrapuram, B., Macdonald, J.R., Mills, R. *et al.* (2011) Comprehensive assessment of array-based platforms and calling algorithms for detection of copy number variants. *Nat Biotechnol*, **29**, 512-520.
162. Zhao, M., Wang, Q., Wang, Q., Jia, P. and Zhao, Z. (2013) Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives. *BMC Bioinformatics*, **14 Suppl 11**, S1.
163. Henrichsen, C.N., Chaignat, E. and Reymond, A. (2009) Copy number variants, diseases and gene expression. *Hum Mol Genet*, **18**, R1-8.
164. Collins, A. (2015) The genomic and functional characteristics of disease genes. *Brief Bioinform*, **16**, 16-23.
165. Marcus, F.B. (2008) *Bioinformatics and Systems Biology Collaborative Research and Resources*. Springer-Verlag Berlin Heidelberg,, Berlin, Heidelberg.
166. Ideker, T., Galitski, T. and Hood, L. (2001) A new approach to decoding life: systems biology. *Annu Rev Genomics Hum Genet*, **2**, 343-372.
167. D'Antonio, M. and Ciccarelli, F.D. (2011) Modification of gene duplicability during the evolution of protein interaction network. *PLoS Comput Biol*, **7**, e1002029.

168. 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.
169. 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**, 859-867.
170. 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**, 5454-5459.
171. Zhang, J. Evolution by gene duplication: an update. *Trends in Ecology & Evolution*, **18**, 292-298.
172. Conrad, B. and Antonarakis, S.E. (2007) Gene duplication: a drive for phenotypic diversity and cause of human disease. *Annu Rev Genomics Hum Genet*, **8**, 17-35.
173. 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**, 427-430.
174. Kent, W.J. (2002) BLAT--the BLAST-like alignment tool. *Genome Res*, **12**, 656-664.
175. An, O., Pendino, V., D'Antonio, M., Ratti, E., Gentilini, M. and Ciccarelli, F.D. (2014) NCG 4.0: the network of cancer genes in the era of massive mutational screenings of cancer genomes. *Database (Oxford)*, **2014**, bau015.
176. Tatusov, R.L., Koonin, E.V. and Lipman, D.J. (1997) A genomic perspective on protein families. *Science*, **278**, 631-637.
177. Fang, G., Bhardwaj, N., Robilotto, R. and Gerstein, M.B. (2010) Getting started in gene orthology and functional analysis. *PLoS Comput Biol*, **6**, e1000703.
178. Altenhoff, A.M., Schneider, A., Gonnet, G.H. and Dessimoz, C. (2011) OMA 2011: orthology inference among 1000 complete genomes. *Nucleic Acids Res*, **39**, D289-294.
179. Waterhouse, R.M., Zdobnov, E.M., Tegenfeldt, F., Li, J. and Kriventseva, E.V. (2011) OrthoDB: the hierarchical catalog of eukaryotic orthologs in 2011. *Nucleic Acids Res*, **39**, D283-288.
180. 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.
181. 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**, D250-254.
182. Chen, F., Mackey, A.J., Stoeckert, C.J., Jr. and Roos, D.S. (2006) OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. *Nucleic Acids Res*, **34**, D363-368.
183. 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.
184. Schreiber, F., Patricio, M., Muffato, M., Pignatelli, M. and Bateman, A. (2014) TreeFam v9: a new website, more species and orthology-on-the-fly. *Nucleic Acids Res*, **42**, D922-925.
185. Vilella, A.J., Severin, J., Ureta-Vidal, A., Heng, L., Durbin, R. and Birney, E. (2009) EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res*, **19**, 327-335.
186. Wapinski, I., Pfeffer, A., Friedman, N. and Regev, A. (2007) Automatic genome-wide reconstruction of phylogenetic gene trees. *Bioinformatics*, **23**, i549-558.

187. van der Heijden, R.T., Snel, B., van Noort, V. and Huynen, M.A. (2007) Orthology prediction at scalable resolution by phylogenetic tree analysis. *BMC Bioinformatics*, **8**, 83.
188. D'Antonio, M., Guerra, R.F., Cereda, M., Marchesi, S., Montani, F., Nicassio, F., Di Fiore, P.P. and Ciccarelli, F.D. (2013) Recessive cancer genes engage in negative genetic interactions with their functional paralogs. *Cell Rep*, **5**, 1519-1526.
189. Powell, S., Forslund, K., Szklarczyk, D., Trachana, K., Roth, A., Huerta-Cepas, J., Gabaldon, T., Rattei, T., Creevey, C., Kuhn, M. *et al.* (2014) eggNOG v4.0: nested orthology inference across 3686 organisms. *Nucleic Acids Res*, **42**, D231-239.
190. Barabasi, A.L. and Oltvai, Z.N. (2004) Network biology: understanding the cell's functional organization. *Nat Rev Genet*, **5**, 101-113.
191. Zhu, X., Gerstein, M. and Snyder, M. (2007) Getting connected: analysis and principles of biological networks. *Genes Dev*, **21**, 1010-1024.
192. Crick, F. (1970) Central dogma of molecular biology. *Nature*, **227**, 561-563.
193. Kuribayashi, K., Krigsfeld, G., Wang, W., Xu, J., Mayes, P.A., Dicker, D.T., Wu, G.S. and El-Deiry, W.S. (2008) TNFSF10 (TRAIL), a p53 target gene that mediates p53-dependent cell death. *Cancer Biol Ther*, **7**, 2034-2038.
194. Nicholson, R.I., Gee, J.M. and Harper, M.E. (2001) EGFR and cancer prognosis. *Eur J Cancer*, **37 Suppl 4**, S9-15.
195. Pickrell, J.K., Marioni, J.C., Pai, A.A., Degner, J.F., Engelhardt, B.E., Nkadori, E., Veyrieras, J.B., Stephens, M., Gilad, Y. and Pritchard, J.K. (2010) Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature*, **464**, 768-772.
196. Stranger, B.E., Forrest, M.S., Dunning, M., Ingle, C.E., Beazley, C., Thorne, N., Redon, R., Bird, C.P., de Grassi, A., Lee, C. *et al.* (2007) Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science*, **315**, 848-853.
197. Rubio-Perez, C., Tamborero, D., Schroeder, M.P., Antolin, A.A., Deu-Pons, J., Perez-Llamas, C., Mestres, J., Gonzalez-Perez, A. and Lopez-Bigas, N. (2015) In silico prescription of anticancer drugs to cohorts of 28 tumor types reveals targeting opportunities. *Cancer Cell*, **27**, 382-396.
198. Luo, J., Solimini, N.L. and Elledge, S.J. (2009) Principles of cancer therapy: oncogene and non-oncogene addiction. *Cell*, **136**, 823-837.
199. Luo, J., Emanuele, M.J., Li, D., Creighton, C.J., Schlabach, M.R., Westbrook, T.F., Wong, K.K. and Elledge, S.J. (2009) A genome-wide RNAi screen identifies multiple synthetic lethal interactions with the Ras oncogene. *Cell*, **137**, 835-848.
200. Steckel, M., Molina-Arcas, M., Weigelt, B., Marani, M., Warne, P.H., Kuznetsov, H., Kelly, G., Saunders, B., Howell, M., Downward, J. *et al.* (2012) Determination of synthetic lethal interactions in KRAS oncogene-dependent cancer cells reveals novel therapeutic targeting strategies. *Cell Res*, **22**, 1227-1245.
201. Mair, B., Kubicek, S. and Nijman, S.M. (2014) Exploiting epigenetic vulnerabilities for cancer therapeutics. *Trends Pharmacol Sci*, **35**, 136-145.
202. Krawczyk, P.M., Eppink, B., Essers, J., Stap, J., Rodermond, H., Odijk, H., Zelensky, A., van Bree, C., Stalpers, L.J., Buist, M.R. *et al.* (2011) Mild hyperthermia inhibits homologous recombination, induces BRCA2 degradation, and sensitizes cancer cells to poly (ADP-ribose) polymerase-1 inhibition. *Proc Natl Acad Sci U S A*, **108**, 9851-9856.
203. Chan, N., Pires, I.M., Bencokova, Z., Coackley, C., Luoto, K.R., Bhogal, N., Lakshman, M., Gottipati, P., Oliver, F.J., Helleday, T. *et al.* (2010) Contextual synthetic lethality of cancer cell kill based on the tumor microenvironment. *Cancer Res*, **70**, 8045-8054.
204. Boone, C., Bussey, H. and Andrews, B.J. (2007) Exploring genetic interactions and networks with yeast. *Nat Rev Genet*, **8**, 437-449.

205. Giaever, G., Chu, A.M., Ni, L., Connelly, C., Riles, L., Veronneau, S., Dow, S., Lucau-Danila, A., Anderson, K., Andre, B. *et al.* (2002) Functional profiling of the Saccharomyces cerevisiae genome. *Nature*, **418**, 387-391.
206. Tong, A.H. and Boone, C. (2006) Synthetic genetic array analysis in Saccharomyces cerevisiae. *Methods Mol Biol*, **313**, 171-192.
207. Tong, A.H., Evangelista, M., Parsons, A.B., Xu, H., Bader, G.D., Page, N., Robinson, M., Raghibizadeh, S., Hogue, C.W., Bussey, H. *et al.* (2001) Systematic genetic analysis with ordered arrays of yeast deletion mutants. *Science*, **294**, 2364-2368.
208. Ooi, S.L., Shoemaker, D.D. and Boeke, J.D. (2003) DNA helicase gene interaction network defined using synthetic lethality analyzed by microarray. *Nat Genet*, **35**, 277-286.
209. Pan, X., Yuan, D.S., Xiang, D., Wang, X., Sookhai-Mahadeo, S., Bader, J.S., Hieter, P., Spencer, F. and Boeke, J.D. (2004) A robust toolkit for functional profiling of the yeast genome. *Mol Cell*, **16**, 487-496.
210. Collins, S.R., Schuldiner, M., Krogan, N.J. and Weissman, J.S. (2006) A strategy for extracting and analyzing large-scale quantitative epistatic interaction data. *Genome Biol*, **7**, R63.
211. Collins, S.R., Miller, K.M., Maas, N.L., Roguev, A., Fillingham, J., Chu, C.S., Schuldiner, M., Gebbia, M., Recht, J., Shales, M. *et al.* (2007) Functional dissection of protein complexes involved in yeast chromosome biology using a genetic interaction map. *Nature*, **446**, 806-810.
212. Schuldiner, M., Collins, S.R., Thompson, N.J., Denic, V., Bhamidipati, A., Punna, T., Ihmels, J., Andrews, B., Boone, C., Greenblatt, J.F. *et al.* (2005) Exploration of the function and organization of the yeast early secretory pathway through an epistatic miniarray profile. *Cell*, **123**, 507-519.
213. Pandey, G., Zhang, B., Chang, A.N., Myers, C.L., Zhu, J., Kumar, V. and Schadt, E.E. (2010) An integrative multi-network and multi-classifier approach to predict genetic interactions. *PLoS Comput Biol*, **6**.
214. Li, B., Cao, W., Zhou, J. and Luo, F. (2011) Understanding and predicting synthetic lethal genetic interactions in Saccharomyces cerevisiae using domain genetic interactions. *BMC Syst Biol*, **5**, 73.
215. Qi, Y., Suhail, Y., Lin, Y.Y., Boeke, J.D. and Bader, J.S. (2008) Finding friends and enemies in an enemies-only network: a graph diffusion kernel for predicting novel genetic interactions and co-complex membership from yeast genetic interactions. *Genome Res*, **18**, 1991-2004.
216. Paladugu, S.R., Zhao, S., Ray, A. and Raval, A. (2008) Mining protein networks for synthetic genetic interactions. *BMC Bioinformatics*, **9**, 426.
217. Szappanos, B., Kovacs, K., Szamecz, B., Honti, F., Costanzo, M., Baryshnikova, A., Gelius-Dietrich, G., Lercher, M.J., Jelasity, M., Myers, C.L. *et al.* (2011) An integrated approach to characterize genetic interaction networks in yeast metabolism. *Nat Genet*, **43**, 656-662.
218. Wong, S.L., Zhang, L.V., Tong, A.H., Li, Z., Goldberg, D.S., King, O.D., Lesage, G., Vidal, M., Andrews, B., Bussey, H. *et al.* (2004) Combining biological networks to predict genetic interactions. *Proc Natl Acad Sci U S A*, **101**, 15682-15687.
219. Kelley, R. and Ideker, T. (2005) Systematic interpretation of genetic interactions using protein networks. *Nat Biotechnol*, **23**, 561-566.
220. Chipman, K.C. and Singh, A.K. (2009) Predicting genetic interactions with random walks on biological networks. *BMC Bioinformatics*, **10**, 17.
221. Tong, A.H., Lesage, G., Bader, G.D., Ding, H., Xu, H., Xin, X., Young, J., Berriz, G.F., Brost, R.L., Chang, M. *et al.* (2004) Global mapping of the yeast genetic interaction network. *Science*, **303**, 808-813.

222. Huang, M.E. and Kolodner, R.D. (2005) A biological network in Saccharomyces cerevisiae prevents the deleterious effects of endogenous oxidative DNA damage. *Mol Cell*, **17**, 709-720.
223. Pan, X., Ye, P., Yuan, D.S., Wang, X., Bader, J.S. and Boeke, J.D. (2006) A DNA integrity network in the yeast Saccharomyces cerevisiae. *Cell*, **124**, 1069-1081.
224. Dixon, S.J., Costanzo, M., Baryshnikova, A., Andrews, B. and Boone, C. (2009) Systematic mapping of genetic interaction networks. *Annu Rev Genet*, **43**, 601-625.
225. Wu, M., Li, X., Zhang, F., Li, X., Kwoh, C.K. and Zheng, J. (2014) In silico prediction of synthetic lethality by meta-analysis of genetic interactions, functions, and pathways in yeast and human cancer. *Cancer Inform*, **13**, 71-80.
226. Kaelin, W.G., Jr. (2005) The concept of synthetic lethality in the context of anticancer therapy. *Nat Rev Cancer*, **5**, 689-698.
227. Guo, J., Liu, H. and Zheng, J. (2015) SynLethDB: synthetic lethality database toward discovery of selective and sensitive anticancer drug targets. *Nucleic Acids Res*.
228. Chan, D.A., Sutphin, P.D., Nguyen, P., Turcotte, S., Lai, E.W., Banh, A., Reynolds, G.E., Chi, J.T., Wu, J., Solow-Cordero, D.E. *et al.* (2011) Targeting GLUT1 and the Warburg effect in renal cell carcinoma by chemical synthetic lethality. *Sci Transl Med*, **3**, 94ra70.
229. Turcotte, S., Chan, D.A., Sutphin, P.D., Hay, M.P., Denny, W.A. and Giaccia, A.J. (2008) A molecule targeting VHL-deficient renal cell carcinoma that induces autophagy. *Cancer Cell*, **14**, 90-102.
230. Jacquemont, C., Simon, J.A., D'Andrea, A.D. and Taniguchi, T. (2012) Non-specific chemical inhibition of the Fanconi anemia pathway sensitizes cancer cells to cisplatin. *Mol Cancer*, **11**, 26.
231. Kranz, D. and Boutros, M. (2014) A synthetic lethal screen identifies FAT1 as an antagonist of caspase-8 in extrinsic apoptosis. *EMBO J*, **33**, 181-197.
232. Josse, R., Martin, S.E., Guha, R., Ormanoglu, P., Pfister, T.D., Reaper, P.M., Barnes, C.S., Jones, J., Charlton, P., Pollard, J.R. *et al.* (2014) ATR inhibitors VE-821 and VX-970 sensitize cancer cells to topoisomerase i inhibitors by disabling DNA replication initiation and fork elongation responses. *Cancer Res*, **74**, 6968-6979.
233. Etemadmoghadam, D., Weir, B.A., Au-Yeung, G., Alsop, K., Mitchell, G., George, J., Australian Ovarian Cancer Study, G., Davis, S., D'Andrea, A.D., Simpson, K. *et al.* (2013) Synthetic lethality between CCNE1 amplification and loss of BRCA1. *Proc Natl Acad Sci U S A*, **110**, 19489-19494.
234. Scholl, C., Frohling, S., Dunn, I.F., Schinzel, A.C., Barbie, D.A., Kim, S.Y., Silver, S.J., Tamayo, P., Wadlow, R.C., Ramaswamy, S. *et al.* (2009) Synthetic lethal interaction between oncogenic KRAS dependency and STK33 suppression in human cancer cells. *Cell*, **137**, 821-834.
235. Sander, J.D. and Joung, J.K. (2014) CRISPR-Cas systems for editing, regulating and targeting genomes. *Nat Biotechnol*, **32**, 347-355.
236. Farmer, H., McCabe, N., Lord, C.J., Tutt, A.N., Johnson, D.A., Richardson, T.B., Santarosa, M., Dillon, K.J., Hickson, I., Knights, C. *et al.* (2005) Targeting the DNA repair defect in BRCA mutant cells as a therapeutic strategy. *Nature*, **434**, 917-921.
237. Bryant, H.E., Schultz, N., Thomas, H.D., Parker, K.M., Flower, D., Lopez, E., Kyle, S., Meuth, M., Curtin, N.J. and Helleday, T. (2005) Specific killing of BRCA2-deficient tumours with inhibitors of poly(ADP-ribose) polymerase. *Nature*, **434**, 913-917.
238. Unni, A.M., Lockwood, W.W., Zejnullahu, K., Lee-Lin, S.Q. and Varmus, H. (2015) Evidence that synthetic lethality underlies the mutual exclusivity of oncogenic KRAS and EGFR mutations in lung adenocarcinoma. *Elife*, **4**, e06907.

239. Srihari, S., Singla, J., Wong, L. and Ragan, M.A. (2015) Inferring synthetic lethal interactions from mutual exclusivity of genetic events in cancer. *Biol Direct*, **10**, 57.
240. Jerby-Arnon, L., Pfetzer, N., Waldman, Y.Y., McGarry, L., James, D., Shanks, E., Seashore-Ludlow, B., Weinstock, A., Geiger, T., Clemons, P.A. *et al.* (2014) Predicting cancer-specific vulnerability via data-driven detection of synthetic lethality. *Cell*, **158**, 1199-1209.
241. Zhang, F., Wu, M., Li, X.J., Li, X.L., Kwoh, C.K. and Zheng, J. (2015) Predicting essential genes and synthetic lethality via influence propagation in signaling pathways of cancer cell fates. *J Bioinform Comput Biol*, **13**, 1541002.
242. Patterson, D. (2009) Molecular genetic analysis of Down syndrome. *Hum Genet*, **126**, 195-214.
243. An, O., Gursoy, A., Gurgey, A. and Keskin, O. (2013) Structural and functional analysis of perforin mutations in association with clinical data of familial hemophagocytic lymphohistiocytosis type 2 (FHL2) patients. *Protein Sci*, **22**, 823-839.
244. Vogelstein, B., Papadopoulos, N., Velculescu, V.E., Zhou, S., Diaz, L.A., Jr. and Kinzler, K.W. (2013) Cancer genome landscapes. *Science*, **339**, 1546-1558.
245. Ciccarelli, F.D. (2010) The (r)evolution of cancer genetics. *BMC Biol*, **8**, 74.
246. 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**, 177-183.
247. 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.
248. Rosenbloom, K.R., Armstrong, J., Barber, G.P., Casper, J., Clawson, H., Diekhans, M., Dreszer, T.R., Fujita, P.A., Guruvadoo, L., Haeussler, M. *et al.* (2015) The UCSC Genome Browser database: 2015 update. *Nucleic Acids Res*, **43**, D670-681.
249. Mermel, C.H., Schumacher, S.E., Hill, B., Meyerson, M.L., Beroukhim, R. and Getz, G. (2011) GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol*, **12**, R41.
250. Cancer Genome Atlas Network. (2012) Comprehensive molecular characterization of human colon and rectal cancer. *Nature*, **487**, 330-337.
251. Ocak, S., Yamashita, H., Udyavar, A.R., Miller, A.N., Gonzalez, A.L., Zou, Y., Jiang, A., Yi, Y., Shyr, Y., Estrada, L. *et al.* (2010) DNA copy number aberrations in small-cell lung cancer reveal activation of the focal adhesion pathway. *Oncogene*, **29**, 6331-6342.
252. Haverty, P.M., Hon, L.S., Kaminker, J.S., Chant, J. and Zhang, Z. (2009) High-resolution analysis of copy number alterations and associated expression changes in ovarian tumors. *BMC Med Genomics*, **2**, 21.
253. Varma, S., Pommier, Y., Sunshine, M., Weinstein, J.N. and Reinhold, W.C. (2014) High resolution copy number variation data in the NCI-60 cancer cell lines from whole genome microarrays accessible through CellMiner. *PLoS One*, **9**, e92047.
254. Park, C.H., Jeong, H.J., Choi, Y.H., Kim, S.C., Jeong, H.C., Park, K.H., Lee, G.Y., Kim, T.S., Yang, S.W., Ahn, S.W. *et al.* (2006) Systematic analysis of cDNA microarray-based CGH. *Int J Mol Med*, **17**, 261-267.
255. Kandoth, C., McLellan, M.D., Vandin, F., Ye, K., Niu, B., Lu, C., Xie, M., Zhang, Q., McMichael, J.F., Wyczalkowski, M.A. *et al.* (2013) Mutational landscape and significance across 12 major cancer types. *Nature*, **502**, 333-339.
256. Quinlan, A.R. and Hall, I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841-842.

257. Bignell, G.R., Greenman, C.D., Davies, H., Butler, A.P., Edkins, S., Andrews, J.M., Buck, G., Chen, L., Beare, D., Latimer, C. *et al.* (2010) Signatures of mutation and selection in the cancer genome. *Nature*, **463**, 893-898.
258. Karolchik, D., Hinrichs, A.S., Furey, T.S., Roskin, K.M., Sugnet, C.W., Haussler, D. and Kent, W.J. (2004) The UCSC Table Browser data retrieval tool. *Nucleic Acids Res*, **32**, D493-496.
259. Schuster-Bockler, B., Conrad, D. and Bateman, A. (2010) Dosage sensitivity shapes the evolution of copy-number varied regions. *PLoS One*, **5**, e9474.
260. D'Antonio, M., Pendino, V., Sinha, S. and Ciccarelli, F.D. (2012) Network of Cancer Genes (NCG 3.0): integration and analysis of genetic and network properties of cancer genes. *Nucleic Acids Res*, **40**, D978-983.
261. 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**, D670-675.
262. An, O., Dall'Olio, G.M., Mourikis, T.P. and Ciccarelli, F.D. (2015) NCG 5.0: updates of a manually curated repository of cancer genes and associated properties from cancer mutational screenings. *Nucleic Acids Res*.
263. Tubio, J.M. (2015) Somatic structural variation and cancer. *Brief Funct Genomics*, **14**, 339-351.
264. Zack, T.I., Schumacher, S.E., Carter, S.L., Cherniack, A.D., Saksena, G., Tabak, B., Lawrence, M.S., Zhsng, C.Z., Wala, J., Mermel, C.H. *et al.* (2013) Pan-cancer patterns of somatic copy number alteration. *Nat Genet*, **45**, 1134-1140.
265. Wu, G., Diaz, A.K., Paugh, B.S., Rankin, S.L., Ju, B., Li, Y., Zhu, X., Qu, C., Chen, X., Zhang, J. *et al.* (2014) The genomic landscape of diffuse intrinsic pontine glioma and pediatric non-brainstem high-grade glioma. *Nat Genet*, **46**, 444-450.
266. Yates, L.R., Gerstung, M., Knappskog, S., Desmedt, C., Gundem, G., Van Loo, P., Aas, T., Alexandrov, L.B., Larsimont, D., Davies, H. *et al.* (2015) Subclonal diversification of primary breast cancer revealed by multiregion sequencing. *Nat Med*, **21**, 751-759.
267. Andersson, A.K., Ma, J., Wang, J., Chen, X., Gedman, A.L., Dang, J., Nakitandwe, J., Holmfeldt, L., Parker, M., Easton, J. *et al.* (2015) The landscape of somatic mutations in infant MLL-rearranged acute lymphoblastic leukemias. *Nat Genet*, **47**, 330-337.
268. Robinson, D., Van Allen, E.M., Wu, Y.M., Schultz, N., Lonigro, R.J., Mosquera, J.M., Montgomery, B., Taplin, M.E., Pritchard, C.C., Attard, G. *et al.* (2015) Integrative clinical genomics of advanced prostate cancer. *Cell*, **161**, 1215-1228.
269. Yang, X., Wood, P.A., Ansell, C. and Hrushesky, W.J. (2009) Circadian time-dependent tumor suppressor function of period genes. *Integr Cancer Ther*, **8**, 309-316.
270. Yang, X., Wood, P.A., Ansell, C.M., Quiton, D.F., Oh, E.Y., Du-Quiton, J. and Hrushesky, W.J. (2009) The circadian clock gene Per1 suppresses cancer cell proliferation and tumor growth at specific times of day. *Chronobiol Int*, **26**, 1323-1339.
271. Gery, S., Komatsu, N., Kawamata, N., Miller, C.W., Desmond, J., Virk, R.K., Marchevsky, A., McKenna, R., Taguchi, H. and Koeffler, H.P. (2007) Epigenetic silencing of the candidate tumor suppressor gene Per1 in non-small cell lung cancer. *Clin Cancer Res*, **13**, 1399-1404.
272. Lund, K., Adams, P.D. and Copland, M. (2014) EZH2 in normal and malignant hematopoiesis. *Leukemia*, **28**, 44-49.
273. Huether, R., Dong, L., Chen, X., Wu, G., Parker, M., Wei, L., Ma, J., Edmonson, M.N., Hedlund, E.K., Rusch, M.C. *et al.* (2014) The landscape of somatic mutations in epigenetic regulators across 1,000 paediatric cancer genomes. *Nat Commun*, **5**, 3630.

274. Kumar, P., Henikoff, S. and Ng, P.C. (2009) Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protoc*, **4**, 1073-1081.
275. Adzhubei, I.A., Schmidt, S., Peshkin, L., Ramensky, V.E., Gerasimova, A., Bork, P., Kondrashov, A.S. and Sunyaev, S.R. (2010) A method and server for predicting damaging missense mutations. *Nat Methods*, **7**, 248-249.
276. Chun, S. and Fay, J.C. (2009) Identification of deleterious mutations within three human genomes. *Genome Res*, **19**, 1553-1561.
277. Schwarz, J.M., Cooper, D.N., Schuelke, M. and Seelow, D. (2014) MutationTaster2: mutation prediction for the deep-sequencing age. *Nat Methods*, **11**, 361-362.
278. Reva, B., Antipin, Y. and Sander, C. (2011) Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic Acids Res*, **39**, e118.
279. Barretina, J., Caponigro, G., Stransky, N., Venkatesan, K., Margolin, A.A., Kim, S., Wilson, C.J., Lehar, J., Kryukov, G.V., Sonkin, D. *et al.* (2012) The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, **483**, 603-607.
280. Garnett, M.J., Edelman, E.J., Heidorn, S.J., Greenman, C.D., Dastur, A., Lau, K.W., Greninger, P., Thompson, I.R., Luo, X., Soares, J. *et al.* (2012) Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature*, **483**, 570-575.
281. Finn, R.D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R.Y., Eddy, S.R., Heger, A., Hetherington, K., Holm, L., Mistry, J. *et al.* (2014) Pfam: the protein families database. *Nucleic Acids Res*, **42**, D222-230.
282. 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**, 2347-2348.
283. Lawrence, M.S., Stojanov, P., Polak, P., Kryukov, G.V., Cibulskis, K., Sivachenko, A., Carter, S.L., Stewart, C., Mermel, C.H., Roberts, S.A. *et al.* (2013) Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature*, **499**, 214-218.
284. Liu, P., Morrison, C., Wang, L., Xiong, D., Vedell, P., Cui, P., Hua, X., Ding, F., Lu, Y., James, M. *et al.* (2012) Identification of somatic mutations in non-small cell lung carcinomas using whole-exome sequencing. *Carcinogenesis*, **33**, 1270-1276.
285. Brennan, C.W., Verhaak, R.G., McKenna, A., Campos, B., Noushmehr, H., Salama, S.R., Zheng, S., Chakravarty, D., Sanborn, J.Z., Berman, S.H. *et al.* (2013) The somatic genomic landscape of glioblastoma. *Cell*, **155**, 462-477.
286. Letunic, I., Doerks, T. and Bork, P. (2015) SMART: recent updates, new developments and status in 2015. *Nucleic Acids Res*, **43**, D257-260.
287. Melé, M., Ferreira, P.G., Reverter, F., DeLuca, D.S., Monlong, J., Sammeth, M., Young, T.R., Goldmann, J.M., Pervouchine, D.D., Sullivan, T.J. *et al.* (2015) Human genomics. The human transcriptome across tissues and individuals. *Science*, **348**, 660-665.
288. Guttman, M., Mies, C., Dudycz-Sulicz, K., Diskin, S.J., Baldwin, D.A., Stoeckert, C.J., Jr. and Grant, G.R. (2007) Assessing the significance of conserved genomic aberrations using high resolution genomic microarrays. *PLoS Genet*, **3**, e143.
289. Klijn, C., Holstege, H., de Ridder, J., Liu, X., Reinders, M., Jonkers, J. and Wessels, L. (2008) Identification of cancer genes using a statistical framework for multiexperiment analysis of nondiscretized array CGH data. *Nucleic Acids Res*, **36**, e13.
290. Walter, V., Nobel, A.B. and Wright, F.A. (2011) DiNAMIC: a method to identify recurrent DNA copy number aberrations in tumors. *Bioinformatics*, **27**, 678-685.

291. Aguirre, A.J., Brennan, C., Bailey, G., Sinha, R., Feng, B., Leo, C., Zhang, Y., Zhang, J., Gans, J.D., Bardeesy, N. *et al.* (2004) High-resolution characterization of the pancreatic adenocarcinoma genome. *Proc Natl Acad Sci U S A*, **101**, 9067-9072.
292. Rouveirol, C., Stransky, N., Hupe, P., Rosa, P.L., Viara, E., Barillot, E. and Radvanyi, F. (2006) Computation of recurrent minimal genomic alterations from array-CGH data. *Bioinformatics*, **22**, 849-856.
293. Shah, S.P., Xuan, X., DeLeeuw, R.J., Khojasteh, M., Lam, W.L., Ng, R. and Murphy, K.P. (2006) Integrating copy number polymorphisms into array CGH analysis using a robust HMM. *Bioinformatics*, **22**, e431-439.
294. Colella, S., Yau, C., Taylor, J.M., Mirza, G., Butler, H., Clouston, P., Bassett, A.S., Seller, A., Holmes, C.C. and Ragoussis, J. (2007) QuantiSNP: an Objective Bayes Hidden-Markov Model to detect and accurately map copy number variation using SNP genotyping data. *Nucleic Acids Res*, **35**, 2013-2025.
295. Negrini, S., Gorgoulis, V.G. and Halazonetis, T.D. (2010) Genomic instability--an evolving hallmark of cancer. *Nat Rev Mol Cell Biol*, **11**, 220-228.
296. Godinho, S.A., Picone, R., Burute, M., Dagher, R., Su, Y., Leung, C.T., Polyak, K., Brugge, J.S., Thery, M. and Pellman, D. (2014) Oncogene-like induction of cellular invasion from centrosome amplification. *Nature*, **510**, 167-171.
297. Iskow, R.C., Gokcumen, O. and Lee, C. (2012) Exploring the role of copy number variants in human adaptation. *Trends Genet*, **28**, 245-257.
298. Liao, D. (2009) Emerging roles of the EBF family of transcription factors in tumor suppression. *Mol Cancer Res*, **7**, 1893-1901.
299. Yu, J., Liang, Q.Y., Wang, J., Cheng, Y., Wang, S., Poon, T.C., Go, M.Y., Tao, Q., Chang, Z. and Sung, J.J. (2013) Zinc-finger protein 331, a novel putative tumor suppressor, suppresses growth and invasiveness of gastric cancer. *Oncogene*, **32**, 307-317.
300. Barbashina, V., Salazar, P., Holland, E.C., Rosenblum, M.K. and Ladanyi, M. (2005) Allelic losses at 1p36 and 19q13 in gliomas: correlation with histologic classification, definition of a 150-kb minimal deleted region on 1p36, and evaluation of CAMTA1 as a candidate tumor suppressor gene. *Clin Cancer Res*, **11**, 1119-1128.
301. Henrich, K.O., Bauer, T., Schulte, J., Ehemann, V., Deubzer, H., Gogolin, S., Muth, D., Fischer, M., Benner, A., Konig, R. *et al.* (2011) CAMTA1, a 1p36 tumor suppressor candidate, inhibits growth and activates differentiation programs in neuroblastoma cells. *Cancer Res*, **71**, 3142-3151.
302. Katoh, M. and Katoh, M. (2003) Identification and characterization of FLJ10737 and CAMTA1 genes on the commonly deleted region of neuroblastoma at human chromosome 1p36.31-p36.23. *Int J Oncol*, **23**, 1219-1224.
303. Keane, M.M., Rivero-Lezcano, O.M., Mitchell, J.A., Robbins, K.C. and Lipkowitz, S. (1995) Cloning and characterization of cbl-b: a SH3 binding protein with homology to the c-cbl proto-oncogene. *Oncogene*, **10**, 2367-2377.
304. Hock, H. (2012) A complex Polycomb issue: the two faces of EZH2 in cancer. *Genes Dev*, **26**, 751-755.
305. Liu, T., Jankovic, D., Brault, L., Ehret, S., Baty, F., Stavropoulou, V., Rossi, V., Biondi, A. and Schwaller, J. (2010) Functional characterization of high levels of meningioma 1 as collaborating oncogene in acute leukemia. *Leukemia*, **24**, 601-612.
306. Lucio-Eterovic, A.K. and Carpenter, P.B. (2011) An open and shut case for the role of NSD proteins as oncogenes. *Transcription*, **2**, 158-161.
307. Ntziachristos, P., Tsirigos, A., Van Vlierberghe, P., Nedjic, J., Trimarchi, T., Flaherty, M.S., Ferres-Marco, D., da Ros, V., Tang, Z., Siegle, J. *et al.* (2012)

Genetic inactivation of the polycomb repressive complex 2 in T cell acute lymphoblastic leukemia. *Nat Med*, **18**, 298-301.
308. Sneeringer, C.J., Scott, M.P., Kuntz, K.W., Knutson, S.K., Pollock, R.M., Richon, V.M. and Copeland, R.A. (2010) Coordinated activities of wild-type plus mutant EZH2 drive tumor-associated hypertrimethylation of lysine 27 on histone H3 (H3K27) in human B-cell lymphomas. *Proc Natl Acad Sci U S A*, **107**, 20980-20985.
309. Arisan, S., Buyuktuncer, E.D., Palavan-Unsal, N., Caskurlu, T., Cakir, O.O. and Ergenekon, E. (2005) Increased expression of EZH2, a polycomb group protein, in bladder carcinoma. *Urol Int*, **75**, 252-257.
310. Raman, J.D., Mongan, N.P., Tickoo, S.K., Boorjian, S.A., Scherr, D.S. and Gudas, L.J. (2005) Increased expression of the polycomb group gene, EZH2, in transitional cell carcinoma of the bladder. *Clin Cancer Res*, **11**, 8570-8576.
311. Weikert, S., Christoph, F., Kollermann, J., Muller, M., Schrader, M., Miller, K. and Krause, H. (2005) Expression levels of the EZH2 polycomb transcriptional repressor correlate with aggressiveness and invasive potential of bladder carcinomas. *Int J Mol Med*, **16**, 349-353.
312. Kleer, C.G., Cao, Q., Varambally, S., Shen, R., Ota, I., Tomlins, S.A., Ghosh, D., Sewalt, R.G., Otte, A.P., Hayes, D.F. *et al.* (2003) EZH2 is a marker of aggressive breast cancer and promotes neoplastic transformation of breast epithelial cells. *Proc Natl Acad Sci U S A*, **100**, 11606-11611.
313. Mimori, K., Ogawa, K., Okamoto, M., Sudo, T., Inoue, H. and Mori, M. (2005) Clinical significance of enhancer of zeste homolog 2 expression in colorectal cancer cases. *Eur J Surg Oncol*, **31**, 376-380.
314. Bachmann, I.M., Halvorsen, O.J., Collett, K., Stefansson, I.M., Straume, O., Haukaas, S.A., Salvesen, H.B., Otte, A.P. and Akslen, L.A. (2006) EZH2 expression is associated with high proliferation rate and aggressive tumor subgroups in cutaneous melanoma and cancers of the endometrium, prostate, and breast. *J Clin Oncol*, **24**, 268-273.
315. Matsukawa, Y., Semba, S., Kato, H., Ito, A., Yanagihara, K. and Yokozaki, H. (2006) Expression of the enhancer of zeste homolog 2 is correlated with poor prognosis in human gastric cancer. *Cancer Sci*, **97**, 484-491.
316. Sudo, T., Utsunomiya, T., Mimori, K., Nagahara, H., Ogawa, K., Inoue, H., Wakiyama, S., Fujita, H., Shirouzu, K. and Mori, M. (2005) Clinicopathological significance of EZH2 mRNA expression in patients with hepatocellular carcinoma. *Br J Cancer*, **92**, 1754-1758.
317. Watanabe, H., Soejima, K., Yasuda, H., Kawada, I., Nakachi, I., Yoda, S., Naoki, K. and Ishizaka, A. (2008) Deregulation of histone lysine methyltransferases contributes to oncogenic transformation of human bronchoepithelial cells. *Cancer Cell Int*, **8**, 15.
318. Varambally, S., Dhanasekaran, S.M., Zhou, M., Barrette, T.R., Kumar-Sinha, C., Sanda, M.G., Ghosh, D., Pienta, K.J., Sewalt, R.G., Otte, A.P. *et al.* (2002) The polycomb group protein EZH2 is involved in progression of prostate cancer. *Nature*, **419**, 624-629.
319. Cohen, I., Poreba, E., Kamieniarz, K. and Schneider, R. (2011) Histone modifiers in cancer: friends or foes? *Genes Cancer*, **2**, 631-647.
320. Stepanenko, A.A., Vassetzky, Y.S. and Kavsan, V.M. (2013) Antagonistic functional duality of cancer genes. *Gene*, **529**, 199-207.
321. Ali Hassan, N.Z., Mokhtar, N.M., Kok Sin, T., Mohamed Rose, I., Sagap, I., Harun, R. and Jamal, R. (2014) Integrated analysis of copy number variation and genome-wide expression profiling in colorectal cancer tissues. *PLoS One*, **9**, e92553.

322. Beroukhim, R., Getz, G., Nghiemphu, L., Barretina, J., Hsueh, T., Linhart, D., Vivanco, I., Lee, J.C., Huang, J.H., Alexander, S. *et al.* (2007) Assessing the significance of chromosomal aberrations in cancer: methodology and application to glioma. *Proc Natl Acad Sci U S A*, **104**, 20007-20012.
323. Dees, N.D., Zhang, Q., Kandoth, C., Wendl, M.C., Schierding, W., Koboldt, D.C., Mooney, T.B., Callaway, M.B., Dooling, D., Mardis, E.R. *et al.* (2012) MuSiC: identifying mutational significance in cancer genomes. *Genome Res*, **22**, 1589-1598.
324. Banerji, S., Cibulskis, K., Rangel-Escareno, C., Brown, K.K., Carter, S.L., Frederick, A.M., Lawrence, M.S., Sivachenko, A.Y., Sougnez, C., Zou, L. *et al.* (2012) Sequence analysis of mutations and translocations across breast cancer subtypes. *Nature*, **486**, 405-409.
325. Li, X.J., Mishra, S.K., Wu, M., Zhang, F. and Zheng, J. (2014) Syn-lethality: an integrative knowledge base of synthetic lethality towards discovery of selective anticancer therapies. *Biomed Res Int*, **2014**, 196034.
326. Mali, P., Yang, L., Esvelt, K.M., Aach, J., Guell, M., DiCarlo, J.E., Norville, J.E. and Church, G.M. (2013) RNA-guided human genome engineering via Cas9. *Science*, **339**, 823-826.
327. Koike-Yusa, H., Li, Y., Tan, E.P., Velasco-Herrera Mdel, C. and Yusa, K. (2014) Genome-wide recessive genetic screening in mammalian cells with a lentiviral CRISPR-guide RNA library. *Nat Biotechnol*, **32**, 267-273.
328. Petsko, G.A. and Ringe, D. (2004) *Protein structure and function*. New Science Press, London Sunderland, MA.
329. Giri, R., Morrone, A., Travaglini-Allocatelli, C., Jemth, P., Brunori, M. and Gianni, S. (2012) Folding pathways of proteins with increasing degree of sequence identities but different structure and function. *Proc Natl Acad Sci U S A*, **109**, 17772-17776.
330. Alexander, P.A., He, Y., Chen, Y., Orban, J. and Bryan, P.N. (2007) The design and characterization of two proteins with 88% sequence identity but different structure and function. *Proc Natl Acad Sci U S A*, **104**, 11963-11968.

Acknowledgement

This thesis carries positive marks from many people, who contributed to its completion either via collaborative work or accompanying me to make it through, or both. I would like to thank all of them for their presence, and special thanks to,

My supervisor, Francesca Ciccarelli, for her continuous support, advices, constructive criticism and guidance throughout my PhD. I am grateful to her for giving me the opportunity to build up my academic development,

My internal and external advisors, Giuseppe Testa and Alexandre Reymond, for the useful and encouraging discussions,

My colleague, Lorena Benedetti, who performed the lab experiments included in my thesis, for her contribution to my work and organising the get-together events,

NCG team, Giovanni M. Dall'Olio and Thanos Mourikis, for the teamwork which led us to a wonderful website, and all the fun behind it,

All the members of the lab, Matteo Cereda, Gennaro Gambardella, Shruti Sinha, Vera Pendino, Fabio Iannelli, Matteo D'Antonio, Elena Gatti and Fiorella Guerra for the scientific discussions, sharing knowledge, dinners, drinks, coffee breaks and much more. They have been the driving force of this work with their friendship.

I would like to acknowledge the financial, academic and technical support of Lifelong Learning Programme-Erasmus Placement provided by the University of Milan, European School of Molecular Medicine (SEMM) and European Institute of Oncology (IEO).

I would like to express my sincere gratitude to my family for their undoubted trust and motivation. I have always had their love and moral support thorough the though times. Last but not least, I would like to thank to my friends from Turkey, Italy, UK and all over the world for the pleasant memories.

Nucleic Acids Research Advance Access published October 29, 2015

Nucleic Acids Research, 2015 **1**
doi: 10.1093/nar/gkv1123

NCG 5.0: updates of a manually curated repository of cancer genes and associated properties from cancer mutational screenings

Omer An, Giovanni M. Dall'Olio, Thanos P. Mourikis and Francesca D. Ciccarelli*

Division of Cancer Studies, King's College London, London SE11UL, UK

Received September 14, 2015; Revised October 12, 2015; Accepted October 14, 2015

ABSTRACT

The Network of Cancer Genes (NCG, http://ncg.kcl.ac.uk/) is a manually curated repository of cancer genes derived from the scientific literature. Due to the increasing amount of cancer genomic data, we have introduced a more robust procedure to extract cancer genes from published cancer mutational screenings and two curators independently reviewed each publication. NCG release 5.0 (August 2015) collects 1571 cancer genes from 175 published studies that describe 188 mutational screenings of 13 315 cancer samples from 49 cancer types and 24 primary sites. In addition to collecting cancer genes, NCG also provides information on the experimental validation that supports the role of these genes in cancer and annotates their properties (duplicability, evolutionary origin, expression profile, function and interactions with proteins and miRNAs).

INTRODUCTION

Cancer genome projects, including The Cancer Genome Atlas (TCGA, https://tcga-data.nci.nih.gov/) and the International Cancer Genome Project (ICGC, https://dcc.icgc.org/) have so far mapped DNA alterations in more than 13 000 cancer samples. These massive sequencing efforts show that somatic modifications vary greatly between and within cancer types (1–3). Only some of the acquired alterations, however, confer a selective advantage that promotes cancer development (*driver alterations*). The large majority of alterations have no or little role in cancer and are fixed in the cancer genome as a by-product of the selection acting on drivers (*passenger alterations*). One of the challenges of cancer genomics is to effectively distinguish between driver and passenger alterations in order to identify the molecular determinants of cancer. Most known driver alterations modify protein-coding genes (*cancer genes*). The ability to identify cancer genes among the wealth of mutated genes is crucial to better understand cancer biology and to empower the development of innovative anti-cancer therapy.

Network of Cancer Genes (NCG) is a database launched in 2010 with the aim to collect cancer genes from the literature. Curators constantly review cancer mutational screenings and annotate altered genes that either have well-established cancer functions (*known cancer genes*) or are putative cancer drivers (*candidate cancer genes*). Originally (4), NCG collected data from only five mutational screenings and annotated most known cancer genes from the Cancer Gene Census (CGC) (5). The last five years have seen the rapid accumulation of cancer genomic data from thousands of samples, with almost all human genes mutated in at least one sample (6,7). Due to this overwhelming amount of data and to avoid the inclusion of mutated genes with no role in cancer, in this release we have substantially reviewed the procedure to identify cancer genes. NCG now collects 1571 cancer genes, 518 of which are known cancer genes. The remaining 1053 genes are candidate cancer genes whose driver role has been predicted in the original publication using a variety of methods (Supplementary Table S1). Given the importance of a robust experimental support for the cancer activity of candidate cancer genes, NCG now collects additional literature describing available orthogonal validations. NCG also annotates various properties of cancer genes such as the presence of extra copies in the genome (gene duplicability), the evolutionary origin, the connectivity of the encoded proteins in the protein–protein and miRNA interaction networks, and the comprehensive gene expression profile across 38 human tissues and 1543 cancer cell lines.

The manual curation of the literature to extract cancer driver genes and the annotation of a large number of additional properties make NCG a comprehensive and updated resource to navigate the overwhelming amount of cancer data with a particular focus on the genetic determinants of cancer.

MANUAL ANNOTATION OF CANCER GENES

In this release of NCG, the procedure for the inclusion of cancer genes in NCG has been reviewed and standardized

*To whom correspondence should be addressed. Tel: +44 20 7848 6616; Fax: +44 20 7848 6220; Email: francesca.ciccarelli@kcl.ac.uk

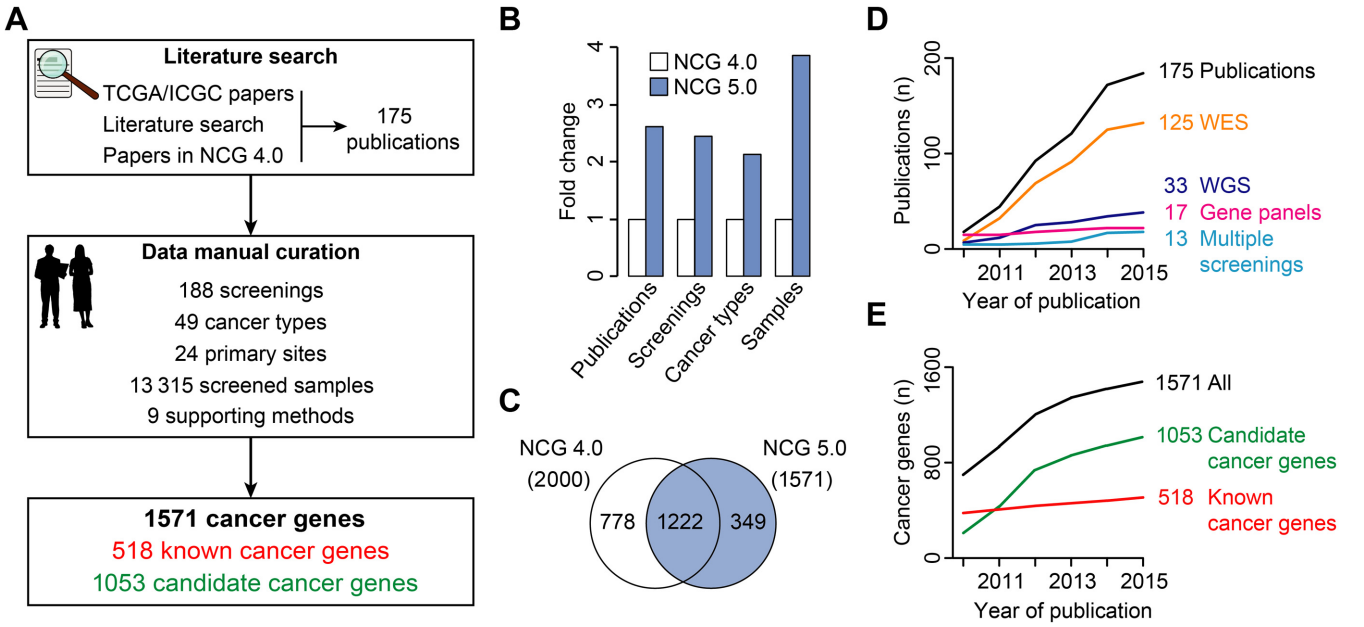


Figure 1. Curation procedure and comparison between NCG 5.0 and NCG 4.0: (**A**) Flowchart of the curation procedure used in NCG. After the identification of relevant publications describing cancer mutational screenings, two independent curators extract cancer genes and related information on types of screening and cancer, primary sites, screened samples and supporting methods. (**B**) Number of publications, screenings, cancer types and screened samples in NCG 5.0 as compared to NCG 4.0. (**C**) Venn diagram of cancer genes in NCG 4.0 and NCG 5.0. The reasons for the removal of 778 genes from the database are detailed in Supplementary Table S2. (**D–E**) Growth of NCG data in time. Shown are the number of publications, screenings and cancer genes starting from 2010, year of the first release of NCG. All screenings that were published prior of 2010, were collapsed.

(Figure 1A). The first difference with previous versions is to restrict the inclusion only to studies that describe mutational screenings of cancer samples and that distinguish between cancer genes and genes with passenger mutations. This led to the identification of 119 new publications. To be consistent with these inclusion criteria, all 68 studies present in the previous release were re-analysed. Twelve of them were excluded because they screened cancer cell lines rather than cancer samples or used no methods to identify cancer genes among all mutated genes. As a result of this extensive literature search, NCG 5.0 currently collects 175 studies (Supplementary Table S1). Two curators reviewed independently each publication to extract cancer genes and complementary information, such as the screening and the cancer types, the primary sites, the number of sequenced samples and the methods that were applied to identify cancer genes (Figure 1A). This manual curation resulted in 1260 cancer genes, 207 of which were annotated as known cancer genes in CGC. The remaining 1053 genes were candidate cancer genes identified in the original study using one or more methods (Supplementary Table S1). Additional known cancer genes were also added from CGC (February 2014), leading to a total of 1571 cancer genes. If information was available, cancer genes were further annotated as dominant (mostly oncogenes) or recessive (mostly tumour-suppressors) genes.

As compared to NCG 4.0 (8), NCG 5.0 now collects information from more than the double number of publications, screenings and cancer types and from four times more cancer samples (Figure 1B). Despite this substantial increase of data, the number of cancer genes decreased from 2000 to 1571 (Figure 1C), because of the more restrictive criteria. In particular, 612 genes were removed because the original publication was excluded and 166 genes because they had no support as cancer drivers (Supplementary Table S2). Overall, the studies in NCG 5.0 describe 188 mutational screenings, including 125 whole exome sequencings, 33 whole genome sequencings, 17 screenings of selected gene panels and 13 screenings based on multiple approaches (Figure 1D). Interestingly, the number of cancer genes with a well-documented role in cancer increases at a much slower pace as compared to candidate cancer genes (Figure 1E). This highlights the currently unmet need of efficient experimental assays that support the predicted role of candidate genes in cancer.

Almost all mutational screenings collected in NCG 5.0 applied only one method to identify cancer genes (Supplementary Table S1). The most common was the recurrence of mutation of a given gene across samples, which was taken as a sign of functional selection (Figure 2A and Supplementary Table S1). Other commonly used methods included MutSig (6) and MuSiC (9) (Figure 2A and Supplementary Table S1). Interestingly, the majority of known cancer genes (67%) had the support of at least two methods (Figure 2B), while most candidate cancer genes (78%) have been predicted by only one method (Figure 2C). In agreement with this, known cancer genes were overall identified as drivers across a higher number of mutational screenings and primary cancer sites as compared to candidate cancer genes (Figure 2D). The tendency of candidate cancer genes to be cancer specific was also reflected by the lower overlap between methods that support them as compared to those that support known cancer genes (Figure 2E). Cases where the overlap was higher (i.e. between MutSig and Invex, Figure

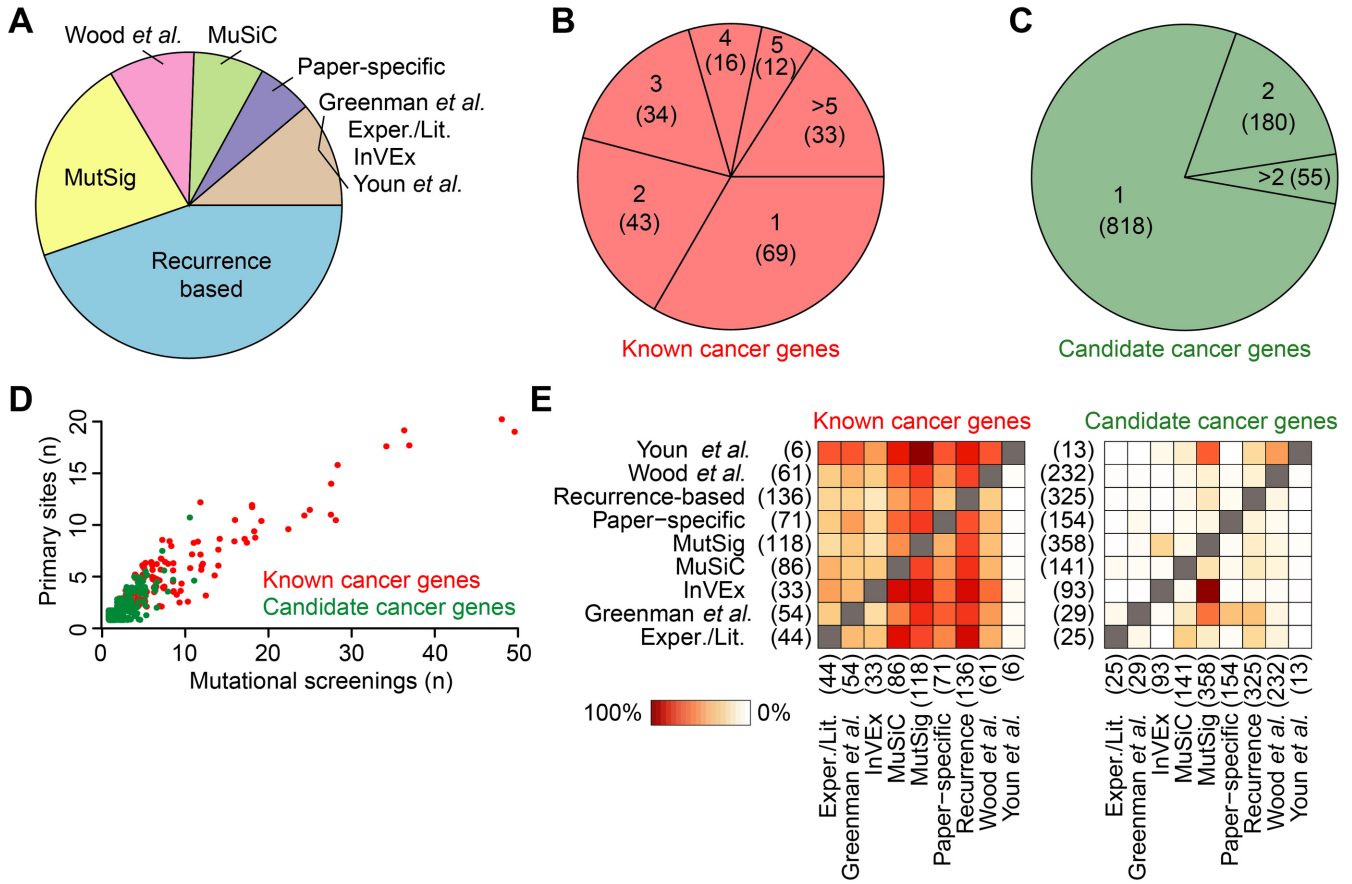


Figure 2. Overview of data in NCG 5.0: (**A**) Cancer mutational screenings divided according to the method that was applied to identify cancer genes in the original publication. Methods and corresponding screenings are described in Supplementary Table S1. (**B–C**) Fractions of known and candidate cancer genes supported by one or more methods. Gene counts are reported in brackets. (**D**) Number of mutational screenings and primary sites where each cancer gene has been reported as a driver. *TP53* is an outlier and has been excluded from the analysis because it has been identified in 113 screenings across 22 primary sites. (**E**) Heatmaps of the overlap between methods identifying known and candidate cancer genes. Each box represents the percentage of cancer genes identified with one method that are also supported by another. For each method, the total number of associated cancer genes is reported in brackets.

2E) corresponded to screenings where both methods were used (Supplementary Table S1).

EXPERIMENTAL VALIDATION OF CANDIDATE CANCER GENES

Candidate cancer genes that are identified using computational methods often lack additional experimental validation of their cancer driver role. The main reason is that functional follow-ups are often cumbersome and require *ad hoc* design for individual genes. The experimental proof of predicted driver role is however crucial for the translatability of potentially relevant discoveries into increased knowledge and novel treatments.

In this release of NCG, we have extensively reviewed the literature to search for experimental validations of candidate cancer genes. NCG now annotates available orthogonal experiments that have been performed in the original study or in follow-up studies for 120 out of 1053 candidate cancer genes (11% of the total, Table 1 and Supplementary Table S3). Most commonly used approaches measure the effect of gene silencing or gene overexpression in cell lines (Figure 3A and Supplementary Table S3) and the majority of candidate genes (83 out of 120) have been validated through multiple assays (Figure 3B).

An interesting case is *CSMD3,* the gene associated with benign adult familial myoclonic epilepsy (10) that encodes a long multi-repeat protein (Figure 3C). *CSMD3* has been found recurrently mutated across several cancer types and, therefore, has been predicted as a cancer driver by several methods (Figure 3D). Because of its length, sequence composition and location in proximity of fragile sites of the genome, *CSMD3* was regarded as a possible false positive in NCG 4.0. The fact that *CSMD3* is constitutionally not expressed in many tissues where it is mutated (Figure 3E) also supports the passenger role of the acquired mutations. Despite this, however, the stable knockout of *CSMD3* in immortalized epithelial cells has been reported to increase cell proliferation (11), thus suggesting a tumour-suppressor role for this gene. This example highlights the difficulty to correctly predict the driver role of mutated genes and the need of multiple independent pieces of evidence to assess the role of mutations in cancer.

Table 1. Experimental validation of candidate cancer genes

Experimental validation	Candidate cancer genes (n)	Publications (n)
Gene overexpression	60	74
Transient RNA interference	58	52
Mutagenesis	31	41
Immunostaining	25	26
Stable gene knockout	23	22
Survival analysis	20	21
Protein activity assay	19	20
Drug response assay	15	17
In silico protein modelling	12	14
Xenograft	10	11
Rhotekin pull-down	2	5
Total	275 (120 unique genes)	303 (166 unique publications)

For each type of experimental validation, the numbers of validated candidate genes and corresponding publications are shown. The complete gene list with references to the original papers is given in Supplementary Table S3.

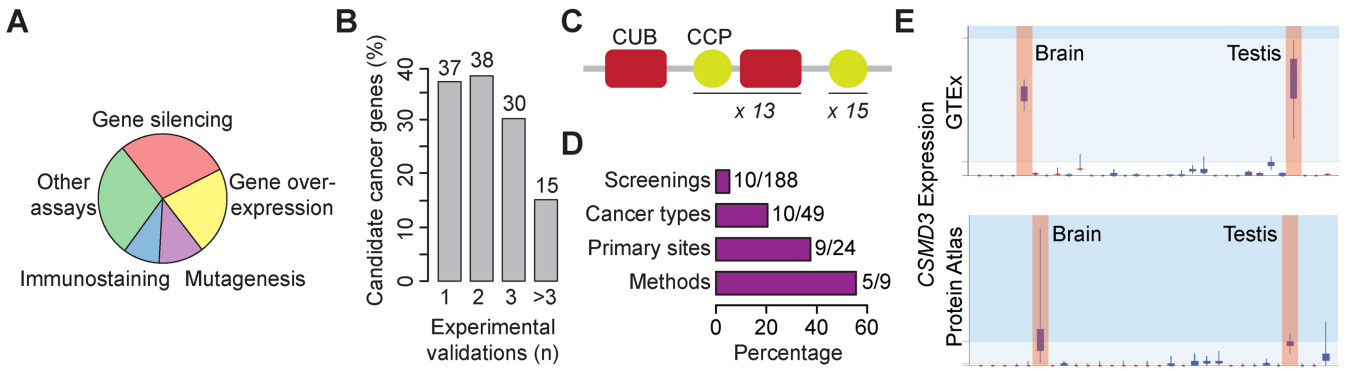


Figure 3. Validation of candidate cancer genes and alteration spectrum of *CSMD3*: (**A**) Fractions of validated candidate cancer genes according to the used experimental assay. Gene silencing refers to stable knockout or transient knockdown via RNA interference. Other assays include *in silico* protein modelling, survival analysis, drug response, protein activity, rhotekin pull-down and xenograft cancer models. (**B**) Percentage of candidate cancer genes that have been validated using one or more experimental approaches. The corresponding number of genes is shown above each bar. The full list of experiments and genes is reported in Supplementary Table S3. (**C**) Protein domain architecture of CSMD3 according to the SMART database (32). (**D**) Percentage of mutational screenings, cancer types, primary sites and methods that support the cancer driver role of *CSMD3*. Corresponding numbers are provided. (**E**) Expression profile of *CSMD3* in normal human tissues. Tissues where the gene is expressed in GTEx and Protein Atlas are highlighted in red.

ANNOTATION OF CANCER GENE PROPERTIES

To annotate the properties of cancer genes, original data on human genes, orthology, protein–protein and miRNA interactions and gene expression have been updated (Table 2).

Applying the previously described method (12), protein sequences from RefSeq v.63 (13) were aligned to the human genome assembly Hg19 to identify unique gene loci. These included 1525 of the 1571 cancer genes (13 cancer genes did not have RefSeq entries and 33 had no match in Hg19 or were gene isoforms). Cancer genes confirm their lower duplicability as compared to non-cancer genes and the signal derives from recessive cancer genes (*P*-value = 0.02, chi-square test, Table 2).

Orthology information from EggNOG v.4 (14) was used to trace the evolutionary origin of 1501 cancer genes, as described earlier (15). In line with previous reports (15–17), a higher fraction of cancer genes have orthologs in premetazoan species as compared to other human genes (*P*-value = 0.03, chi-square test, Table 2).

Four sources of primary interaction data (BioGRID v.3.4.125 (18); MIntAct v.190 (19); DIP (April 2015) (20); HPRD v.9 (21)) were integrated to rebuild the human protein–protein interaction network. This network included 1332 cancer proteins, which encode a higher fraction of hubs (defined as 25% most connected nodes of the network) as compared to other human proteins (*P*-value = 2.7×10^{-56}, chi-square test, Table 2). We verified that cancer genes encode a higher fraction of protein hubs also in the network derived from high-throughput screenings (*P*-value = 7.7×10^{-13}, chi-square test, Table 2). This excludes biases due to the higher number of single-gene experiments involving cancer proteins.

To complete the annotation of protein–protein interactions, NCG now collects also information on 752 cancer proteins involved in complexes as gathered from three resources (CORUM (February 2012) (22), HPRD v.9 (21), Reactome v.53 (23)). Supporting the signal from the overall protein–protein interaction network, a higher percentage of cancer proteins engage in complexes as compared to non-cancer proteins (*P*-value = 3.0×10^{-67}, chi-square test, Table 2).

Interactions between 324 miRNAs and 1101 cancer genes were derived from miRTarBase v.4.5 (24) and miRecords (April 2013) (25). Similarly to the protein–protein interaction network, also in the miRNA network a significantly larger fraction of cancer genes are target of miRNAs as

Table 2. Data and properties of cancer genes in NCG 5.0

Data sets in NCG 5.0		All cancer genes (1571)	Known cancer genes (518)		Candidate cancer genes (1053)	Other human genes
			Dominant (395)	Recessive (112)		
Human genes	All genes	1525	382	112	1020	17 489
	Duplicated genes (%)	280 (18%)	76 (20%)	12 (11%)	187 (18%)	3520 (20%)
Orthology	All genes	1501	379	110	1001	16 618
	Pre-metazoan genes (%)	992 (66%)	233 (61%)	80 (72%)	672 (67%)	10 516 (63%)
Protein–protein interactions	All nodes	1332	371	110	840	13 262
	Hubs (%)	558 (42%)	213 (57%)	78 (71%)	257 (31%)	2970 (22%)
	All nodes in HT network	1177	339	108	720	11 481
	Hubs in HT network (%)	386 (33%)	148 (44%)	52 (48%)	177 (25%)	2681 (23%)
Protein complexes	Proteins (%)	752 (49%)	238 (62%)	87 (78%)	418 (41%)	4917 (28%)
miRNA interactions	miRNA target genes (%)	1101 (72%)	332 (87%)	99 (88%)	662 (65%)	10 643 (61%)
	miRNAs	324	247	163	250	438
Expression in normal tissues	All genes in GTEx	1513	379	111	1012	16 818
	Ubiquitous genes (%)	965 (64%)	301 (79%)	98 (88%)	555 (55%)	11 077 (66%)
	Tissue-specific genes (%)	62 (4%)	5 (1%)	0 (0%)	57 (6%)	726 (4%)
	All genes in Protein Atlas	1517	378	112	1016	16 889
	Ubiquitous genes (%)	831 (55%)	278 (74%)	95 (85%)	447 (44%)	9492 (56%)
	Tissue-specific genes (%)	90 (6%)	11 (3%)	1 (1%)	78 (8%)	1042 (6%)
Expression in cancer cell lines	Cancer cell line encyclopedia	1426	367	106	942	15 158
	COSMIC Cancer Lines	1398	358	105	924	14 788
	Genentech data set	1524	381	112	1020	17 164

Of the 518 known cancer genes derived from CGC, 391 are annotated as dominant (mostly oncogenes), 108 as recessive (mostly tumour-suppressors), four as both as dominant and recessive and 15 have no specified mode of action. Duplicated genes have one or more duplicated loci in the genome covering ≥60% of their length (12). Pre-metazoan genes originated in the Last Universal Common Ancestor, Eukaryotes or Opisthokonts. Ubiquitously expressed genes are expressed in ≥95% tissues (29 tissues in GTEx and 30 tissues in Protein Atlas). HT = high throughput (publications reporting ≥100 interactions).

compared to other human genes (P-value = 3.0×10^{-18}, chi-square test, Table 2).

This release of NCG provides information on the expression of cancer genes in normal tissues and in cancer cell lines. For normal tissues, NCG relies on GTEx v.1.1.8 (26) and Protein Atlas (April 2015) (27), which both derive gene expression from RNASeq data in a total of 38 tissues. Expression values (FPKM for GTEx and RPKM for Protein Atlas) were used to derive expression categories (low, medium and high expression) for each gene and to calculate the distribution of gene expression across samples in each tissue. In both data sets, larger fractions of known cancer genes, but not of candidate cancer genes, are ubiquitously expressed (expression in >95% of all tissues) as compared to other genes (P-value = 1.3×10^{-13} and $P = 1.3 \times 10^{-19}$ for GTEx and Protein Atlas, respectively, chi-square test, Table 2). Conversely, significantly lower fractions of known cancer genes, but not of candidate cancer genes, are tissue specific (P-value = 4.2×10^{-4} and P-value = 6.9×10^{-4}, for GTEx and Protein Atlas, respectively, chi-square test, Table 2).

Three data sets (Cancer Cell Lines Encyclopedia (28), COSMIC Cancer Lines Project (29) and the recently released Genentech data set (30)) were used to derive gene expression in a total of 1543 cancer cell lines (Table 2). For each cancer gene, NCG provides the original expression value in each cell line as well as the normalized expression score, calculated as previously reported (31).

DATA ACCESS

NCG web interface has been reorganized, with particular focus on the summary of gene information and on the visualization of gene expression profiles. The gene summary now includes additional cross-references to external resources on protein domain architecture (32), drug and compound interactions (33,34) and protein druggability (35). For each cancer gene, the type of mutational screenings, the supporting methods and any experimental validation are detailed. Gene expression profiles are now shown as interactive graphs reporting the distribution of expression levels in each normal tissue and as summary tables in cancer cell lines.

NCG website provides overview statistics of the data contained in the database, including the list of 49 cancer types and corresponding 24 primary sites, the distribution of known and candidate cancer genes per primary sites, and information on 48 possible false positives. These include 14 genes derived from the literature (6), 4 additional genes that likely accumulate a high number of alterations due to their length and 30 olfactory receptor genes. All data contained in the database can be exported in batch using the advanced search option.

NCG USAGE

NCG offers a multi-level annotation of cancer genes that can be queried to gain insights on mutation status, properties, function and expression profiles of cancer genes (Figure 4A). This information facilitates the characterization of cancer genes and associated features. For example, gene duplicability has been exploited to extract duplicated tumour suppressor genes and to verify the occurrence of negative epistasis between them and their paralogs (36). Another useful feature of NCG is the comprehensive overview of gene expression profiles across a vast range of normal tissues and cancer cell lines. This can guide the selection of the most adequate cell systems for planning *in vitro* experiments (Figure 4B).

NCG is exploited widely as a repository of cancer genes (17,37–50). Examples include the use of NCG to test for the proximity of cancer genes to retrovirus insertion sites (48) and to evaluate the features of cancer classification methods (41). NCG also facilitates the interpretation of cancer mutational screenings by annotating the properties of mutated genes (Figure 4C) overall and in selected cancer types

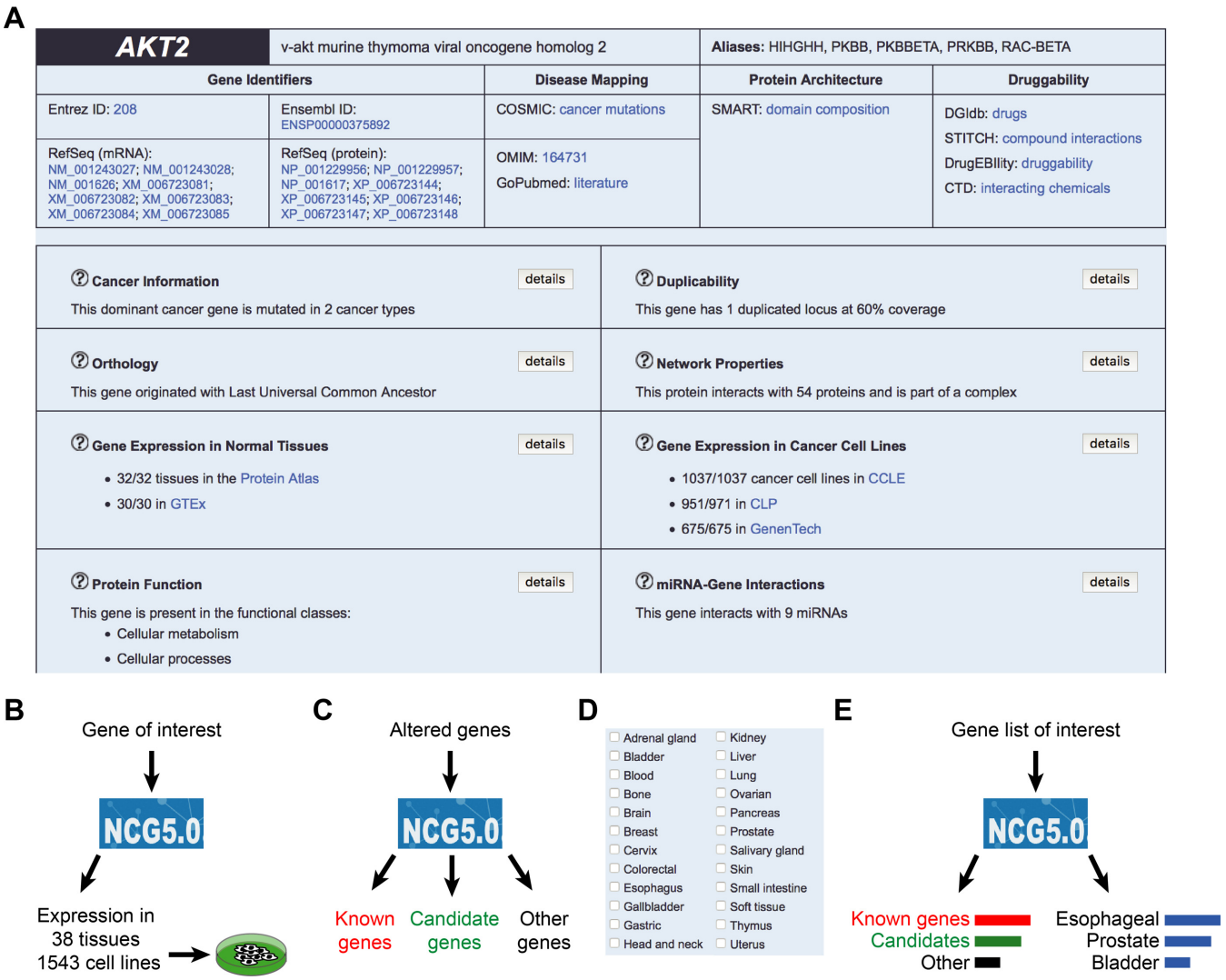


Figure 4. Examples of NCG usage: (**A**) Example of information available in NCG for a given cancer gene, in this case the oncogene *AKT2*. NCG summarizes the gene mutation profile across cancer types, information on duplicability, orthology, protein–protein and miRNA interactions and gene expression (**B**) NCG can facilitate the selection of the best cell systems for experimental assays by providing the expression profile of the gene of interest in several tissues and cell lines. (**C**) NCG can be used to annotate altered genes from mutational screenings. (**D**) The advanced search interface of NCG allows the identification of drivers in a variety of cancer types. (**E**) NCG can be integrated in gene enrichment analysis pipelines as a source of cancer genes.

(Figure 4D). For example, NCG has been used to verify whether genes undergoing copy number variations in familial breast cancer were already known cancer genes (49). Finally, NCG can be easily integrated into more complex analytical pipelines (Figure 4E). In the method developed by Zeller *et al.*, NCG provides a source of true cancer genes to prioritize drivers (50). In the DOSE bioconductor package, NCG is implemented as a source of cancer genes to perform enrichment analysis (51).

FUTURE WORK

It is expected that mutational screenings of cancer samples will continue to produce large amounts of data in the next years. The launch of personal genome initiatives ((52) and www.genomicsengland.co.uk) and the delivery of pan-cancer projects will substantially enlarge the spectrum of cancer types and samples with available mutational profiles. This will allow the discovery of novel cancer genes, particularly of those that recur in few samples and are currently difficult to identify. In parallel, the development of novel approaches for high-throughput functional screenings (e.g. based on the CRISPR-Cas technology (53–56)) promises to improve the efficiency of experimental validation assays.

In this exciting scenario, NCG will continue in its commitment to manually curate the literature to extract cancer genes and annotate available orthogonal supports. NCG will also expand to include other types of cancer driver alterations, such as copy number variations, gene rearrangements and non-coding modifications (57,58). In addition to enlarge the repertoire of cancer drivers, NCG will integrate new properties, e.g. the epigenetic regulation of cancer genes and their germline mutations.

As data become available, NCG will include the clinical relevance of cancer genes, such as their actionability

as pharmacological targets (59) and their applicability as biomarkers of cancer progression. All these efforts will contribute towards a more complete characterization of the molecular determinants of cancer.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENT

The authors thank Alex Mastrogiannopoulos for his help in the manual curation of the experimental validation of candidate cancer genes and all members of the Ciccarelli lab for providing suggestions to improve NCG.

FUNDING

European Union's Seventh Framework Programme [(FP7/2007-2013) under grant agreement No. 259743] (MODHEP consortium). The authors acknowledge support from the National Institute for Health Research (NIHR) Biomedical Research Centre based at Guy's and St Thomas' NHS Foundation Trust and King's College London.
Conflict of interest statement. None declared.

REFERENCES

1. Stratton,M.R., Campbell,P.J. and Futreal,P.A. (2009) The cancer genome. *Nature*, **458**, 719–724.
2. Garraway,L.A. and Lander,E.S. (2013) Lessons from the cancer genome. *Cell*, **153**, 17–37.
3. Vogelstein,B., Papadopoulos,N., Velculescu,V.E., Zhou,S., Diaz,L.A. Jr and Kinzler,K.W. (2013) Cancer genome landscapes. *Science*, **339**, 1546–1558.
4. 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–D675.
5. 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**, 177–183.
6. Lawrence,M.S., Stojanov,P., Polak,P., Kryukov,G.V., Cibulskis,K., Sivachenko,A., Carter,S.L., Stewart,C., Mermel,C.H., Roberts,S.A. *et al.* (2013) Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature*, **499**, 214–218.
7. Kandoth,C., McLellan,M.D., Vandin,F., Ye,K., Niu,B., Lu,C., Xie,M., Zhang,Q., McMichael,J.F., Wyczalkowski,M.A. *et al.* (2013) Mutational landscape and significance across 12 major cancer types. *Nature*, **502**, 333–339.
8. An,O., Pendino,V., D'Antonio,M., Ratti,E., Gentilini,M. and Ciccarelli,F.D. (2014) NCG 4.0: the network of cancer genes in the era of massive mutational screenings of cancer genomes. *Database*, **2014**, bau015.
9. Dees,N.D., Zhang,Q., Kandoth,C., Wendl,M.C., Schierding,W., Koboldt,D.C., Mooney,T.B., Callaway,M.B., Dooling,D., Mardis,E.R. *et al.* (2012) MuSiC: identifying mutational significance in cancer genomes. *Genome Res.*, **22**, 1589–1598.
10. Shimizu,A., Asakawa,S., Sasaki,T., Yamazaki,S., Yamagata,H., Kudoh,J., Minoshima,S., Kondo,I. and Shimizu,N. (2003) A novel giant gene CSMD3 encoding a protein with CUB and sushi multiple domains: a candidate gene for benign adult familial myoclonic epilepsy on human chromosome 8q23.3-q24.1. *Biochem. Biophys. Res. Commun.*, **309**, 143–154.
11. Liu,P., Morrison,C., Wang,L., Xiong,D., Vedell,P., Cui,P., Hua,X., Ding,F., Lu,Y., James,M. *et al.* (2012) Identification of somatic mutations in non-small cell lung carcinomas using whole-exome sequencing. *Carcinogenesis*, **33**, 1270–1276.
12. 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**, 427–430.
13. Pruitt,K.D., Brown,G.R., Hiatt,S.M., Thibaud-Nissen,F., Astashyn,A., Ermolaeva,O., Farrell,C.M., Hart,J., Landrum,M.J., McGarvey,K.M. *et al.* (2014) RefSeq: an update on mammalian reference sequences. *Nucleic Acids Res.*, **42**, D756–D763.
14. Powell,S., Forslund,K., Szklarczyk,D., Trachana,K., Roth,A., Huerta-Cepas,J., Gabaldon,T., Rattei,T., Creevey,C., Kuhn,M. *et al.* (2014) eggNOG v4.0: nested orthology inference across 3686 organisms. *Nucleic Acids Res.*, **42**, D231–D239.
15. D'Antonio,M. and Ciccarelli,F.D. (2011) Modification of gene duplicability during the evolution of protein interaction network. *PLoS Comput. Biol.*, **7**, e1002029.
16. Domazet-Loso,T. and Tautz,D. (2008) An ancient evolutionary origin of genes associated with human genetic diseases. *Mol. Biol. Evol.*, **25**, 2699–2707.
17. 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.
18. Chatr-Aryamontri,A., Breitkreutz,B.J., Oughtred,R., Boucher,L., Heinicke,S., Chen,D., Stark,C., Breitkreutz,A., Kolas,N., O'Donnell,L. *et al.* (2015) The BioGRID interaction database: 2015 update. *Nucleic Acids Res.*, **43**, D470–D478.
19. Orchard,S., Ammari,M., Aranda,B., Breuza,L., Briganti,L., Broackes-Carter,F., Campbell,N.H., Chavali,G., Chen,C., del-Toro,N. *et al.* (2014) The MIntAct project–IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res.*, **42**, D358–D363.
20. 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–D451.
21. 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–D772.
22. Ruepp,A., Waegele,B., Lechner,M., Brauner,B., Dunger-Kaltenbach,I., Fobo,G., Frishman,G., Montrone,C. and Mewes,H.W. (2010) CORUM: the comprehensive resource of mammalian protein complexes–2009. *Nucleic Acids Res.*, **38**, D497–D501.
23. Milacic,M., Haw,R., Rothfels,K., Wu,G., Croft,D., Hermjakob,H., D'Eustachio,P. and Stein,L. (2012) Annotating cancer variants and anti-cancer therapeutics in reactome. *Cancers*, **4**, 1180–1211.
24. Hsu,S.D., Tseng,Y.T., Shrestha,S., Lin,Y.L., Khaleel,A., Chou,C.H., Chu,C.F., Huang,H.Y., Lin,C.M., Ho,S.Y. *et al.* (2014) miRTarBase update 2014: an information resource for experimentally validated miRNA-target interactions. *Nucleic Acids Res.*, **42**, D78–D85.
25. 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–D110.
26. Melé,M., Ferreira,P.G., Reverter,F., DeLuca,D.S., Monlong,J., Sammeth,M., Young,T.R., Goldmann,J.M., Pervouchine,D.D., Sullivan,T.J. *et al.* (2015) Human genomics. The human transcriptome across tissues and individuals. *Science*, **348**, 660–665.
27. Uhlén,M., Fagerberg,L., Hallstrom,B.M., Lindskog,C., Oksvold,P., Mardinoglu,A., Sivertsson,A., Kampf,C., Sjostedt,E., Asplund,A. *et al.* (2015) Proteomics. Tissue-based map of the human proteome. *Science*, **347**, 1260419.
28. Barretina,J., Caponigro,G., Stransky,N., Venkatesan,K., Margolin,A.A., Kim,S., Wilson,C.J., Lehar,J., Kryukov,G.V., Sonkin,D. *et al.* (2012) The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, **483**, 603–607.
29. Garnett,M.J., Edelman,E.J., Heidorn,S.J., Greenman,C.D., Dastur,A., Lau,K.W., Greninger,P., Thompson,I.R., Luo,X., Soares,J. *et al.* (2012) Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature*, **483**, 570–575.
30. Klijn,C., Durinck,S., Stawiski,E.W., Haverty,P.M., Jiang,Z., Liu,H., Degenhardt,J., Mayba,O., Gnad,F., Liu,J. *et al.* (2015) A comprehensive transcriptional portrait of human cancer cell lines. *Nat. Biotechnol.*, **33**, 306–312.

31. D'Antonio,M. and Ciccarelli,F.D. (2013) Integrated analysis of recurrent properties of cancer genes to identify novel drivers. *Genome Biol.*, **14**, R52.
32. Letunic,I., Doerks,T. and Bork,P. (2015) SMART: recent updates, new developments and status in 2015. *Nucleic Acids Res.*, **43**, D257–D260.
33. Griffith,M., Griffith,O.L., Coffman,A.C., Weible,J.V., McMichael,J.F., Spies,N.C., Koval,J., Das,I., Callaway,M.B., Eldred,J.M. *et al.* (2013) DGIdb: mining the druggable genome. *Nat. Methods*, **10**, 1209–1210.
34. Kuhn,M., Szklarczyk,D., Pletscher-Frankild,S., Blicher,T.H., von Mering,C., Jensen,L.J. and Bork,P. (2014) STITCH 4: integration of protein-chemical interactions with user data. *Nucleic Acids Res.*, **42**, D401–D407.
35. Bento,A.P., Gaulton,A., Hersey,A., Bellis,L.J., Chambers,J., Davies,M., Kruger,F.A., Light,Y., Mak,L., McGlinchey,S. *et al.* (2014) The ChEMBL bioactivity database: an update. *Nucleic Acids Res.*, **42**, D1083–D1090.
36. D'Antonio,M., Guerra,R.F., Cereda,M., Marchesi,S., Montani,F., Nicassio,F., Di Fiore,P.P. and Ciccarelli,F.D. (2013) Recessive cancer genes engage in negative genetic interactions with their functional paralogs. *Cell Rep.*, **5**, 1519–1526.
37. Cheng,F., Jia,P., Wang,Q., Lin,C.C., Li,W.H. and Zhao,Z. (2014) Studying tumorigenesis through network evolution and somatic mutational perturbations in the cancer interactome. *Mol. Biol. Evol.*, **31**, 2156–2169.
38. Haemmerle,R., Phaltane,R., Rothe,M., Schroder,S., Schambach,A., Moritz,T. and Modlich,U. (2014) Clonal dominance with retroviral vector insertions near the ANGPT1 and ANGPT2 genes in a human xenotransplant mouse model. *Mol. Ther. Nucleic Acids*, **3**, e200.
39. Liu,W. and Xie,H. (2013) Predicting potential cancer genes by integrating network properties, sequence features and functional annotations. *Sci. China Life Sci.*, **56**, 751–757.
40. Liu,Y., Tian,F., Hu,Z. and DeLisi,C. (2015) Evaluation and integration of cancer gene classifiers: identification and ranking of plausible drivers. *Sci. Rep.*, **5**, 10204.
41. List,M., Hauschild,A.C., Tan,Q., Kruse,T.A., Mollenhauer,J., Baumbach,J. and Batra,R. (2014) Classification of breast cancer subtypes by combining gene expression and DNA methylation data. *J. Integr. Bioinform.*, **11**, 236.
42. Nayak,L., Tunga,H. and De,R.K. (2013) Disease co-morbidity and the human Wnt signaling pathway: a network-wise study. *OMICS*, **17**, 318–337.
43. Phaltane,R., Haemmerle,R., Rothe,M., Modlich,U. and Moritz,T. (2014) Efficiency and safety of O(6)-methylguanine DNA methyltransferase (MGMT(P140K))-mediated in vivo selection in a humanized mouse model. *Hum. Gene Ther.*, **25**, 144–155.
44. Saadatian,Z., Masotti,A., Nariman Saleh Fam,Z., Alipoor,B., Bastami,M. and Ghaedi,H. (2014) Single-nucleotide polymorphisms within micrornas sequences and their 3' UTR target sites may regulate gene expression in gastrointestinal tract cancers. *Iran Red Crescent Med. J.*, **16**, e16659.
45. Srivastava,A., Kumar,S. and Ramaswamy,R. (2014) Two-layer modular analysis of gene and protein networks in breast cancer. *BMC Syst. Biol.*, **8**, 81.
46. Shangguan,H., Tan,S.Y. and Zhang,J.R. (2015) Bioinformatics analysis of gene expression profiles in hepatocellular carcinoma. *Eur. Rev. Med. Pharmacol. Sci.*, **19**, 2054–2061.
47. Yu,H., Mitra,R., Yang,J., Li,Y. and Zhao,Z. (2014) Algorithms for network-based identification of differential regulators from transcriptome data: a systematic evaluation. *Sci. China Life Sci.*, **57**, 1090–1102.
48. Olszko,M.E., Adair,J.E., Linde,I., Rae,D.T., Trobridge,P., Hocum,J.D., Rawlings,D.J., Kiem,H.P. and Trobridge,G.D. (2015) Foamy viral vector integration sites in SCID-repopulating cells after MGMTP140K-mediated in vivo selection. *Gene Ther.*, **22**, 591–595.
49. Masson,A.L., Talseth-Palmer,B.A., Evans,T.J., Grice,D.M., Hannan,G.N. and Scott,R.J. (2014) Expanding the genetic basis of copy number variation in familial breast cancer. *Hered. Cancer Clin. Pract.*, **12**, 15.
50. Zeller,M., Magnan,C.N., Patel,V.R., Rigor,P., Sender,L. and Baldi,P. (2014) A genomic analysis pipeline and its application to pediatric cancers. *IEEE/ACM Trans. Comput. Biol. Bioinformatics*, **11**, 826–839.
51. Yu,G., Wang,L.G., Yan,G.R. and He,Q.Y. (2015) DOSE: an R/Bioconductor package for disease ontology semantic and enrichment analysis. *Bioinformatics*, **31**, 608–609.
52. Collins,F.S. and Varmus,H. (2015) A new initiative on precision medicine. *N. Engl. J. Med.*, **372**, 793–795.
53. Shalem,O., Sanjana,N.E., Hartenian,E., Shi,X., Scott,D.A., Mikkelsen,T.S., Heckl,D., Ebert,B.L., Root,D.E., Doench,J.G. *et al.* (2014) Genome-scale CRISPR-Cas9 knockout screening in human cells. *Science*, **343**, 84–87.
54. Gilbert,L.A., Horlbeck,M.A., Adamson,B., Villalta,J.E., Chen,Y., Whitehead,E.H., Guimaraes,C., Panning,B., Ploegh,H.L., Bassik,M.C. *et al.* (2014) Genome-scale CRISPR-mediated control of gene repression and activation. *Cell*, **159**, 647–661.
55. Wang,T., Wei,J.J., Sabatini,D.M. and Lander,E.S. (2014) Genetic screens in human cells using the CRISPR-Cas9 system. *Science*, **343**, 80–84.
56. Ran,F.A., Hsu,P.D., Wright,J., Agarwala,V., Scott,D.A. and Zhang,F. (2013) Genome engineering using the CRISPR-Cas9 system. *Nat. Protoc.*, **8**, 2281–2308.
57. Borah,S., Xi,L., Zaug,A.J., Powell,N.M., Dancik,G.M., Cohen,S.B., Costello,J.C., Theodorescu,D. and Cech,T.R. (2015) Cancer. TERT promoter mutations and telomerase reactivation in urothelial cancer. *Science*, **347**, 1006–1010.
58. Vinagre,J., Almeida,A., Populo,H., Batista,R., Lyra,J., Pinto,V., Coelho,R., Celestino,R., Prazeres,H., Lima,L. *et al.* (2013) Frequency of TERT promoter mutations in human cancers. *Nat. Commun.*, **4**, 2185.
59. McGranahan,N. and Swanton,C. (2015) Biological and therapeutic impact of intratumor heterogeneity in cancer evolution. *Cancer Cell*, **27**, 15–26.

Database, Vol. 2014, Article ID bau015, doi:10.1093/database/bau015

Database update

NCG 4.0: the network of cancer genes in the era of massive mutational screenings of cancer genomes

Omer An[1], Vera Pendino[1], Matteo D'Antonio[1], Emanuele Ratti[1], Marco Gentilini[1] and Francesca D. Ciccarelli[1,2,*]

[1]Department of Experimental Oncology, European Institute of Oncology, IFOM-IEO Campus, Via Adamello 16, 20139 Milan, Italy and [2]Division of Cancer Studies, King's College London, London SE1 1UL, UK

***Corresponding author:** Tel: +44 (0)20 7848 6616; Fax: +44 (0)20 7848 6220; Email: francesca.ciccarelli@kcl.ac.uk

Submitted 29 November 2013; Revised 10 January 2014; Accepted 2 February 2014

Citation details: An,O., Pendino,V., D'Antonio,M., *et al.* NCG 4.0: the network of cancer genes in the era of massive mutational screenings of cancer genomes. *Database* (2014) Vol. 2014: article ID bau015; doi:10.1093/database/bau015.

NCG 4.0 is the latest update of the Network of Cancer Genes, a web-based repository of systems-level properties of cancer genes. In its current version, the database collects information on 537 known (i.e. experimentally supported) and 1463 candidate (i.e. inferred using statistical methods) cancer genes. Candidate cancer genes derive from the manual revision of 67 original publications describing the mutational screening of 3460 human exomes and genomes in 23 different cancer types. For all 2000 cancer genes, duplicability, evolutionary origin, expression, functional annotation, interaction network with other human proteins and with microRNAs are reported. In addition to providing a substantial update of cancer-related information, NCG 4.0 also introduces two new features. The first is the annotation of possible false-positive cancer drivers, defined as candidate cancer genes inferred from large-scale screenings whose association with cancer is likely to be spurious. The second is the description of the systems-level properties of 64 human microRNAs that are causally involved in cancer progression (oncomiRs). Owing to the manual revision of all information, NCG 4.0 constitutes a complete and reliable resource on human coding and non-coding genes whose deregulation drives cancer onset and/or progression. NCG 4.0 can also be downloaded as a free application for Android smart phones.

Database URL: http://bio.ieo.eu/ncg/

Introduction

Sequencing of exomes and genomes from thousands of cancer samples led to the identification of an increasing number of mutated genes that may contribute to driving human cancer (1–3). Owing to the massive amount of information derived from these studies, it is often difficult to get an overview of the genes that play a driver role in cancer on mutation (cancer genes). Since 2010, the Network of Cancer Genes (NCG) has been collecting information on a manually curated list of known and candidate cancer genes (4, 5). Known cancer genes have robust experimental support on their role in cancer onset and progression. Candidate cancer genes instead derive from large-scale mutational screenings of cancer samples and have been identified using statistical methods with poor or no experimental follow-up. Candidate cancer genes are thus prone to include false positives as a consequence of the difficult discrimination between passenger and driver mutations (6, 7). To account for this, NCG 4.0 reports a list of candidate cancer genes whose association with cancer is likely to be spurious owing to function, length and literature evidence.

For each known and candidate cancer gene, NCG 4.0 annotates a series of systems-level properties, defined as features that distinguish a group of genes (in this case, cancer-related genes) from the rest, and that cannot be

Database, Vol. 2014, Article ID bau015, doi:10.1093/database/bau015

ascribed to the function of the single gene alone (8). Systems-level properties currently reported in NCG are of evolutionary origin and duplicability, primary and secondary interaction network of the encoded proteins and miRNA regulatory networks. In addition, NCG 4.0 provides information on gene expression in 109 human tissues and on their functional characterization based on Gene Ontology (9). Owing to the increasing evidence of the primary role of microRNA (miRNA) deregulation in the onset of human cancer (10, 11), NCG 4.0 also annotates the systems-level properties of 64 cancer-related miRNAs (oncomiRs) manually derived from the literature.

Compared with other databases collecting all cancer mutations, such as COSMIC (12), ICGC (13) and CGAP (14), NCG 4.0 provides the community with a manually reviewed and constantly updated repository only of cancer drivers. In addition, it also annotates the properties of these genes, thus resulting useful to address different types of questions regarding cancer determinants (Figure 1) and to mine the increasing amount of information on cancer mutations.

Database Description and Updates

Manual collection of cancer genes

NCG 4.0 annotates the properties of 2000 cancer genes, defined as genes that contribute in promoting the onset and/or the development of human cancer. This list is derived from the union of two datasets. The first combined a literature-based repository of 484 genes from the Cancer Gene Census (377 dominant, 111 recessive and 4 genes that can act as both dominant and recessive, as frozen in January 2013) (15) with 77 genes whose amplification is causally implicated in cancer (16). This led to 537 experimentally supported cancer genes, which we defined as 'known cancer genes'. The second dataset consisted of 1463 genes that are likely to be involved in cancer development on mutation, which we defined as 'candidate

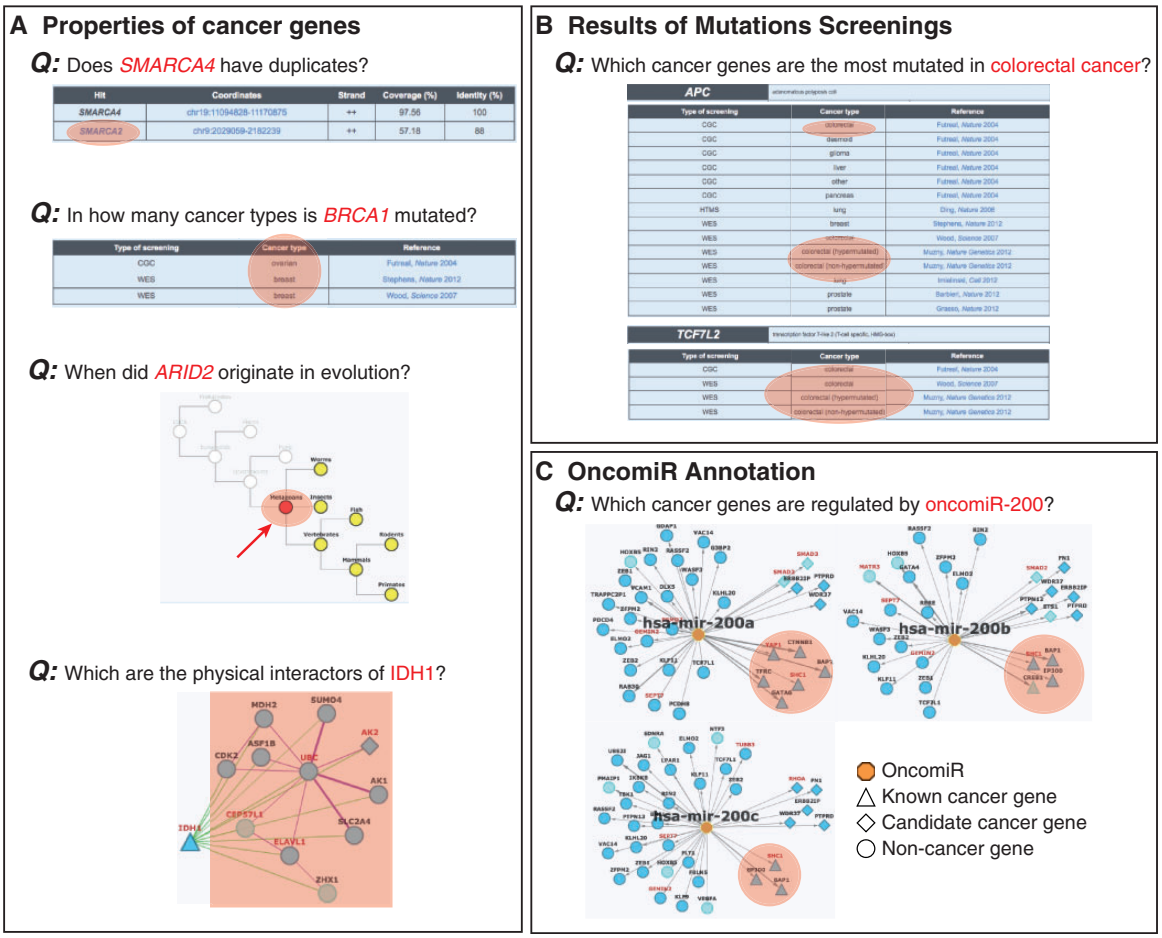


Figure 1. Examples of queries that can be done in NCG. Information stored in NCG can be used to address different queries regarding the properties of (**A**) individual cancer genes, (**B**) cancer types and (**C**) oncomiRs. Relevant information to address the specific queries is highlighted in orange.

cancer genes'. These genes derived from the manual revision of 67 publications corresponding to 77 re-sequencing screenings of the whole exomes (49 screenings), the whole genomes (19 screenings) and selected gene sets (9 screenings), conducted on 3640 samples from 23 cancer types (Supplementary Table S1) (17–83). These papers represented a comprehensive set of high-throughput cancer re-sequencing screenings.

Compared with the previous version, NCG 4.0 appreciably increased the number of cancer genes, particularly candidates, and of sequenced samples (Figure 2A). Such accretion of knowledge reflects the current massive worldwide efforts to characterize cancer mutational landscapes in detail. Although we are expected to reach a plateau in the discovery of new driver genes because genes frequently (and significantly) mutated in some cancer types are also mutated at low frequency in other cancer types (1), our data show that we are still in the growing phase. In particular, for most cancer types the number of new candidate cancer genes increases with the number of sequenced samples (Figure 2B). As already noticed (1, 6), most cancer genes, and in particular candidates, are specific for a given cancer type, and only few known cancer genes recurrently mutate in several cancers (Figure 2C). This observation once again confirms the heterogeneity of cancer mutation landscape (3).

Human gene set and orthology information

To identify the list of unique human genes, we aligned 33 427 protein sequences from RefSeq v.51 (84) to the reference human genome Hg19, using a method previously developed by our group (5, 8). This led to the identification of 19 045 unique gene loci, including 1961 of the 2000 cancer genes. Of the remaining 39 cancer genes, 29 did not have RefSeq protein entries and 10 were discarded because their protein sequences aligned to the genome for <60% of their length. For each cancer gene we retrieved

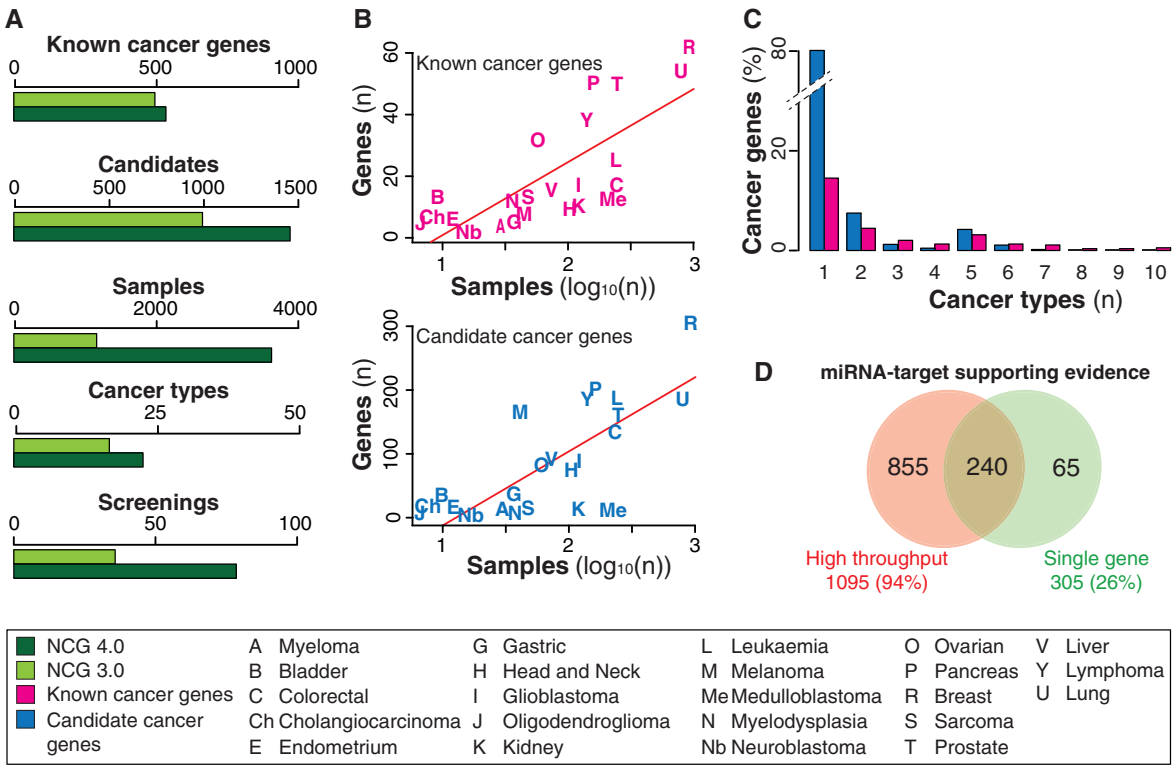


Figure 2. Overview of the data collected in NCG 4.0. (**A**) Comparison of data stored in NCG 3.0 and NCG 4.0. (**B**) Linear regression curves between the number of known and candidate cancer genes and the number of sequenced samples in each cancer type. Some cancer types deviate from linearity and this can be due to different reasons. For example, melanoma has a high number of candidate cancer genes (169) despite the low number of sequenced samples (41). In this case, the most likely explanation is that most of these candidate genes derive from two screenings (61, 75) that did not apply any methods to identify cancer drivers (Table 1, Supplementary Table S1). In the case of medulloblastoma, candidate and known cancer genes are only 25 despite 211 samples having been screened. This likely depends on the low mutation frequency of medulloblastoma [<1 mutation/Mb (40, 57, 64, 67)]. (**C**) Recurrence of known and candidate cancer genes in different cancer types. The only cancer genes that have been found mutated in more than 10 different cancer types are *TP53* (20 cancer types), *PIK3CA* (13 cancer types) and *PTEN* (12 cancer types). (**D**) Comparison of cancer miRNA targets that have been identified using single gene (i.e. reporter assay, western blot) and high throughput approaches (i.e. microarray, proteomic experiments and next-generation sequencing).

Database, Vol. 2014, Article ID bau015, doi:10.1093/database/bau015

duplicability, evolutionary origin, functional annotation, gene expression profile, protein–protein interaction and gene-microRNA interaction.

We assessed gene duplicability by the presence of one or more additional hits on the genome covering at least 60% of cancer protein length (8). Of the 1961 cancer genes, 325 (17%) had at least one extra copy on the genome. This was a significantly lower fraction compared with the rest of human genes (21%, P-value = 7.8×10^{-06}, chi-square test), thus confirming the tendency of cancer genes to preserve a singleton status in the genome (8).

We assessed orthology relationships for 1978 of the 2000 cancer genes annotated in EggNOG v.3.0 (85) and used this information to infer the evolutionary origin of each cancer gene, defined as the most ancient node of the tree of life where the ortholog for that gene could be found (86). As already reported (86, 87), we confirmed that the fraction of old cancer genes that originated in prokaryotes and unicellular eukaryotes (1500, 76% of the total) was higher than in the rest of human genes (68%, P-value = 6.1×10^{-13}, chi-square test). Moreover, we also confirmed that recessive cancer genes are older than dominant cancer genes (4). The vast majority of recessive cancer genes (87/111, 78%) originated already with the last universal common ancestor or with unicellular eukaryotes, compared with only 67% of dominant cancer genes (P-value = 0.03, chi-square test).

Protein–protein and miRNA-target interaction networks

We rebuilt the human protein–protein interaction network integrating direct experimental evidence from five sources: HPRD (frozen on 13 April 2010) (88), BioGRID v.3.2.96 (89), IntAct v.159 (frozen on 14 December 2012) (90), MINT (frozen on 26 October 2012) (91) and DIP (frozen on 10 October 2010) (92). This resulted in a global network of 16 241 proteins (nodes) and 164 008 binary interactions (edges), supported by 33 497 independent publications. Interaction data were available for 1706 cancer proteins, and hubs (defined as proteins with at least 15 interactions) constituted 45% of all cancer genes, compared with 30% of the rest of human genes (P-value = 3.60 10^{-38}, chi-square test).

The interaction network between miRNAs and cancer genes relied on experimental data extracted from three different sources: TarBase v.5.0 (93), miRecords v.4.0 (94) and miRTarBase v.4.4 (95). The integration of these data led to 1160 cancer targets of miRNAs (58% of the total). This was a significantly higher proportion compared with the rest of human genes (48%, P-value = 1.02×10^{-17}, chi-square test) and confirmed the tendency of cancer genes to be regulated by miRNAs (4). This enrichment may reflect the fact that cancer genes are overall better characterized and thus more information is available on them. However, >70% of miRNA targets have been identified through high-throughput screenings (such as microarray, mass spectrometry and sequencing, Figure 2D), thus partially reducing the bias. Finally, we also updated the list of cancer genes that host miRNAs within their genomic loci (87 genes, 4.4% of the total).

Novel Features of NCG 4.0

Identification of possible false cancer genes

With the increasing evidence of an overwhelming number of mutations acquired during cancer progression (most of which with no role in the disease), a number of statistical methods have been developed to identify cancer drivers within the whole set of mutated genes. These methods take into account several features including the tendency of the same gene to be mutated across many samples, the cancer-specific background mutation rate, the gene length and expression and the mutation effect on the encoded protein (Table 1, Supplementary Table S1). Despite all efforts to refine the identification of driver mutations, current approaches are still prone to false positives, i.e. mutated genes that are erroneously identified as cancer drivers (6, 7). For example, genes encoding olfactory receptors are often included in the list of candidates, because they tend to mutate although the biological function and expression pattern of these genes strongly dismiss a possible functional role in the disease. Similarly, overly long genes are also probable false positives because their recurrent mutations in several samples are most likely due to their length more than to their function (6, 7). Because the main goal of NCG is to annotate the properties of cancer genes, we decided to collect all putative cancer genes from primary data without removing possible false positives. However, we added a warning concerning the possible spurious cancer associations for 60 genes (39 olfactory receptors, 14 genes with long exons and/or introns and 7 additional false positives derived from literature (7) (Figure 3A, Table 1). Although gene length by itself does not imply spurious associations, we derived the length distributions of all candidate cancer genes and considered genes with long introns (Figure 3B) or long exons (Figure 3C) as possible false positives.

Gene expression profiles

To complete the functional annotation of cancer genes, we derived expression levels for 1528 of them from two high-throughput gene expression experiments on 109 human tissues (99, 100). We normalized and processed the raw CEL files obtained from the corresponding Gene Expression Omnibus series (GSE2361 and GSE1133) using the MAS5 algorithm of the R *affy* package (101, 102). Because more than one probe can be associated with one gene, the expression level of each cancer gene in a given

Table 1. Methods used to identify candidate cancer genes and possible false positives

Method	MuSiC (96)	Mutsig (7)	Wood *et al.* (80, 97)	Greenman *et al.* (98)	Paper-specific	Recurrence-based	None
Candidate cancer genes	Genes that mutate with significantly higher rate than the background, considering multiple mutational mechanisms. It allows for pathway and proximity analysis, clinical correlation test and PFAM/OMIM query	Genes that mutate more often than expected, given the background mutation rate. It clusters mutations in hotspots and considers the functional impact and the conservation of the genomic site. The latest version takes into account patient and genomic mutation patterns	Genes that (a) mutate in both discovery and validation screens; (b) whose mutations exceed a certain threshold and; (c) mutate at a frequency higher than the passenger mutation rate	Genes that mutate at higher frequency than expected. Expectation is estimated using silent mutations	*Ad hoc* methodology developed for the specific set of samples and cancer type analyzed in the paper	Recurrence of mutations in a gene within samples is taken as evidence of its causal involvement in disease onset. Particularly used when few samples and/or cancer types with low mutation instability are analyzed	Often associated to whole genome screening, when only one or very few samples are sequenced. In such cases, all mutated genes are retained as possible candidates
Number of screenings	5	17	13	3	10	17	12
Possible false positives	*6 (LRP1B, OR6A2, OR11L1, OR5B17, OR10G7, RYR2)*	*17 (CNTNAP2, CSMD3, LRP1B, ORC4C15, OR8H2, OR8K1, OR6K3, OR5L2, OR2T33, PCLO, LRP2, MUC4, NEB, RYR2, SYNE1, SYNE2, TTN)*	*9 (CCDC168, CNTNAP2, CSMD3, EYS, LRP2, MUC16, OR2L13, OR51E1, TTN)*	*None*	*15 (CSMD3, CNTNAP2, OR4L1, OR10G9, OR5L1, OR4K14, OR4C13, OR4C6, OR51L2, OR1M1, OR2A42, OR10AG1, OR2A2, OR4K1, OR52E8)*	*13 (CNTN5, CSMD3, DMD, LRP1B, FSIP2, OR2M4, OR10R2, OR1L8, OR4C46, PCLO, RYR2, SYNE1, TTN)*	*17 (CTNNA3, CNTN5, CSMD1, OR2T11, OR2T34, OR5M5P, OR4S2, OBSCN, OR1J2, OR4D11, OR5H2, OR4Q3, OR4N5, OR52A5, RYR3, SYNE2, TTN)*

For each method used to identify candidate cancer genes (i.e. new possible cancer drivers) in the 77 screenings, reported are a brief description of the procedure, the number of screenings that relied on it and the associated possible false positives.

Database, Vol. 2014, Article ID bau015, doi:10.1093/database/bau015

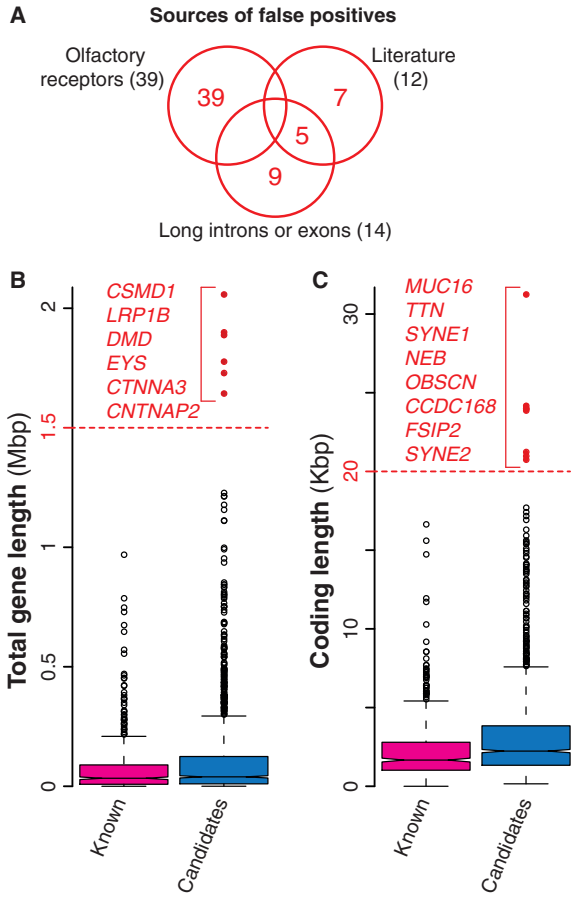


Figure 3. Possible false positives among candidate cancer drivers. (**A**) Venn diagram of the three groups of possible false positives. In total, we identified 60 genes, 65% of which were olfactory receptors, 23% were long genes and the remaining 20% were derived from literature (7). (**B**) Distribution of the total length for known and candidate cancer genes. Total gene length was measured as total number of nucleotides spanning the entire gene locus, including exons and introns. Red dots indicate possible false positives (gene longer than 1.5 Mb). (**C**) Length distribution of the coding regions for known and candidate cancer genes computed as the number of nucleotides covering the coding exons. Genes longer than 20 Kb (red dots) were considered as possible false positives.

tissue was defined as the mean expression levels of all probes with detection $P < 0.05$. If all probes for a gene had detection $P > 0.05$, the gene was considered as not expressed.

To make a comparative assessment of the expression levels of a cancer gene i in a given tissue t with those of all other genes in the same tissue, we first calculated the expression levels of all human genes in that tissue. We then derived the normalized expression level n of the cancer gene i in the tissue t, measured as:

$$n_{i,t} = \frac{(e_{i,t} - E_t)}{(e_{i,t} + E_t)}$$

where $e_{i,t}$ was the expression level of the cancer gene i in tissue t and E_t was the median expression level of all genes in tissue t. Normalized expression levels allowed a direct comparison of the expression of all genes in each given tissue.

Manual collection of miRNAs involved in human cancer (oncomiRs)

We manually gathered the list of oncomiRs from the literature and included only miRNA families (i.e. miRNAs with high sequence similarity) and miRNA clusters (i.e. miRNAs that are neighbors in the genome and co-transcribed) whose role in cancer was well described and experimentally supported (103–108). This led to 64 oncomiRs involved in 27 cancer types. Similarly to protein-coding genes we retrieved details on duplicability, evolutionary origin and interaction network for all these oncomiRs.

To infer oncomiR duplicability, we downloaded 1424 human miRNAs from miRBase v.17 (109) and considered all mature miRNAs with the same seed (i.e. the 6–8 nt-long region at the 5′-end of the sequence) as duplicated miRNAs. The rationale for this choice was that, because seeds determine the specificity in target recognition, their sequences are the most conserved among homologous miRNAs (110). Among 64 oncomiRs, 51 (79%) were duplicated compared with 33% other duplicated human miRNAs ($P = 4.5 \times 10^{-16}$, chi-square test). Therefore, unlike protein-coding cancer genes that maintain a singleton status in the genome, oncomiRs tend to have additional copies that share the site of recognition of the RNA targets.

To pinpoint when oncomiRs appeared in evolution, we developed a procedure similar to that used for protein-coding genes and traced the most ancient miRNA ortholog. We first retrieved the orthologs of 835 human miRNAs for which miRNA families were available in miRBase (including all 64 oncomiRs). We then assigned the origin of each miRNA as the most ancient ortholog within the corresponding family. Sixty oncomiRs (94% of the total) had orthologs in vertebrates, compared with only 19% of the rest of human miRNAs, thus suggesting that oncomiRs originated earlier than the rest of human miRNAs. It is worth noticing that the marked differences in duplicability and origin between oncomiRs and other human miRNAs are at least partly inflated by the high interest in oncomiRs that boosted the search of their paralogs and orthologs in other species.

Web Interface, Implementation and Data Availability

NCG 4.0 runs on an Apache web server and data are stored in a MySQL database. The web interface was developed in

PHP and network visualization was implemented in Cytoscape Web (http://cytoscapeweb.cytoscape.org/) (111).

We modified NCG 4.0 web interface to enhance functionalities and facilitate the retrieval of the properties of cancer genes and oncomiRs. In addition to searching for single genes or list of genes of interest, the user can now visualize and browse all 2000 cancer genes, as well as retrieve cancer genes based on specific filters. NCG 4.0 also provides a detailed report on the cancer types and the corresponding publications where it was found mutated. Similar types of searches can be done on the 64 oncomiRs.

All data stored in NCG 4.0 are summarized in the statistics section that provides an overview on the properties of cancer genes and oncomiRs. For example, it is possible to compare mutation frequency, number of cancer genes and oncomiRs as well as their recurrence across the different cancer types and screenings. The bulk content of the database as well as the list of cancer genes, false positives and oncomiRs can be downloaded as text files. We developed a mobile phone application for NCG 4.0 that is freely available from the Web site.

Supplementary Data

Supplementary data are available at *Database* online.

Acknowledgements

The authors thank all members of the Ciccarelli laboratory for testing the database and providing useful suggestions to improve it, and Alessandro Ogier for his help in implementing the web interface and the mobile application.

Funding

Associazione Italiana Ricerca sul Cancro [AIRC-IG 12742] and Italian Ministry of Health [Grant Giovani Ricercatori 2010] to F.D.C. Funding for open access charge: Associazione Italiana Ricerca sul Cancro [AIRC-IG 12742].

Conflict of interest. None declared.

References

1. Garraway,L.A. and Lander,E.S. (2013) Lessons from the cancer genome. *Cell.*, **153**, 17–37.
2. Stratton,M.R., Campbell,P.J. and Futreal,P.A. (2009) The cancer genome. *Nature.*, **458**, 719–724.
3. Vogelstein,B., Papadopoulos,N., Velculescu,V.E. *et al*. (2013) Cancer genome landscapes. *Science.*, **339**, 1546–1558.
4. D'Antonio,M., Pendino,V., Sinha,S. *et al*. (2012) Network of cancer genes (NCG 3.0): integration and analysis of genetic and network properties of cancer genes. *Nucleic Acids Res.*, **40**, D978–D983.
5. 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–D675.
6. D'Antonio,M. and Ciccarelli,F.D. (2013) Integrated analysis of recurrent properties of cancer genes to identify novel drivers. *Genome Biol.*, **14**, R52.
7. Lawrence,M.S., Stojanov,P., Polak,P. *et al*. (2013) Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature.*, **499**, 214–218.
8. Rambaldi,D., Giorgi,F.M., Capuani,F. *et al*. (2008) Low duplicability and network fragility of cancer genes. *Trends Genet.*, **24**, 427–430.
9. The Gene Ontology Consortium, (2013) Gene ontology annotations and resources. *Nucleic Acids Res.*, **41**, D530–D535.
10. Calin,G.A. and Croce,C.M. (2006) MicroRNA-cancer connection: the beginning of a new tale. *Cancer Res.*, **66**, 7390–7394.
11. Croce,C.M. (2009) Causes and consequences of microRNA dysregulation in cancer. *Nat. Rev. Genet.*, **10**, 704–714.
12. Forbes,S.A., Bindal,N., Bamford,S. *et al*. (2011) COSMIC: mining complete cancer genomes in the catalogue of somatic mutations in Cancer. *Nucleic Acids Res.*, **39**, D945–D950.
13. Zhang,J., Baran,J., Cros,A. *et al*. (2011) International cancer genome consortium data portal—a one-stop shop for cancer genomics data. *Database.*, **2011**, bar026.
14. Riggins,G.J. and Strausberg,R.L. (2001) Genome and genetic resources from the cancer genome anatomy project. *Hum. Mol. Genet.*, **10**, 663–667.
15. Futreal,P.A., Coin,L., Marshall,M. *et al*. (2004) A census of human cancer genes. *Nat. Rev. Cancer.*, **4**, 177–183.
16. Santarius,T., Shipley,J., Brewer,D. *et al*. (2010) A census of amplified and overexpressed human cancer genes. *Nat. Rev. Cancer.*, **10**, 59–64.
17. Agrawal,N., Frederick,M.J., Pickering,C.R. *et al*. (2011) Exome sequencing of head and neck squamous cell carcinoma reveals inactivating mutations in NOTCH1. *Science.*, **333**, 1154–1157.
18. Banerji,S., Cibulskis,K., Rangel-Escareno,C. *et al*. (2012) Sequence analysis of mutations and translocations across breast cancer subtypes. *Nature.*, **486**, 405–409.
19. Barbieri,C.E., Baca,S.C., Lawrence,M.S. *et al*. (2012) Exome sequencing identifies recurrent SPOP, FOXA1 and MED12 mutations in prostate cancer. *Nat. Genet.*, **44**, 685–689.
20. Barretina,J., Taylor,B.S., Banerji,S. *et al*. (2010) Subtype-specific genomic alterations define new targets for soft-tissue sarcoma therapy. *Nat. Genet.*, **42**, 715–721.
21. Berger,M.F., Hodis,E., Heffernan,T.P. *et al*. (2012) Melanoma genome sequencing reveals frequent PREX2 mutations. *Nature.*, **485**, 502–506.
22. Berger,M.F., Lawrence,M.S., Demichelis,F. *et al*. (2011) The genomic complexity of primary human prostate cancer. *Nature.*, **470**, 214–220.
23. Bettegowda,C., Agrawal,N., Jiao,Y. *et al*. (2011) Mutations in CIC and FUBP1 contribute to human oligodendroglioma. *Science.*, **333**, 1453–1455.
24. Biankin,A.V., Waddell,N., Kassahn,K.S. *et al*. (2012) Pancreatic cancer genomes reveal aberrations in axon guidance pathway genes. *Nature.*, **491**, 399–405.
25. Chapman,M.A., Lawrence,M.S., Keats,J.J. *et al*. (2011) Initial genome sequencing and analysis of multiple myeloma. *Nature.*, **471**, 467–472.

26. Clark,M.J., Homer,N., O'Connor,B.D. *et al*. (2010) U87MG decoded: the genomic sequence of a cytogenetically aberrant human cancer cell line. *PLoS Genet.*, **6**, e1000832.
27. Dalgliesh,G.L., Furge,K., Greenman,C. *et al*. (2010) Systematic sequencing of renal carcinoma reveals inactivation of histone modifying genes. *Nature.*, **463**, 360–363.
28. Ding,L., Ellis,M.J., Li,S. *et al*. (2010) Genome remodelling in a basal-like breast cancer metastasis and xenograft. *Nature.*, **464**, 999–1005.
29. Ding,L., Getz,G., Wheeler,D.A. *et al*. (2008) Somatic mutations affect key pathways in lung adenocarcinoma. *Nature.*, **455**, 1069–1075.
30. Fujimoto,A., Totoki,Y., Abe,T. *et al*. (2012) Whole-genome sequencing of liver cancers identifies etiological influences on mutation patterns and recurrent mutations in chromatin regulators. *Nat. Genet.*, **44**, 760–764.
31. Grasso,C.S., Wu,Y.M., Robinson,D.R. *et al*. (2012) The mutational landscape of lethal castration-resistant prostate cancer. *Nature.*, **487**, 239–243.
32. Greif,P.A., Eck,S.H., Konstandin,N.P. *et al*. (2011) Identification of recurring tumor-specific somatic mutations in acute myeloid leukemia by transcriptome sequencing. *Leukemia.*, **25**, 821–827.
33. Gui,Y., Guo,G., Huang,Y. *et al*. (2011) Frequent mutations of chromatin remodeling genes in transitional cell carcinoma of the bladder. *Nat. Genet.*, **43**, 875–878.
34. Guichard,C., Amaddeo,G., Imbeaud,S. *et al*. (2012) Integrated analysis of somatic mutations and focal copy-number changes identifies key genes and pathways in hepatocellular carcinoma. *Nat. Genet.*, **44**, 694–698.
35. Guo,G., Gui,Y., Gao,S. *et al*. (2012) Frequent mutations of genes encoding ubiquitin-mediated proteolysis pathway components in clear cell renal cell carcinoma. *Nat. Genet.*, **44**, 17–19.
36. Cancer Genome Atlas Research Network. (2012) Comprehensive genomic characterization of squamous cell lung cancers. *Nature.*, **489**, 519–525.
37. Huang,J., Deng,Q., Wang,Q. *et al*. (2012) Exome sequencing of hepatitis B virus-associated hepatocellular carcinoma. *Nat. Genet.*, **44**, 1117–1121.
38. Imielinski,M., Berger,A.H., Hammerman,P.S. *et al*. (2012) Mapping the hallmarks of lung adenocarcinoma with massively parallel sequencing. *Cell.*, **150**, 1107–1120.
39. Jiao,Y., Shi,C., Edil,B.H. *et al*. (2011) DAXX/ATRX, MEN1, and mTOR pathway genes are frequently altered in pancreatic neuroendocrine tumors. *Science.*, **331**, 1199–1203.
40. Jones,D.T., Jager,N., Kool,M. *et al*. (2012) Dissecting the genomic complexity underlying medulloblastoma. *Nature.*, **488**, 100–105.
41. Jones,S., Zhang,X., Parsons,D.W. *et al*. (2008) Core signaling pathways in human pancreatic cancers revealed by global genomic analyses. *Science.*, **321**, 1801–1806.
42. Kan,Z., Jaiswal,B.S., Stinson,J. *et al*. (2010) Diverse somatic mutation patterns and pathway alterations in human cancers. *Nature.*, **466**, 869–873.
43. Cancer Genome Atlas Research Network. (2012) Comprehensive molecular portraits of human breast tumours. *Nature.*, **490**, 61–70.
44. Le Gallo,M., O'Hara,A.J., Rudd,M.L. *et al*. (2012) Exome sequencing of serous endometrial tumors identifies recurrent somatic mutations in chromatin-remodeling and ubiquitin ligase complex genes. *Nat. Genet.*, **44**, 1310–1315.
45. Lee,W., Jiang,Z., Liu,J. *et al*. (2010) The mutation spectrum revealed by paired genome sequences from a lung cancer patient. *Nature.*, **465**, 473–477.
46. Ley,T.J., Mardis,E.R., Ding,L. *et al*. (2008) DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. *Nature.*, **456**, 66–72.
47. Li,M., Zhao,H., Zhang,X. *et al*. (2011) Inactivating mutations of the chromatin remodeling gene ARID2 in hepatocellular carcinoma. *Nat. Genet.*, **43**, 828–829.
48. Lilljebjorn,H., Rissler,M., Lassen,C. *et al*. (2012) Whole-exome sequencing of pediatric acute lymphoblastic leukemia. *Leukemia.*, **26**, 1602–1607.
49. Lohr,J.G., Stojanov,P., Lawrence,M.S. *et al*. (2011) Discovery and prioritization of somatic mutations in diffuse large B-cell lymphoma (DLBCL) by whole-exome sequencing. *Proc. Natl Acad. Sci. USA.*, **109**, 3879–3884.
50. Love,C., Sun,Z., Jima,D. *et al*. (2012) The genetic landscape of mutations in Burkitt lymphoma. *Nat. Genet.*, **44**, 1321–1325.
51. Mardis,E.R., Ding,L., Dooling,D.J. *et al*. (2009) Recurring mutations found by sequencing an acute myeloid leukemia genome. *N. Engl. J. Med.*, **361**, 1058–1066.
52. Morin,R.D., Mendez-Lago,M., Mungall,A.J. *et al*. (2011) Frequent mutation of histone-modifying genes in non-Hodgkin lymphoma. *Nature.*, **476**, 298–303.
53. Cancer Genome Atlas Research Network. (2012) Comprehensive molecular characterization of human colon and rectal cancer. *Nature.*, **487**, 330–337.
54. Cancer Genome Atlas Research Network. (2008) Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature.*, **455**, 1061–1068.
55. Ong,C.K., Subimerb,C., Pairojkul,C. *et al*. (2012) Exome sequencing of liver fluke-associated cholangiocarcinoma. *Nat. Genet.*, **44**, 690–693.
56. Parsons,D.W., Jones,S., Zhang,X. *et al*. (2008) An integrated genomic analysis of human glioblastoma multiforme. *Science.*, **321**, 1807–1812.
57. Parsons,D.W., Li,M., Zhang,X. *et al*. (2010) The genetic landscape of the childhood cancer medulloblastoma. *Science.*, **331**, 435–439.
58. Pasqualucci,L., Trifonov,V., Fabbri,G. *et al*. (2011) Analysis of the coding genome of diffuse large B-cell lymphoma. *Nat. Genet.*, **43**, 830–837.
59. Peifer,M., Fernandez-Cuesta,L., Sos,M.L. *et al*. (2012) Integrative genome analyses identify key somatic driver mutations of small-cell lung cancer. *Nat. Genet.*, **44**, 1104–1110.
60. Piazza,R., Valletta,S., Winkelmann,N. *et al*. (2013) Recurrent SETBP1 mutations in atypical chronic myeloid leukemia. *Nat. Genet.*, **45**, 18–24.
61. Pleasance,E.D., Cheetham,R.K., Stephens,P.J. *et al*. (2010) A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature.*, **463**, 191–196.
62. Pleasance,E.D., Stephens,P.J., O'Meara,S. *et al*. (2010) A small-cell lung cancer genome with complex signatures of tobacco exposure. *Nature.*, **463**, 184–190.
63. Puente,X.S., Pinyol,M., Quesada,V. *et al*. (2011) Whole-genome sequencing identifies recurrent mutations in chronic lymphocytic leukaemia. *Nature.*, **475**, 101–105.
64. Pugh,T.J., Weeraratne,S.D., Archer,T.C. *et al*. (2012) Medulloblastoma exome sequencing uncovers subtype-specific somatic mutations. *Nature.*, **488**, 106–110.
65. Quesada,V., Conde,L., Villamor,N. *et al*. (2011) Exome sequencing identifies recurrent mutations of the splicing factor SF3B1 gene in chronic lymphocytic leukemia. *Nat. Genet.*, **44**, 47–52.

66. Richter,J., Schlesner,M., Hoffmann,S. *et al*. (2012) Recurrent mutation of the ID3 gene in Burkitt lymphoma identified by integrated genome, exome and transcriptome sequencing. *Nat. Genet.*, **44**, 1316–1320.
67. Robinson,G., Parker,M., Kranenburg,T.A. *et al*. (2012) Novel mutations target distinct subgroups of medulloblastoma. *Nature.*, **488**, 43–48.
68. Rudin,C.M., Durinck,S., Stawiski,E.W. *et al*. (2012) Comprehensive genomic analysis identifies SOX2 as a frequently amplified gene in small-cell lung cancer. *Nat. Genet.*, **44**, 1111–1116.
69. Sausen,M., Leary,R.J., Jones,S. *et al*. (2012) Integrated genomic analyses identify ARID1A and ARID1B alterations in the childhood cancer neuroblastoma. *Nat. Genet.*, **45**, 12–17.
70. Schwartzentruber,J., Korshunov,A., Liu,X.Y. *et al*. (2012) Driver mutations in histone H3.3 and chromatin remodelling genes in paediatric glioblastoma. *Nature.*, **482**, 226–231.
71. Shah,S.P., Morin,R.D., Khattra,J. *et al*. (2009) Mutational evolution in a lobular breast tumour profiled at single nucleotide resolution. *Nature.*, **461**, 809–813.
72. Stephens,P.J., Tarpey,P.S., Davies,H. *et al*. (2012) The landscape of cancer genes and mutational processes in breast cancer. *Nature.*, **486**, 400–404.
73. Stransky,N., Egloff,A.M., Tward,A.D. *et al*. (2011) The mutational landscape of head and neck squamous cell carcinoma. *Science.*, **333**, 1157–1160.
74. Totoki,Y., Tatsuno,K., Yamamoto,S. *et al*. (2011) High-resolution characterization of a hepatocellular carcinoma genome. *Nat. Genet.*, **43**, 464–469.
75. Turajlic,S., Furney,S.J., Lambros,M.B. *et al*. (2012) Whole genome sequencing of matched primary and metastatic acral melanomas. *Genome Res.*, **22**, 196–207.
76. Varela,I., Tarpey,P., Raine,K. *et al*. (2011) Exome sequencing identifies frequent mutation of the SWI/SNF complex gene PBRM1 in renal carcinoma. *Nature.*, **469**, 539–542.
77. Wang,K., Kan,J., Yuen,S.T. *et al*. (2011) Exome sequencing identifies frequent mutation of ARID1A in molecular subtypes of gastric cancer. *Nat. Genet.*, **43**, 1219–1223.
78. Wang,L., Tsutsumi,S., Kawaguchi,T. *et al*. (2012) Whole-exome sequencing of human pancreatic cancers and characterization of genomic instability caused by MLH1 haploinsufficiency and complete deficiency. *Genome Res.*, **22**, 208–219.
79. Wei,X., Walia,V., Lin,J.C. *et al*. (2011) Exome sequencing identifies GRIN2A as frequently mutated in melanoma. *Nat. Genet.*, **43**, 442–446.
80. Wood,L.D., Parsons,D.W., Jones,S. *et al*. (2007) The genomic landscapes of human breast and colorectal cancers. *Science.*, **318**, 1108–1113.
81. Yan,X.J., Xu,J., Gu,Z.H. *et al*. (2011) Exome sequencing identifies somatic mutations of DNA methyltransferase gene DNMT3A in acute monocytic leukemia. *Nat. Genet.*, **43**, 309–315.
82. Yoshida,K., Sanada,M., Shiraishi,Y. *et al*. (2011) Frequent pathway mutations of splicing machinery in myelodysplasia. *Nature.*, **478**, 64–69.
83. Zang,Z.J., Cutcutache,I., Poon,S.L. *et al*. (2012) Exome sequencing of gastric adenocarcinoma identifies recurrent somatic mutations in cell adhesion and chromatin remodeling genes. *Nat. Genet.*, **44**, 570–574.
84. Pruitt,K.D., Tatusova,T., Brown,G.R. *et al*. (2012) NCBI reference sequences (RefSeq): current status, new features and genome annotation policy. *Nucleic Acids Res.*, **40**, D130–D135.
85. Powell,S., Szklarczyk,D., Trachana,K. *et al*. (2012) eggNOG v3.0: orthologous groups covering 1133 organisms at 41 different taxonomic ranges. *Nucleic Acids Res.*, **40**, D284–D289.
86. D'Antonio,M. and Ciccarelli,F.D. (2011) Modification of gene duplicability during the evolution of protein interaction network. *PLoS Comput. Biol.*, **7**, e1002029.
87. 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.
88. Keshava Prasad,T.S., Goel,R., Kandasamy,K. *et al*. (2009) Human protein reference database–2009 update. *Nucleic Acids Res.*, **37**, D767–D772.
89. Chatr-Aryamontri,A., Breitkreutz,B.J., Heinicke,S. *et al*. (2013) The BioGRID interaction database: 2013 update. *Nucleic Acids Res.*, **41**, D816–D823.
90. Kerrien,S., Aranda,B., Breuza,L. *et al*. (2012) The IntAct molecular interaction database in 2012. *Nucleic Acids Res.*, **40**, D841–D846.
91. Ceol,A., Chatr Aryamontri,A., Licata,L. *et al*. (2010) MINT, the molecular interaction database: 2009 update. *Nucleic Acids Res.*, **38**, D532–D539.
92. Salwinski,L., Miller,C.S., Smith,A.J. *et al*. (2004) The database of interacting proteins: 2004 update. *Nucleic Acids Res.*, **32**, D449–D451.
93. Papadopoulos,G.L., Reczko,M., Simossis,V.A. *et al*. (2009) The database of experimentally supported targets: a functional update of TarBase. *Nucleic Acids Res.*, **37**, D155–D158.
94. Xiao,F., Zuo,Z., Cai,G. *et al*. (2009) miRecords: an integrated resource for microRNA-target interactions. *Nucleic Acids Res.*, **37**, D105–D110.
95. Hsu,S.D., Lin,F.M., Wu,W.Y. *et al*. (2011) miRTarBase: a database curates experimentally validated microRNA-target interactions. *Nucleic Acids Res.*, **39**, D163–D169.
96. Dees,N.D., Zhang,Q., Kandoth,C. *et al*. (2012) MuSiC: identifying mutational significance in cancer genomes. *Genome Res.*, **22**, 1589–1598.
97. Sjoblom,T., Jones,S., Wood,L.D. *et al*. (2006) The consensus coding sequences of human breast and colorectal cancers. *Science.*, **314**, 268–274.
98. Greenman,C., Wooster,R., Futreal,P.A. *et al*. (2006) Statistical analysis of pathogenicity of somatic mutations in cancer. *Genetics.*, **173**, 2187–2198.
99. Ge,X., Yamamoto,S., Tsutsumi,S. *et al*. (2005) Interpreting expression profiles of cancers by genome-wide survey of breadth of expression in normal tissues. *Genomics.*, **86**, 127–141.
100. Su,A.I., Wiltshire,T., Batalov,S. *et al*. (2004) A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl Acad. Sci. USA.*, **101**, 6062–6067.
101. Gautier,L., Cope,L., Bolstad,B.M. *et al*. (2004) affy—analysis of affymetrix genechip data at the probe level. *Bioinformatics.*, **20**, 307–315.
102. Hubbell,E., Liu,W.M. and Mei,R. (2002) Robust estimators for expression analysis. *Bioinformatics.*, **18**, 1585–1592.
103. Esquela-Kerscher,A. and Slack,F.J. (2006) Oncomirs—microRNAs with a role in cancer. *Nat. Rev. Cancer.*, **6**, 259–269.
104. Kent,O.A. and Mendell,J.T. (2006) A small piece in the cancer puzzle: microRNAs as tumor suppressors and oncogenes. *Oncogene.*, **25**, 6188–6196.
105. Lujambio,A. and Lowe,S.W. (2012) The microcosmos of cancer. *Nature.*, **482**, 347–355.

Database, Vol. 2014, Article ID bau015, doi:10.1093/database/bau015

106. Manikandan,J., Aarthi,J.J., Kumar,S.D. *et al*. (2008) Oncomirs: the potential role of non-coding microRNAs in understanding cancer. *Bioinformation.*, **2**, 330–334.

107. Spizzo,R., Nicoloso,M.S., Croce,C.M. *et al*. (2009) Snapshot: microRNAs in cancer. *Cell.*, **137**, 586–586.

108. Yang,D., Sun,Y., Hu,L. *et al*. (2013) Integrated analyses identify a master microRNA regulatory network for the mesenchymal subtype in serous ovarian cancer. *Cancer Cell.*, **23**, 186–199.

109. Kozomara,A. and Griffiths-Jones,S. (2011) miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Res.*, **39**, D152–D157.

110. Bartel,D.P. (2004) MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell.*, **116**, 281–297.

111. Lopes,C.T., Franz,M., Kazi,F. *et al*. (2010) Cytoscape web: an interactive web-based network browser. *Bioinformatics.*, **26**, 2347–2348.