



**UNIVERSITY  
OF TRENTO**

**International PhD Program in Biomolecular Sciences**

**Department of Cellular, Computational  
and Integrative Biology – CIBIO  
34<sup>th</sup> Cycle**

**Computational analysis of effects and interactions  
among human variants in complex diseases**

**Advisor:**

Alessandro Romanel

**Tutor:**

Alberto Inga

**Ph.D. Thesis of**

Samuel Valentini

Academic Year 2021/2022



## Declaration

I, Samuel Valentini, confirm that this is my own work and the use of all material from other sources has been properly and fully acknowledged.

Trento, 2<sup>nd</sup> June 2022

*Samuel Valentini*



# Abstract

In the last years, Genome-Wide Associations Studies (GWAS) found many variants associated with complex diseases. However, the biological and molecular links between these variants and phenotypes are still mostly unknown. Also, even if sample sizes are constantly increasing, the associated variants do not explain all the heritability estimated for many traits.

Many hypotheses have been proposed to explain the problem: from variant-variant interactions, the effect of rare and ultra-rare coding variants and also technical biases related to sequencing or statistic on sexual chromosomes. In this thesis, we mainly explore the hypothesis of variant-variant interaction and, briefly, the rare coding variants hypothesis while also considering possible molecular effects like allele-specific expression and the effects of variants on protein interfaces. Some parts of the thesis are also devoted to explore the implementation of efficient computational tools to explore these effects and to perform scalable genotyping of germline single nucleotide polymorphisms (SNPs) in huge datasets.

The main part of the thesis regards the development of a new resource to identify putative variant-variant interactions. In particular, we integrated ChIP-seq data from ENCODE, transcription factor binding motifs from several resources and genotype and transcript level data from GTeX and TCGA. This new dataset allows us to formalize new models, to make hypothesis and to find putative novel associations and interactions between (mainly non-coding) germline variants and phenotypes, like cancer-specific phenotypes. In particular, we focused on the characterization of breast cancer and Alzheimer's Disease GWAS risk variants, looking for putative variants' interactions.

Recently, the study of rare variants has become feasible thanks to the biobanks that made available genotypes and clinical data of thousands of patients. We characterize and explore the possible effects of rare coding inherited polymorphisms on protein interfaces in the UKBioBank trying to understand if the change in structure of protein can be one of the causes of complex diseases.

Another part of the thesis explores variants as causal molecular effect for allele-specific expression. In particular, we describe UTRs variants that can alter the post-transcriptional regulation in mRNA leading to a phenomenon where an allele is more expressed than the other. Finally, we show those variants can have prognostic significance in breast cancer.

This thesis work introduces results and computational tools that can be useful to a broad community of researcher studying human polymorphisms effects.

# Contents

Abstract .....	5
List of abbreviations .....	10
Introduction .....	12
Background .....	12
Thesis aims .....	16
Main techniques and data resources .....	17
Next generation sequencing .....	17
Variant Genotyping .....	18
GWAS analysis .....	19
eQTL analysis .....	20
The encode project .....	21
Chapter 1: PaCBAM: a tool for variants genotyping .....	23
Introduction .....	23
Results .....	23
Materials and methods .....	24
Discussion.....	24
Article .....	25
Abstract .....	25
Background .....	26
Implementation.....	27
Results .....	28
Conclusion.....	31
Availability and requirements.....	31
Supplementary Material.....	31
Chapter 2: Identification of variants affecting mRNA translation potential.....	55
Introduction .....	55
Results .....	55
Materials and methods .....	56
Discussion.....	56
Article .....	57
Summary .....	57
Introduction .....	58
Results .....	59
Discussion.....	69

Limitation of Study .....	72
Methods .....	72
Supplementary Material.....	79
Chapter 3: Finding functional relations among common human genetic variants.....	84
Introduction .....	84
Results .....	84
Materials and methods .....	85
Discussion.....	86
Article .....	88
Abstract .....	88
Introduction .....	89
Materials and Methods .....	90
Results .....	96
Discussion.....	114
Resource Availability.....	118
Supplementary Material.....	118
Chapter 4: Somatic mutations and rare variants in protein interfaces.....	128
Introduction .....	128
Results .....	129
Identification of somatic mutations in protein interfaces in TCGA.....	129
cBioPortal enrichment.....	132
Rare variants in cancer genes interface enrichment .....	134
Material and methods.....	141
Protein Interfaces .....	141
TCGA dataset .....	141
cBioPortal dataset .....	142
Enrichment in protein interfaces and multiple test correction .....	142
Interface filtering for germline analysis .....	142
UKBioBank dataset .....	142
Proportion tests.....	143
Discussion.....	144
Discussion, conclusion and future works .....	147
Bibliography.....	151





# List of abbreviations

**ASE:** Allele Specific Expression

**CNV:** Copy Number Variant

**eQTL:** Expression Quantitative Trait Locus

**GWAS:** Genome Wide Association Study

**INDEL:** Small Insertion or Deletion

**MAF:** Minor Allelic Frequency

**NGS:** Next-Generation Sequencing

**PCR:** Polymerase chain reaction

**PFM:** Positional Frequency Matrix

**PRS:** Polygenic Risk Score

**PWM:** Positional Weight Matrix

**QTL:** Quantitative Trait Locus

**RBP:** RNA Binding Protein

**SNP:** Single Nucleotide Polymorphism

**TF:** Transcription Factor

**VAF:** Variants Allele Frequency

**WES:** Whole-Exome Sequencing

**WGS:** Whole-Genome Sequencing



# Introduction

## Background

Almost all the human traits are influenced by the genetic background of individuals, to some degree ranging from Mendelian monogenic disorders to highly polygenic complex traits. In the 80s and 90s, thanks to linkage analysis and Sanger sequencing, the first variants causing diseases were identified. In particular, variations in the *CFTR* (1) for cystic fibrosis and multiple CAG repeats in the *HTT* gene for Huntington Disease (2). However, it was with the first human genome (3) that the study of genetic variability really started. In few years, the community developed the first Genome-Wide Associations Study (GWAS) that linked age related macular degeneration to variants (4). During years, GWAS have grown in sample size reaching more than half a million individuals for several traits like schizophrenia, breast cancer (5) and Alzheimer's Disease (6). GWAS studies have been extended to cover many phenotypes and, in cancer, variants have been shown to shape cancer evolution, cancer molecular subtypes and cancer patients immune response (7,8). Usually, a GWAS provides only a statistical and not a mechanistic link between a variant and a trait, but they showed that almost every trait is influenced by a genetic component. However, even with a growing number of samples, fractions of heritability estimated from twin studies for traits are still missing (9). Many hypotheses have been formulated to explain this phenomenon.

The first hypothesis was related to the low power of the first GWAS studies that were unable to detect all the associations due to low sample sizes and more recent studies are detecting more and more associations but with lower and lower effects and heritability. Also, most GWAS only tested additive models and excluded sexual chromosomes from the analysis leading to technical biases (10).

Another hypothesis to explain the missing heritability is related to rare variants. In particular, SNP array and early GWAS sample sizes were inadequate to detect variant with minor allelic frequency smaller than 5%. Whole genome sequencing is rapidly closing

the gaps and it is allowing to detect many variants and the consequent analysis of rare variants in a GWAS settings.

Unfortunately, the study of rare variants is crippled by their rarity and the consequent sample size required to perform associations studies. To overcome those limitations several variants can be combined into functional units like genes or protein interfaces to gain the required statistical power.

(11). Recently, thanks to BioBanks the first Coding Wide Associations Studies on rare variants have been created (12) showing many associations of coding variants with phenotypes and also providing one of the likely causal gene related to the phenotype.

Other even less studied possible contributor to the missing heritability are structural variants like Copy-Number Variants (CNV), structural rearrangements, inversions and translocations. CNVs and structural variants have been studied less with respect to SNPs and INDELS because CNVs are technically more difficult to detect and characterize.

Recent studies found that many healthy individuals carry CNV and that African populations carry 10% more DNA with respect of the current reference genome suggesting that, in fact, CNVs can account for a part of the missing heritability (13). First association studies found that associated CNVs are usually located in regions where a tag SNP for the trait have already been detected and thus they concluded that common CNVs are not likely to contribute much to the missing heritability (14). Other studies focused on rarer CNVs and found that associated CNV are not randomly distributed and are more likely to impact causal genes or GWAS loci relevant to the trait (15). Studies on BioBanks have only recently started to perform CNV-GWAS studies but with great technical limitations given by the CNV calling using microarrays, CNV detection tools and the complex nature of CNVs (16). In conclusion, CNVs effects on some phenotypes, are an understudied genetic variant type that will greatly benefit from novel sequencing technology and variant calling techniques and where possibly some of the missing heritability is still hidden.

Another very recent hypothesis on the source of the missing heritability is epigenetics through a phenomenon called Transgenerational Epigenetic Inheritance. Epigenetic inheritance has been proposed in plants since most plants can be propagated without requiring germline cells and in fact it has already been observed in plants (17), but its effect on animals and humans are still uncertain and unclear. Mathematical models have

been proposed to estimate the heritability given by epigenetic inheritance but the contribution of epigenetics is still unclear (18).

Finally, the last hypothesis on the missing heritability is related to epistasis between variants and other variants or between variants and the environment. Epistasis was firstly introduced by William Bateson in (19) on qualitative phenotypes and was defined as the effect of some variants that can block others. Then, the term evolved to describe the phenomenon where the additive combination of two or more variants deviates from the sum of the single effects (20). The effects of epistasis are extremely difficult to detect since if only pairs of variants are investigated the number of tests required scales quadratically leading to an enormous amount of tests that can be performed only using the most performant supercomputers but, more importantly, multiple tests correction is the main burden in finding interacting loci given the amount of tests performed. To explore epistasis different functional studies are required to approach the problem (21). In the past decades many variants have been associated to complex traits but, even if a lot of hits have been replicated in different studies, in very few cases biological links have been found to describe the molecular effects of GWAS polymorphisms. The main difficulty in linking a GWAS hit to a mechanistic link is that, usually, the causal variant is in Linkage Disequilibrium (LD) with the GWAS hit, making extremely difficult to dissect the haplotype to identify the link. LD makes flanking non causal variants to have the same statistical association as the true causal variant making the whole haplotype associated to the phenotype. The first observation made was that, even if the GWAS variants are located in non-coding part of the genome, they are enriched for regulatory elements and their effect can be related to gene regulation (22). The next step was to try to explain the effects of variant through gene expression. Expression Quantitative Trait Loci (eQTL) is a genomic locus that can modulate the transcript level of a gene. Usually, eQTLs are located near the modulated gene in a window of 1 megabase for humans (*cis*- eQTL), less studied but equally important are *trans*- eQTL which are eQTL acting on genes outside the 1 megabase window or on other chromosomes.

eQTLs can start a phenomenon called Allele-Specific Expression (ASE). ASE is the effect of when an allele is more expressed than the other. ASE has been extensively studied the X chromosome in females where its inactivation can shape the phenotype in case of heterozygosity and has been showed to be linked to several diseases (23). When an eQTL

is present, the allele with the variant can be differentially expressed with respect to the other allele, an effect that has already been linked to cancer (24).

Given the great amount of associated variants and the variety of possible effects that polymorphisms can have, the research community tried to develop methods to stratify the population and to identify people at risk of developing pathologies. Ideally, GWAS would have changed how we approach common diseases granting a personalized therapy and screening tailored on the genetic background of each individual. However, the complexity, the small effect of every variant and the interactions between them greatly limited the application of GWAS in medicine. Research focused on an aggregative tool called Polygenic Risk Score. A PRS is the weighted sum of the risk alleles present in an individual multiplied by the risk effect of each variant. PRS have been shown to be able to stratify individuals in function of their lifetime risk of diseases, but they are still lacking the ability to correctly predict the onset of a disease. In particular, the clinical utility of PRS is still unclear since they are able to stratify the population but they do not provide enough information at an individual level (25). New machine learning and deep learning techniques are now being developed trying to improve the stratification and the associations metrics for single individuals (26) and in the future thanks to biggest datasets and better methods, personalized screening based on the genetic backgrounds will be available and deployed in clinical settings.

In the last 20 years, since the release of the first human genome draft, researchers started to identify loci associated to complex diseases and traits. Many tools to analyze, detect and explore those associations were developed, but most of the mechanistic links and part of the heritability are still elusive. Hopefully, in the next years we will be able to refine our knowledge on germline variants finding the molecular mechanisms behind their effects, close the gaps still present in the heritability and to develop tools to stratify the population and personalize therapies based on the genetic background of individuals.

## Thesis aims

This thesis aims to explore human germline variants in term of their identification and characterization, their effects and interactions on the expression levels of genes and possible molecular mechanisms involved.

We start by introducing a novel command line tool PaCBAM for Whole-Exome/Targeted sequencing scalable analysis and variant genotyping. This tool allows to scale variant genotyping across large-scale sequencing datasets.

Then, we explore possible effects of variants in terms of allele specific expression and their interactions. Specifically, we developed new methods to detect variants that can affect mRNA translation and explore variant-variant interactions trying to identify patterns of epistatic/additive effects among polymorphisms.

Finally, we analyze the effects of rare coding variants in protein interface across cancer genes.

At the end of the thesis, we introduced new models and tools to characterize and explore novel functional effects of variants.



# Main techniques and data resources

## Next generation sequencing

Next-generation sequencing is a paradigm in sequencing that started in the mid-2000s. NGS main idea is to sequence simultaneously many DNA fragments to greatly reduce sequencing time and costs. The technique was firstly introduced in (27) as a proof of concept to improve sequencing time with respect to Sanger sequencing. Further, thanks to next-generation sequencing, we have been able to increase enormously also the efficiency in sequencing with respect to the previous sequencing techniques.

The underlying technology used in NGS is sequencing by synthesis, which detects and identifies single nucleotides while they are added base by base for the synthesis of a new DNA strand, complementary to the strand of which we want to read the sequence. Sequencing by synthesis starts with an amplification of the DNA sample we want to read using PCR. After amplification, strands are forced to divide and only one of the two strands of DNA is kept. An admixture of labelled nucleotides is injected allowing the growth of the complementary strand by only one nucleotide. Every nucleotide is bound to a fluorophore, so that it is labelled with a different color. The machine can detect the color and identify which nucleotide bounds to the strand. After detection, fluorophores are cleaved, and another cycle is initiated allowing to sequence the next base. Sequencing by synthesis usually is able to sequence small fragments up to 100 base with high accuracy (28). Conversely, in the last years, new long-read technologies have been developed to overcome the short reads sequencing limitations: sequencing of highly repeated regions and complex structural variants calling. The first technique is called Single-Molecule Real-Time, where a polymerase is fixed on the bottom of a zero-mode waveguide. The zero-mode waveguide works together with an DNA polymerase where the zero-mode waveguide detects the fluorescent tag of the last incorporated nucleotide by the DNA polymerase. (29).

Currently, the most advanced developed technology is NanoPore. It passes a DNA strand through a pore where nucleotides are detected one by one through the modification of the electric voltage in the medium (30). Unfortunately, long read technologies have a higher error rate with respect to their short read counterpart. This disadvantage can be limited by sequencing the same DNA fragment in a circular fashion allowing multiple

reads of the same DNA molecule and using a consensus approach to call bases (31). Long read sequencing is still in its infancy but has already been proved to be a game changer in the development of the Telomere-to-Telomere human reference genome where, thanks to long read sequencing, the gaps made of long repeated elements in telomeres and near centrosome have been sequenced giving us the best human reference to date (32).

## Variant Genotyping

Variant genotyping allows to detect the polymorphisms in the DNA of an individual. In the past, variant calling was performed using SNP arrays but recently NGS data is becoming the gold standard of variant genotyping.

A SNP array consists in an array of allele specific probes to which specific fragments of DNA can bind. To genotype using a SNP array, a library of DNA must be created by DNA fragmentation. Then, a fluorophore is bound to each fragment to allow fluorescence detection. The labeled DNA is then injected in the array where it binds the variant probes. The genotype can call variants by the fluorescence detected across each probe (33). SNP arrays are a cheap and effective way to genotype an individual and the most advanced SNP arrays incorporate about 1,000,000 probes for genotyping as many SNPs. Unfortunately, SNP arrays allow to detect only the variants included in it, which limits the genotyping resolution. Other variants can then be imputed using a population reference panel (34). However, rare variants are difficult to impute since they could not be included in the reference panel. SNP arrays are still a cost-effective method to genotype individuals but now NGS allows to a better genotyping resolution that is needed in GWAS studies.

NGS and, in particular, Whole-Genome Sequencing allow to detect more variants with respect to the SNP arrays. Variant calling from NGS data starts with a quality control procedure that consists in the removal of duplicated reads, local realignment, to limit alignment errors due to INDELS, and base quality recalibration. After those quality control processes, it is possible to genotype the variant from the NGS data using one of the available tools and methods (35).

The easiest SNPs calling methods are based on heuristic. Heuristics use base coverage, read and base quality and allelic frequency thresholds to detect SNPs. They are a simple

and efficient way to genotype a huge number of loci, however, they can be less precise in calling SNPs where one of the parameters is close to the threshold.

Another class of SNPs callers are probabilistic based variant caller. Those callers rely on probabilistic tests such as binomial test (36) or Bayes Theorem (37). Probabilistic callers provide a more robust genotyping but are computationally more expensive than heuristic methods.

INDELS calling is more complex and different methodologies have been proposed for a correct genotyping. The first method, implemented by many tools like SAMtools(38) and GATK(35), uses an alignment-based method where they analyze the alignment errors. Other tools, like PEMer (39), exploit paired end sequencing trying to detect deviations across paired read distances. Most advanced methods use local de-novo assembly to generate the haplotype like GATK HaplotypeCaller (40). Then, the reads are realigned to the novel contig that models the haplotype and a likelihood is computed given the read. The haplotype called will be the one with the highest likelihood.

Recently, machine learning and deep learning methods have been employed for SNPs and INDELS calling using pileups images (41) and so introducing a novel genotyping technique.

## GWAS analysis

Genome-wide Association studies try to identify genetic loci that are more frequent in a group of individuals that express a phenotype with respect to another group without the phenotype of interest. During the last years, GWAS were able to identify and replicate many SNPs and INDELS associated with thousands of traits and diseases.

A GWAS starts with the selection of a population and a trait to study. Individuals are then genotyped using a SNP array or through WES/WGS. Other information such as age, sex and ethnicity are collected since they can be used to reduce confounder effects. Secondly, variants can be imputed using a reference panel if they come from a SNP array. A quality control is then applied to remove variants with a low call rate, not in Hardy-Weinberg equilibrium or variants on sexual chromosomes.

Then, the associations are performed using linear regression, if the phenotype is continuous, or logistic regression, if the phenotype is binary. Usually, the models include the confounding factors to correct for samples stratification.

After testing, usually, all variants associated with a p-value  $< 5 \times 10^{-8}$ , which is a genome wide Bonferroni threshold on the 1,000,000 haploblocks in the human genome, are counted as significant. Here, it should be taken into account that, usually, GWAS find many variants in linkage disequilibrium as associated with the trait. Linkage disequilibrium introduces a correlation in the study and fine mapping is required to better understand and identify the causal variant inside the block.

The first and easiest fine mapping technique is to simply take the most associated variant inside the block. While doing this can lead to reasonable hits, the real causal variant can be confounded if the effects come from multiple independent variants inside an haploblock. Fine mapping and causal variant identification are extremely complex and they are still an open problem in GWAS even if many statistical tools based on conditional or Bayesian statistic have been developed like FINEMAP(42), where they build a likelihood model of the variants in an haploblock and they find the one that has the highest likelihood . Another promising line is try to functionally prioritize variants that are located in functional regions of the genome, that can alter TF binding motifs or that are eQTLs (43).

GWAS shaped the investigation of complex traits revealing many associations between variants and complex traits. However, the molecular links are still mostly unknown and post-GWAS functional analysis are required to link polymorphisms to genes to improve our knowledge of complex traits.

## eQTL analysis

Since most of the variants identified in GWAS are located in the non-coding part of the genome, their effect on the phenotype was mostly unclear. One of the first links between GWAS and biological functions were eQTLs. eQTLs are polymorphisms that explain part of the transcript level of genes. eQTLs are typically divided in two major classes *cis*-eQTLs, when the variants are comprised in a window from 1 to 5 megabases around the affected gene, and *trans*-eQTLs when they located outside of the genomic window or even on other chromosomes.

An eQTL analysis start with the quantification of the mRNAs in the tissue of interest usually in term of Read Per Kilobase of transcript per Million mapped reads. Then,

similarly to a GWAS analysis, individuals are genotyped and confounding variables like age, sex and ethnicity are recorded. Finally, a linear model association is performed between the transcript level of the gene of interest against the genotype and the confounding variables. Usually, only variants inside the *cis* window are tested to limit the burden of multiple testing.

*Trans*-eQTLs are more difficult to detect since, potentially, they require testing the associations using all the variants in the genome which increases the number of statistical tests required. The largest association project GTEx (44) focuses only on *cis*-eQTL while other studies performs *trans*-eQTL association only on variants of interest or already know GWAS variants (45) to reduce multiple testing burden. The interest on *trans*-eQTLs is rising since it is an understudied field and can provide some of the missing links in GWAS. Recently, a new model that links weak *trans* to direct genetic contributions to traits has been developed (46). In the model, the effect of genes is partitioned into a group of core genes, that directly affect a trait, and peripheral genes that can act on core genes through a gene regulatory network. The analysis on the model concluded that a lot of the missing heritability in traits can be explained by a great amount of small *trans* effects on peripheral genes that can be amplified by the gene network.

In conclusion, eQTLs are a powerful resource to investigate the effect of inherited variants and to detect the missing mechanistic links between variants and traits.

## The encode project

The ENCODE project started in 2003 with the aim to detect sequence of DNA that are functional in the genome. ENCODE tries to functionally evaluate all the regions inside the human genome. Several techniques are used to generate data and the consortium developed a uniformed protocol for data processing, improving the standardization and reproducibility of the dataset. The effort of ENCODE lead to the creation of more than 1,640 datasets annotating the human and the mouse genomes. The ENCODE consortium found that the 80.4% of the human genome has some regulatory feature in at least one tested tissue. They were able to classify the chromatin state into seven different statuses and they annotated about 400,000 regions as enhancers and about 70,000 as promoters. ENCODE showed that many GWAS SNPs are located in regions annotated as functional,

giving strength to the hypothesis that non-coding GWAS variants are involved in regulatory processes in cells. In the end, the ENCODE project produced the most complete dataset, to date, of human functional elements. The whole ENCODE project is described in (47).

# Chapter 1: PaCBAM: a tool for variants genotyping

## Introduction

Per base genotyping is still one of the most time-consuming step in a NGS analysis pipeline. The analysis of large cohorts of patients, usually consisting of thousands of whole-exome or targeted sequencing files, needed to characterize common and rare variants, requires scalable tools in terms of computing cores usage to be deployed and integrated in pipelines running on modern computers and clusters. Moreover, the standard pileup file format is cumbersome, taking a huge amount of memory and is difficult to use. To improve the running time and memory usage of pileups, region coverage and variant calling, we implemented PaCBAM, a new multithreaded tool used to compute the pileup of targeted and whole-exome sequencing data.

## Results

In this paper we presented PaCBAM, a tool for the analysis of the Whole-Exome and targeted sequencing data. The tool improves the execution time of per-base and region coverage computation in Whole-Exome and targeted sequencing data with respect to the state of the art tool available. The tool also implements an on-the-fly deduplication strategy which allows to skip an intermediate processing step in an NGS pipeline. PaCBAM is easy to deploy in an NGS pipeline on workstations or remote servers since it's available as binary and it is containerized in Docker/Singularity.

Finally, PaCBAM implements a reporting tool used for a visual quality control of the results.

## Materials and methods

PaCBAM was developed using the C programming language while the visual reports are implemented in python.

The performance evaluation has been implemented using a WES BAM file from the 1000 Genomes Project in a containerized environment allowing full reproducibility of the analysis. PaCBAM was tested against the main state of the art pileup tools available both in pileup and dedup mode.

I have contributed to the project by designing and implementing the performance evaluation and the visual reporting tool.

## Discussion

PaCBAM is a new tool that allows to compute pileups and coverage statistics from whole-exome and targeted sequencing data. PaCBAM showed a great improvement in execution time when computing single-base and region statistics and when using the remove duplicate option which allows to skip an entire preprocessing step and we believe that it is a useful tool for large datasets analysis.



## Article

PaCBAM: fast and scalable processing of whole exome and targeted sequencing data

Samuel Valentini<sup>1</sup>, Tarcisio Fedrizzi<sup>1</sup>, Francesca Demichelis<sup>1</sup>, Alessandro Romanel<sup>1#</sup>

<sup>1</sup> Department of Cellular, Computational and Integrative Biology (CIBIO), University of Trento, Trento, Italy

#Corresponding author

**Journal:** BMC Genomics volume 20, Article number: 1018 (2019)

**Publisher:** Springer Nature

**Doi:** <https://doi.org/10.1186/s12864-019-6386-6>

## Abstract

**Background:** Interrogation of targeted sequencing NGS data is rapidly becoming a preferred approach for the exploration of large cohorts in the research setting and importantly in the context of precision medicine. Single-base and genomic region level data retrieval and processing still constitute major bottlenecks in NGS data analysis. Fast and scalable tools are hence needed.

**Results:** PaCBAM is a command line tool written in C and designed for the characterization of genomic regions and single nucleotide positions from whole exome and targeted sequencing data. PaCBAM computes depth of coverage and allele-specific pileup statistics, implements a fast and scalable multi-core computational engine, introduces an innovative and efficient *on-the-fly* read duplicates filtering strategy and provides comprehensive text output files and visual reports. We demonstrate that PaCBAM exploits parallel computation resources better than existing tools, resulting in important reductions of processing time and memory usage, hence enabling an efficient and fast exploration of large datasets.

Conclusions: PaCBAM is a fast and scalable tool designed to process genomic regions from NGS data files and generate coverage and pileup comprehensive statistics for downstream analysis. The tool can be easily integrated in NGS processing pipelines and is available from Bitbucket and Docker/Singularity hubs under the MIT license.

## Background

Genomic region and single-base level data retrieval and processing, which represent fundamental steps in genomic analyses such as copy number estimation, variant calling and quality control, still constitute one of the major bottlenecks in NGS data analysis. To deal with the computationally intensive task of calculating depth of coverage and pileup statistics at specific chromosomal regions and/or positions, different tools have been developed. Most of them, including specific modules of SAMTools (38) and BEDTools (48) and the most recent Mosdepth (49), only measure and optimize the computation of depth of sequencing coverage. Few others, like the *pileup* modules of SAMTools, Sambamba (50), GATK (35) and ASEQ (36) provide instead statistics at single-base resolution, which is essential to perform variant calling, allele-specific analyses and exhaustive quality control. Although most of these tools offer parallel computation options, scalability in terms of memory and multiple processes/threads usage is still limited. To enable an efficient exploration of large scale NGS datasets, here we introduce PaCBAM, a tool that provides fast and scalable processing of targeted re-sequencing data of varying sizes, from WES to small gene panels. Specifically, PaCBAM computes depth of coverage and allele-specific pileup statistics at regions and single-base resolution levels and provides data summary visual reporting utilities. PaCBAM introduces also an innovative and efficient *on-the-fly* read duplicates filtering approach. While most tools for read duplicates filtering work on SAM/BAM files sorted by read name (38,51) or read position (50, [broadinstitute.github.io/picard](https://broadinstitute.github.io/picard)) and generate new SAM/BAM files, PaCBAM performs the filtering directly during the processing, not requiring the creation of intermediate BAM/SAM files and fully exploiting parallel resources.

## Implementation

PaCBAM is a command line tool written in C programming language that combines multi-threaded computation, SAMTools APIs, and an ad-hoc data structures implementation. PaCBAM expects as input a sorted and indexed BAM file, a sorted BED file with the coordinates of genomic regions (namely the *target*, e.g. captured regions of a WES experiment), a VCF file specifying a list of SNPs of interest within the *target* and a reference genome in FASTA format. PaCBAM implements a multi-threaded solution that optimizes the execution time when multiple cores are available. The tool splits the list of regions provided in the BED file and spawns different threads to execute parallel computations using a shared and optimized data structure. The shared data structure collects both region and single-base level information and statistics which are processed and finally exposed through four different output options. Each output mode provides the user with only the statistics of interest, generating a combination of the following text output files: a) *depth of coverage of all genomic regions*, which for each region provides the mean depth of coverage, the GC content and the mean depth of coverage of the sub-region (user specified, default 0.5 fraction) that maximizes the coverage peak signal, to account for the reduced coverage depth due to incomplete match of reads to the captured regions (**Figure S1.1**); b) *single-base resolution pileup*, which provides for each genomic position in the target the read depth for the 4 possible bases (A, C, G and T), the total depth of coverage, the variants allelic fraction (VAF), the strand bias information for each base; c) *pileup of positions with alternative base support*, which extracts the pileup statistics only for positions with positive VAF, computed using the alternative base with highest coverage (if any); d) *pileup of SNPs positions*, which extracts the pileup statistics for all SNPs specified in the input VCF file and uses the alternative alleles specified in the VCF file for the VAF calculation and the genotype assignment (**Supplementary Material** for details). All output files are tab-delimited text files and their format details are provided in the **Supplementary Material**.

PaCBAM allows the user to specify the minimum base quality score and the minimum read mapping quality to filter out reads during the pileup processing.

In addition, we implemented an efficient *on-the-fly* duplicated reads filtering strategy which implements an approach that is similar to the Picard MarkDuplicates method but

that applies the filter during region and single-base level information retrieval and processing without the need of creating new BAM files (**Supplementary Material**). The filtering strategy, which fully exploits multi-core capabilities, uses single or paired read alignment positions (corrected for soft-clipping at the 5' end) and total mapping size information to identify duplicates and implements ad-hoc data structures to obtain computational efficiency.

PaCBAM package also includes a Python script to generate visual data reports which can be directly used for quality control. Reports include plots summarizing distributions of regions and per-base depth of coverage, SNPs VAF distribution and genotyping, strand bias distribution, substitutions spectra, regions GC content (**Figure S1.4-S1.8**).

## Results

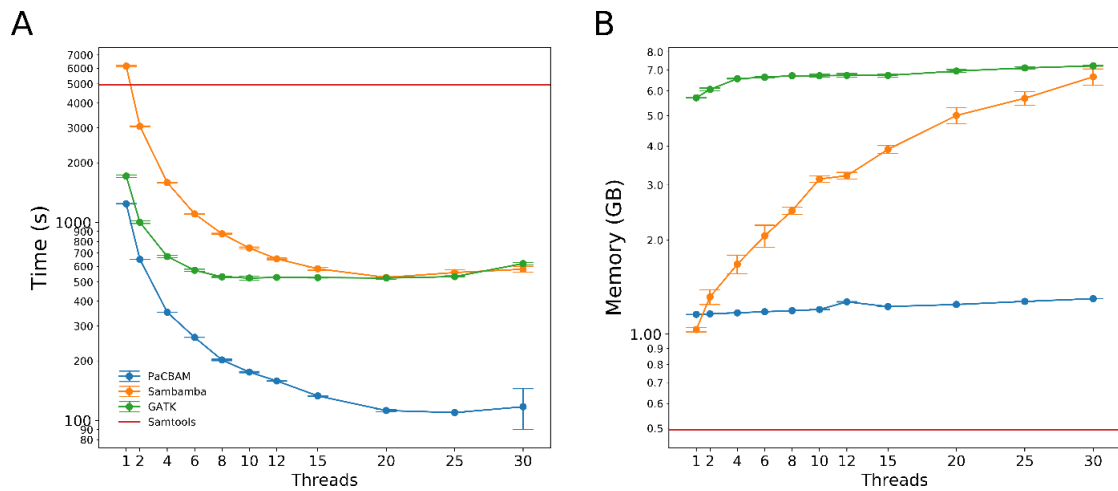
PaCBAM performances were tested on an AMD Opteron 6380 32-cores machine with 256 GB RAM. To mimic different application scenarios, we measured the execution time and memory used by PaCBAM to compute pileups from multiple input BAM files spanning different depth of coverage and different target sizes (**Supplementary Material, Table S1.1**) using an increasing number of threads. We compared PaCBAM performances against pileup modules of SAMTools, Sambamba and GATK (SAMTools offer no parallel pileup option).

In terms of runtime, as shown in **Figure 1.1A** and **Figure S1.9-S1.11**, PaCBAM and Sambamba are the only tools that scale with the number of threads used. PaCBAM outperforms all other tools in all tested conditions. Of note, while PaCBAM pileup output files are of constant size, output files of SAMTools, Sambamba and GATK have a size that is function of the coverage; among all the experiments we run in the performance analyses, PaCBAM output is up to 17.5x smaller with respect to outputs generated by the other tested tools.

While GATK and PaCBAM, as shown in **Figure 1.1B** and **Figure S1.12-S1.14**, have a memory usage that depends only on the target size, Sambamba usage depends on both target size and number of threads and SAMTools usage is constant. Above 8 cores, PaCBAM beats both GATK and Sambamba in all tested conditions in memory usage.

As an example of performance comparison, when analyzing a BAM file with ~300X mean coverage and ~30Mbp target size using 30 threads (**Figure 1.1A-B**), PaCBAM improves

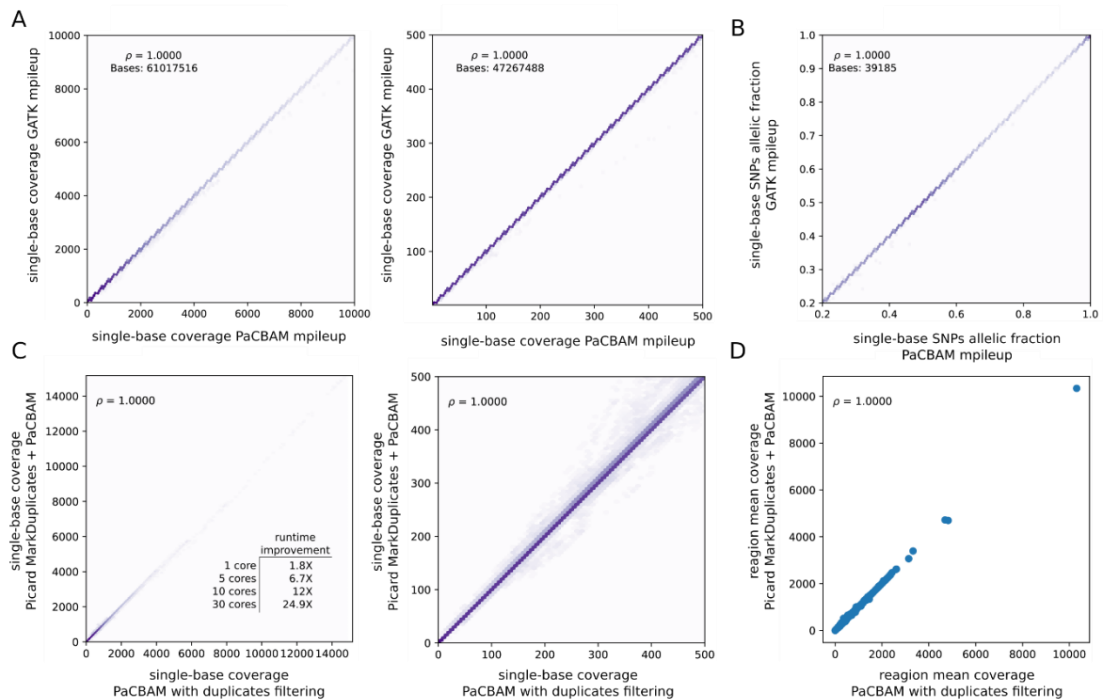
execution time of 4.9x/5.27x and requires 80%/82% less memory compared to Sambamba/GATK.



**Figure 1.1 PaCBAM performances.** Time (A) and memory (B) required by PaCBAM to perform a pileup compared to SAMtools, GATK and Sambamba, using increasing number of threads. The figure focuses on the analysis of a BAM file with  $\sim 300x$  mean coverage and  $\sim 30\text{Mbp}$  target size using 30 threads. Note that parallel pileup option is not available for SAMtools and red lines in panel A and B refer to the average of single thread executions.

Of note, in the sequencing scenarios here considered, PaCBAM demonstrates up to 100x execution time improvement and up to 90% less memory usage with respect to the single-base pileup module of our previous tool ASEQ (Figure S1.15).

Depth of coverage and pileup statistics of PaCBAM pileup were compared to GATK results on a BAM file with  $\sim 300X$  average coverage and  $\sim 64\text{Mbp}$  target size observing almost perfect concordance (Figure 1.2A-B).



**Figure 1.2 Comparison of PaCBAM results with other tools.** A) Comparison of PaCBAM and GATK depth of coverage (left) with zoom in the coverage range [0,500] (right); number of positions considered in the analysis and correlation results are reported. B) Comparison of allelic fraction of ~40K positions annotated as SNPs in dbSNP database v144 and having an allelic fraction >0.2 in both PaCBAM and GATK pileup output. C) Single-base coverage obtained by running either Picard MarkDuplicates + PaCBAM pileup or PaCBAM pileup with duplicates filtering option active (left) with zoom in the coverage range [0,500] (right). D) Regional mean depth of coverage obtained by running either Picard MarkDuplicates + PaCBAM pileup or PaCBAM pileup with duplicates filtering option active.

PaCBAM duplicates removal strategy was tested by comparing PaCBAM pileups obtained from a paired-end BAM file first processed with Picard MarkDuplicates or parallel Sambamba markup, to PaCBAM pileups obtained from the same initial BAM file but using the embedded on-the-fly duplicates filtering. As shown in **Figure 1.2C-D** and **Figure S1.16**, both single-base and region level statistics results are strongly concordant, with single-base total coverage difference (with respect to Picard) that in 99.94% of positions is <10X, single-base allelic fraction difference that in 99.95% of positions is <1% and region mean coverage difference that in 99.96% of regions is <10X (**Figure S1.17A-B-C**). In addition, PaCBAM strategy improves overall execution time of 2.5x/1.7x with a single thread and of 25x/3x with 30 threads compared PaCBAM to Picard and parallel Sambamba, respectively (**Table S1.2, Figure 1.2C, Figure S1.16A**).

Overall, these analyses demonstrate that PaCBAM exploits parallel computation resources better than existing tools, resulting in evident reductions of processing time and memory usage, that enable a fast and efficient coverage and allele-specific characterization of large WES and targeted sequencing datasets. The performance analysis is completely reproducible using an ad-hoc Debian-based Singularity container (**Supplementary Material**).

## Conclusion

We presented PaCBAM, a fast and scalable tool to process genomic regions from NGS data files and generate coverage and pileup statistics for downstream analysis such as copy number estimation, variant calling and data quality control. Although designed for targeted re-sequencing data, PaCBAM can be used to characterize any set of genomic regions of interest from NGS data. PaCBAM generates both region and single-base level statistics and provides a fast and innovative *on-the-fly* read duplicates filtering strategy. The tool is easy to use, can be integrated in any NGS pipeline and is available in source/binary version on Bitbucket and containerized from Docker and Singularity hubs.

## Availability and requirements

Project name: PaCBAM

Project home page: [bcglab.cibio.unitn.it/PaCBAM](http://bcglab.cibio.unitn.it/PaCBAM)

Operating system(s): Platform independent

Programming language: C, Python

License: MIT

## Supplementary Material

### 1. Assignment of genotypes to input SNPs

PaCBAM provides an option to assign genotype calls to all SNPs listed in the input VCF file. The tool implements two approaches. The first one is based on the VAF value and specifically assigns genotype 0/0 when  $VAF \leq 0.2$ , assigns genotype 0/1 when VAF is in the range (0.2,0.8) and assigns genotype 1/1 when VAF is  $\geq 0.8$ . The second approach instead implements a binomial test with probabilities  $p$  and  $q$  for the reference and the

alternative allele, respectively, and a significance cutoff at 1%. To account for the reference bias mapping, we apply default probabilities  $p=0.55$  and  $q=0.45$ . High precision and recall of these two approaches in genotyping common SNPs, compared to SNP array genotype calls, was previously shown in (52).

## 2. *On-the-fly* read duplicates filtering

PaCBAM duplicates filtering strategy is applied while computing single base and region level statistics and fully exploits parallel computation. To allow duplicated reads identification at a specific captured region  $R$  with genomic coordinates  $chr:start-end$ , a preliminary fetching of reads is performed. Specifically, reads in the extended region  $chr:(start-W)-(end+W)$  are retrieved (with  $W$  tunable by the user with default value equal to 1,000) and a hash-map collecting positional and mapping information on paired (or single) end reads is populated using the read name as key string. For each read, using the CIGAR value, the alignment position is corrected for soft-clipping at the 5' end and the size of mapped region is calculated. Duplicates filtering is then performed by searching for paired (or single) end reads with same corrected positions and selecting the one with largest total mapped region size as representative for each duplicated reads group. Only selected reads are kept in the hash-map and considered during the second and standard fetching of reads that is used to compute single base and region level statistics. As shown in **Figure S1.17**, default  $W$  parameter value equal to 1,000 represents a good trade-off balance between computational performances and duplicates filtering effectiveness when paired-end reads are used. When single-end alignment files are processed  $W$  can be set to 0.

## 3. Creation of BAM files for performance experiments

PaCBAM has been tested on BAM files representing different target sizes. BAM files were created using different BED files representing the different target sizes. Starting from the original Nimblegen SeqCap EZ Exome v3 kit BED file of size 64,190,747bp containing genomic coordinates of exonic captured regions, new BED files of sizes 6,424,707bp, 16,053,802bp, 32,102,630bp, and 48,144,328bp (corresponding to 10%, 25%, 50% and 75% of the original BED file, respectively) have been generated using a gene-level random



sampling strategy. Random sampling was performed by first annotating the original Nimblegen BED file adding the HUGO symbols of the genes corresponding to the captured regions. Random genes were then uniformly sampled from the set of all genes (without replacement) and all overlapping captured regions were incrementally added to a BED file until the desired target size was reached.

BAM files corresponding to all combinations reported in **Table S1.2** were created from the 1,000 Genomes Project HG02057 individual FASTQ files. Alignment was performed with BWA (53), SAMtools were used to create BAM files, GATK was used to perform realignment and recalibration and SAMtools were finally used to fix MD tags. BAM file subsampling was done using SAMtools view command, specifying the fraction of the reads (using -s option) to obtain the desired mean depth of coverage and providing pre-designed BED files (using -L option) to get the desired target representation.

#### 4. Design of performance experiments and reproducibility

To allow reproducibility of our performance analysis, we created a Debian-based Singularity container. The container provides a standardized and configured environment with all the dependencies required to run all tools and replicate the overall analysis. All used BAM, BED and VCF files and all implemented scripts are available in the container. The analysis pipeline runs each of tested tools three times for each combination of input parameters. For each run the elapsed real time (wall clock) and the peak memory usage of the process was measured using the GNU time command. No other user process was running on the test machine while collecting performance data. The pipeline also produces automatically all tables and images related with performance evaluation here reported.

#### 5. Output files specification

##### 2.1 File \*.pileup

For each genomic position specified in the input BED file it provides:

- Contig (e.g. chromosome)
- Genomic coordinate of the position
- Read depth of the 4 possible bases A, C, G and T
- Variant allelic fraction (VAF, considering all alternative bases)
- Total depth of coverage

- Strand bias information for each base (when run option “strandbias” is used)

### 2.2 File \*.snps

For each genomic position specified in the input VCF file and present in the regions specified in the input BED file it provides:

- Contig (e.g. chromosome)
- Genomic coordinate of the position
- Position ID (e.g. rsID) specified in the VCF input file
- Reference and alternative bases
- Read depth of the 4 possible bases A, C, G and T
- Variant allelic fraction (VAF, computed with alternative base specified in the VCF file)
- Total depth of coverage
- Genotype (when run option “genotype” or “genotypeBT” is used)

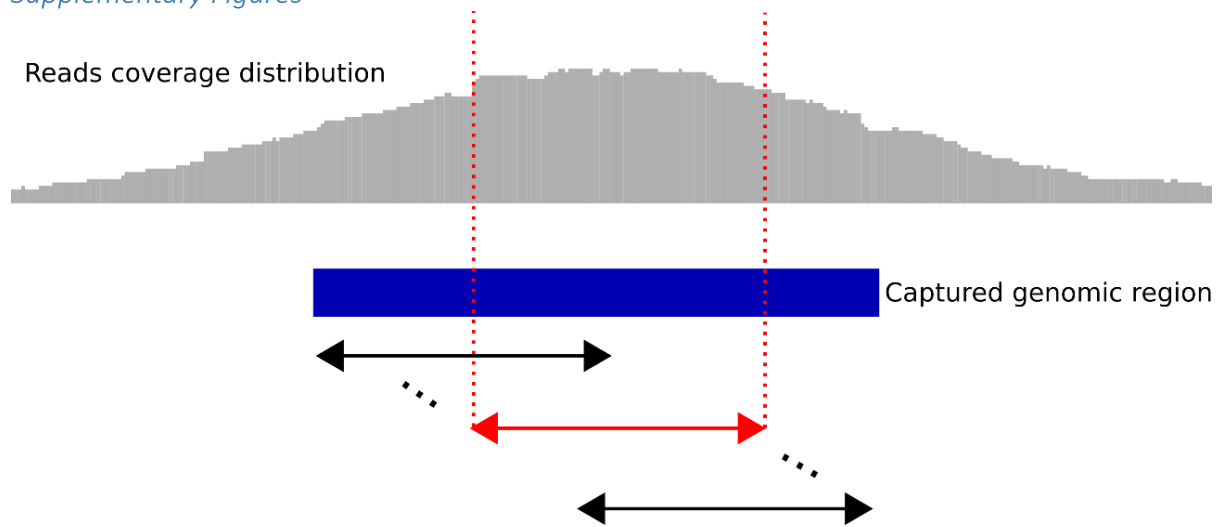
### 2.3 File \*.rc

- Contig (e.g. chromosome)
- Genomic start/end coordinates of the region
- Genomic start/end coordinates of the *peaked* region
- Mean read depth of the region
- Mean read depth of the *peaked* region
- GC content in the region

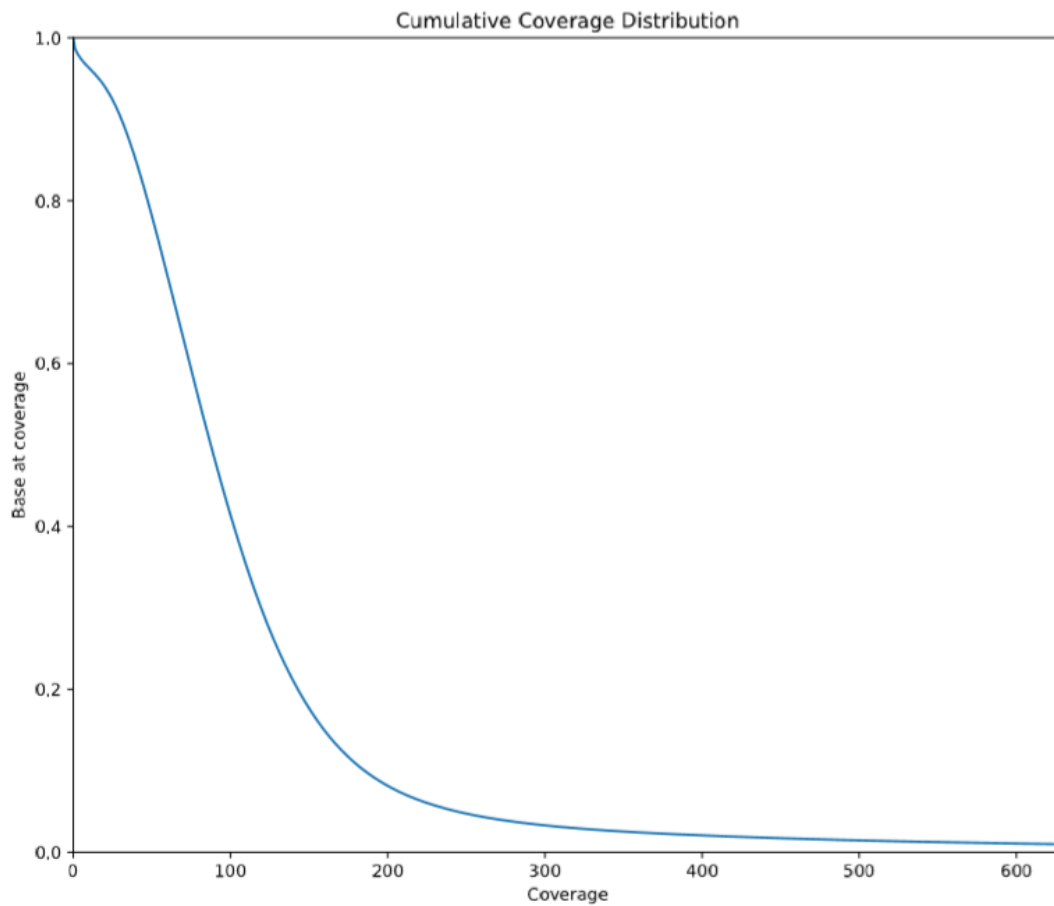
### 2.4 File \*.pabs

- Contig (e.g. chromosome)
- Genomic coordinate of the position
- Reference and alternative bases
- Read depth of the 4 possible bases A, C, G and T
- Variant allelic fraction (VAF, considering all alternative bases)
- Total depth of coverage
- Strand bias information for each base (when run option “strandbias” is used)

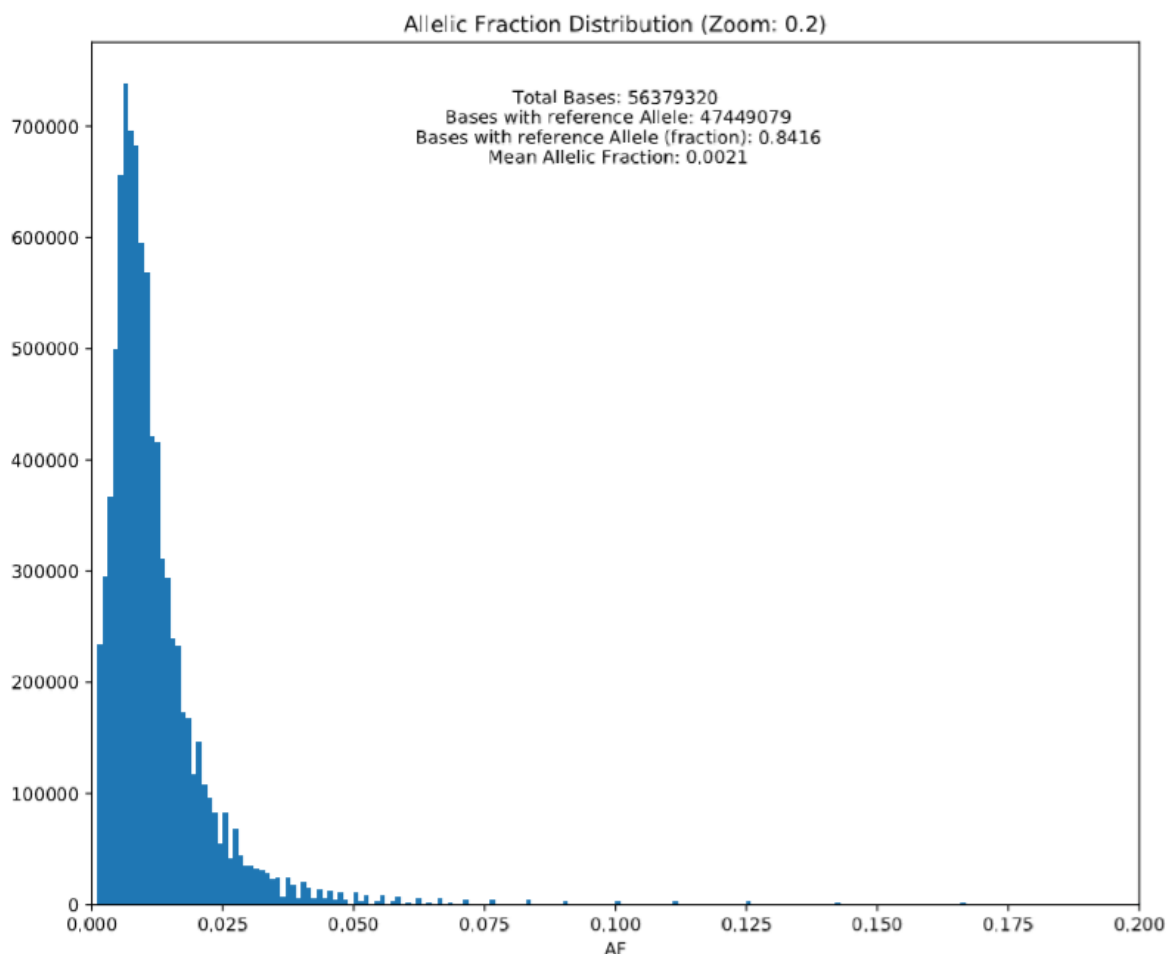
Supplementary Figures



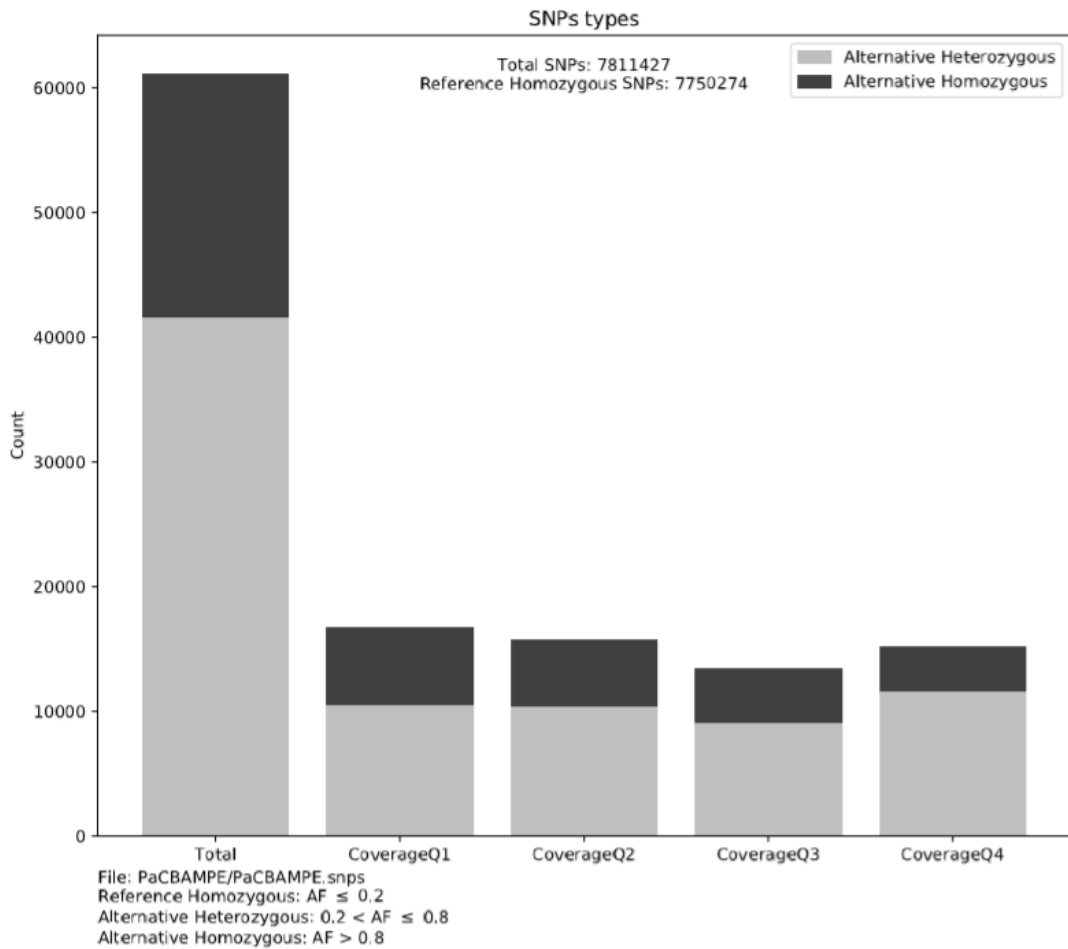
*Figure S1.1: Genomic region mean coverage computation. Example of region mean coverage computation using a user specified region fraction equal to 0.5. In red the 0.5 fraction of the region supporting the maximum mean depth of coverage for that region. This value (along with corresponding genomic coordinates) is reported in the PaCBAM output together with the overall region mean depth of coverage.*



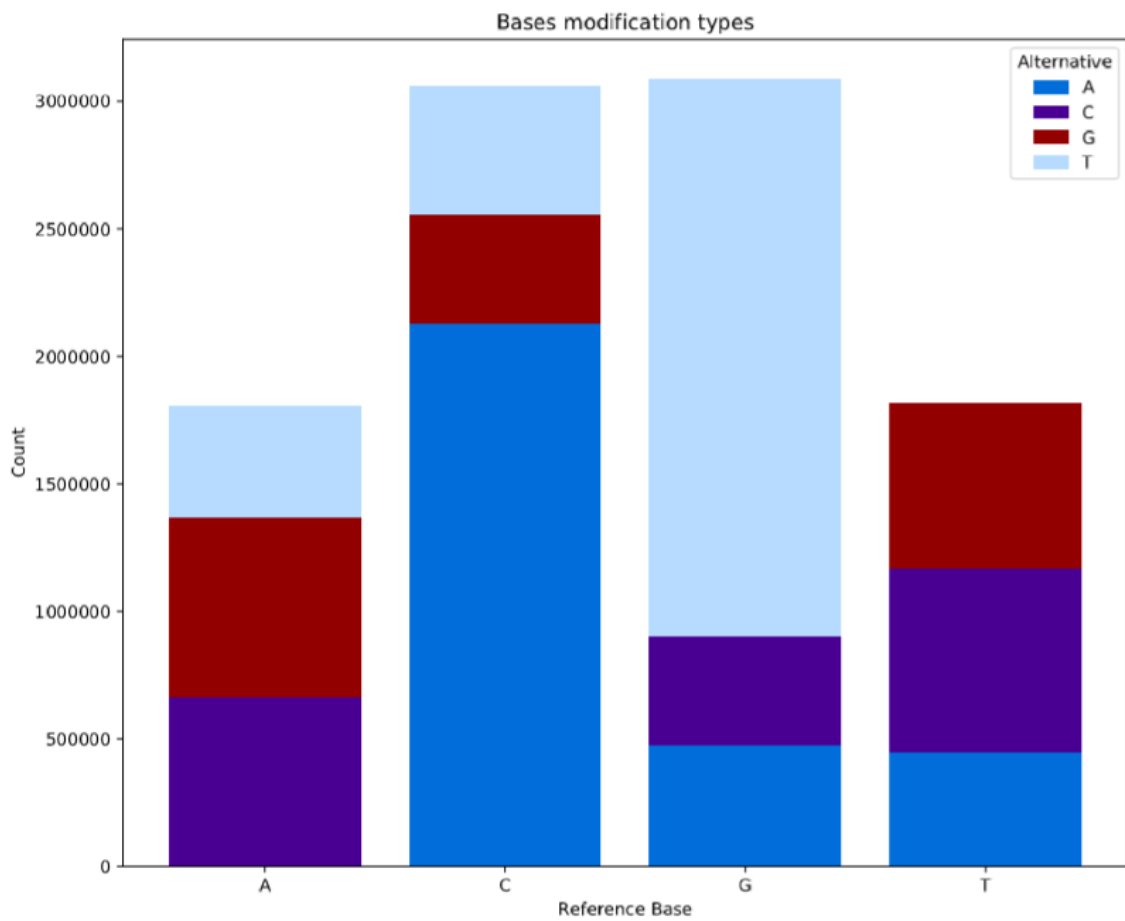
*Figure S1.2: **Cumulative coverage distribution report.** Example of visual report of the cumulative coverage distribution for all positions reported in the PaCBAM pileup output file.*



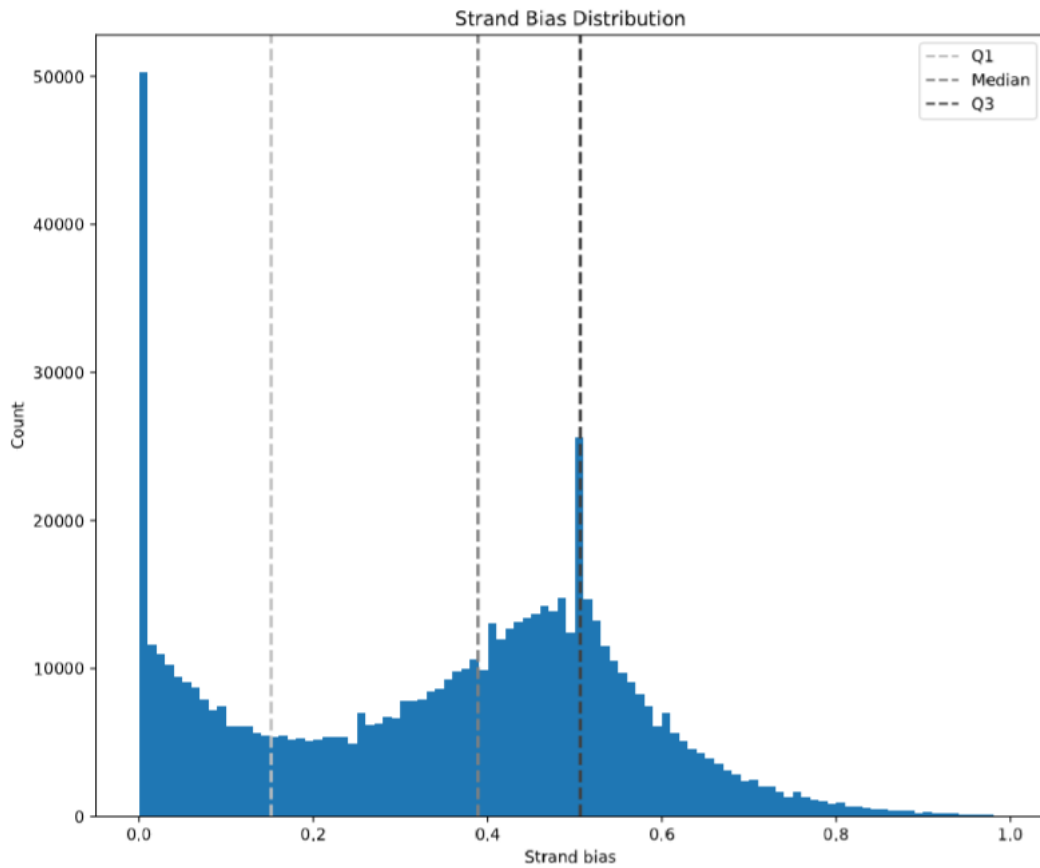
*Figure S1.3: Variant allelic fraction distribution report. Example of visual report of variants allelic fraction distribution for all positions in the PaCBAM output pileup. The image shows a detail in the variants allelic fraction range (0-0.2]. This data could provide information regarding the sequencing error distribution for the specific experiment.*



*Figure S1.4: SNP allelic fraction distribution report. Example of visual report of the allelic fraction (AF) distribution of all positions contained in the PaCBAM \*.snps output file. SNPs are classified as heterozygous or alternative homozygous based on standard AF thresholds. Classification is also reported stratified by coverage quartiles.*



*Figure S1.5: Alternative bases distribution report. Example of visual report of the distribution of alternative bases found for each reference base across all positions reported in the \*.pabs PaCBAM output file (i.e. all positions with non-zero variant allelic fraction).*

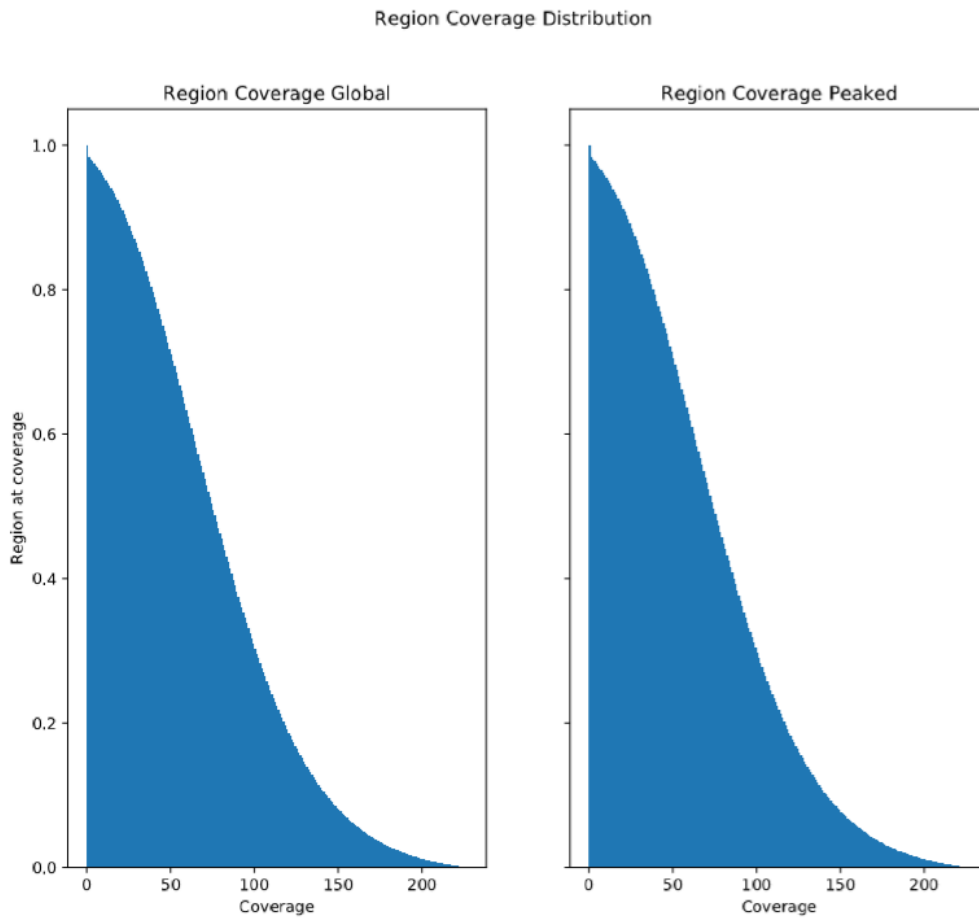


**Figure S1.6: Strand bias distribution report.** Example of PaCBAM visual report of the distribution of strand bias computed across all positions reported in the \*.snvs PaCBAM output file (i.e. all positions with non-zero variant allelic fraction). Strand bias is computed at each position using the formula:

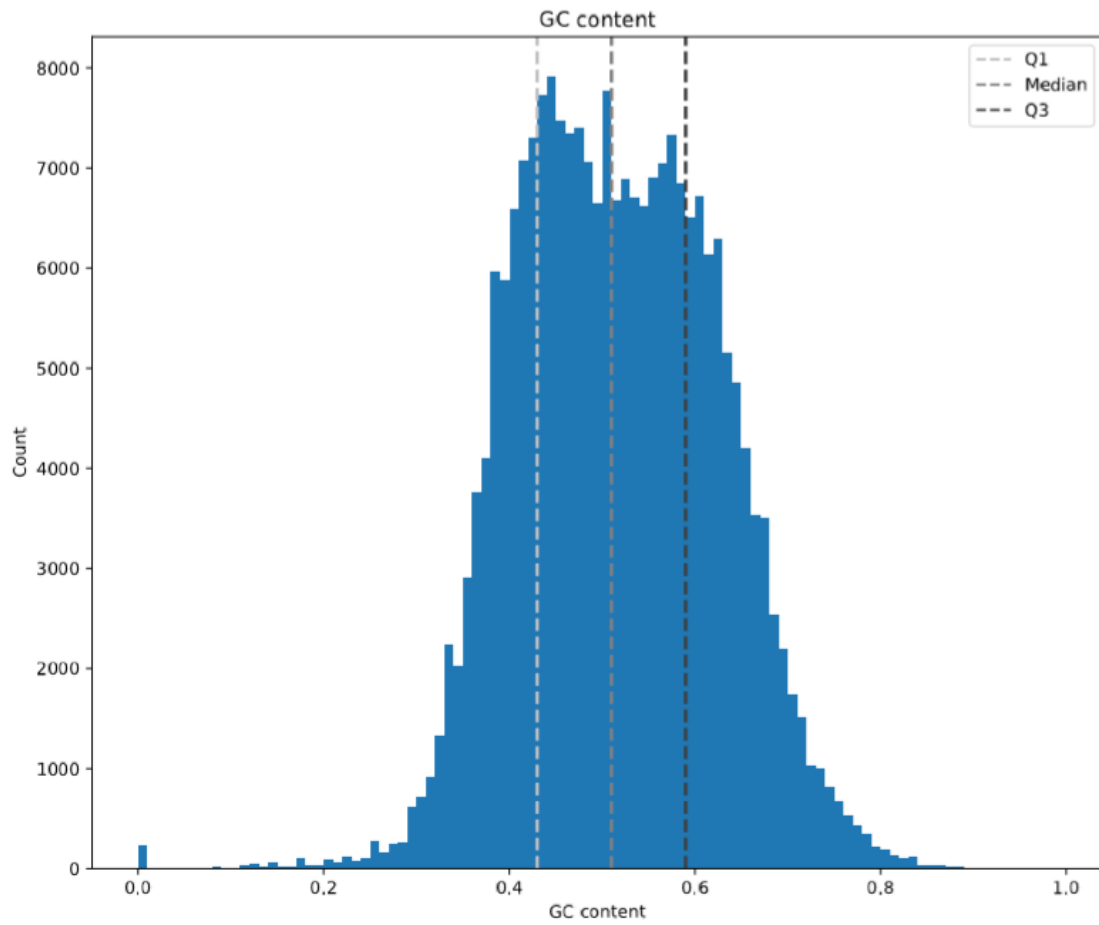
$$\text{abs}\left(\frac{\text{REFERENCE}_{\text{reverse}}}{\text{REFERENCE}_{\text{tot}}} - \frac{\text{ALTERNATIVE}_{\text{reverse}}}{\text{ALTERNATIVE}_{\text{tot}}}\right)$$

which computes the absolute difference between ratio of the number of reverse reads supporting the reference base over the total number of reads supporting the reference base and the ratio of the number of reverse reads supporting the alternative base over the total number of reads supporting the alternative base. Values towards 1 represent strong strand bias for the corresponding position.

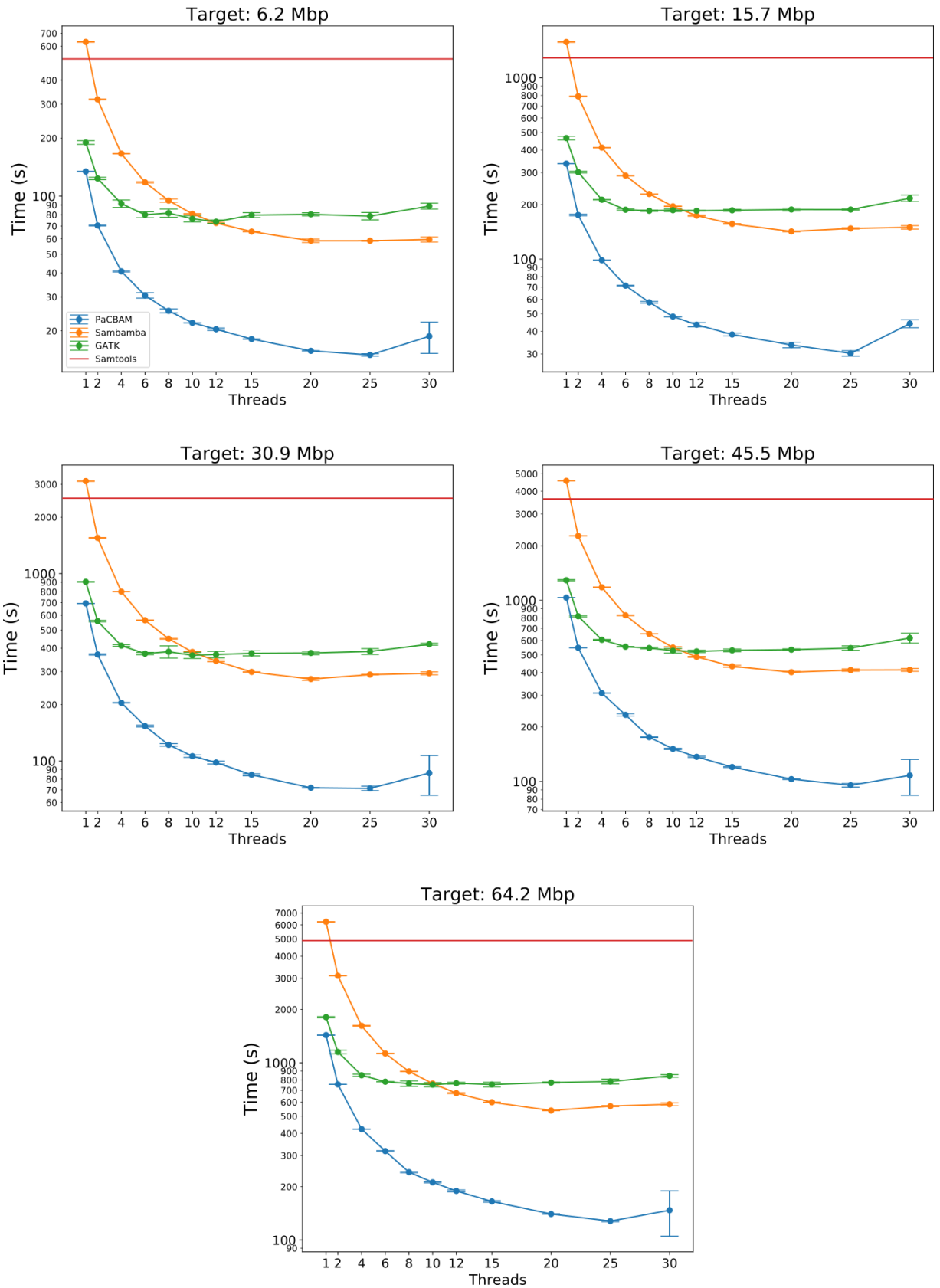




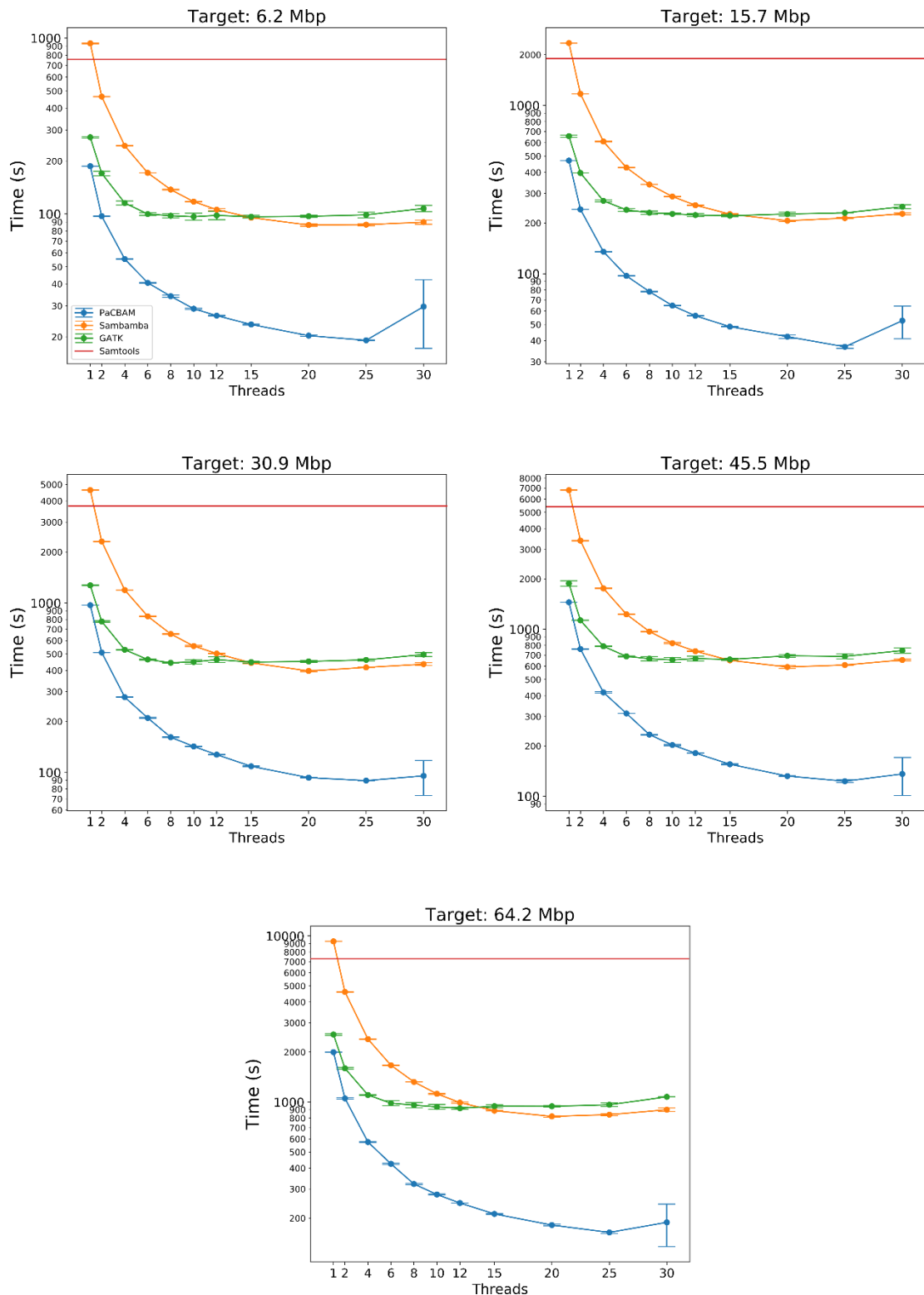
*Figure S1.7: Genomic regions depth of coverage distribution report. Example of visual report of the mean depth of coverage distribution computed across all regions reported in the PaCBAM output file. Distribution is reported both for regions overall mean coverage and for regions fractions maximizing mean coverage.*



*Figure S1.8: **Genomic regions GC content distribution report.** Example of visual report of the distribution of GC content computed across all regions reported in the \*.rc PaCBAM output file.*



**Figure S1.9: Run time comparison at 150X depth of coverage.** Run time comparison among PaCBAM pileup and pileup module of SAMtools, GATK and Sambamba. Comparison is performed on BAM files at mean depth of coverage ~150X, at different target sizes and by increasing the number of threads.



*Figure S1.10: Run time comparison at 230X depth of coverage. Run time comparison among PaCBAM pileup and pileup module of SAMtools, GATK and Sambamba. Comparison is performed on BAM files at mean depth of coverage ~230X, at different target sizes and by increasing the number of threads.*

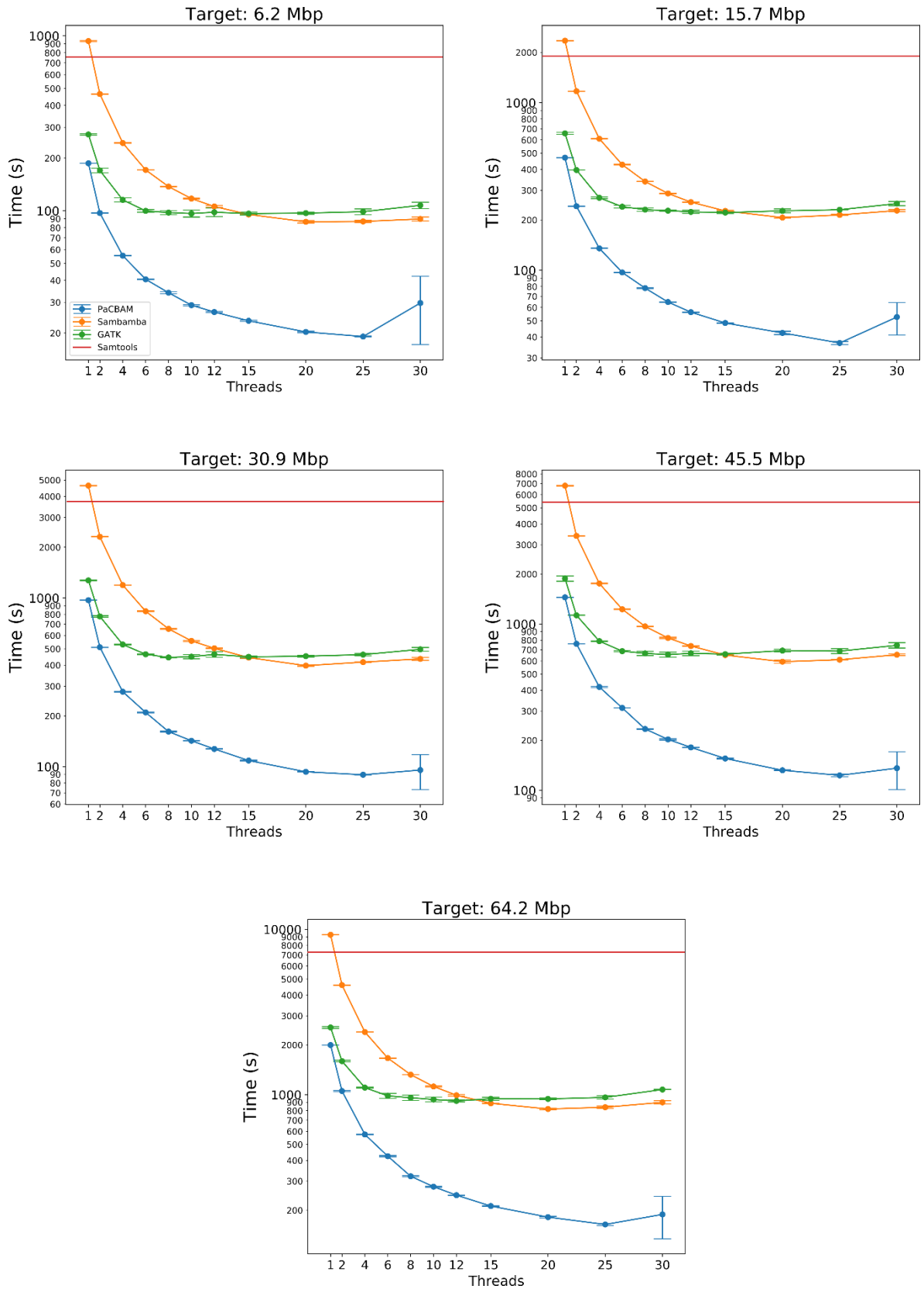


Figure S1.11: **Run time comparison at 300X depth of coverage.** Run time comparison among PaCBAM pileup and pileup module of SAMtools, GATK and Sambamba. Comparison is performed on BAM files at mean depth of coverage ~300X, at different target sizes and by increasing the number of threads.

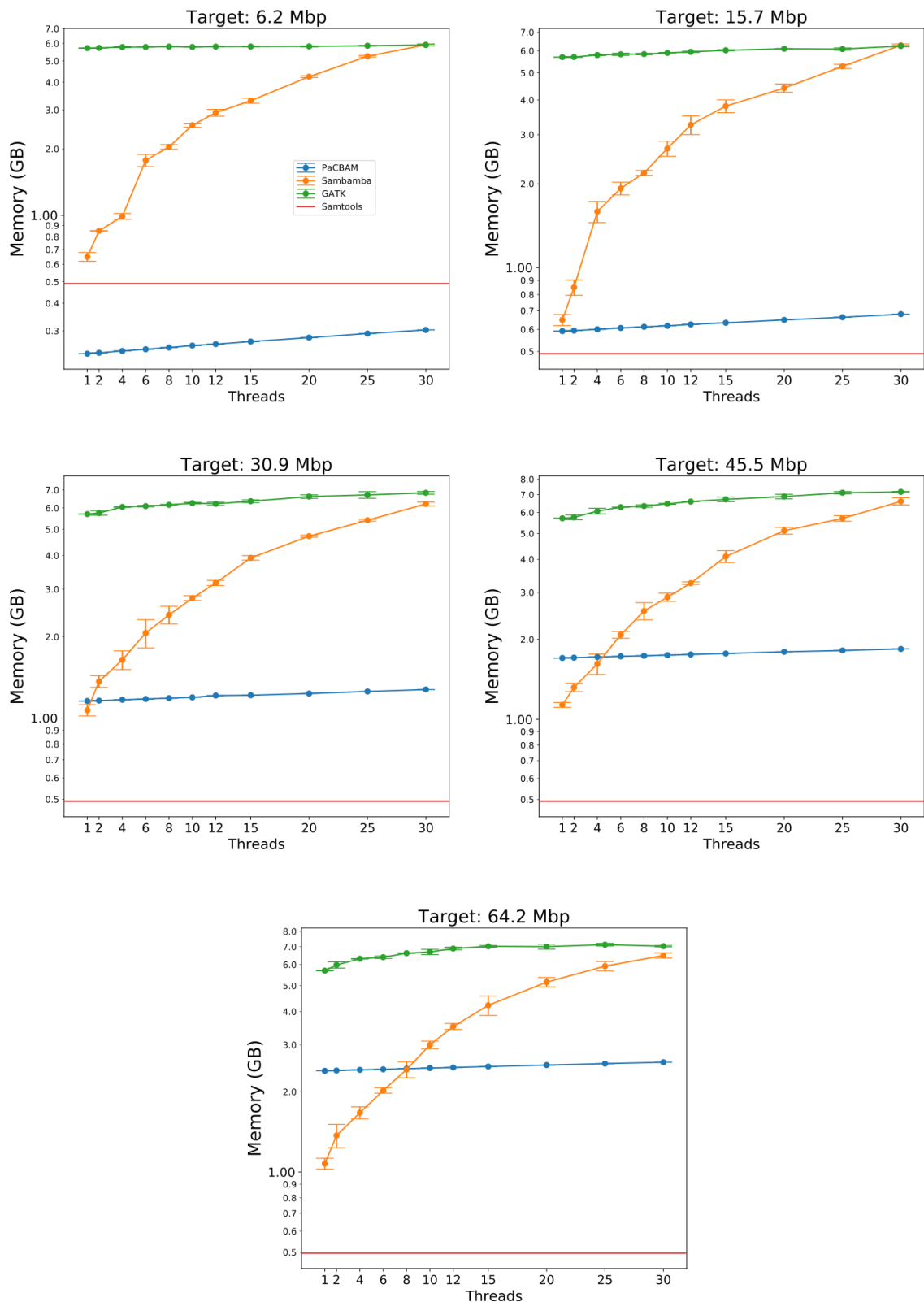


Figure S1.12: *Memory usage comparison at 150X depth of coverage. Memory usage comparison among PaCBAM pileup and pileup module of SAMtools, GATK and Sambamba. Comparison is performed on BAM files at mean depth of coverage ~150X, at different target sizes and by increasing the number of threads.*

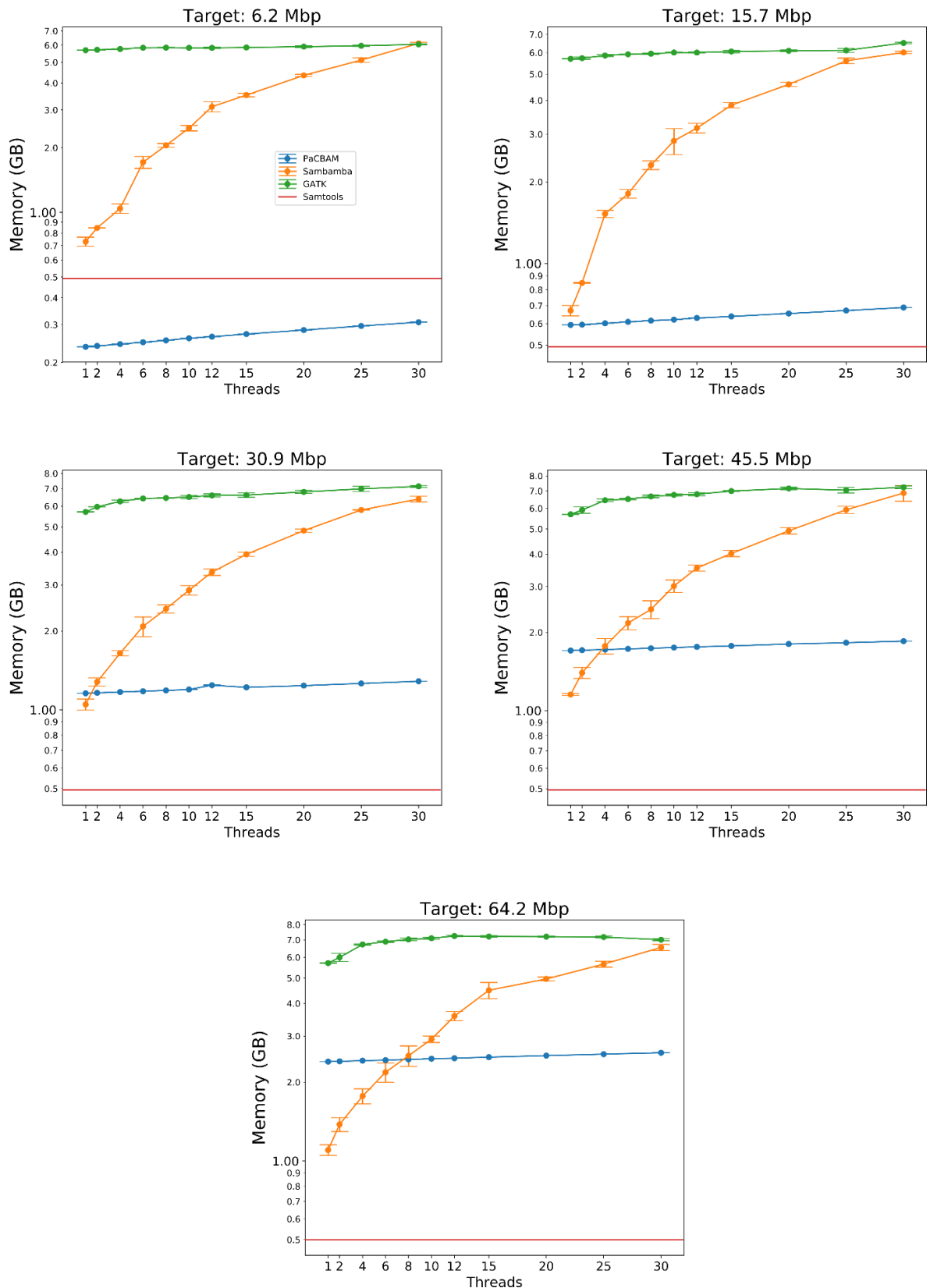


Figure S1.13: **Memory usage comparison at 230X depth of coverage.** Memory usage comparison among PaCBAM pileup and pileup module of SAMtools, GATK and Sambamba. Comparison is performed on BAM files at mean depth of coverage ~230X, at different target sizes and by increasing the number of threads.

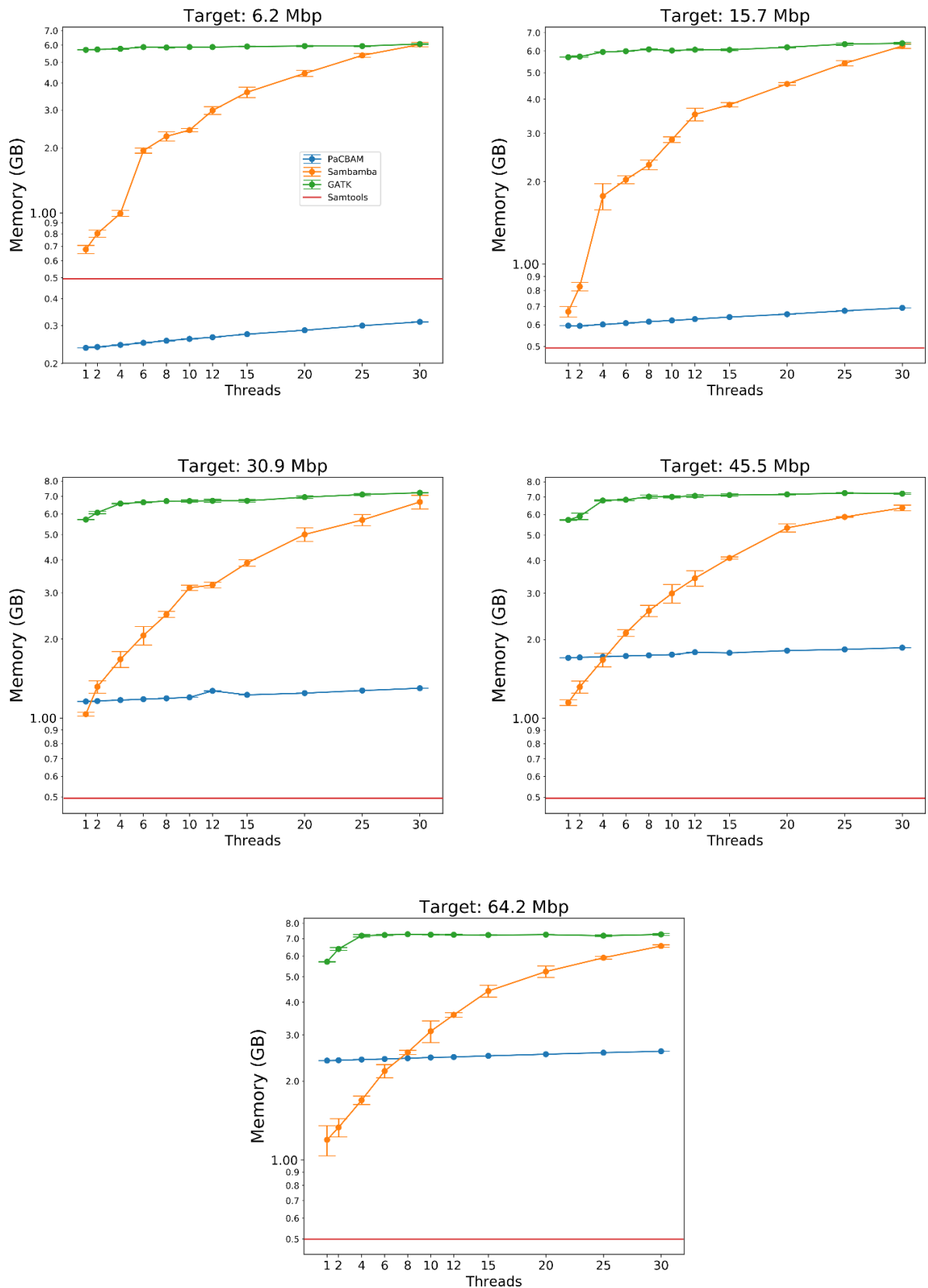
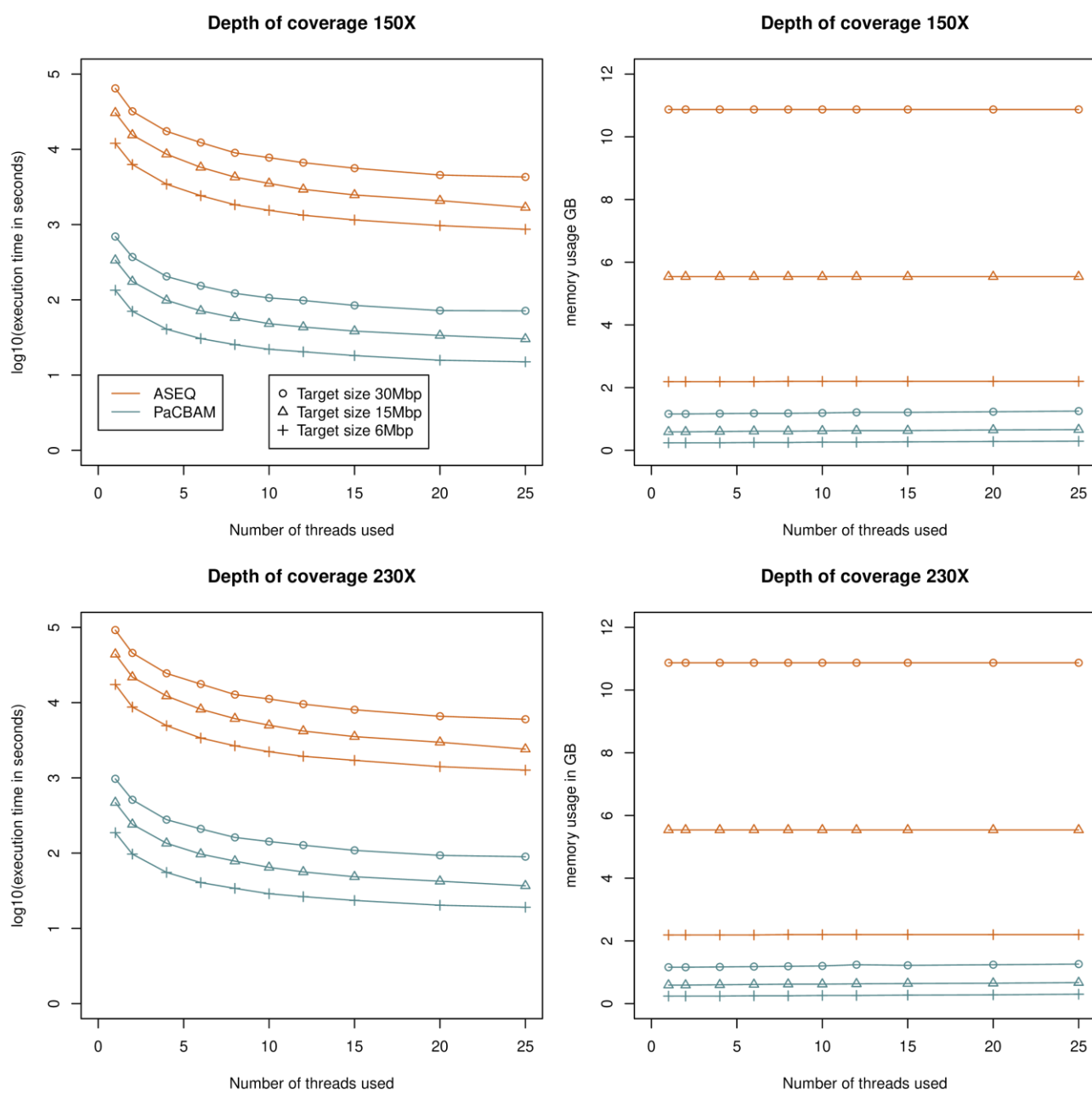
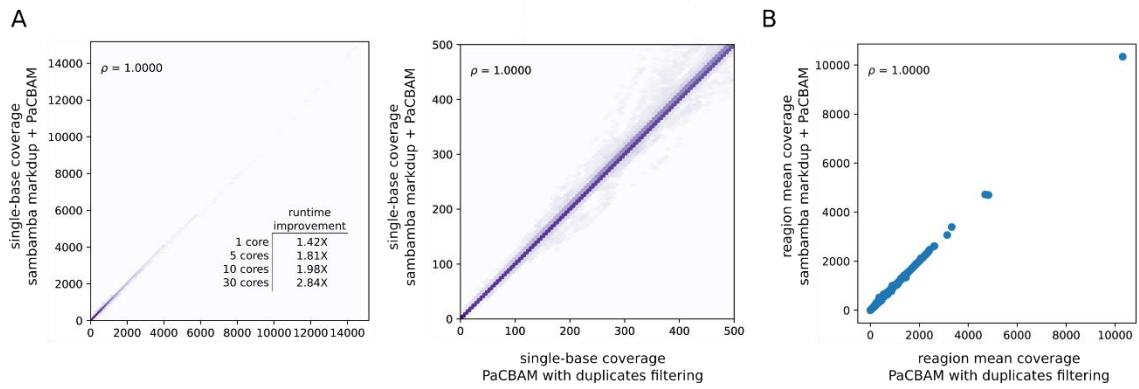


Figure S1.14: **Memory usage comparison at 300X depth of coverage.** Memory usage comparison among PaCBAM pileup and pileup module of SAMtools, GATK and Sambamba. Comparison is performed on BAM files at mean depth of coverage ~300X, at different target sizes and by increasing the number of threads.

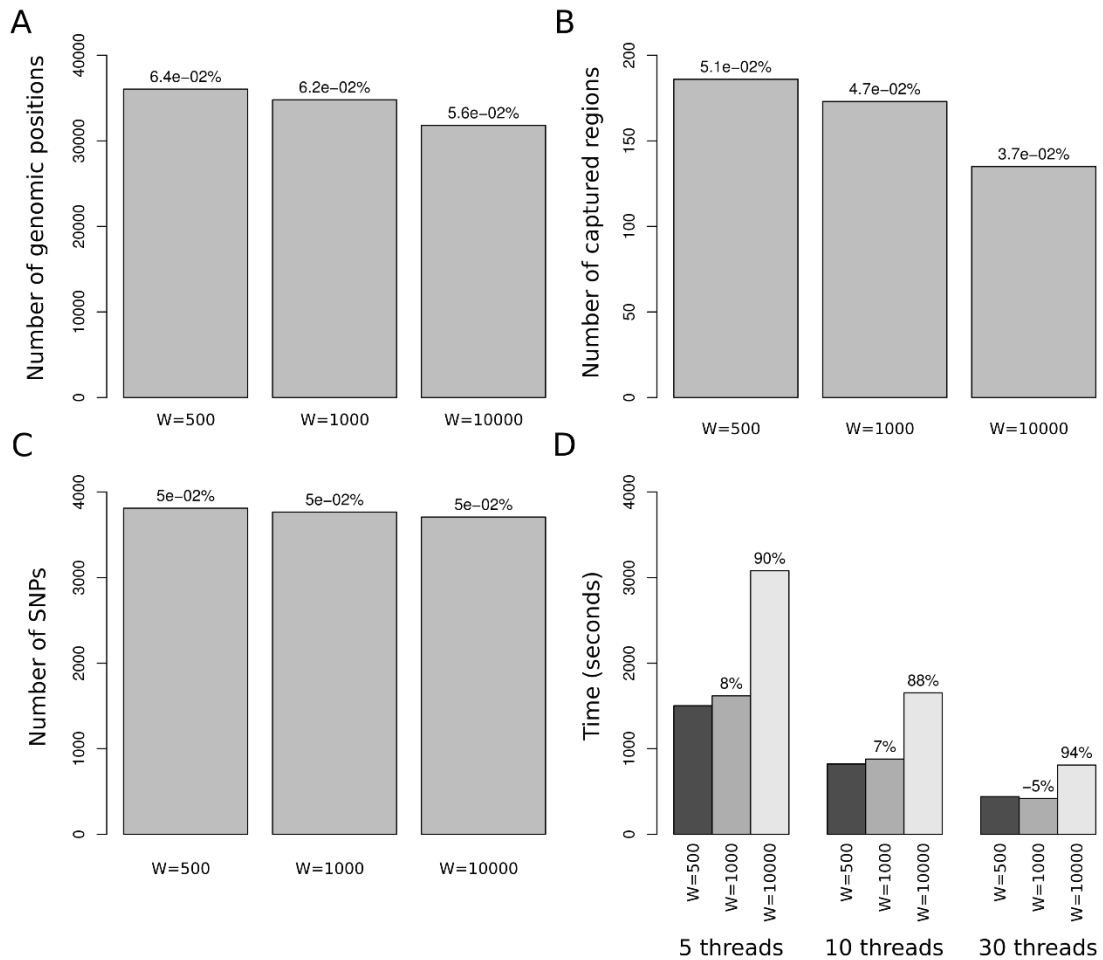




*Figure S.1.15: Memory usage comparison among PaCBAM pileup and pileup module of ASEQ. Comparison is performed on BAM files at mean depth of coverage ~150X and ~230X, at different target sizes and by increasing the number of threads.*



**Figure S1.16: Comparison of PaCBAM duplicates filtering strategy to Sambamba markdup and Picard MarkDuplicates modules.** A) Read duplicates filtering comparison between Sambamba markdup and PaCBAM (left) with zoom at smaller coverage interval (right). B) Regional mean depth of coverage of a BAM obtained by running either sambamba markdup + PaCBAM pileup or PaCBAM pileup with duplicates filtering option active. All results are highly concordant with correlation equal to 1.



**Figure S1.17: Performance of PaCBAM duplicated reads filtering.** A) Number of positions that have a difference in coverage with respect to Picard results  $\geq 10$  reads across different values for the  $W$  PaCBAM parameter; percentages on the bars are calculated with respect to the total number of genomic positions captured in the considered WES kit ( $N=56,379,320$ ), excluding SNVs annotated in dbSNP v151. B) Number of regions that have a difference in coverage with respect to Picard results  $\geq 10$  reads across different values for the  $W$  PaCBAM parameter; percentages on the bars are calculated with respect to the total number captured regions in the considered WES kit ( $N=368.146$ ). C) Number of SNPs that have an allelic fraction difference with respect to Picard results  $\geq 1\%$  across different  $W$  PaCBAM parameters; percentages on the bars are calculated with respect to the total number of SNVs annotated in dbSNP v151 and present in the WES kit ( $= 7,811,427$ ). Results in panels A, B and C are computed comparing runs of PaCBAM with duplicated reads filtering and runs of PaCBAM on files output of Picard MarkDuplicates. D) Execution time of PaCBAM with duplicated reads filtering across different values for the  $W$  PaCBAM parameter and number of threads; percentages on the bar represent the time increase with respect to the previous bar value.

Supplementary Tables

Mean depth of coverage	Target size (bp)
306.02	64,190,747
229.51	64,190,747
152.99	64,190,747
314.78	45,535,680
236.09	45,535,680
157.37	45,535,680
315.59	30,893,295
236.70	30,893,295
157.79	30,893,295
314.03	15,744,773
235.54	15,744,773
157.02	15,744,773
314.28	6,229,847
235.70	6,229,847
157.13	6,229,847

*Table S1.1: Mean depth of coverage and target sizes of all BAM files used to test PaCBAM performance.*

Tools	Threads	Time (s)	Memory (MB)
DedupPaCBAM Duplicate Window 500	1	6557.00	3823.49
DedupPaCBAM Duplicate Window 500	5	1502.66	3878.53
DedupPaCBAM Duplicate Window 500	10	821.73	3937.54
DedupPaCBAM Duplicate Window 500	30	440.45	4182.63
DedupPaCBAM Duplicate Window 1000	1	7082.33	3823.46
DedupPaCBAM Duplicate Window 1000	5	1618.61	3887.65
DedupPaCBAM Duplicate Window 1000	10	879.37	3959.21
DedupPaCBAM Duplicate Window 1000	30	417.96	4236.96
DedupPaCBAM Duplicate Window 10000	1	13105.67	3853.50
DedupPaCBAM Duplicate Window 10000	5	3080.07	3941.18
DedupPaCBAM Duplicate Window 10000	10	1652.61	4052.34
DedupPaCBAM Duplicate Window 10000	30	809.34	4569.39
Sambamba markdup + PaCBAM pileup	1	9308.40	3802.29
Sambamba markdup + PaCBAM Pileup	5	2712.57	5290.98
Sambamba markdup + PaCBAM Pileup	10	1626.27	7441.42
Sambamba markdup + PaCBAM Pileup	30	1251.32	18473.83
Picard markdup + PaCBAM Pileup	1	12422.76	28994.42
Picard markdup + PaCBAM Pileup	5	10764.17	28994.42
Picard markdup + PaCBAM Pileup	10	10541.08	28994.42
Picard markdup + PaCBAM Pileup	30	10396.25	28994.42

*Table S1.2: Time and memory usage of duplicates filtering performance analyses. When combining MarkDuplicates and PaCBAM the memory usage is the peak memory usage of the entire pipeline.*

<b>Tool</b>	<b>Version</b>
Sambamba	0.6.8-pre1 compiled with LDC 1.8.0 and LLVM 5.0.1
GATK	3.8-0-ge9d806836
SAMtools	1.7
Picard MarkDuplicates	2.17.4

**Table S1.3: Versions of the tools used in performance evaluation analysis.**

# Chapter 2: Identification of variants affecting mRNA translation potential

## Introduction

Most of the studies on variants have been focusing on coding or non-coding variants in regulatory regions. Less focus has been put in understanding the effects of variants in genes UTRs where post-transcriptional regulation can happen. In particular, variants can alter microRNA binding sites or can modify the RNA translation potential leading to a phenomenon where one allele is expressed more than the other.

In this paper, we analyzed the MCF7 cell line under different treatments total and polysomal RNA and we developed a new method to identify possible SNPs associated with allele specific expression.

## Results

In this paper we developed a new method to identify SNPs that can alter the mRNA translation potential. We analyzed the MCF7 cell line in three different conditions: mock, doxorubicin and Nutlin. We sequenced the total and the polysomal fraction of the RNA in the treated cells. We identified 11,544 heterozygous SNPs in the MCF7 cell line and, of those, 1,802 are in 3' UTR while 729 are in the 5'UTR using public data then, using a pileup approach we identified 3,974 variants analyzable in our experiments. We then identified 147 unique variants that show unbalance between the polysomal and the total fraction of RNA. We experimentally validated two SNPs in UTRs that are close to genes related to p53 and we showed that, in fact, the variants could alter the translation efficiency. We then explored the effect of the identified variants on the survival data of the TCGA dataset where we found both protective and hazardous SNPs.

Finally, we analyzed the changes in RNA Binding Proteins motifs scores in variants showing a change in the survival. We identified several motifs disruption. We validate one of these by using a RIP assay.

## Materials and methods

Variants in the MCF7 cell line have been retrieved from two public studies and the two dataset have been merged keeping only the concordant heterozygous SNPs in common. The allelic imbalance between the two fractions of RNA has been computed using a pileup approach keeping reads and bases with at least a quality of 20 and a coverage of 10.

Starting from the 147 transSNPs, we retrieved all the variants in linkage disequilibrium with a  $r^2 > 0.8$ . We then computed the Kaplan-Meier survival curves under dominant and recessive models on the variables: Overall Survival, Disease-Specific Survival, Disease-Free Interval and Progression-Free Interval. We considered only variants with a p value  $< 0.05$  and we aggregate the result in the LD block requiring at least 5% of the variant to be associated with the trait.

Starting from the 33 UTR transSNPs identified in the survival analysis we performed a motif analysis. For each variant we analyzed the reference sequence (hg19) and the sequence altered with the variant.

We treated MCF7 cells with doxorubicin or Nutlin for 16 hours. From the cytoplasmic lysates we fractionated the polysomes. We sequenced two biological replicates for each RNA type for each condition. Sequencing was performed on an Illumina HiSeq.

I contributed to the article by analyzing and characterizing the identified transSNPs, by performing the survival analyses using TCGA data and by developing and analyzing the RNA Binding Proteins motifs analysis.

## Discussion

Using polysomial and total RNA fractions in MCF7 cell lines we were able to model and quantify small difference in the relative amount of RNA fractions.

We identified 147 variants (about 4% of the total variants analyzed) that exhibit allele specific expression pattern among at least one treatment condition. In our transSNPs catalog we identified variants associated to cancer risk and some other that have a prognostic. More specifically, variants with a prognostic value usually fall in genes that are related to cancer.



Our approach allowed us to detect SNPs that affect the RNA translation efficiency showing a small but significant prognostic value. Those variants can be helpful in stratifying patient for clinical outcome.

## Article

TranSNPs: A class of functional SNPs affecting mRNA translation potential revealed by fraction-based allelic imbalance

Samuel Valentini<sup>1\*</sup>, Caterina Marchioretto<sup>1,2\*</sup>, Alessandra Bisio<sup>1\*</sup>, Annalisa Rossi<sup>1</sup>, Sara Zaccara<sup>1,3</sup>, Alessandro Romanel<sup>1#</sup>, Alberto Inga<sup>1#</sup>

<sup>1</sup> Department of Cellular, Computational and Integrative Biology (CIBIO), University of Trento, Trento, Italy

<sup>2</sup> Department of Biomedical Sciences (DBS), University of Padova, Padova, Italy

<sup>3</sup> Weill Medical College, Cornell University, New York 10065, New York, USA

\*Equal contribution

#Corresponding authors

**Journal:** iScience Volume 24, Issue 12, 17 December 2021, 103531

**Publisher:** Elsevier

**Doi:** <https://doi.org/10.1016/j.isci.2021.103531>

## Summary

Few studies have explored the association between SNPs and alterations in mRNA translation potential. We developed an approach to identify SNPs that can mark allele-specific protein expression levels and could represent sources of inter-individual variation in disease risk. Using MCF7 cells under different treatments, we performed polysomal profiling followed by RNA-sequencing of total or polysome-associated mRNA fractions and designed a computational approach to identify SNPs showing a significant change in

the allelic balance between total and polysomal mRNA fractions. We identified 147 SNPs, 39 of which located in UTRs. Allele-specific differences at translation level were confirmed in transfected MCF7 cells by reporter assays. Exploiting Breast Cancer data from TCGA we identified UTR SNPs demonstrating distinct prognosis features and altering binding sites of RNA binding proteins. Our approach produced a catalog of *tranSNPs*, a class of functional SNPs associated with allele-specific translation and potentially endowed with prognostic value for disease risk.

## Introduction

Single Nucleotide Polymorphisms (SNPs) represent one of the largest classes of genetic variations. They underlie and are responsible for inter-individual variations in complex-disease phenotypes, including cancer risk or aggressiveness. Wide attention has been given to SNPs that can lead to allele-specific changes in gene expression, for instance, by modifying the affinity of transcription factor binding sites in promoter or enhancer regions directly, or indirectly via influencing epigenetic regulation. Included in these examples are functionally relevant SNPs affecting p53 function as a transcription factor or p53 protein expression by altering target binding sites (54–57), or association-driven studies that have candidate SNPs as modulators of major cancer drivers, such as *AR* (58,59), *ER* (60,61) or *cMYC* (62,63).

A fraction of SNPs identified in the human population is located within coding regions or UTRs. In this case, both mechanism- and association-driven studies have pursued functional SNPs that can modify aspects of post-transcriptional gene regulation, for example by altering microRNA binding sites (64).

Specific tools have been implemented to mine the available wealth of RNA-seq-based gene expression data and identify and pursue instances of allele-specific gene expression (65–72). The same cannot be said for annotating candidate SNPs driving or being associated with alterations in mRNA translation potential. Recently, we showed that allelic imbalance restricted to polysome-bound mRNAs can be exploited to investigate the functional significance of 5'UTR Single Nucleotide Variants (SNVs) at the *CDKN2A* gene (73), since specific SNVs affected *CDKN2A* translation. Those results led us to hypothesize that a comparative analysis of allelic-imbalance from total and polysomal mRNAs extracted and sequenced starting from the same cell sample, that is independently

genotyped for heterozygous SNPs, could overcome the intrinsic noise in SNP calling and coverage from RNA-seq data. If so, a catalogue of coding and UTRs SNPs associated and potentially causative of alterations in mRNA translation potential could be obtained. Here we describe such an approach starting from a commonly used breast cancer cell line and two disease-relevant treatments. About 4% of the heterozygous SNPs that could be followed showed a significant level of imbalance within polysomal RNAs. Nearly 25% of these are located in UTR sequences, and some appear to stratify cancer patients for distinct clinical outcomes.

Thus, our approach led us to distinguish both constitutive and treatment-dependent SNPs that are associated with or can cause changes in mRNA translation efficiency. We propose that our approach can lead to identify a class of functional SNPs, that we call *tranSNPs*, endowed with prognostic value for cancer or other diseases.

## Results

### *Using polysome profiling to identify SNPs exhibiting allelic imbalance*

In order to develop an approach to map SNPs associated with allelic imbalance within translating ribosomes, we took advantage of a well-characterized cancer cell line with available genotyping data. We thus chose MCF7 cells, a p53 wild type breast adenocarcinoma-derived cells we had previously used for polysomal profiling studies (74). After determining the set of heterozygous SNPs using DNA data, we exploited RNA-seq reads to compare relative allelic representation between the transcripts isolated in the cytoplasm because of their association with polysomes or lack thereof. Besides a mock condition, we included two treatments, 1 $\mu$ M doxorubicin or 10 $\mu$ M Nutlin, that we found can reduce global translation, engage p53 responses, and lead to a significant uncoupling between transcriptome and translome changes. Indeed, many transcripts were significantly modulated by the treatments only at the level of polysomal RNA fractions (74,75). We reasoned that these treatments could uncover specific allelic imbalance events due to changes in translation specificity mediated, for instance, by the action of RNA binding proteins (74,75). Polysomal profiling was performed using cytoplasmic lysates fractionated by a linear 10-50% sucrose gradient, as previously described (74–76). Cytoplasmic mRNA associated with light fractions, with ribosome subunits, or with the 80S monosome-assumed to be not actively translating- were pooled

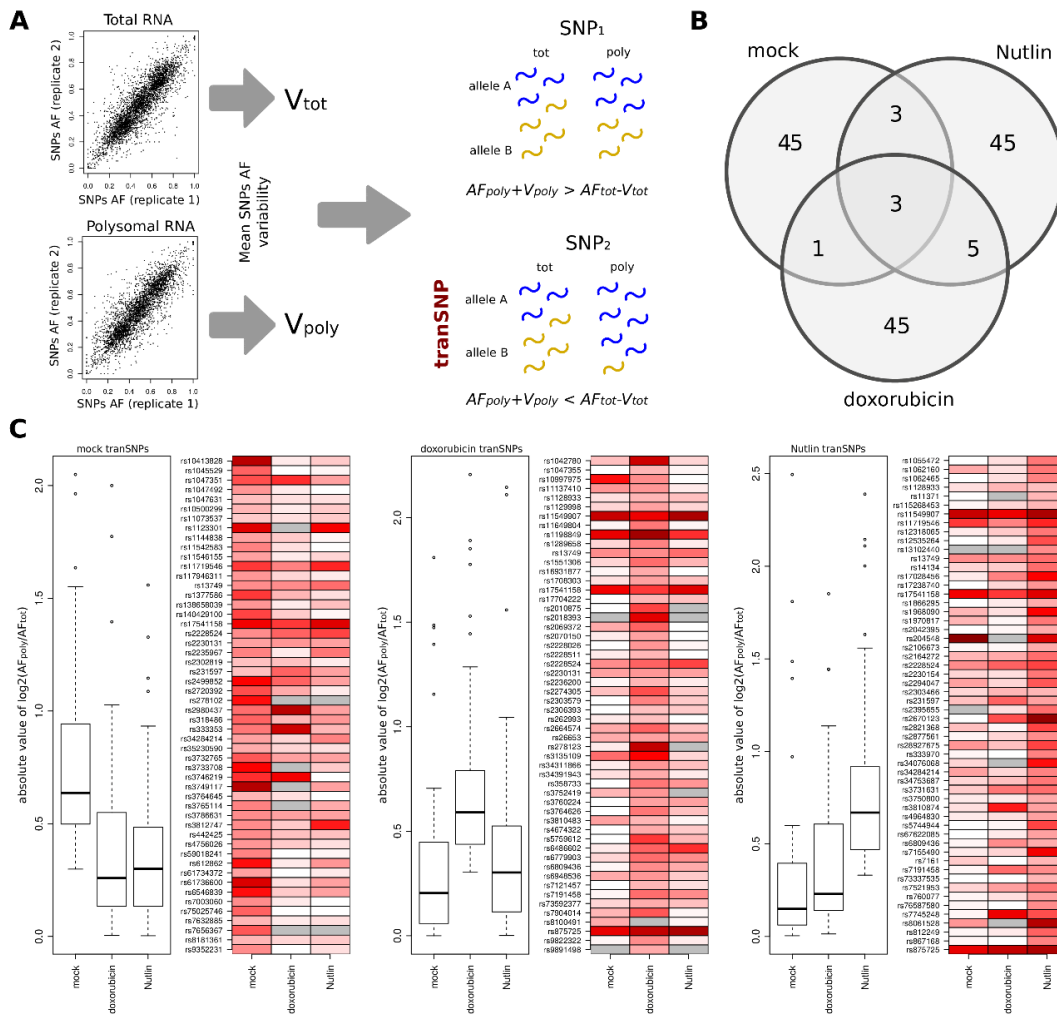
together and sequenced separately from mRNAs associated with two or more ribosomes. Total RNA was also collected and sequenced. The RNA-sequencing of two biological replicates for each treatment and fraction produced on average 32 million unique mapped reads. Differential gene expression analysis and pathway enrichment analysis confirmed that the two treatments activated a canonical p53 response (**Figure S2.1**).

#### *Identifying instances of allelic imbalance in polysomal RNA*

To identify SNPs exhibiting allelic imbalance in one or more RNA fractions, we first retrieved from public databases an exhaustive list of SNPs that are heterozygous in the MCF7 cell line. We integrated SNP-array and exome-based genotype calls and overall identified 11,544 heterozygous SNPs of which 1,802 in 3'UTRs and 729 in 5' UTRs (**Table S2.1**). Then, we determined the allelic fractions of these SNPs in our RNA-seq data using a pileup approach based on high quality reads and bases and requiring a minimum local depth of coverage of 10X. We found an average of 4,100 of the expected 11,544 SNPs per sample with confirmed coverage signal and, overall, we found 3,974 SNPs analyzable in at least one condition.

Our approach focuses on the comparison of SNP allelic counts across biological replicates of the two RNA fractions that were sequenced for each of the three treatments. We first determined which is the distribution of SNP allelic fractions across all experiments and measured to what extent they are variable across biological replicates. Interestingly, the variability of SNP allelic fractions was limited across biological replicates (**Figure S2.2**) with a mean replicate's difference of about 7% that was consistent among the different RNA fractions and treatments (**Table S2.2**). In addition, the range of SNP allelic fractions was quite wide (**Figure S2.2**). This suggests that heterozygous SNP allelic counts from RNA-seq data are potentially biased by position-specific sequencing properties, indicating that divergence from an expected 0.5 allelic fraction should not, as commonly done, be directly linked to putative allele-specific expression phenomena, nor to a specific expression imbalance direction. Hence, we designed and implemented a computational approach that first calculates a condition-specific variability of SNP allelic fractions exploiting the available biological replicates, and then uses it to identify significant SNP allelic imbalances across different conditions.

We used this approach to identify allelic imbalances between polysomal and total RNA fractions. Specifically, considering SNPs allelic fraction (AF) intervals calculated by summing and subtracting condition-specific AF variabilities from SNPs AFs, we searched for SNPs presenting polysomal and total RNA non-overlapping AF intervals (**Figure 2.1A**). The analysis led to 162 imbalance instances, involving 147 SNPs that were identified in both biological replicates preserving the imbalance direction. This conservative list of 147 SNPs represents the first catalogue of *tranSNPs*: polymorphisms exhibiting allelic imbalance specifically in polysome-bound mRNA fractions (**Table S2.3** and **S2.4**). The list is almost equally divided between constitutive, doxorubicin-dependent, Nutlin-dependent polysomal allelic imbalances (**Figure 2.1B**). Only 3 *tranSNPs* were common to all conditions, and only 8 to the two different p53-activating treatments. Despite the limited observed intersection among our conservative calls, the comparison of condition-associated *tranSNPs* allelic fractions across all conditions showed not only an expected clear shift in the condition-associated *tranSNPs* AF distribution, but also a fraction of SNPs with comparable imbalance across 2 or 3 conditions (**Figure 2.1C**). The condition-associated lists of genes harboring the *tranSNPs* showed mild enrichment for specific biological processes or molecular functions, including regulation of spindle organization, mitotic cytokinesis, protein kinase regulator activity and catalytic activity (**Figure S2.3**). Characteristics of *tranSNPs* are shown in **Table 2.1** and are compared with those in the larger set of 11,544 initial SNPs and those in the set of 3,974 analyzable SNPs. Of the 147 *tranSNPs* we identified, 39 are located in UTRs and represent top candidates for a direct role on the observed allelic imbalance.

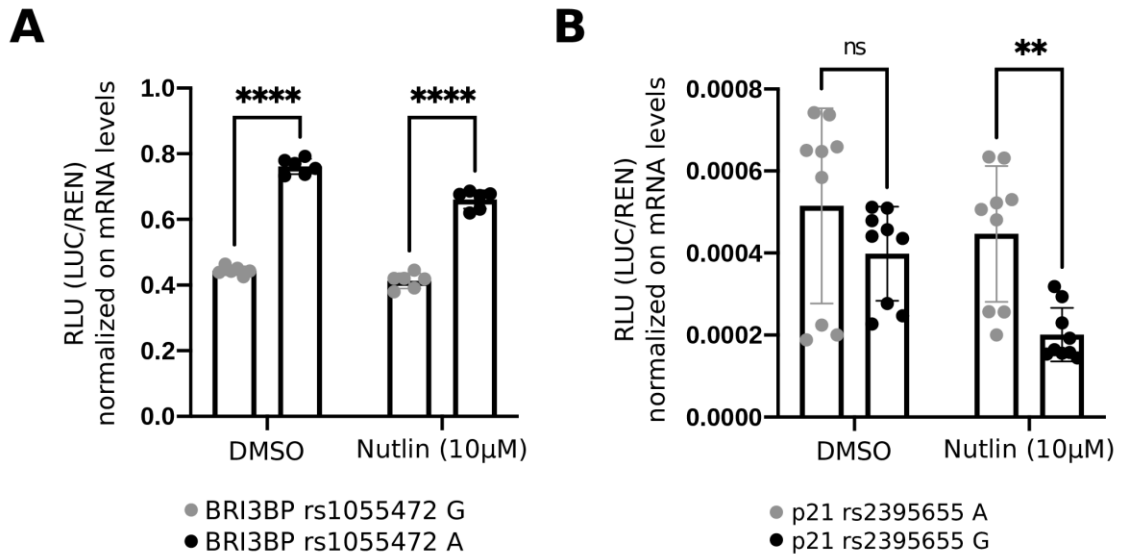


**Figure 2.1: Identification of SNPs allelic imbalance across different RNA fractions. A)** Schematic representation of the approach developed to identify RNA fraction-specific SNP allelic imbalances. RNA-seq based SNP allelic fraction variability is estimated both in total and polysomal RNA fractions. Then variability extended SNP allelic fractions are compared and only non-overlapping total versus polysomal imbalances are retained as transSNPs. In the example, SNP2 satisfies the condition and is hence nominated as transSNP. AF = allelic fraction; V = mean AF variability among replicates. **B)** Venn diagram showing private and shared transSNPs identified across the three analyzed conditions. **C)** Allelic imbalance distribution of condition-associated transSNPs is shown across the different conditions. Aggregate distribution is shown using boxplots, while single SNPs distribution is shown using a heatmap, where red intensity represents the level of imbalance. In the boxplot, the imbalance is shown as absolute  $\log_2$  ratio of allelic fraction in polysomal RNA and allelic fraction in total RNA. In the heatmap, red intensity is proportional to this value; grey represents no value.

*BRI3BP 3'UTR rs1055472 and CDKN1A 5'UTR rs2395655 alleles are functionally distinct.* We chose two UTR SNPs showing polysomal allelic imbalance for validation and opted for a gene reporter assay to evaluate the functional impact of the two pairs of *transSNPs*. We selected two Nutlin-dependent *transSNPs* and two genes harboring them whose functions are related to p53.

*BRI3BP*, also known as human cervical cancer oncogene 1 (*HCCR-1*) might act as negative modulator of p53 (77,78). The two rs1055472 alleles of the *BRI3BP* 373nt 3'UTR were cloned downstream of the Firefly cDNA. MCF7 cells were transiently transfected with each of the two alleles, and the activity of the reporter relative to the Renilla control luciferase was measured in untreated cells or in cells treated with Nutlin. In both cases we observed that the alternative allele led to a relative increase in the reporter gene activity (**Figure 2.2A**), suggesting an overall differential translation efficiency. Although the SNP was computationally classified as polysomally imbalanced only after Nutlin treatment (**Table S2.3** and **S2.4**), computational and experimental data combined suggest that the two alleles are functionally distinct, impacting on gene expression.

*CDKN1A* (p21) is an important cyclin-dependent kinase inhibitor and one of the major direct p53 transcriptional targets mediating cell cycle arrest (79,80). The gene is highly regulated at transcriptional and post-transcriptional level (76,81). *CDKN1A* transcripts were highly induced by both doxorubicin and Nutlin treatment and the coverage of rs2395655 alleles in the untreated condition was very low (<10X). We evaluated the *CDKN1A* 5'UTR rs2395655 alleles cloned upstream of the Firefly cDNA. We observed an overall bimodal distribution in the reporter gene activity, whose amplitude is reduced when the alternative allele is present (**Figure 2.2B**, left panel). The treatment with Nutlin further mitigated this bimodal distribution with an overall reporter gene activity that is strongly reduced when the alternative allele is present (**Figure 2.2B**, right panel). Since the SNP was concordantly classified as polysomally imbalanced after Nutlin treatment, in this case we highlight a scenario with functionally distinct alleles where differential translation efficiency may impact protein expression only under specific conditions.



**Figure 2.2: *TransSNPs* results in functionally distinct alleles.** **A)** MCF7 cells were transiently transfected with pGL4.13-based vectors containing BRI3BP 3'UTR fragments differing for the indicated BRI3BP SNP allele and the control pRLSV40 Renilla vector. After 24 hours of transfection, cells were treated with Nutlin for 24 hours before performing dual-luciferase assays. Firefly luciferase signals were normalized to Renilla to control for transfection efficiency and to relative Firefly mRNA levels to take into account differences in reporter's transcript levels. Individual values from independently transfected wells are plotted. **B)** Same as A), except that the p21-5'UTR was cloned in the low-expression pGL3-basic vector. \*\*  $p < 0.01$ ; \*\*\*\*  $p < 0.0001$ , adjusted  $p$ -value based on a 2-way ANOVA with Sidak's multiple comparison test. Data are represented as mean  $\pm$  SD.

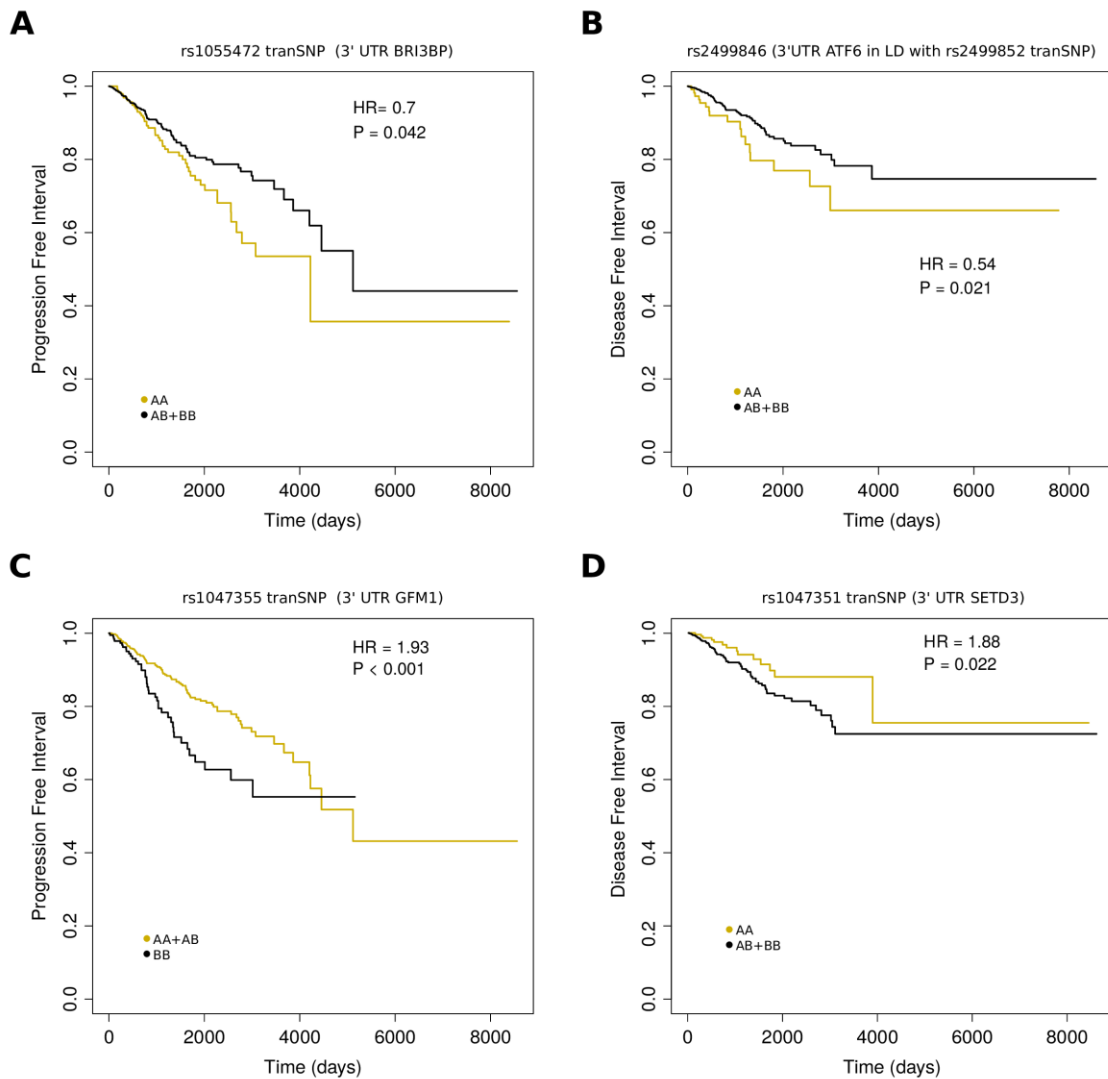
*TransSNPs can have prognostic significance in Breast Cancer TCGA data.*

Considering that rs1055472 and rs2395655 are *transSNPs* with functionally distinct alleles and the two genes harboring them are related to p53, we next investigated their potential clinical impact. We focused on breast cancer for consistency with the cell line models used and exploited the richness of data available in TCGA (82). Interestingly, exploring TCGA survival data (83) and using Kaplan-Meier curves, we observed that patients harboring the rs1055472 alternative allele show a statistically significant increase in progression-free interval time ( $p$ -value=0.042, **Figure 2.3A**). Of note, *BRI3BP* and *TP53* transcript levels were not correlated (**Figure S2.4A**) and variant rs1055472 was not associated with patients' *TP53* somatic aberration status (**Figure S2.4B**), nor with the utilization of DNA-damage agents (**Figure S2.4C**). Further, the analysis repeated on *TP53* aberrant or *TP53* wild-type patients only (**Figure S2.4D**) showed similar trends. Overall, this indicates that rs1055472 signal is not dependent on *TP53* status. In addition, the



signal also persisted when a multivariate model including breast cancer subtype as covariate was used (p-value=0.028), also highlighting an independence of the signal from ER status (**Figure S2.4E**).

Motivated by this result, we then explored more systematically whether *transSNPs* could be associated with distinct clinical variables in cancer patients. To this end, our *transSNPs* list was extended including all common SNPs in strong linkage disequilibrium (LD) ( $r^2 > 0.8$ ) with them, obtaining 3,003 SNPs distributed across 120 LD blocks with genotype imputable from TCGA data. The extended SNP list was then used to interrogate clinical data in the breast cancer cohort from TCGA (83), using Overall Survival (OS), Disease and Progression Free Intervals (DFI, PFI), and Disease Specific Survival (DSS) endpoints. Kaplan-Meier curves were built stratifying patients for the presence of the minor allele (AA vs AB+BB) or the presence of the homozygous genotype for the alternative allele (AA+AB vs BB). Cox proportional hazards regression models were built to perform the analysis. To limit false positive results, for each considered outcome, association results demonstrating a p-value<0.05 were aggregated at the level of LD blocks and >5% of SNPs reproducing the association signal in a block were required to nominate the block as associated. On average, we found 10.7% [min 8.3%, max 14.2%] associated blocks across all outcomes with an average fraction of SNPs associated in a block equal to 56.8% and an average number of SNPs per block of 23.4 (**Table S2.5**). Across the associated blocks, we identified 33 SNPs in the UTR sequences of 17 genes demonstrating significant prognostic effects (**Figure 2.3B-D**, and **Table S2.6**). Both protective and risk alternative SNP alleles were found.



**Figure 2.3: Prognostic significance of transSNPs in Breast Cancer.** *A)* Progression Free Interval analysis of *BRI3BP* related tranSNP. Kaplan-Meier curves along with summary statistics are reported. *B-D)* Examples of tranSNPs presenting prognostic significance. Kaplan-Meier curves along with summary statistics are reported.

*UTR TransSNPs could alter RNA binding protein target sites.*

To search for a mechanism that could underlie polysomal imbalance and the observed differential translation, we examined the potential impact on RNA binding proteins (RBP) binding sites for the 33 UTR *transSNPs* resulting from the survival analysis. These polymorphisms might indeed directly cause the observed allelic imbalance of SNPs located in UTRs by, for example, impacting on UTR structure or binding sites for RBPs or also microRNAs. RNA binding consensus motifs were retrieved from the RBPDB database (84). TESS software (85) was used to compute motifs scores starting from 60bp sequences with the SNP reference or alternative allele in the center. Overall, we identified 61

putative motif disruptions involving 27 UTR variants and RBPs associated with translation control and mRNA stability (**Table S2.7**), such as ELAVL1/HuR, FUS, YBX1, and PABPC1 (86–90).

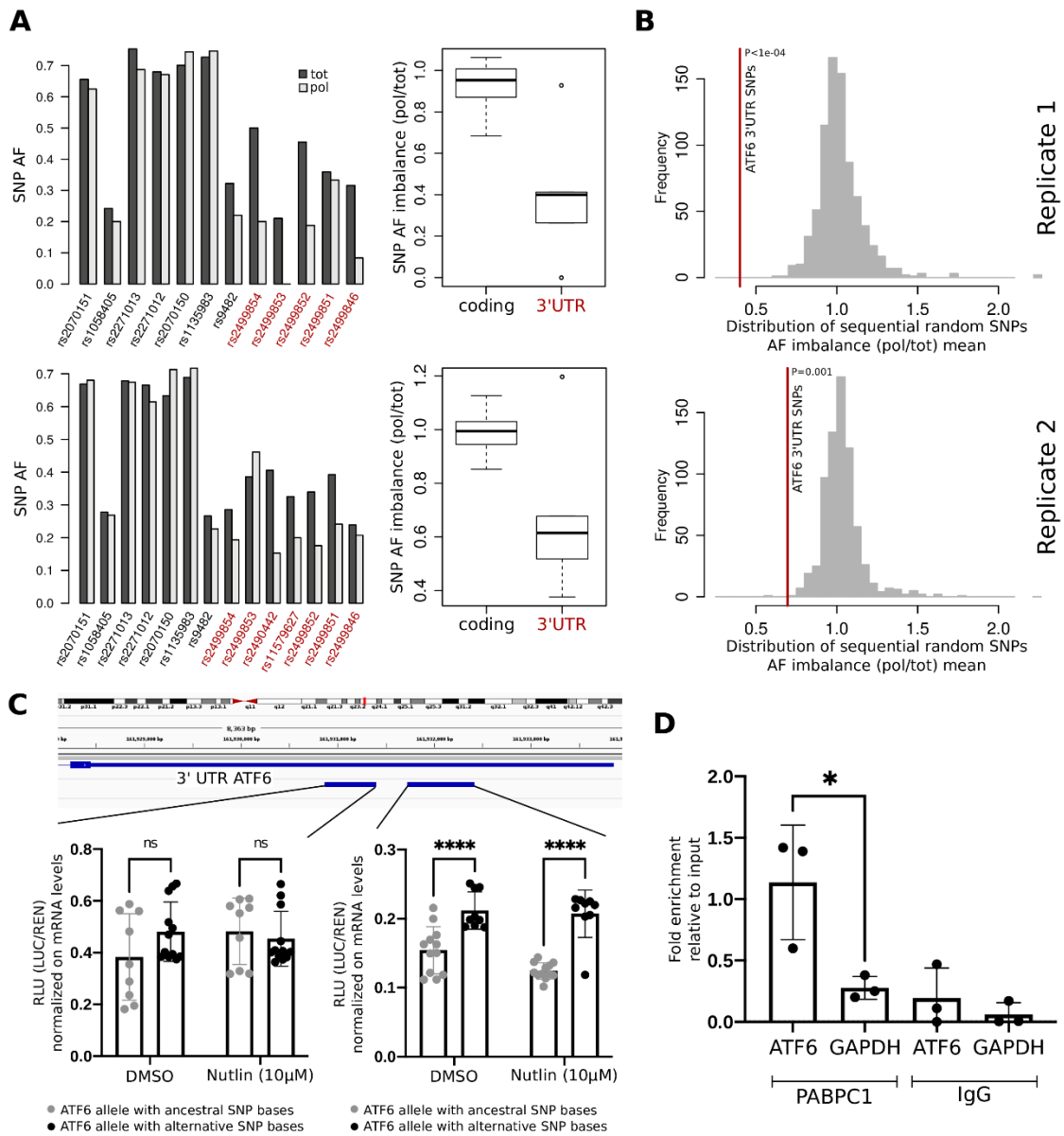
#### *UTR specific imbalance in ATF6 transcript*

Among the UTR SNPs showing potential prognostic significance, a peculiar allelic imbalance pattern was observed for polymorphisms in the *ATF6* transcript. Although only one *ATF6* SNP was present in our initial *tranSNPs* list, an overall trend of imbalance in polysomal-associated RNAs was observed and restricted to SNPs located in its relatively long 3'UTR (**Figure 2.4A**). This imbalance trend was consistent across most *ATF6* 3'UTR SNPs that are heterozygous in MCF7, and the overall imbalance distribution was strongly statistically significant ( $P < 1e-04$ ) when compared to imbalance distributions obtained from 1,000 sets of random sequential heterozygous SNPs of the same cardinality (**Figure 2.4B**). On the contrary, the *ATF6* heterozygous coding SNPs did not show any significant imbalance. Furthermore, alternative alleles were confirmed to be all in phase using 1,000 Genomes Project phased genotype data, *ATF6* transcript level was not associated with clinical outcomes in both univariate and multivariate models, and no differential *ATF6* transcript level was observed for patients carrying the clinically distinct haplotype structure.

A gene reporter assay was used to evaluate the functional impact of two different sub-regions of the long *ATF6* 3'UTR, corresponding to chr1:161930870-161931403 and chr1:161931723-161932422 (hg19) genomic coordinates (**Figure 2.4C**). For the downstream region, we observed a consistent statistically significant increase in the reporter gene activity for the allele harboring the alternative SNPs bases (**Figure 2.4C**), both in untreated and Nutlin treated cells. For the upstream region, we observed a similar trend of increased reporter gene activity (although not reaching statistical significance) for the allele harboring the alternative SNPs bases in the untreated cells but not in the Nutlin treated cells.

Computational predictions suggest that the allele harboring the alternative SNPs bases loses two PABPC1 binding sites that have strong scores when considering the allele harboring the reference SNPs bases (**Table S2.7**), suggesting an allele dependent distinct processing of the 3'UTR. Indeed, using RIP assays, we demonstrate that the *ATF6*

transcript, and specifically the 3'UTR region, is a PABPC1 target in MCF7 cells (Figure 2.4D).



**Figure 2.3: Haplotype structure and allelic imbalance along the ATF6 gene and impact of UTR TransSNPs.** **A)** RNA-seq based allelic fractions of ATF6 heterozygous SNPs are reported for both coding and 3'UTR (in red) SNPs. On the top we show the distribution observed in the first biological replicate while on the bottom we show the distribution observed in the second biological replicate. **B)** Significance of ATF6 3'UTR SNPs allelic imbalance (red line) versus distribution of sequential random SNPs imbalances. On top using heterozygous SNPs data from the first biological replicate while on the bottom using data from the second biological replicate. **C)** Dual-luciferase assays in MCF7 cells transiently transfected with reporter vectors containing ATF6 3'UTR SNP alleles. Experiments were developed as described in Figure 2. \*\*\*\*  $p < 0.0001$ , adjusted  $p$ -value based on a 2-way ANOVA with Sidak's multiple comparison test. Data are represented as mean  $\pm$  SD. **D)** RIP experiment probing the interaction of PABPC1 with the ATF6

*transcript. Bars plot the average fold enrichment relative to the input sample. Individual average values from three biological replicates are also shown. Results obtained with an IgG control antibody are included. \*  $p < 0.05$ , two-tailed, unpaired t-test. Data are represented as mean  $\pm$  SD.*

## Discussion

The relative abundance of an mRNA in polysomal versus subpolysomal fraction is frequently used as a proxy for protein synthesis and can capture changes both in global and specific translation efficiency, driven by structural or sequence elements in the context of cell stress responses. It is common knowledge that cancer cells experience chronic stress as well as acute stress responses that converge on translation controls, unfolded protein responses, oxidative stress, and proteasome functions. Recent studies, including some from our group, are placing wild-type p53 in an important position to respond and influence these pathways, and also are indicating that the gain of function properties of several mutant p53 proteins may converge on the dysregulation of controls on translation quality, protein folding, and proteasome functionality. Hence, in this study, we chose a p53 wild-type cancer cell line well characterized in terms of p53-induced responses (91), including post-transcriptional and translational changes (74). We used both a commonly used chemotherapeutic drug that activates the DNA damage response and a selective MDM2 inhibitor that results in acute p53 activation without apparent genotoxic response (75,92). Doses and treatment time points were selected based on previous characterization of both p53 activation pathway and cell outcome (93,94). Furthermore, for both treatments, we have previously shown a global as well as a specific impact on translation (74).

Since MCF7 cells have been genotyped both by SNP arrays and by exome-sequencing as part of the cell-line encyclopedia and NCI-60 studies, we could leverage RNA-seq data of total and polysome-bound mRNAs to focus on allele counts for all transcribed SNPs in the heterozygous state to reveal an imbalance in one fraction over the other. We hypothesized that this approach, although restrictive, could overcome limitations deriving from the quality and quantity of sequence reads that can be attributed to each allele of a SNP pair. Computational approaches that use allele counts of heterozygous SNPs to identify imbalance events have been implemented and widely used to investigate

allele-specific expression patterns both in healthy and disease tissues (65–68,70,71). However, only few approaches (69,72) have been developed to compare allelic imbalances across matched samples, and all of the available methods have been specifically designed in the context of cancer (e.g. to compare cancer samples with their matched normal counterparts), limiting hence their broad applicability.

Using polysomal profiling and RNA relative quantification from either pooled or individual fractions typically show relatively small differences, that can, however, underlie an important effect at the level of protein amounts. The approach we developed to identify *tranSNPs* is focused on computing and comparing the relative change of allele counts across mRNA fractions. To account for technical noise, the approach embeds in the calculation the amount of allele counts variability observed in a given condition exploiting data from biological replicates. Of note, such variability was not affected by the cellular treatment employed and was consistent across different RNA fractions. We identified 147 *tranSNPs*, representing nearly 4% of the overall analyzed SNPs, a fraction that although conservative, can be considered comparable with the recent fraction of about 2% of allele-specific expressed genes computed across tumor versus normal matched samples (72). Indeed, in (72), differential allele-specific expression of genes was computed by aggregating at gene level heterozygous SNPs data and no statistically significant difference between genes with differential ASE and those without with respect to the number of heterozygous SNPs across the length of the genes was found. We checked the intersection of our *tranSNP* list with the current list of GWAS variants from the GWAS catalog (95) that are related to cancer (N=2,657), although it is possible that *tranSNPs* might not be directly related to the risk of developing a tumor but be related to a distinct clinical outcome. Only an intersection between the LD block defined by our *tranSNP* rs760077 and association signal with gastric cancer (variant rs760077 and LD variants rs140081212, rs4072037) and breast cancer (LD variants rs2075570, rs2974935) was identified. We further explored the intersection between the *tranSNPs* and a set of GWAS-implicated cancer risk SNPs identified in 41 genes of the p53 response pathway, of which those mapping at *KITLG*, *CDKN2A*, and *TEX9* genes were reported to potentially impact drug sensitivity (96,97). Given that MCF7 cells express wild type p53 and that the doxorubicin or Nutlin treatments elicit a p53-dependent response, we were motivated to examine the potential overall with our *tranSNP* list. However, only 13 of those 41 GWAS-implicated

cancer risk SNPs could be evaluated in our MCF7 RNA-seq data. They map at six genes (*CERSS*, *ISYNA1*, *CFLAR*, *SESN1*, *AKAP9*, *CYP51A1*) that were not reported to impact drug sensitivity (96,97). Finally, we considered the results of a very recent study where over 12,000 3'UTR variants, including more than 2,000 disease-associated SNPs from the GWAS catalog, were cloned within 100nt-long fragments and tested in a massively parallel reporter assay measuring relative RNA expression (98). Surprisingly, about 25% of the tested GWAS SNPs resulted in steady-state changes in reporter transcripts' abundance. Among the GWAS SNPs included in that screen, 131 are heterozygous in MCF7 and could be analyzed by our method. We found no intersection with our *tranSNP* list, consistent with our approach being focused on allelic differences that are detected among polysomal-associated mRNAs but not total mRNAs.

A subset of the identified UTR *tranSNPs* showed apparent prognostic value. Some of these are located in genes that have been already established to play a role in cancer in general and in breast cancer or in relation to p53 functions, in particular. These include *BRI3BP*, also known as human cervical cancer oncogene 1 (*HCCR-1*) that might act as a negative modulator of p53 (77,78), the *ATF6* gene, an important modulator of endoplasmic reticulum stress that can promote cell survival (99,100) and the protein lysine methyltransferase *SETD3* gene, that was recently reported as a prognostic marker in breast cancer patients (101), and also shown to promote apoptosis in response to DNA damage, in part through the modulation of p53 (102). Other genes implicated by our results are *GFM1*, a mitochondrial translation elongation factor that has been linked to oxidative stress and cancer cell survival (103), and *PLEC*, an important member of a large family of scaffolding proteins that can modulate various aspects of cell function. The plectin gene is frequently found co-amplified with *MYC* in BRCA1-deficient breast cancer (104). Sequence variants predicted to be deleterious have also been identified in primary breast cancers (105). Physical interaction of plectin with BRCA2 has also been reported (106). In all these cases, the *tranSNPs* we uncovered may be modifiers of protein expression, influencing mRNA translation efficiency, with consequences on cancer phenotypes.

To begin exploring a more precise mechanism of action, motifs analysis based on eCLIP or RIP-seq data was exploited to identify UTR *tranSNPs* that may alter or lead to gain or loss of RNA binding sites. Based on these predictions and focusing on a peculiar imbalance in *ATF6* transcript that we observed only across 3' UTR SNPs, we performed RIP assays and confirmed that *ATF6* is a target of PABPC1, a well-known RNA binding protein. This, combined with predicted putative loss of two PABPC1 binding sites across one of the two *ATF6* alleles, suggests an allele-dependent processing of the 3'UTR. Overall, our approach led us to identify a class of SNPs that are associated with changes in mRNA translation efficiency, potentially acting at post-transcriptional level, and that show prognostic value for cancer, hence implicating the potential identification or stratification of cancer patients based on genetic markers that are helpful in the prediction of prognosis in regard to death, progression or recurrence.

#### Limitation of Study

This study considers a single cell-line and two biological replicates for each considered treatment/condition. Survival analysis is focused on breast cancer data. Our catalog of *tranSNPs* is hence limited to the data we analyzed.

## Methods

#### Data and code availability

RNA-seq BAM files have been deposited at BioProject under the accession number PRJNA693005.

#### Experimental Model and Subject Details

##### Cell lines and culture conditions

MCF7 cells were cultured in standard RPMI (Lonza) supplemented with 10% FBS (brand), 100 units/ml penicillin, 100 mg/ml streptomycin antibiotic mix, and 2mM glutamine. Cells were periodically tested to ascertain the absence of mycoplasma contamination. The authenticity of the parental cells was confirmed by a commercial genotypic service (BMR



genomics, Padova, Italy). Doxorubicin was purchased from Sigma. Nutlin was purchased from Cayman chemicals.

## Method Detail

### SNP status in MCF7 cells

We retrieved from the literature genotype assays for MCF7. Two different public datasets (GEO and NCI-60) were used containing information about homozygosity or heterozygosity conditions for SNP alleles in MCF7, the unique identifier, the relative and alternative base in specific single positions. One SNP-array based file (GSM888366) contains 756,260 SNPs with the information about the reference and alternative bases, while the second exome-based file (107) contains 85,593 SNPs. The SNPs in common ( $N=6,454$ ) were compared to verify the consistency among the two datasets: more than 91% of the SNP calls were consistent. This concordance suggested the possibility to merge the two files removing the data with opposite information. A first screening was performed through the elimination of SNPs called as homozygous in the two datasets, but with a different base call: *i.e.* reference “0/0” vs alternative “1/1”. Together, the final file presents 11,544 heterozygous SNPs of which 1,802 in 3’UTR and 729 in 5’UTR. Variants were annotated using VariantAnnotation R package (108).

### Differential allele-specific expression analysis

To identify allelic imbalances between polysomal and total RNA fractions a computational approach that exploits SNP allelic fractions variability across biological replicates was implemented. Given an experimental condition  $C$ , SNP allelic fractions variability for condition  $C$  is calculated as:

$$V_c = \text{mean} \left( \left\{ \left| \text{snp}_i^{R_j} - \text{snp}_i^{R_k} \right| \mid s. t. 1 \leq i \leq N \text{ and } 1 \leq j, k \leq M \text{ and } j < k \right\} \right)$$

where  $N$  is the number of considered SNPs,  $M$  is the number of biological replicates, and  $\text{snp}_i^{R_j}$  is the allelic fraction of SNP  $i$  computed from the RNA-seq BAM file representing the biological replicate  $R_j$  using ASEQ pileup module (70) and defined as:

$$snp_i^{R_j} = \frac{\#ALT}{\#ALT + \#REF}$$

where  $\#ALT$  and  $\#REF$  correspond to, respectively, the number of reads supporting the SNP  $i$  alternative base and the number of reads supporting the SNP  $i$  reference base. Of note, in the pileup computation only reads and bases with quality above 20 and a minimum of local depth of coverage of 10X were considered.

Now, given variabilities values  $V_{poly}$  and  $V_{tot}$  computed, respectively, from RNA-seq replicate samples of polysomal and total RNA fractions, we define as *tranSNPs* all SNPs satisfying the following condition:

$$\forall_{j \in \{1, \dots, M\}} (snp_i^{poly_j} + V_{poly} < snp_i^{tot_j} - V_{tot}) \vee \forall_{j \in \{1, \dots, M\}} (snp_i^{poly_j} - V_{poly} > snp_i^{tot_j} + V_{tot})$$

#### *Survival analysis using TCGA data*

From our list of 147 *tranSNPs* an extended list of variants was built retrieving variants in strong linkage disequilibrium. We queried the Ensembl rest API version GRCh37 (109) and retrieved all the variants with  $r^2 > 0.8$  in a genomic window of 500kbp using general population data. For a total of 3,003 variants, distributed among 120 LD blocks, we were able to retrieve genotype calls from TCGA data. Specifically, raw TCGA genotype calls were downloaded from TCGA legacy data portal ([portal.gdc.cancer.gov/legacy-archive](http://portal.gdc.cancer.gov/legacy-archive)) and only genotypes with confidence score larger than 0.1 were retained. Genotypes were analyzed with SHAPEIT v2 (110) to infer haplotype structure and optimize genotype content information for the imputation process. Variants were imputed using IMPUTE v2.3.2 (34) against a reference panel built from 1,000 Genomes Project data. The final extended list of variants was used to perform variant-specific survival analysis using TCGA breast cancer survival data (83). Overall Survival (OS), Disease-Specific Survival (DSS), Disease-Free Interval (DFI) and Progression-Free Interval (PFI) were considered in our analysis. Dominant and recessive models were built, respectively, grouping together patients with homozygous reference and heterozygous genotypes and homozygous alternative and heterozygous genotypes. Kaplan-Meier survival curves were built for each variant and Cox proportional hazards regression models were used and inspected. To limit false positive results, for each considered outcome, association results demonstrating a

p-value <0.05 (minimum between Log rank test p-value and Wald test p-value was considered) were aggregated at the level of LD blocks and >5% of variants reproducing the association signal in a block were required to nominate the block as associated. Survival analyses were performed using *survival* R package (111).

#### *Identification of SNPs disrupting putative RBP consensus motifs*

We retrieved 552 human RNA binding consensus motifs from RBPDB database (84). Motif analysis was performed on the list of 33 UTR *trans*SNPs resulting from the survival analysis. For each variant, two sequences of length 60bp with the SNP reference or alternative allele in the middle were built using GRCh37 reference genome. All sequences were then collected in a FASTA file and TESS software (85) was ran on this file to compute motifs scores. TESS tool provides a set of log-likelihood-ratio-based scores, among which we used the score *La*, which represent the log-odds ratio of the match, and the score *Lm*, which represents the maximum possible log-odds ratio for a match from a given consensus motif. For each motif overlapping a variant, results were collected only if the score *La* for the reference or the alternative allele was at least 3 and it was computed on the forward (transcribed) strand. Confident results were filtered by considering only entries having a *La/Lm* value greater than 0.5 in at least one of the two conditions. In addition, only variants presenting an absolute ratio between reference and alternative *La* scores greater than 10% or presence of a motif with the reference but not with the alternative (or vice-versa) were retained.

#### *Polysomal profiling and RNA-sequencing.*

MCF7 parental cells were seeded in P100 dishes and treated at about 70% confluence by 1 $\mu$ M doxorubicin or 10 $\mu$ M Nutlin, for 16 hours. DMSO was used for treatment control. Cytoplasmic lysates were obtained and polysome fractionations performed as described in (74–76,112,113). Briefly, cytoplasmic lysates were loaded on a 15ml, 10-50% sucrose gradient, ultra-centrifuged (40k rpm for 100') and fractionated with an automated fraction collector (1ml per fraction Teledyne ISCO), tracing the 254<sub>nm</sub> absorbance. All the lighter fractions containing subpolysomal fractions (from the top to the gradient up to the fractions corresponding to the 80S monosome) were pooled in a tube. Heavier fractions corresponding to polysomal RNAs (two or more ribosomes) were also pooled in

a separate tube. RNA was purified by extraction with 1 volume of phenol-chloroform. The aqueous fraction was recovered after centrifugation, transferred to a new tube where RNA precipitation was induced by the addition of 1 volume of isopropanol and a subsequent centrifugation step. Pellets were washed once using 70% V/V ethanol to remove phenol contaminations and resuspended in sterile, RNase free water. Total RNA samples were instead obtained by TRIzol (ThermoFisher) extraction of a separate cell population prepared in parallel. RNA preparations were quantified by NanoDrop microvolume spectrophotometer. Purity and integrity were checked by Agilent 2100 Bionalyzer electrophoretic runs exploiting the Agilent RNA 6000 Nano kit. Sampling with RIN over 6 were used to prepare libraries for RNA sequencing. Two biological replicates for each RNA type (total and polysomal) and condition (mock, doxorubicin, Nutlin) were sequenced. Libraries were prepared following the TruSeq RNA Library preparation kit v2 protocol (Illumina), starting from 1 µg of input RNA. Paired end (100bp) sequencing was performed on a HiSeq 2500 (Illumina). FASTA files were aligned to the reference genome GRCh37 using TopHat v2.0.10 (114), resulting in BAM files with an average of ~32 million unique mapping (mapping quality > 5) reads. Differential expression analysis was performed with Cuffdiff (115) comparing Nutlin and doxorubicin samples against mock samples, considering both total and polysomal RNA fractions. Pathway enrichment analysis of deregulated genes in all tested comparisons was performed using clusterProfiler R package (116) exploiting KEGG database (117). Comparison of multiple gene lists obtained from condition-associated *transSNPs* was performed with Metascape (118).

#### *Cloning 5'UTR or 3'UTR allelic variants for gene reporter assays.*

The commercial Firefly luciferase plasmid pGL4.13 was used to clone the two rs1055472 alleles *BR13BP* 3'UTR as well as two regions of the *ATF6* 3'UTR containing different set of SNPs, using a PCR based approach that introduced the XbaI restriction enzyme site. MCF7 cDNA was used as template, given the heterozygous state for the SNPs of interest. The cloned regions correspond to genomic coordinates chr12:125509977-125510349 (hg19) for *BR13BP* and chr1:161930870-161931403 or chr1:161931723-161932422 for *ATF6*. Primers used for the cloning are listed in the Key Resource Table, where the lowercase bases are the restriction site sequence included in the primers after two initial adenines.

The pGL3 basic vector was instead used to clone the two rs2395655 alleles of the *CDKN1A* 5'UTR exploiting the unique NcoI site that overlaps with the firefly start codon and using complementary primers containing the 5'UTR sequence of interest for each SNP allele that were annealed prior to ligation to generate NcoI sites at both ends. After in vitro ligation (T4 DNA ligase, NEB), *E. coli* competent cells were transformed and colonies picked to recover plasmids based first on a colony PCR approach, followed by plasmid recovery from positive colonies followed by electrophoresis migration and digestion pattern analysis. The successful cloning of both SNP alleles for the two targets was then verified by Sanger sequencing (Eurofins Genomics) of independent candidate colonies.

#### *Dual-Luciferase assays*

pGL4.13- or pGL3-derivative plasmids were transiently transfected along with the pRL-SV40 Renilla control luciferase in a 3:1 ratio (250ng plus 50ng for 24 well format) using Mirus-LTI transfection reagents following the manufacturer's protocol. When needed, cells were treated with doxorubicin or Nutlin for 16 hours, starting 24 hours after transfection. Luciferase assays were performed using Dual-Luciferase™ Reporter (DLR™) Assay System (Promega), following the provided protocol. Light units were measured using a Tecan M200pro multilable plate reader through three technical replicates of well. Presented in the graphs are the average Relative Light Units (Firefly/Renilla) and standard deviations of three biological experiments, compared using a two-tailed, equal variance Student's t-test.

#### *RIP assays*

Ribonucleoprotein particles immunoprecipitation (RIP) experiments were performed following a published protocol with some modifications described recently (119,120), starting from  $4 \times 10^7$  MCF7 cells. After lysis, the supernatants were collected and 1% of each sample was set aside as input while the remaining was incubated for 4 hours at 4°C with protein A magnetic beads (Thermo Fisher Scientific) coated either with 3 µg of an anti-PABPC1 antibody (Abcam) or with 3 µg of normal Rabbit IgG (Santa Cruz). After the washing steps, RNA was isolated from Input and IP samples using TRIzol (Thermo Fisher Scientific) and converted to cDNA using the RevertAid First Strand cDNA Synthesis Kit and standard protocol (Thermo Fisher Scientific). qPCR reactions were performed using the

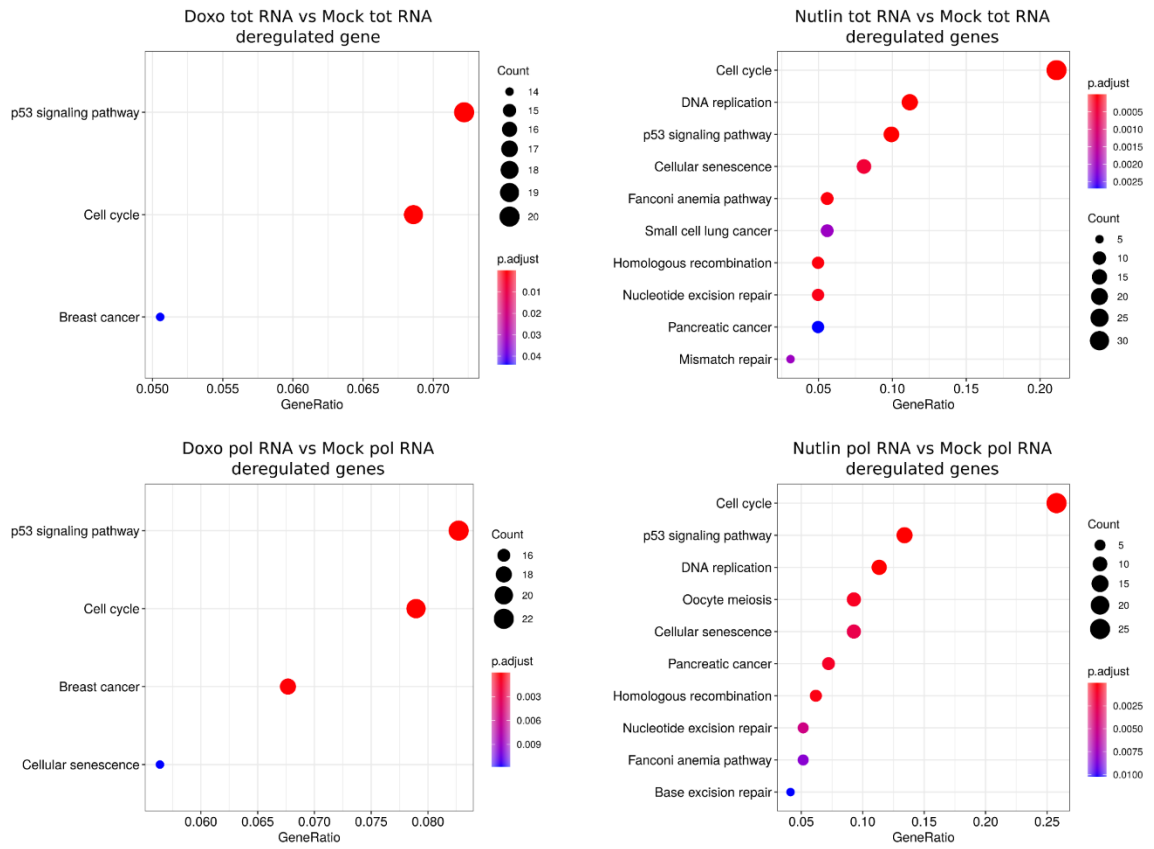
qPCRBIO SyGreen Mix (PCR Biosystems) master mix on a CFX384 thermal cycler (Biorad). GAPDH was tested as a comparison. Two pairs of primers were used for ATF6, of which one targeting exon 9, and the second targeting the 3'UTR sequence comprising the rs2499847 SNP whose alleles are predicted to impact a PABPC1 binding site (**Table S2.7**). The primers sequences can be found in the Key Resource Table.

#### *Quantification and Statistical Analysis*

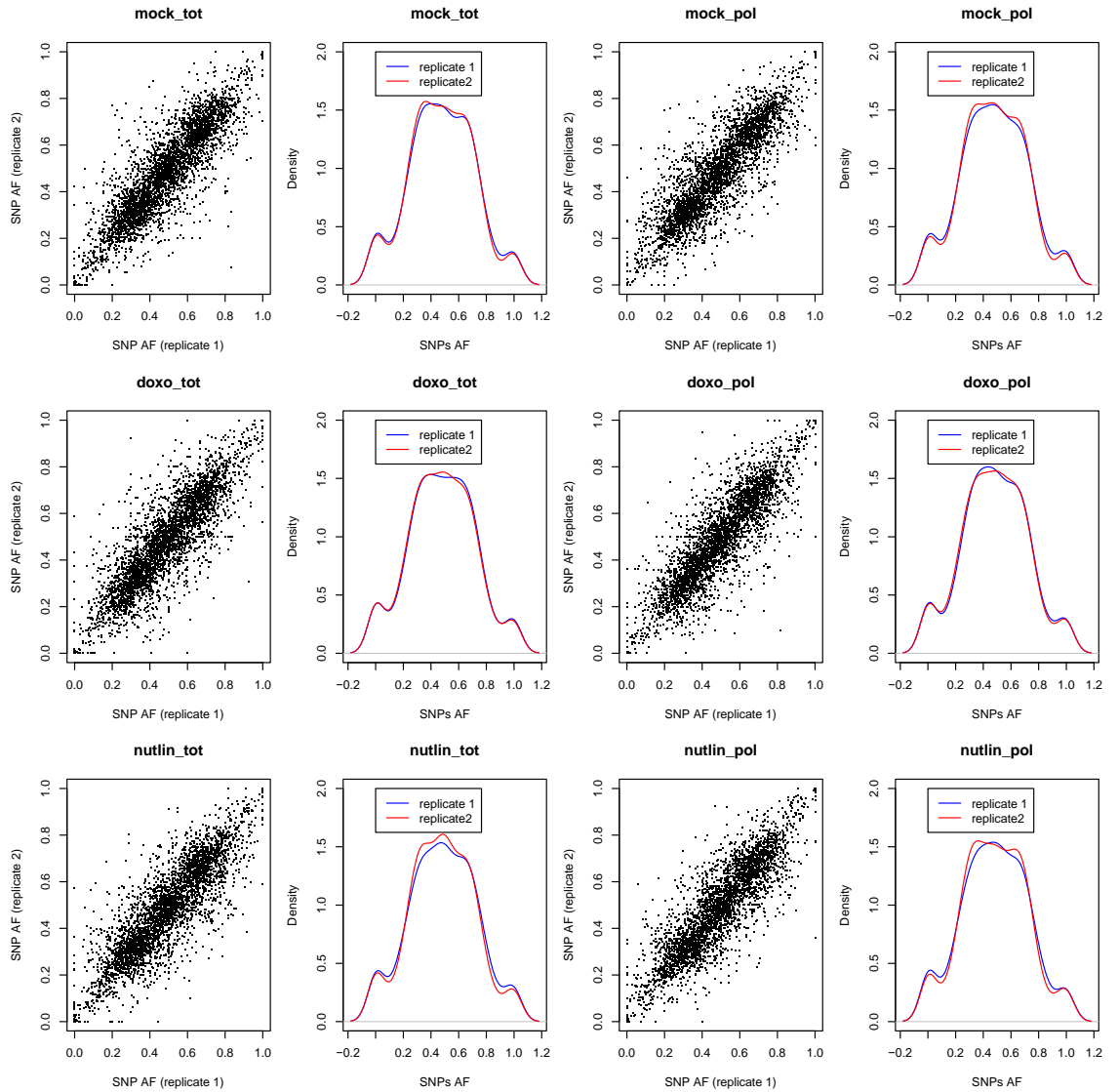
Statistical tests applied throughout the study are specified in results, figure legends and in the methods accordingly. Given a list of variants, LD blocks were determined identifying the number of cliques in a graph where nodes are variants and edges are present between two nodes when the corresponding variants are in LD with an  $r^2 > 0.8$ . In all considered cases the number of cliques corresponded to the number of connected components. Luciferase assay data were analyzed using a 2-way ANOVA with Sidak's multiple comparison test. RIP assays were analyzed using a two-tailed unpaired t-test.

# Supplementary Material

## Supplementary Figures

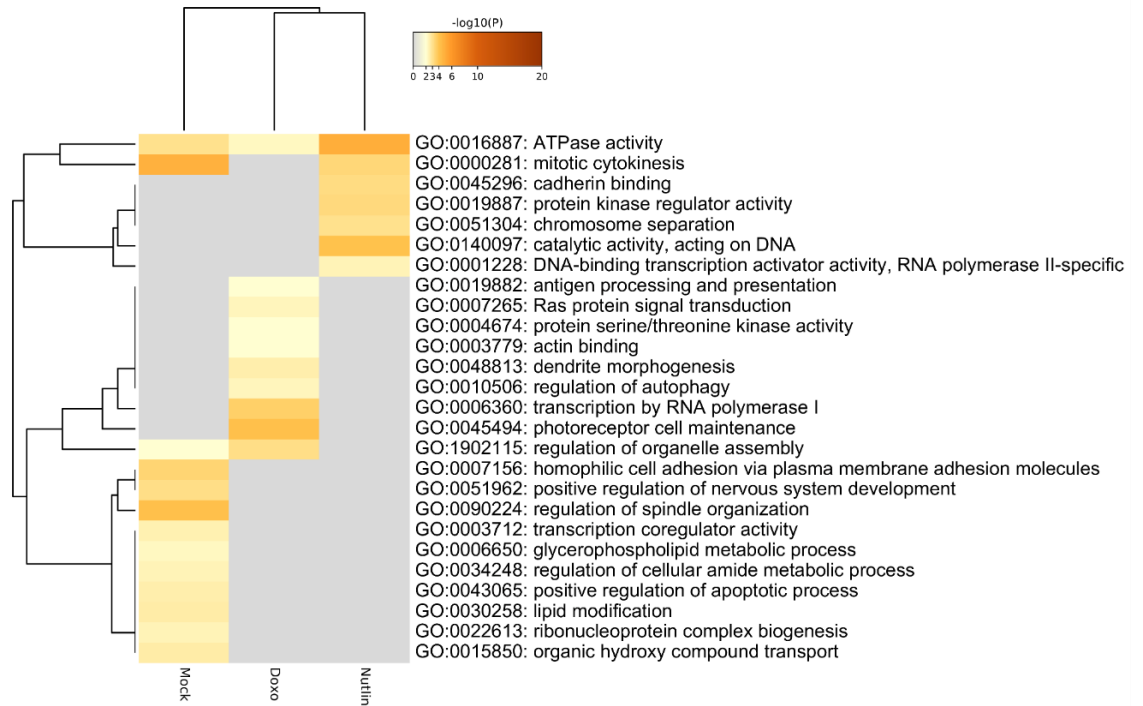


*Figure S2.1: Dotplots of gene set enrichment analysis results. Significant terms are shown together with adjusted p-values, gene ratios and gene counts.*

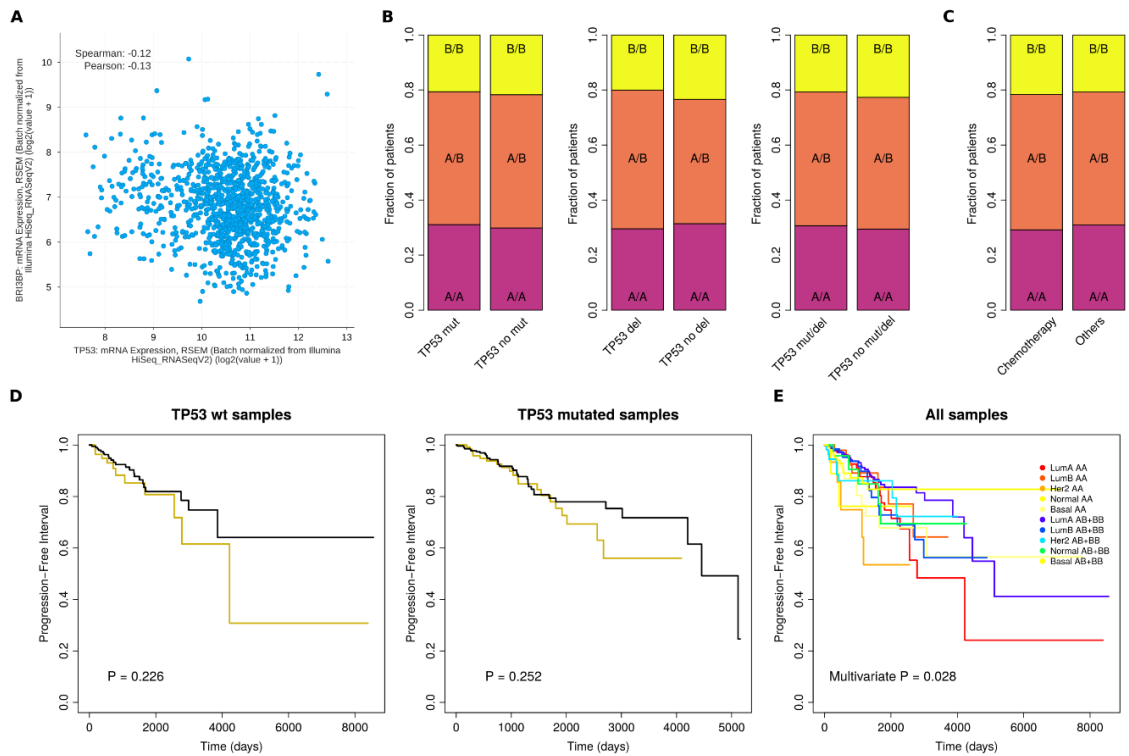


**Figure S2.2. Distribution of SNPs allelic fractions in RNA-seq data across mRNA fractions and treatments.** For each RNA fraction and treatment we show the comparison of RNA-seq based allelic fraction of putative MCF7 heterozygous SNPs across different biological replicates and the distribution of these RNA-seq based allelic fractions.





*Figure S2.3 Metascape analysis. Comparison of the top GO terms enriched by the multiple gene lists obtained from the condition-associated tranSNPs.*



**Figure S2.4 Characterization of rs1055472 association.** A) Association between BRI3BP and TP53 transcript levels (cbioportal data); B) Distribution of rs1055472 genotype across TCGA patients stratified by TP53 mutational status; from the left stratification by point mutations, somatic copy number aberrations (deletions) and integration of point mutations and deletions; C) Distribution of rs1055472 genotype across TCGA patients stratified by DNA-damage treatment (chemotherapy); D) PFI analysis considering TP53 mutated (left) and TP53 wild type (right) patients' only; E) Multivariate analysis considering breast cancer subtype

*Supplementary Tables*

Supplementary Tables are available at

Supplementary Table S2.1:

<https://www.cell.com/cms/10.1016/j.isci.2021.103531/attachment/e59ceca7-0a18-458b-827b-d410c369e584/mmc2.xlsx>

Supplementary Table S2.2:

<https://www.cell.com/cms/10.1016/j.isci.2021.103531/attachment/d1140315-9167-4f24-9773-f902e4dd8524/mmc3.xlsx>

Supplementary Table S2.3:

<https://www.cell.com/cms/10.1016/j.isci.2021.103531/attachment/7fbfab2f-642b-428d-9c8f-6c1fbadc56f9/mmc4.xlsx>

Supplementary Table S2.4:

<https://www.cell.com/cms/10.1016/j.isci.2021.103531/attachment/92ee460d-9535-42e7-9017-c85f8905f4c4/mmc5.xlsx>

Supplementary Table S2.5:

<https://www.cell.com/cms/10.1016/j.isci.2021.103531/attachment/6dae1eec-aeec-4d10-ae3e-b49fe7fe37f6/mmc6.xlsx>

Supplementary Table S2.6:

<https://www.cell.com/cms/10.1016/j.isci.2021.103531/attachment/927facde-b99e-4949-89f9-a8121b7444cc/mmc7.xlsx>

Supplementary Table S2.7:

<https://www.cell.com/cms/10.1016/j.isci.2021.103531/attachment/5e7a9d03-a881-4f83-a893-aadb1843167b/mmc8.xlsx>

# Chapter 3: Finding functional relations among common human genetic variants

## Introduction

In recent years, many resources have been developed to explore the effects of germline variants. However, these tools don't allow to explore relations among multiple variants. To tackle these limitations, we developed Polypact, a novel web resource that allows the analysis of multiple variants in terms of their effects on transcript levels and changes in motifs binding scores. Polypact provides two new network frameworks that allow to detect putative variant-variant and variant-gene interactions.

Variant-variant interactions on gene expression, epistatic or additive, have not been extensively studied since most eQTL dataset relies on *cis*-eQTL. Resources like GTEx (44) don't perform *trans*-eQTLs associations since their calling is burden by the multiple test correction required. Also, *cis*-eQTLs effects are usually local and linked to the modification of transcription factor binding site, chromatin interactions or some other regulatory element making difficult to find other variants associated that are not in linkage disequilibrium making the identification of interactions extremely difficult given the huge amount of correlation between variants.

## Results

Polypact is a new web resource that characterizes over 18 million common human germline variants and that aims to identify putative interactions between common human germline variants. Polypact characterizes variants by integrating ChIP-seq data from ENCODE and Roadmap projects, transcription factor binding motifs from Homer, Hocomoco, Jaspas and Transfac, and genotype and gene expression data from GTEx and TCGA databases.

Polym pact builds on the data by defining and introducing two novel network models: a similarity network model used to identify variants that can have a common effect on transcript level of genes or commonly alter a transcription factor binding motif and a variant-gene network model that aims to identify interactions between variants and genes.

Using Polym pact, we used the similarity network framework to identify putative interactions between variants in the Alzheimer's diseases and breast cancer GWAS variants dataset. Finally, we analyzed those interactions in detail using the variant gene network model.

Polym pact is freely available and can be queried using a web interface and a REST API for programmatic access.

## Materials and methods

Variants SNPs and INDELS have been retrieved from dbSNP version 151 and they have been filtered using a custom script keeping only the ones that have a minor allelic frequency >1% in the general population. We integrated CHIP-seq data from ENCODE and Roadmap projects to identify variants that fall in a functional peak of a transcription factor or a histone mark. We retrieved regulatory elements from CONREL to find variants falling in a regulatory region. Then, we performed a motifs scan over to identify the changes introduced in transcription factor binding scores. We defined four cases: match when the variant does not change significantly the binding score of a motif, change when it does, addition when a binding motif is created and deletion when a binding motif is deleted. Finally, we integrated the GTEX and TCGA germline and computed linear associations between genotypes and gene expression levels under additive, dominant and recessive models in fifteen different tissues.

Using the data in Polym pact we formalized two novel network models to identify variants that have potential cooperative effects on gene regulation and to model the landscape of interactions between variants and genes.

The first model, called similarity network, defines a network where variants with a similar putative effect on gene expression or on motif binding scores are linked together and, using community detections algorithms, we are able to identify groups of variants that

may exhibit complex patterns on genes regulation when considered together. The second network model links the effects of variants to genes and allows to explore in detail multiple relations.

Finally, we retrieved GWAS variants associated to cancer and Alzheimer's Disease in the GWAS catalog, and we analyzed them using Polym pact.

I contributed to the project by creating the Polym pact database and by performing all the data integrations analyses apart from the merging of the TCGA and GTex datasets. I have formalized, implemented and analyzed the networks models introduced in the resource. I have supervised and contribute to the development of the web interface. I have developed the case studies on variants interactions and supervised and implemented parts of the case study on the functional element enrichment in cancer variants.

## Discussion

Polym pact is a new web resource that allows the exploration of putative variant-variant and variant-gene interactions. We integrated data from many public databases and we developed new models to characterize variants. In particular, we introduced and performed a motif scan on over 18 million variants using more than 5,000 transcription factors and RNA binding protein motifs. Secondly, we introduced a new genotype transcript level analysis that can help in identifying medium/strong *cis* effects and new *trans* effects.

Using the Polym pact data we implemented a new similarity network model between variants. This model allows to discover groups of variants that show a common behavior in modulating gene expression or in altering transcription factor binding motifs.

Then, we developed another network model to analyze putative interactions between variants and genes. This model, called variant-gene network, showed a complex topology when built using all the variants modulating at least one gene in a tissue. In particular, it showed putative regulatory hubs and putative feedback loops that can model dynamic instability.

Analyzing the Polym pact data, we showed that GWAS variants are enriched for regulatory elements with respect to random variants.

Then, we analyzed the breast cancer and Alzheimer's Disease GWAS variants using our framework finding many examples of putative variant-variant interactions in the modulation of transcript level of genes where variants are modulating transcripts in a cooperative or additive way.

We believe that Polypact can be used to analyze the effects of multiple variants and to identify putative interactions.

## Article

Polypact: Exploring functional relations among common human genetic variants

Samuel Valentini<sup>1</sup>, Francesco Gandolfi<sup>1</sup>, Mattia Carolo<sup>1</sup>, Davide Dalfovo<sup>1</sup>, Lara Pozza<sup>1</sup>, Alessandro Romanel<sup>1#</sup>

<sup>1</sup> Department of Cellular, Computational and Integrative Biology (CIBIO), University of Trento, Trento, Italy

#Corresponding author

**Journal:** Nucleic Acids Research, Volume 50, Issue 3, 22 February 2022, Pages 1335-1350

**Publisher:** Oxford University Press

**Doi:** <https://doi.org/10.1093/nar/gkac024>

## Abstract

In the last years, many studies were able to identify associations between common genetic variants and complex diseases. However, the mechanistic biological links explaining these associations are still mostly unknown. Common variants are usually associated with relatively small effect size, suggesting that interactions among multiple variants might be a major genetic component of complex diseases. Hence, elucidating the presence of functional relations among variants may be fundamental to identify putative variants' interactions. To this aim, we developed Polypact, a web-based resource that allows to explore functional relations among human common variants by exploiting variants' functional element landscape, impact on transcription factor binding motifs, and effect on transcript levels of protein-coding genes. Polypact characterizes over 18 million common variants and allows to explore putative relations by combining clustering analysis and innovative similarity and interaction network models. The



properties of the network models were studied and the utility of Polym pact was demonstrated by analysing the rich sets of Breast Cancer and Alzheimer’s GWAS variants. We identified relations among multiple variants, suggesting putative interactions. Polym pact is freely available at [bcglab.cibio.unitn.it/polym pact](http://bcglab.cibio.unitn.it/polym pact).

## Introduction

Common genetic variants in the form of Single Nucleotide Polymorphisms (SNPs) and Small Insertions and Deletions (INDELs) are the most frequent forms of DNA polymorphisms. SNPs and INDELs are supposed to be the largest source of phenotypic variation across individuals (121). Although common variants are mostly located outside of gene coding regions and seem to have no direct consequences on protein sequences and phenotypes, genome-wide association studies (GWAS) identified thousands of them associated with complex traits and diseases (95). Despite expression quantitative trait loci (eQTL) studies have broadly shown that non-coding variants modulate gene expression (122), there are still limited examples of clear mechanistic models linking common variants and biological functions (123,124) and the functional role of most of them remains largely unknown. Indeed, most variants identified in GWAS studies have low effect size, suggesting that individual variants have a small impact on the heritability of complex traits and diseases (125). In addition, complex traits and diseases are often affected by many genes. Overall, this suggest that the interaction among common variants may play an important role and could represents a major genetic component of complex diseases (126).

Advances in high-throughput technologies, especially those based on next-generation sequencing (NGS), have generated a huge amount of genomic datasets of different types. Several databases and web applications have been developed upon these datasets to annotate genetic variants, providing effective platforms for the exploration of their functional properties. Some of these resources are focused on specific aspects of SNPs and INDELs like SNP2TFBS (127), which annotates how variants’ alleles may affect transcription factors (TFs) motifs, or HACER (128), which allows to explore how non-coding variants in active enhancers may modulate gene expression. Other resources instead, like RegulomeDB (129), Haploreg (130) and the recent VARAdb (131), provide extensive annotations of common variants by integrating ChIP-seq data, chromatin

accessibility and interaction data, TFs motif changes, eQTLs and GWAS data. Although these resources provide important information to investigate the functional role of single variants, none of them allows to aggregate information of multiple variants, limiting hence their applicability to investigate to what extent different variants may be involved in the modulation of same genes or genes involved in same biological processes. Tools and frameworks to explore functional relations and links among multiple variants are indeed needed and fundamental to help identifying putative variants' interactions.

To overcome these limitations, we developed Polymact, a computational resource and framework which allows to investigate the presence of functional relations among multiple variants. On the one side, Polymact characterizes over 18 million common, mainly non-coding, variants by combining: (i) cell line and tissues regulatory elements data; (ii) the landscape of changes observed in transcription factors binding sites (TFBS) scores; (iii) the association of genetic variants genotype with the expression of protein coding genes in various healthy human tissues. On the other side, Polymact provides a novel framework to explore functional relations among a set of queried variants, combining clustering analysis, a network model describing similarities which also includes community detection features, and an additional network model which integrates all functional annotations computed and collected in Polymact to explore in detail interactions among variants and genes.

We believe that Polymact could become a useful and effective computational platform to investigate the potential impact of multiple common genetic variants in human diseases and biological processes.

## Materials and Methods

### *Collection of genetic variants*

We collected genetic variants information from dbSNP version 151 (132) using version hg19 as human reference genome. We kept all common variants with Minor Allele Frequency greater or equal than one percent, considering the general population frequencies available from 1,000 Genomes Project (133) or the TOPMed (134) data. Overall, we collected 18,683,752 genetic variants composed by 14,810,175 SNPs and 3,873,577 INDELS.

### *Functional annotation of genetic variants*

ChIP-seq data from ENCODE (135) and RoadMap (136) projects (as available in March 2018) were retrieved. We collected data for 9,074 narrow peak experiments and 1,395 broad peak experiments on 42 tissues and 238 cell lines, annotating 755 functional elements divided between 724 transcription factors and 31 histone marks. Then, using the BEDTools intersect module (137) with default parameters, we checked, for each collected variant, if its genomic position fell within a functional peak in all replicates of a given TF/histone mark specific experiment. Overall, we annotated all the variants by the number of experiments that supports a TF or histone mark in a cell line/tissue. We also annotated variants based on functional marker data available from CONREL (138), a resource we recently developed which provides an extensive collection of consensus promoters, enhancers and active enhancers across 38 tissue types.

### *Impact of genetic variants on binding motifs*

We retrieved 5,424 TFBS consensus motifs (in the form of position frequency matrices PFM) from Transfac Professional (139), Hocomoco (140), Homer (141), and Jaspar (142) and 552 human RNA binding protein consensus motifs from RBPDB database (143). Extending an approach we previously proposed and used in (144), for each variant we performed an extensive motif search using a pattern matching approach, considering a 30bp flanking window around the variant and using the TESS computational tool (145). RBP motifs were used to characterize only UTR variants.

Among the log-likelihood-ratio-based scores provided by TESS we used the score  $L_a$ , which represents the log-odds ratio of the match, and the score  $L_m$ , which represents the maximum possible log-odds ratio. TFBS and RBP scores (hereafter referred to as *binding motifs scores*) were computed considering both the reference genome sequence and the sequence modified with the variant alternative allele. For each motif, significance of scores was determined comparing the calculated scores against a motif-specific reference distribution of scores computed across random genomic sequences. For motifs shorter than 11 nucleotides we enumerated all the possible nucleotide combinations, while for longer motifs we extracted 1,000,000 random sequences from the hg19 human reference genome. Considering all positive scores obtained across all motifs tested at the specific

genetic variants, score ratios  $L_a/L_m$  were calculated and normalized considering the average of  $L_a/L_m$  ratios and the average of length-specific  $L_a/L_m$  ratios.

Overall, motif matches at the specific genetic variant locus were retained when: i) the match overlaps the genetic variant; ii) the score for the reference allele or the score for the alternative allele was at least 6 (TESS default parameter) for TFBSs and 2 for RBPs (which motifs are generally smaller); iii) the score p-value for the reference or the alternative allele calculated against the motif-specific reference distribution is smaller than 0.001; iv) the normalized  $L_a/L_m$  score ratio for the reference or alternative allele is greater than 0.5.

Retained variants were classified as a “match”, when the difference between the score computed on the reference sequence and the alternative sequence was less than 10%, and as a “change”, when this difference was equal or greater than 10%. Instead, we call an “addition” when the alternative allele score is positive (and respects all the other thresholds) and the reference allele score is negative, while a “deletion” is called in the opposite case. When the analysed genetic variant is a small insertion and the motif match starts inside the added genomic sequence, we call it an “addition” in all cases. Examples of considered cases are provided in **Figure S3.1**.

#### *Integration of TCGA and GTEx projects data*

Genotype and transcriptomic information from either TCGA (146) and GTEx (147) datasets were collected and examined. We conducted the analysis across 15 different human tissues for which genotype-expression normal matching samples were provided, including breast, brain, uterus, lung, liver, cervix, prostate, pancreas, stomach, esophagus, thyroid, skin, ovary, colon and bladder. Specifically, genotype and normal RNA-seq samples from each tissue were processed and analysed separately according to the tissue-specific data availability from TCGA and GTEx, generating a unique (GTEx/TCGA) combined dataset when data from both resources were present. A comprehensive list of all processed tissues and the amounts of tissue-specific samples is reported in **Table S3.1**.

#### *RNA-seq data from TCGA and GTEx projects*

Tissue-specific RNA-seq data from either TCGA normal (non-tumor) samples or GTEx samples were downloaded from the Recount2 (148) project data portal and processed as

follows: raw count matrices were extracted and filtered to retain only protein-coding genes according to GRCh38-v25 Human Gencode annotation ([www.gencodegenes.org](http://www.gencodegenes.org)). Tissue-specific TCGA and GTEx RNA-seq samples were combined into a unique matrix and genes having RPKM  $\geq 0.5$  in at least the 10% of the samples were considered as expressed and hence retained in the downstream analyses. Normalized gene counts were then obtained using edgeR (149) followed by a voom-quantile normalization function (150) to correct for both technical and biological variability across samples. Tissue-specific TCGA and GTEx combined data was further normalized using ComBat (151) to adjust for the source-specific batch effect generated in the merging step.

#### *Genotype data from TCGA and GTEx projects*

Tissue-specific raw TCGA genotype calls were downloaded from TCGA legacy data portal ([portal.gdc.cancer.gov/legacy-archive](http://portal.gdc.cancer.gov/legacy-archive)) and converted into the common PLINK (152) file format (MAP/PED) retaining only genotypes with confidence score larger than 0.1. PED files underwent a first pre-filtering step to remove duplicate SNPs and discard variants with a call rate smaller than 0.75. MAP/PED files were then converted into more readable GEN/SAMPLE format using Gtool ([well.ox.ac.uk/~cfreeman/software/gwas/gtool.html](http://well.ox.ac.uk/~cfreeman/software/gwas/gtool.html)). Chromosome-separated GEN files were then analysed with SHAPEIT v2 (110) to infer haplotype structure and optimize genotype content information for the imputation process. Variants were imputed using IMPUTE v2.3.2 (34) against a reference panel built from 1,000 Genomes Project data. Imputed TCGA genotype calls were intersected with imputed GTEx genotype data obtained from dbgap (phs000424.p7.v2) and samples with overall call rate below 0.9 were excluded. Only variants with MAF greater or equal than 1% were finally retained.

#### *Ancestry analysis*

Ancestry analysis was performed using EthSEQ (153). For each tissue-specific TCGA/GTEx integrated genotype data, a selection of random 10% common variants with MAF  $>5\%$  (about 700,000) were selected and used to run EthSEQ using a reference model built from 1,000 Genomes Project data. The first three principal components of the PCA analysis performed by EthSEQ, which effectively describe the major populations structure (154), were extracted from EthSEQ results.

### *Association of genetic variants genotype with transcript levels*

Tissue-specific associations between variants genotypes and genes transcripts were calculated using the following model of linear correlation:

$$E \sim \beta_0 + \beta_1 G + \beta_2 A + \beta_3 S + \beta_4 PC1 + \beta_5 PC2 + \beta_6 PC3$$

where, E is the transcript level of a gene,  $\beta_0$  is the intercept coefficient, G is the genotype of a genetic variant, A is the individual's age, S is the individual's sex and PC1, PC2 and PC3 are the first three EthSEQ principal components. Each genetic variant was tested against all the genes expressed in the tested tissue using three different association models: additive, dominant, and recessive. In the additive model we grouped samples in three different genotype classes: reference homozygous, heterozygous and alternative homozygous. In the dominant model we combined the heterozygous samples with alternative homozygous while in the recessive model heterozygous are combined with reference homozygous. Age, sex, and the three PCA terms were included to correct biases toward genes whose expression changes during life, variants that are more common in a sex with respect to the other and effects on transcript that are due to individuals' ancestry. We tested a variant for the association only if the genotype had at least 3 samples in each genotype class. P-values for the associations were obtained by a two-tailed t-test on the genotype coefficient  $\beta_1$  under the null hypothesis that  $\beta_1$  is equal to zero. For each variant and model, p-values were corrected using Benjamini-Hochberg method considering all tested genes as multiple hypothesis. Only associations with a corrected p-value less than 0.005 were considered for further analysis.

### *Variants Similarity Network*

Given a variant  $v$ , let  $G$  be the set of genes annotated in Polympact having transcript levels associated with  $v$ . Now, let  $I_v$  be the set of pairs such that:

$$I_v \subseteq G \times \{+, -\}$$

where a gene  $g$  is associated with "+" when the variant alternative allele increases the transcript level of  $g$  and with "-" when the variant alternative allele decreases it. Now, the similarity of two variants  $v_1$  and  $v_2$  in terms of transcript level associations (denoted also as *variants transcript similarity*) is defined as:

$$S_{transcript}(v_1, v_2) = Jaccard(I_{v_1}, I_{v_2}) = \frac{|I_{v_1} \cap I_{v_2}|}{|I_{v_1} \cup I_{v_2}|}$$

The function  $S_{motifs}(v_1, v_2)$  is defined by applying the same idea to binding motif alteration PolymPact data (*variants motifs similarity*). Specifically, a gene is associated with “+” when the variant alternative allele increases the binding motif score and with “-” when the variant alternative allele decreases it.

Based on these definitions, we can define a similarity network where the nodes are variants and two variants are connected if and only if their transcript (or motifs) similarity is greater than zero. Since connected variants may have relationships with common genes, we can use community detection algorithms to identify groups of variants presenting similar functional impact.

To study the similarity degree of variants’ pairs, we selected for each tissue all variants associated with the transcript level of at least one gene. Then, we computed with PLINK the sets of variants that are not in linkage disequilibrium using a genomic window of size 250kb and using 0.1, 0.5 and 0.8 as  $r^2$  thresholds. Finally, for each threshold, we built both variants transcripts and variants motifs similarity networks.

### *Variant-Gene Network*

Combining all data available in PolymPact, we finally developed a model to describe the complex interaction landscape between a set of common genetic variants and genes. We formalized this model as a variant-gene network, defined as a directed bipartite graph where nodes are variants or genes, and edges are relations we found between variants and genes. Edges have a variant to a gene direction when the variant associates to the transcript level of the gene, while an edge has a gene to a variant direction when the gene binds at the variant locus.

To analyse PolymPact variant-gene networks, we constructed for each tissue the networks using all variants associated with at least one transcript level. Then, we analysed the networks structures by finding the strongly connected components and exploring centrality measures like degree, betweenness and closeness. Finally, we enumerated every possible 2-cycle in the network. All the analyses were conducted using NetworkKit (155).

### *Polym pact database and web interface implementation*

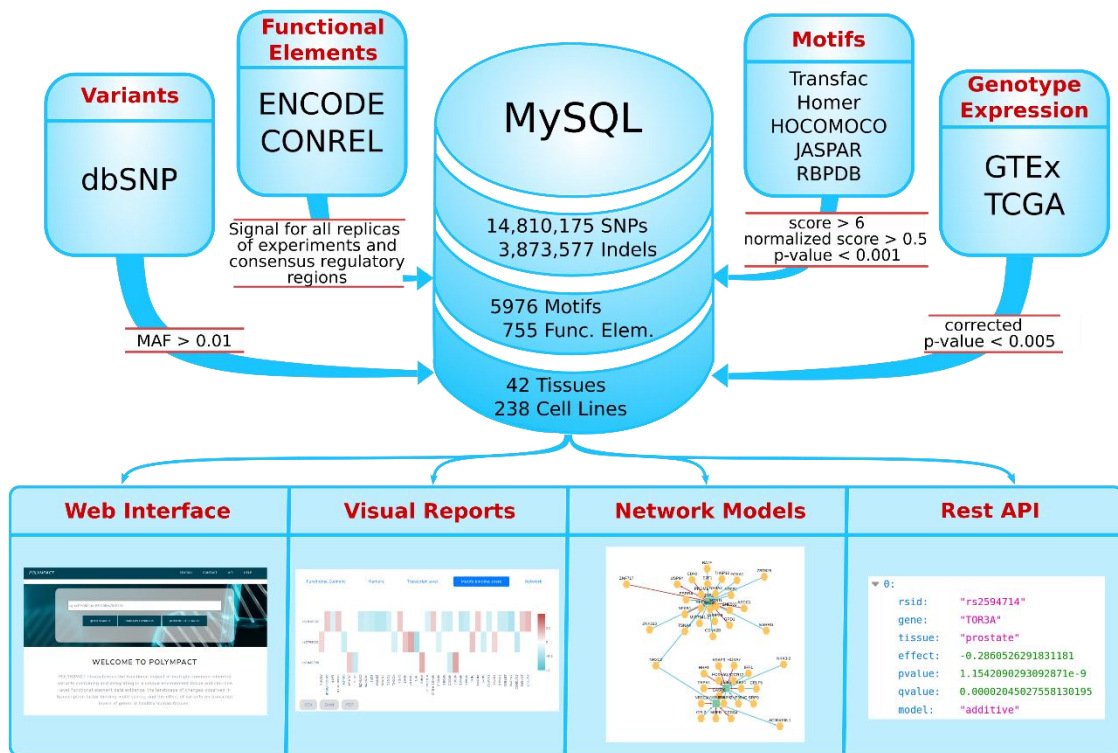
Polym pact database is hosted on a MySQL version 5.7 containerized with singularity version 3.4. The web interface is implemented in Python3 using the Django framework version 3.0.5. The data visualization is obtained using Plotly-Dash version 3.1 for the heatmaps and Cytoscape-js 0.1.1 for the networks. Community detection in similarity networks is performed using the Louvain algorithm (156).

## Results

### *Overview of Polym pact data*

Polym pact characterizes (**Figure 3.1**) more than 18 million common human genetic variants and allows for the exploration of: i) their functional properties, exploiting more than 10,000 cell lines and tissue ChIP-seq experiments; ii) their impact on binding motifs scores, exploiting about 6,000 TFBS/RBP consensus motifs; iii) their tissue-specific association with transcript levels, exploiting genotype and RNA-seq data of more than 5,000 human individuals. A summary of the data contained in Polym pact is reported in **Figure 3.2**.





**Figure 3.1. Polym pact data and services.** Polym pact is implemented integrating common variants information and genotypes, ChIP-seq data, TFBS and RBP motifs and genotype/transcript level data retrieved and integrated from several databases. Data are filtered for high quality characteristics and stored in a MySQL database. Polym pact offers an intuitive web interface providing visual reports and an innovative network visualization. It is also accessible programmatically through a REST API.

More than 95% of the variants characterized by Polym pact are non-coding and annotated as intergenic or intronic variants (**Figure 3.2A**). Specifically, 143,725 are annotated as variants in the 3' UTR, 12,362 in the 5' UTR, 10,018,421 are intergenic, 210,232 are located in a transcription termination site, 82,826 are exonic, 7,899,469 are intronic, 84,127 in non-coding RNAs, and 232,590 in promoters.

Regarding the functional characterization of the variants, we found that 18,545,354 of 18,683,752 (99.26%) fall within at least one ChIP-seq peak (18,485,601 fall in at least one histone mark peak and 18,409,488 in at least one TF peak considering both broad and narrow peak data) in at least one tissue. As shown in **Figure 3.2B**, the majority of variants fall within few peaks across all tissues (**Figure 3.2B**) with half variants falling in 2 to 3 peaks in every tissue. In addition, 170,239 (0.9%) variants have a promoter annotation in at least

one cell-line/tissue, whereas 7,839,972 (42%) have an enhancer annotation and 4,357,136 (23%) have an active enhancer annotation.

With respect to the TFBS motifs analysis landscape, we observed that more than 99.9% (18,678,853) of the variants cause at least one putative change, addition or deletion of a transcription factor. More specifically, 17,277,379 variants cause at least a putative change, 7,724,608 cause at least one addition and 8,859,076 cause at least one deletion. About 59% of the motifs analysis results are annotated as match, while 32% show a change in the score. Additions and deletions account, respectively, for the 3% and 6% of the overall motifs analysis results (**Figure 3.2C left**). The distribution of the number of variants matching or altering a certain number of motifs show that we have more than 2.5 million variants overlapping 40 to 50 putative motifs that are annotated as matches, changes, additions or deletions. The distribution is slightly asymmetrical with very few variants that are associated with only 1 to 10 motifs (**Figure 3.2C right**). Putative change, addition or deletion of RBP motifs was observed in 95,265 UTR variants (66%), with 66,654 variants causing at least a putative change, 17,488 causing at least one addition and 27,412 causing at least one deletion.

Moving to the association with transcript levels, unlike eQTL analysis and similarly to what we have previously proposed in (144), association of genetic variant genotypes and transcript levels is here performed by testing each variant against all protein-coding transcripts, to search for association patterns that might be similarly shared across different variants. We found 3,653,655 variants with a total of 6,451,090 associations across fifteen tissues and the three different association models. Of these, 1,037,712 were additive associations, 2,555,425 were dominant and 2,857,953 were recessive. As shown in **Figure 3.2D**, thyroid was the tissue with the highest number of variants displaying associations (N=873,525) and bladder the one with the lowest number (N=63,448). Although the median value of associations per variant is one, the mean value is pretty variable across tissues (minimum 5.6 for uterus and maximum 104.2 for skin) indicating the presence of variants strongly enriched for associations. Indeed, as shown in **Table S3.2**, while the 75<sup>th</sup> percentile of the tissue-specific variants associations distributions indicates an average number of associations per variant that is less than 3, when considering the 95<sup>th</sup> percentile, we observe an average value of 97 associations, which rapidly increases to 1,171 associations when we consider the 99<sup>th</sup> percentile of the

tissue-specific variants association distributions. Skin was the tissue with the highest number of total identified associations, while uterus was the tissue with the lowest number (Figure S3.2). As expected, most (~99%) of these associations are putative *trans* associations. Although our approach differs from standard eQTL analysis, we checked to what extent the putative *cis* associations we found are similar to the landscape of *cis* associations reported by GTEx eQTL data. Focusing for simplicity on a subset of tissues, we took the intersection between the variants characterized in GTEx and Polypact, and computed the fraction of *cis* associations in Polypact by selecting, similarly to GTEx, variants within one megabase of distance from the modulated gene in the selected tissues. We found a good concordance with about 60% of our *cis* associations that are also reported in GTEx and preserving in all cases the association direction (Figure S3.3).

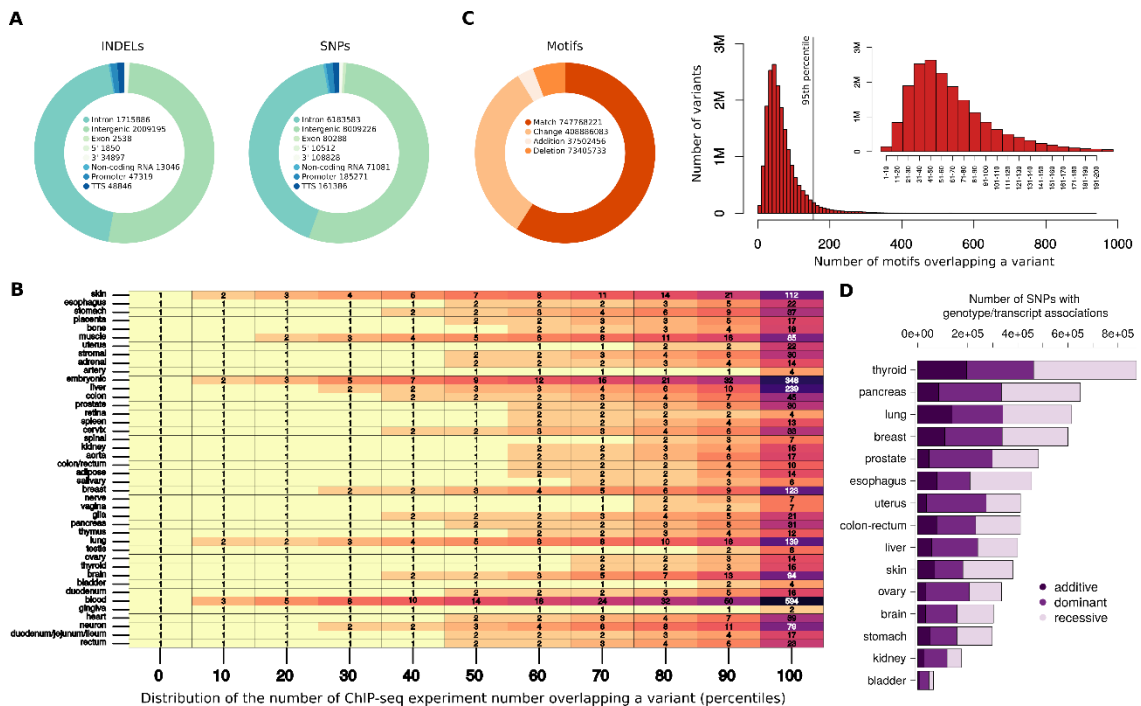


Figure 3.2: **Summary of the data contained in Polypact.** A) Annotations for the different types of variants stored in Polypact. B) Percentiles of the distribution of the number of variants overlapping a ChIP-seq peak in various tissues. C) Types of binding motifs results and distribution of the number of motifs overlapping a variant for match, change, addition and deletion types. In small, a zoom of the major distribution part. D) Number of variants associated with a transcript level in additive, dominant and recessive models in different tissues.

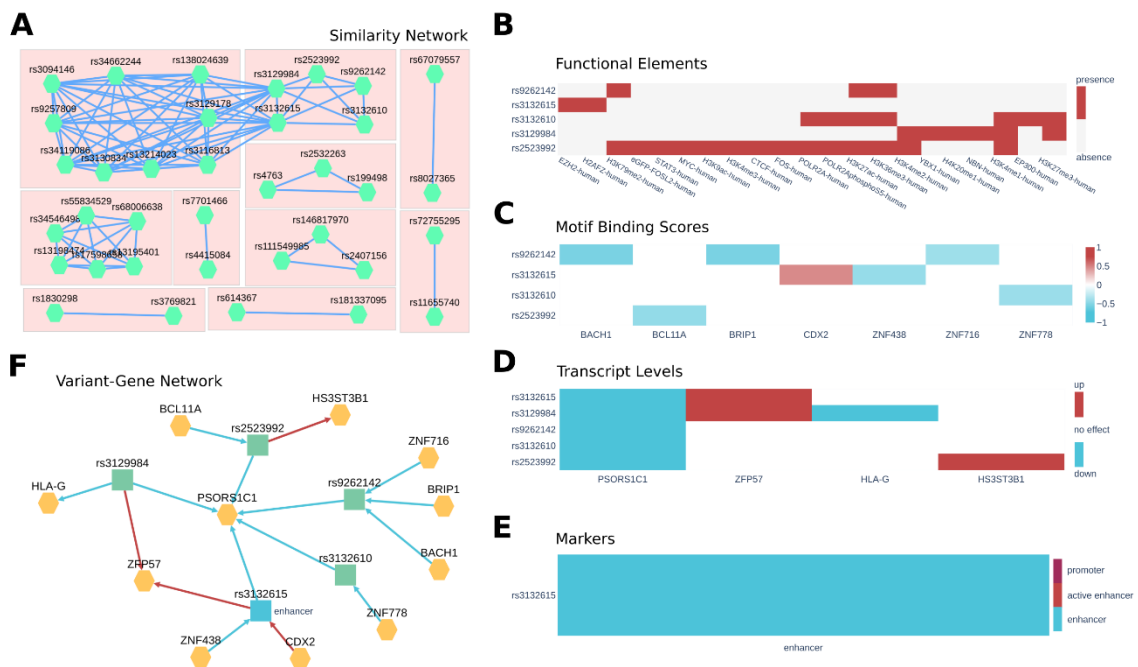
### *Database and web interface*

Polypact offers a web interface accessible through a web browser that can be used to query the resource by selecting the variants of interest and the preferred parameters setting. The only mandatory parameter is the list of variants IDs (in the form of rsids or strings with the variant position, reference and alternative alleles) while all the others are optional. The resource offers two search modes: quick and advanced search for both similarity and interaction analysis modes. The quick search is available from the home page (**Figure S3.4A**) and retrieves the data for the requested variants using the default parameters (all tissues, all motifs effects and all associations models). In the advanced search page (**Figure S3.4B**) a selection tree can be used to select a specific tissue of interest or a selection of specific cell lines available for that tissue. Using checkboxes, it is possible to specify the peak file format for the ChIP-seq data (narrow and/or broad peaks), the model used in the genotype/transcript association analysis (additive, dominant and/or recessive models) and the type of binding motifs results (match, change, addition and/or deletion). In addition, the corrected p-value threshold for the genotype/transcript association analysis (default 0.005) and the normalized difference in binding motifs scores (default 0) can be set to further filter displayed genetic variants results. Of note, only for a subset of selectable tissues the genotype/transcript association analysis data is available, and the cell-line selection is exploited only for the analysis of functional elements.

Polypact similarity analysis provides an interactive interface to explore the similarity network of queried variants' effects on transcript levels (**Figure 3.3A**) or on binding motifs scores. Computed network communities are highlighted, and each single community can be selected to perform an in-depth interaction analysis.

Polypact interaction analysis provides first a graphical representation, in the form of a heatmap, to explore functional relationships among the queried variants, separately for the functional elements, the binding motifs score alterations, and the transcript level associations. The heatmaps are accessible through, respectively, the "Functional Elements", "Transcript Levels" and "Motif Binding Scores" tabs (**Figure 3.3BCD**) and are clustered using hierarchical clustering in a way that variants with similar characteristics are represented closer in the visualization. All the data is also reported in a tabular form and can be filtered and downloaded in various file formats. The "Markers" (**Figure 3.3E**)

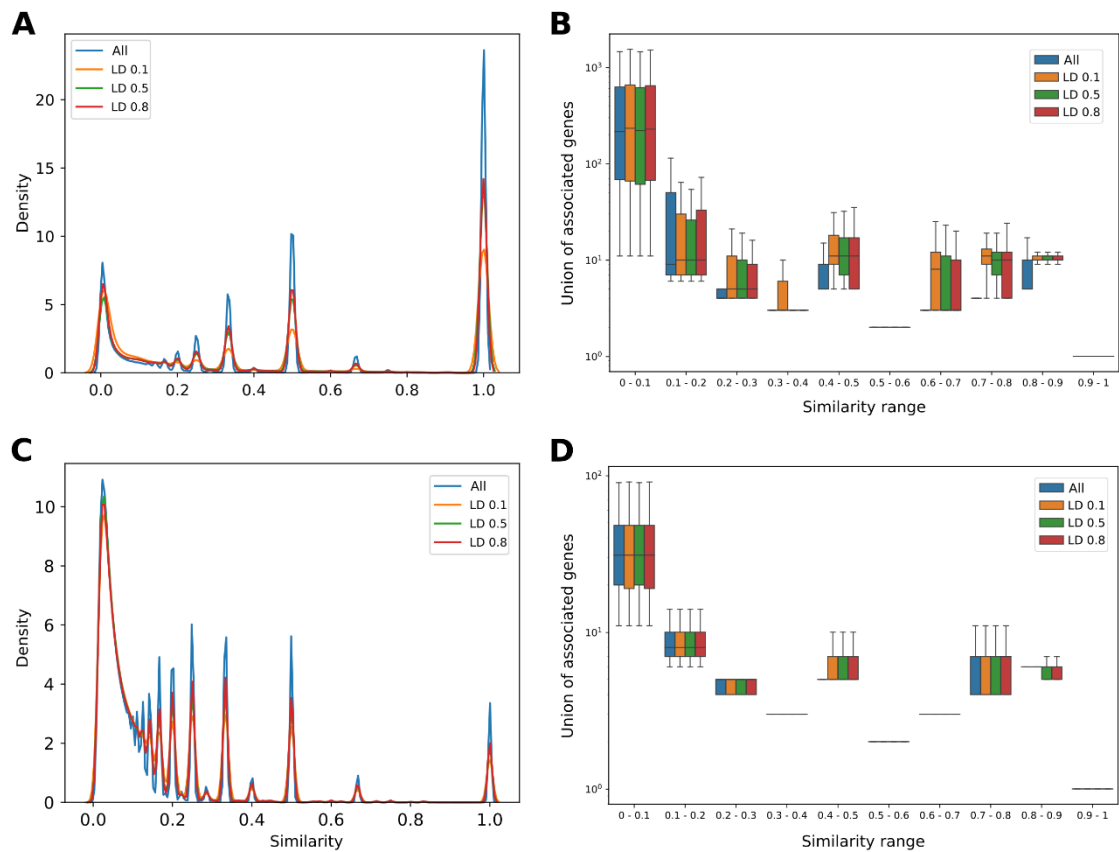
tab provides additional insights into the regulatory role of the genomic regions spanning the variants and highlights links to our external resource CONREL to visualize the variant and its genomic context into a genome browser view. The variant-gene network model is accessible from the “Network” tab (**Figure 3.3F**) where genes are reported in yellow and variants have colours representing their functional marker annotations across the cell lines/tissues selected. Edges are red if the variant alternative allele increases the binding motif score or is associated with increased transcript level; they are blue if the variant alternative allele decreases the binding motif score or is associated with decreased transcript level.



**Figure 3.3 Polypact web interface.** **A)** Polypact similarity network model built from transcript association data. Each node represents a variant and two variants are connected if and only if they have a similarity greater than zero. **B)** Histone marks and transcription factor ChIP-seq peaks overlapping in the genomic region of the variant. **C)** Binding motifs scores. Red annotates variants that are increasing the motif binding score of the motif while blue annotates variants reducing it. **D)** Genotype to transcript level associations. Alternative alleles lowering the transcript level are depicted in blue while alternative alleles increasing it are depicted in red. **E)** CONREL marker annotations in the genomic region of the variant. **F)** Polypact variant-gene network model. Genes are coloured in yellow while variants are in blue if they are annotated as putative active enhancers or in green otherwise. Edges from a variant to a gene represents an association to transcript levels and are blue if the transcript are reduced and red if transcript levels are increased. Edges from a gene to a variant represent binding motifs changes and are red if the binding score is increased and blue if it is decreased.

### *Properties of Similarity Networks*

Using PolymPact data, similarity networks considering all variants' pairs were created for 15 tissues. Networks based on binding motifs scores focused only on effects classified as addition or deletion, considered more relevant from a functional perspective. On average 1.5% (N= 1,241,691,365) of all possible variants' pairs had a positive  $S_{transcript}$  similarity and 0.4% (N= 367,267,669) had a positive  $S_{motifs}$  similarity. Comparable results were obtained when high linkage disequilibrium (LD) variants were filtered (**Table S3.3** and **Table S3.4**). Analysis of similarity values distributions across the networks revealed specific properties. Focusing for example on the breast tissue transcript similarity network, but comparably for all other tissues, the distribution of  $S_{transcript}$  values was multimodal with a range of peaks located across the overall range [0,1] and the highest peak located in value one, representing perfect similarity (**Figure 3.4A**). As shown in **Figure 3.4B**, most similarities located in the highest peak were, as expected, from variants' pairs associated with a single gene; in spite of that, we observed a tail of pairs involving tens of genes. Concordant distributions were obtained when correcting for linkage disequilibrium, demonstrating that a large fraction of similarities are not due to LD. Results obtained considering the distribution of  $S_{motifs}$  similarity values were comparable (**Figure 3.4CD**), further demonstrating the presence of a vast range of variants' pairs not in LD sharing common functional relations.



**Figure 3.4. Analysis of variants similarity networks.** **A)** Distribution of variants transcript similarity values in the breast similarity network. **B)** Cardinality of the union of associated transcripts of each interacting pair divided by similarity thresholds. **C)** Same as A) but with variants motifs similarity values. **D)** Same as B) but with variants motifs similarity values.

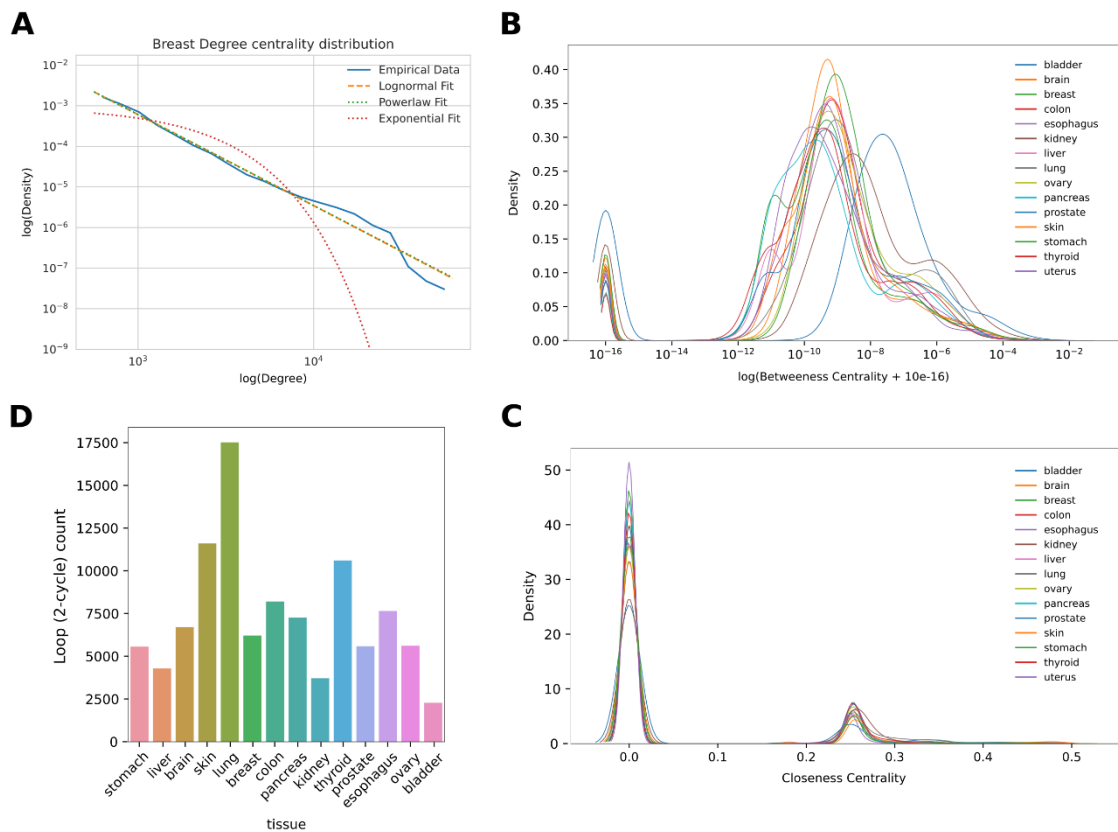
#### Properties of Variant-Gene Networks

For each tissue, we created a variant-gene network considering all the variants associated with at least one transcript level in that tissue, and studied the topology of the network by exploring the number of connected components, the distribution of centrality metrics and the embedded 2-cycles. Each tissue network showed a similar topology, consisting of a single strongly connected component and many isolated nodes (Table S3.5). Degree centrality distributions highlighted across all tissues a heavy tailed power law or a log-normal distribution with a likelihood ratio test propending for the log-normal distribution (Figure 3.5A, Figure S3.5). Betweenness centrality distribution showed instead that, for each tissue, a large number of nodes do not participate in the network connections being the nodes outside the strongly connected component (Figure 3.5B). In particular, most tissues follow a similar distribution suggesting a conserved topological structure with

bladder tissue showing a shift in the distribution, probably due to the low number of nodes in the network. Inspection of closeness centrality also showed a conserved distribution among tissues with a peak in zero, where all the nodes not belonging to the strongly connected component are located, and a second peak near 0.25, which contains the nodes in the main connected component (**Figure 3.5C**).

We then focused on variant gene network cycles, which are structures involving relations between variants and genes. Specifically, we focused on variants associated with the transcript level of a TF that are also modifying the binding motif score of the same TF, forming a 2-cycle in the network. Cycles are of particular interest because they may underlie the presence of positive or negative feedback loops between variants and transcription factors. We observed 2-cycles in every tissue (**Figure 3.5D**) from a maximum of 17,522 in lung to a minimum of 2,283 in bladder. By investigating the possible functional impact of 2-cycles we found that variants involved in 2-cycles are enriched in functional markers (p-value=1.7e-76, **Table S3.6**).





**Figure 3.5 Analysis of variant-gene networks.** **A)** Log-log plot of the degree centrality distribution of the breast variant-gene network. **B)** Betweenness centrality distribution of the variant-gene network across tissues. **C)** Closeness centrality distribution of the variant-gene network across tissues. **D)** Number of 2-cycles across tissues.

### Case Studies

To explore the utility of Polym pact, we considered a first case study based on cancer risk GWAS common variants and a second case study based on Alzheimer’s disease risk GWAS common variants (Table S3.7, Table S3.8).

### Cancer risk GWAS variants

2,657 variants related to cancer were retrieved from the GWAS catalogue (95), of which 2,370 were present in Polym pact.

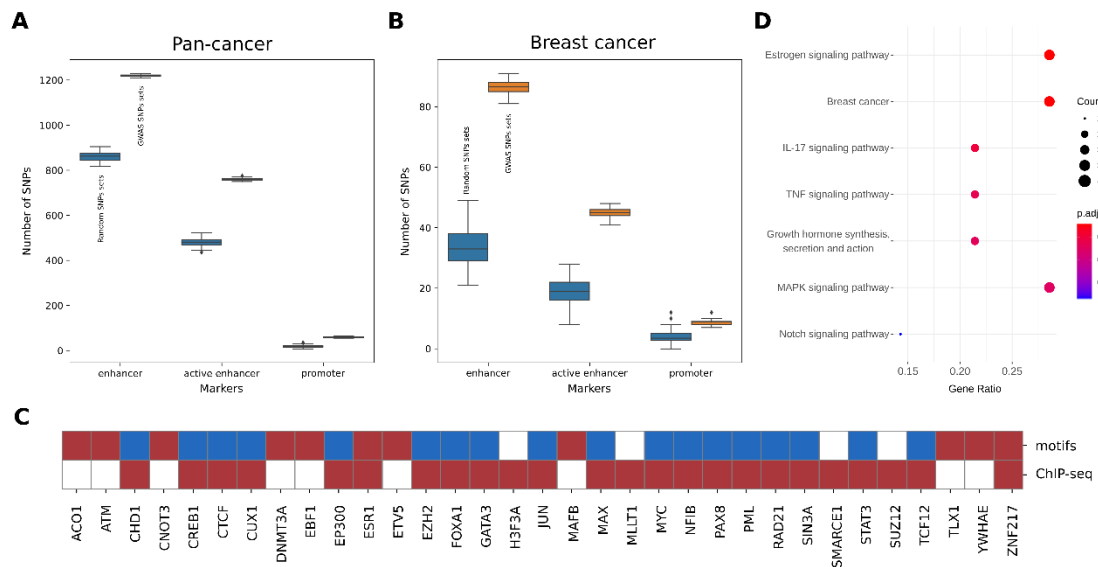
We first explored the landscape of functional annotations across the loci identified by the GWAS risk variants. After computing the extent of linkage disequilibrium across the 2,370 variants using the Ensembl REST API (157), we identified 1,958 LD blocks; we considered pairs of variants with an  $r^2$  greater than 0.5 to be LD. Then we built 100 random GWAS

sets where a single variant is randomly selected from each LD block, hence obtaining 100 sets of 1,958 GWAS variants that are not in LD. We also created 100 sets of 1,958 random variants selected among all variants in Polym pact (excluding the 2,370 GWAS variants) and preserving the distribution of minor allele frequency of the GWAS variants.

We then counted for each GWAS and random variant the number of overlapping marker regions and compared the distribution of counts in the GWAS variants sets with respect to the random variants sets. As shown in **Figure 3.6A** and **Figure S3.6A**, markers of promoters, enhancers, active enhancers together with a subset of histone marks result enriched in the GWAS sets with respect to the random sets ( $p$ -value $<0.01$ ), clearly supporting the observation that variants associated with cancer risk have an active functional role.

Being the number of risk variants reported in the GWAS catalogue not uniformly distributed across the different cancer types, we decided to further explore risk variants functional properties by focusing only on a single cancer type. Specifically, we selected the richest subset of 853 GWAS variants that are associated to breast cancer risk, 808 of which are characterized in Polym pact. Of those, 58 variants are associated with at least one transcript level with 445 total unique associations, of which 71 (~16%) are *cis*-associations. Out of the 808 variants, we identified 653 LD blocks and we generated as previously 100 random sets of 653 GWAS variants (not in LD) and 100 random sets of 653 random variants. Also in this case, markers of promoters, enhancers, active enhancers and a selection of histone marks resulted enriched in the GWAS variants sets with respect to the random variants sets ( $p$ -value $<0.01$ , **Figure 3.6B** and **Figure S3.6B**). In addition, more than 30 genes known to be implicated in cancer were found to have enriched transcription factor functional peaks in the GWAS variants sets with respect to the random variants sets and/or binding motifs that are changed, added or deleted by GWAS variants alternative alleles (**Figure 3.6C**). In particular, the estrogen receptor *ESR1* and the oncogene *ZNF217* are both enriched for functional peaks in the GWAS variants and have binding motifs that are significantly impacted by a subset of the same variants. Interestingly, focusing more generally on all transcription factors (not only cancer genes) that have this dual characteristic, we found a set of genes that enrich for pathways related to hormone synthesis, estrogen signalling and breast cancer (**Figure 3.6D**), overall

supporting the implication of breast cancer GWAS variants in cancer relevant biological processes.



**Figure 3.6 Pan-cancer and breast cancer GWAS analysis.** *A) GWAS variants associated to all cancer types annotated as enhancers, active enhancers and promoters compared with random sets of variants. B) Same as A) but with GWAS variants associated to breast cancer. C) Cancer transcription factor over-represented in breast cancers GWAS variants with respect to random sets in ChIP-seq data and TFBS motif alterations. Red p-values < 0.01, blue p-values > 0.01, white no data. D) Cancer related terms resulting from the enrichment analysis on all transcription factors over-represented for both ChIP-seq and TFBS motif alterations.*

To further characterize the functional role of breast cancer risk GWAS we analysed all the 808 variants using the Polypact transcript similarity network created with standard parameters and focusing on breast tissue. We found 10 unique network communities (Figure 3.3A). Out of them, we selected 4 communities associated with genes *CASP8*, *MAN2C1*, *BTN3A2* and *ARL17A* which are all reported in literature as possibly involved in cancer. The *CASP8* network community contains the two variants rs1830298 and rs3769821. In particular, the variant rs1830298 is 60kbp far away from the variant rs3769821, which is annotated as an intron variant of *CASP8*. Variant rs1830298 alternative allele reduces the binding score of *NR2C2* hormone receptor while rs3769821 decreases the binding score of the tumor suppressor *IRF1* (Figure 3.7A); both variants have the GWAS catalogue reported risk allele (allele C) that is strongly associated with a decrease in the *CASP8* transcript levels (Figure 3.7BC). Of note, our integrated TCGA and GTEx dataset from which the associations were computed is composed by individuals

with mainly European (75%) and African (16%) ancestry, populations were the two variants have respectively moderate ( $r^2 \sim 0.5$ ) and low ( $r^2 \sim 0.2$ ) linkage disequilibrium. Inspection of transcript levels at all variants genotype combinations (**Figure 3.7D**) revealed that homozygous alternative genotype (TT genotype for both variants) is needed to sustain on average high *CASP8* transcript level. In particular, a first decrease of *CASP8* levels is observed in presence of the risk allele C in at least one variant (e.g. when at least one variant has heterozygous genotype) and a further decrease is observed when at least one variant has homozygous CC genotype. These results, together with the genotype combinations observed from phased data retrievable from the 1,000 Genomes Project data (**Figure S3.7**) strongly suggest a putative interaction effect that rs1830298 and rs3769821 have in maintaining a high level of *CASP8* transcript, which is lost in individuals carrying the breast cancer risk allele in at least one of the two variants.

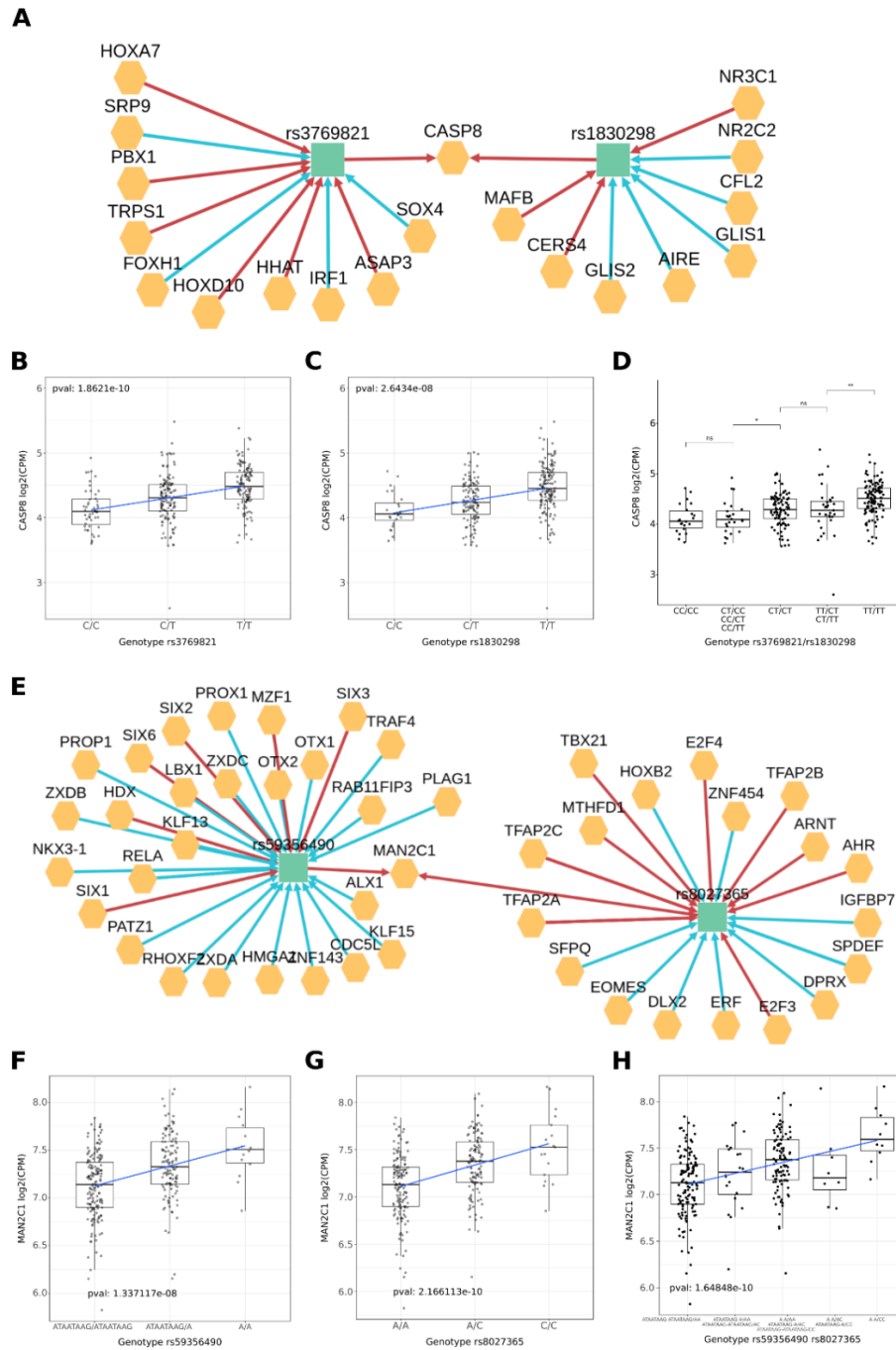
The *MAN2C1* network community (**Figure 3.7E**) includes SNP rs8027365 (*PTPN9* intron variant, risk allele A) and the small deletion rs59356490 (intergenic variant, risk allele deletion not present) located 120kb away and they are not reported to be in LD. Both variants modulate additively the transcript level of *MAN2C1* (**Figure 3.7F-H**) and variant rs59356490 has a functional annotation for *POLR2A* and *ESR1* and overall it deletes 23 TFBS motifs. The combination of the two variants shows an additive trend where the highest transcript level is reached when both variants are present.

The *BTN3A2* network community contains 6 variants: rs13195401, rs13198474, rs17598658, rs34546498, rs55834529 and rs68006638 and they are all associated with a decrease in transcript level for *BTN3A2* gene in the dominant model. Among them we selected the pair rs13195401 (annotated as *BTN2A1* non-sense variant, risk allele G) and rs13198474 (annotated as *SLC17A3* 5' UTR variant, risk allele G) having the lowest LD ( $r^2=0.49$ ) in the general population (**Figure S3.8**). The combination of the two effects shows a trend where the decrease is small when only the variant rs13195401 is present, the decrease is higher when only the variant rs13198474 is present, and the highest decrease in the transcript level is reached when both variants are present.

Finally, we analyzed the network community of *ARL17A*. The two variants in this community are rs2532263 (*KANSL1* intron variant, risk allele G) and rs4763 (*ARHGAP27* 3' UTR variant, risk allele G) and both are associated with an increase of *ARL17A*,

*LRRC37A*, *LRRC37A2* and *CRHR1* genes transcript levels. For all genes the variants have a full additive effect similarly to *MAN2C1* (**Figure S3.9**).

Interestingly, we also found that the variant rs8050871, located in a region annotated as active enhancer, has a *cis* effect on the transcript level of gene *ZNF23* causing a decrease in its transcript level. The variant also deletes a binding motif for the same TFs creating a loop (a 2-cycle) in the variant gene network. Overall, this suggests that the variant is possibly involved in a regulatory positive feedback loop, potentially inducing dynamic instability.



**Figure 3.7 Figure 7. Effect of variants on CASP8 and MAN2C1 transcripts. A) Variant-gene network of the two variants rs3769821 and rs1830298. B) Effect of the variant rs3769821 on the transcript level of CASP8 gene under the additive model. C) Effect of the variant rs1830298 on the transcript level of CASP8 gene under the additive model. D) Combined effect of the two variants rs3769821 (first pair of nucleotides in the label) and rs1830298 (second pair) on CASP8. E) Variant-gene network of the two variants rs59356490 and rs8027365. F) Effect of the variant rs59356490 on the transcript level of MAN2C1 under the additive model. G) Effect of the variant rs8027365 on the transcript level of MAN2C1 under the additive model. H) Combined effect of the two variants rs59356490 (first pair of nucleotides in the label) and rs8027365 (second pair) on MAN2C1.**

### *Alzheimer's disease GWAS variants*

1044 common variants related to Alzheimer's disease were retrieved from the GWAS catalogue, 810 of which were present in Polympact.

To highlight the utility of Polympact in identifying more putative functional relations, we analysed the 810 variants exploiting the similarity network computed on the brain tissue. We identified 4 network communities and focused only on the 3 ones composed by variants that are not in high LD.

The first community contains variants rs199499 and rs7207400, that are reported in the GWAS catalogue as associated to trait *Alzheimer's disease in APOE ε4- carriers* and that present a low LD in the general population ( $r^2=0.18$ ) and moderate LD ( $r^2=0.52$ ) in the European population. The second community is formed by three variants, rs113260531, rs7225151 and rs80257887, that are reported as associated to trait *Alzheimer's disease or family history of Alzheimer's disease*; the first two variants are located on chromosome 17 and are in moderate LD ( $r^2=0.67$ ) while the third variant is located on chromosome 19. Finally, the last community is formed by variants rs7963314 and rs79926713, associated with trait *Alzheimer's disease and Late-onset Alzheimer's disease* and are located on two different chromosomes.

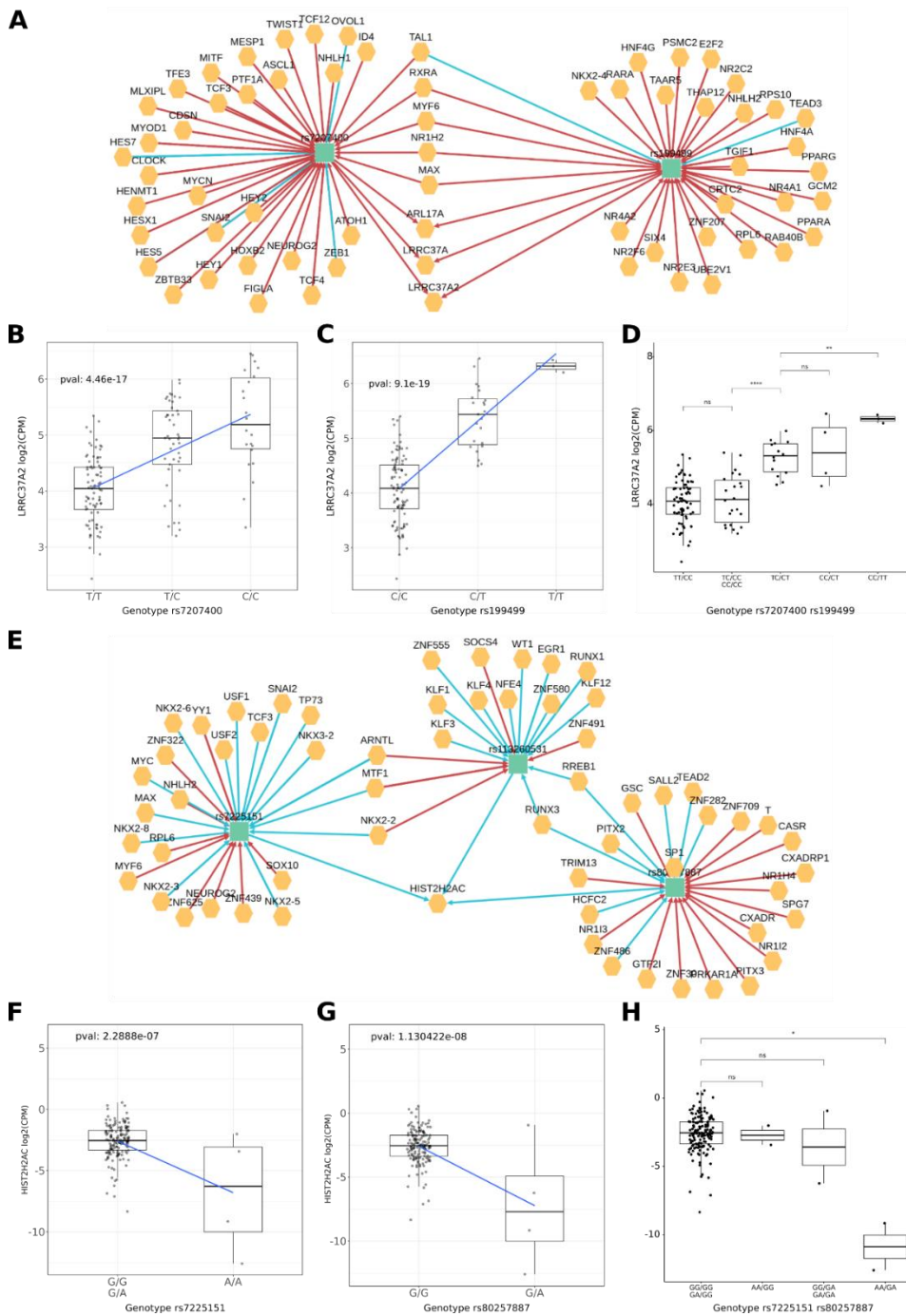
The first community variants are located on chromosome 17 and are about 1Mbp afar. Variant rs199499 is annotated as intron variant of gene *LRRC37A2* and is located about 800kB downstream to the gene *MAPT*, while rs7207400 is annotated as intron variant of *LINC02210-CRHR1*. The Polympact network (**Figure 3.8A**) shows that both variants are increasing the transcript levels of *LRRC37A*, *LRRC37A2* and *ARL17A*. Interestingly, both variants are increasing the binding of MYF6 and have an opposite effect on the binding of TAL1. Also, rs7207400 notably creates new binding for MYCN and TCF4. Risk alleles (C for rs199499, T for rs7207400) are strongly associated with decreased levels of gene transcripts *LRRC37A*, *LRRC37A2* and *ARL17A* (**Figure 3.8BC** and **Figure S3.10ABDE**). Genotype combinations (**Figure 3.8D** and **Figure S3.10CF**) show that absence of risk allele for both variants is needed to guarantee the highest transcripts levels. A first decrease in transcript levels is indeed observed when one of the two variants has heterozygous genotype and a second decrease is observed when one of the two variants has homozygous genotype for the risk allele. In addition, 1,000 Genomes Project phased genotypes indicate that risk variants are almost always present on the same allele (**Figure**

**S3.11).** Overall, the data suggest that both rs199499 and rs7207400 non-risk alleles are required in phase to sustain the highest levels of *LRRC37A*, *LRRC37A2* and *ARL17A* transcripts.

The second community is composed by variants located on different chromosomes: chromosome 17 for variants rs113260531 and rs7225151 (both annotated as upstream variants for gene *SCIMP*, risk allele A) and chromosome 19 for variant rs80257887 (annotated as intron variant of *CEACAM20*, risk allele A). The PolymPact variant-gene network (**Figure 3.8E**) shows that variants rs113260531 and rs80257887 are significantly decreasing the binding scores of RUNX3 and RREB1 and have alternative alleles associated with decreased *HIST2H2AC* gene transcript levels. We selected variants rs7225151 over rs113260531 having more alternative homozygous samples for further analysis. Specifically, a recessive effect is observed for variant rs7225151 (**Figure 3.8F**) with AA genotype associated with lower *HIST2H2AC* transcript level, while a dominant effect is observed for variant rs80257887 (**Figure 3.8G**) with AA or AG genotype associated with lower *HIST2H2AC* transcript level. Notably, a reduction of *HIST2H2AC* transcript level (**Figure 3.8H**) is evident in individuals carrying both AA risk genotype for variant rs113260531 and AA or AG risk genotype for variant rs80257887.

Finally, we analysed variants rs7963314 and rs79926713. Variant rs79926713 is located on chromosome 6 (annotated as intron variant of *SYNGAP1*, risk allele T) while rs7963314 is located on chromosome 12 (annotated as intergenic, risk allele A). Variant rs79926713 is annotated as promoter and is associated with an increase in transcript of gene *PPP1R12A* in the recessive models. Variants rs7963314 is instead associated in the modulation of 19 genes in the recessive model, including *PPP1R12A* gene (**Figure S3.10GHIJ**).





**Figure 3.8 Alzheimer's disease GWAS variants analysis.** *A)* Variant-gene network of the two variants *rs7207400* and *rs199499*. *B)* Effect of the variant *rs7207400* on the transcript level of *LRRC37A2* gene under the additive model. *C)* Effect of the variant *rs199499* on the transcript level of *LRRC37A2* gene under the additive model. *D)* Combined effect of the two variants *rs7207400* (first pair of nucleotides in the label) and *rs199499* (second pair) on *LRRC37A2*. *E)* Variant-gene network of the community formed by variants *rs7225151*, *rs113260531* and *rs80257887*. *F)* Effect of the variant *rs7225151* on the transcript level of *HIST2H2AC* gene under the recessive model. *G)* Effect of the variant *rs80257887* on the transcript level of *HIST2H2AC* gene under the dominant model. *H)* Combined effect of the two variants *rs7225151* (first pair of nucleotides in the label) and *rs80257887* (second pair) on *HIST2H2AC*.

## Discussion

The study of common human genetic variants can provide insights into the biological cause of complex traits and diseases. Although several databases and web applications have been developed in the last decade to annotate and characterize genetic variants, the aggregation of these information to identify variants links and interactions has been largely unexplored. To this aim we developed Polympact, a tool that enables the exploration and the analysis of common genetic variants and their potential interactions by exploiting the integration of a large variety of biological data and analyses. Reasoning that variants' interaction could be identified by characterizing their impact and involvement in the modulation of same genes or same biological pathways and processes, we first designed a workflow to uniformly characterize a large amount of common variants based on specific functional properties retrieved from well-known public databases. Then, on top of this uniform and homogenous annotations we developed a framework to represent and explore variants functional relations. More specifically, we combined genotype data together with transcription factor and histone marks ChIP-seq peak data, TFBS and RBP motifs data and transcriptomic profiling via RNA-seq across multiple human tissues, and we implemented a framework, provided as a dedicated web-server, to systematically characterize variants and to explore the landscape of variants functional relations through the combination of clustering analysis and novel network models.

While the uniform characterization of variants provided by Polympact was tailored with respect to the built clustering and network models, the resource we provide extends and complements annotations provided by other databases. Indeed, Polympact binding motifs data were determined both for an extended number of variants and an extended number/type of motifs. The recent SNP2TFBS tool (127), for example, characterizes only around 3 million SNPs and uses only Jaspar database (142). Of note, provided that variants in UTRs can alter mRNA translation potential (158) also RBP consensus motifs were included to characterize UTR variants. In addition, our functional characterization in terms of regulatory elements uses our recent CONREL tool (138), exploiting hence a novel tissue level functional annotation of variants. Further, our genotype/transcript association analysis approach well complements eQTL interaction data and was already proven successful in characterizing and prioritizing variants in terms of their impact on

specific genes or biological processes (144). Although we recognize that this approach could limit the identification of moderate/weak *cis* effects, a good concordance with GTEx *cis*-eQTL data is shown, and overall we believe that enabling the identification of *trans* effects is fundamental to unravel key features of the architecture of complex diseases (46).

To characterize variants' functional relations, we first introduced the notion of similarity network, which allows for the identification of variants that have common effects on the level of the same transcripts or the binding score of the same TFs/RBPs. In-depth analysis of the distribution of variants' pairs similarities across networks built from different tissue data, revealed how these distributions are highly conserved also when keeping only variants not in linkage disequilibrium, supporting hence the presence of many independent variants that can possibly interact and further highlighting a landscape of complex patterns in gene regulation.

Additionally, we introduced the notion of variant-gene network, which provides a detailed network view of variants and genes interactions across different tissues integrating all Polymact data. In-depth analysis of these networks built from different tissue data, revealed heavy-tailed degree distribution highlighting the presence of regulatory hubs (variants or TFs) in the network. We also analyzed the 2-cycles present in the networks showing that variants forming these types of loops are enriched for regulatory markers, suggesting hence the possible presence of positive and negative feedback loops related to specific TFs. Overall, the analysis unravelled a complex topology and highlighted that our variant-gene network can be a useful tool to detect and analyse complex interaction patterns. Additional mesoscale and group-centric metrics could be considered to further explore properties of these large networks (159).

Using the exhaustive list of common genetic risk variants available from the GWAS catalog, we then showed that Polymact is able to highlight important features and functional relations among disease risk variants in terms of their functional genomic context, binding motifs alterations and transcript level modulations.

Exploiting Polymact data we first showed that cancer GWAS risk variants are enriched for regulatory elements annotations, in line with previous studies (160–162). In addition, we found that the set of transcription factors with functional peaks enriched for GWAS variants and having binding motifs modified by the same variants have a statistically

significant role in cancer-related pathways, suggesting that GWAS variants may hence modulate downstream effects of oncogenic pathways. These results were also confirmed when we focused on about 800 GWAS breast cancer risk variants, were using the same selection criteria we found a set of transcription factors enriched in pathways specific to breast cancer and to response to hormone related pathways. This, in line with previous observations made by us (144) and others (58) in the context of other hormone driven cancers, suggests that common genetic variants may modulate downstream effects of hormone signalling by altering the binding of hormone receptors or hormone regulated genes, potentially favouring the risk of developing cancer in only a subset of individuals carrying a specific genetic makeup. Notably, our analysis highlighted *ESR1* (estrogen receptor alpha), *GATA3* gene which is known to influence response to estrogen (163) and the well-known oncogene *MYC*. In particular, our data shows that GWAS variants associated with breast cancer risk not only are enriched in regions that are bound by the estrogen receptor but also tend to alter the way in which ESR1 binds these regions.

Further inspection of the network models built by Polympact on breast cancer GWAS variants revealed a putative interaction between two variants that, when present on the same allele, synergistically modulate the transcript level of *CASP8* gene, a key regulator of apoptotic response already shown to be downregulated in breast cancer (164,165) and involved in cancer initiation when deficiently expressed (166,167). Specifically, we have shown that *CASP8* transcript level is reduced when GWAS risk allele for at least one of the two variants is present, with the lowest expression that is observed when at least one of the two variants has a risk allele homozygous genotype. This suggests that the presence of the GWAS risk allele may favour the evasion from apoptosis, a well-known cancer hallmark, increasing hence the risk of breast cancer initiation. Our findings are in line with (168) where the authors show that the strongest associations with breast cancer risk in the region come from variant rs1830298 and that variant rs3769821 is an eQTL for *CASP8*. Our results are consistent with the authors' hypothesis that one or more variants in the region are responsible for the reduced expression in *CASP8*.

With respect to our results related with *MAN2C1* gene, it has been shown that the gene may inhibit the function of tumor suppressor gene *PTEN* in breast and prostate cancer (169) and another study found that the gene may have a protective role in cancer initiation with respect of progression (170). In our analysis, each risk allele of variants

rs8027365 and rs67079557 contribute to a reduction in the expression of *MAN2C1* transcript, suggesting hence a protective role of *MAN2C1* in breast cancer initiation.

In the context of breast cancer GWAS variants we also found variant rs8050871, involved in a 2-cycle. The variant is located in a putative active enhancer and simultaneously associated with decreased transcript level of *ZNF23* and decreased *ZNF23* binding at the variant locus. Provided that *ZNF23* is a gene downregulated in cancer and associated to inhibition of cell-cycle progression (171,172), the identified feedback loop could potentially contribute to an enhanced cellular proliferation and potentially an increased cancer risk.

Searching for additional examples of multiple variants functional relations, we studied GWAS variants associated to Alzheimer's disease and showed that absence of risk alleles for variants rs199499 and rs7207400 is necessary to sustain the transcript level of several genes (*LRRC37A*, *LRRC37A2* and *ARL17A*) in the complex genomic region 17q21.31. This region, which hosts the Alzheimer related *MAPT* gene (173,174), is known to have undergone an inversion event during evolution (175) and to be associated with abnormal tau protein deposit (176). Both rs199499 and rs7207400 variants were observed to modify the binding motif of *TAL1* gene, which is known for its effects on GABAergic neurogenesis (177). Variant rs7207400 also creates a binding motif for the TCF4 transcription factor, involved in synaptic plasticity (178), and the well-known *MYCN* gene, essential in neurogenesis.

We also found that specific rs113260531, rs7225151 and rs80257887 variants risk allele patterns reduce the transcript level of *HIST2H2AC*, a histone protein shown to be downregulated in brain blood vessels of Alzheimer's disease mouse model (179). Variants rs7225151 and rs80257887 are in moderate LD ( $r^2 = 0.6$ ) while rs113260531 is located on a different chromosome. Variants rs113260531 and rs80257887 were also observed to decrease the binding score of *RUNX3*, a transcription factor that is essential in the development and fundamental formation of axons (180), and *RREB1*, a regulator of glutamatergic axons death (181).

Overall, we have shown that Polympact represents a useful tool to explore functional annotations and properties of common genetic variants, leading not only to an effective characterization of single variants but also to an effective investigation of putative functional relations and potential interactions among multiple variants. We hence believe

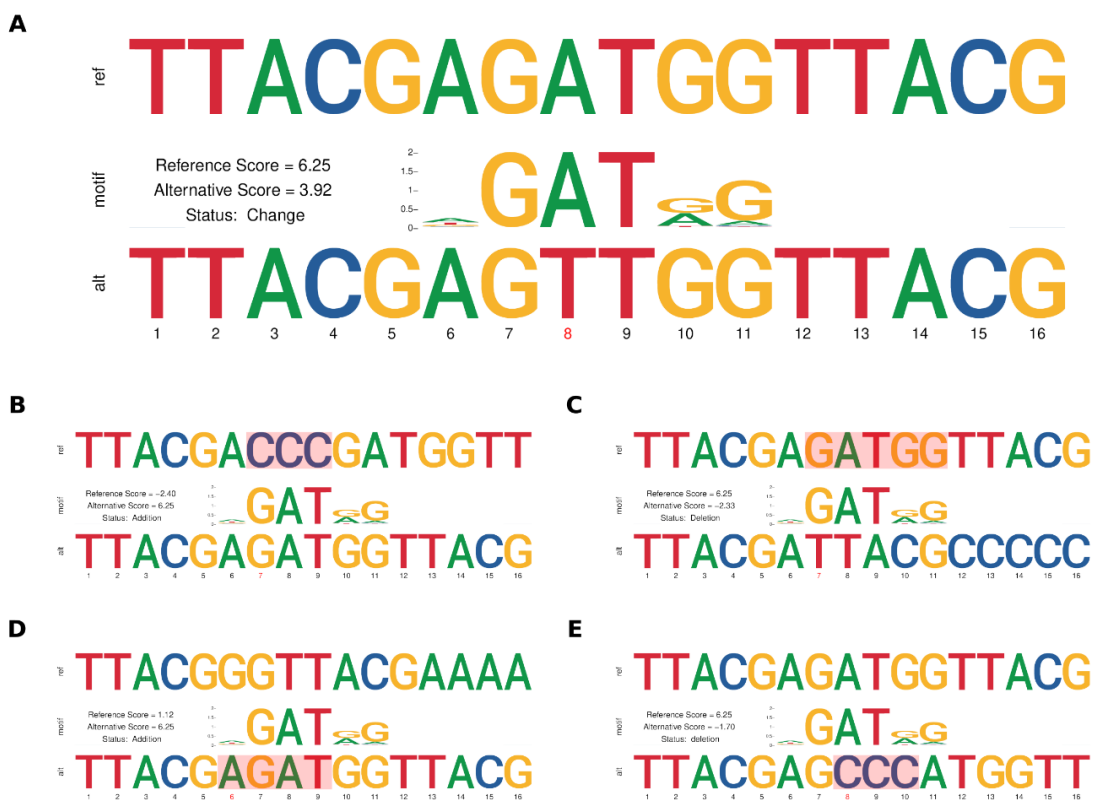
Polypact might be broadly applied and used to generate hypothesis about the biological causes of complex diseases.

## Resource Availability

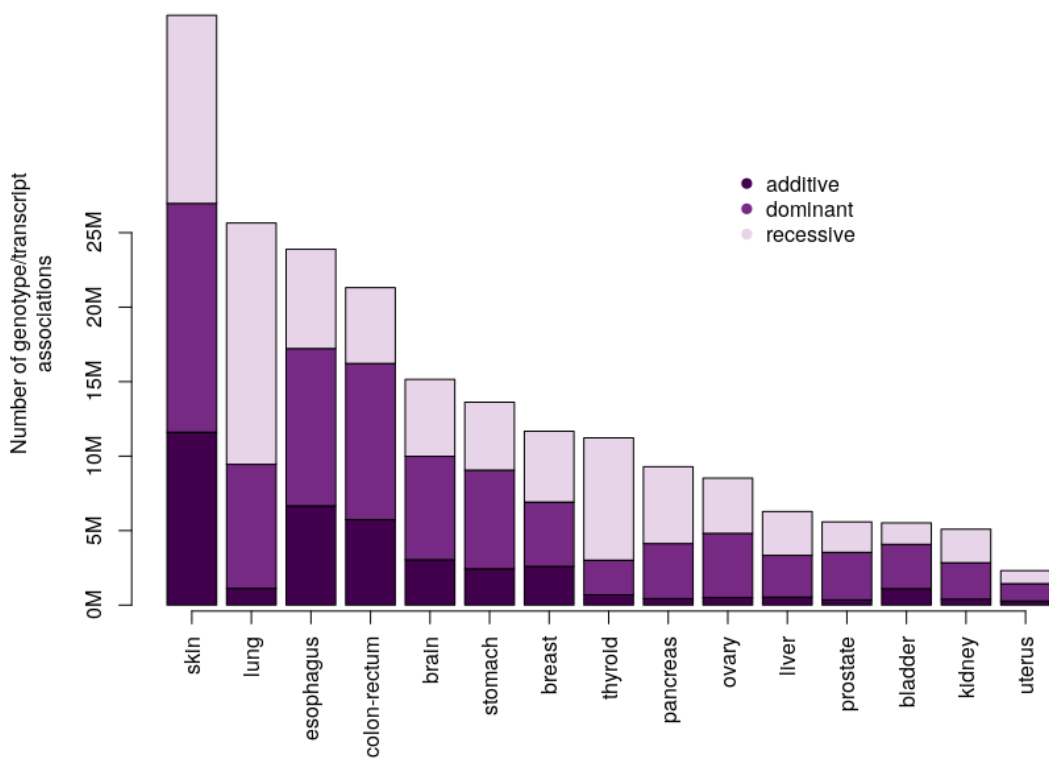
Polypact is available at [bcglab.cibio.unitn.it/polypact](http://bcglab.cibio.unitn.it/polypact)

## Supplementary Material

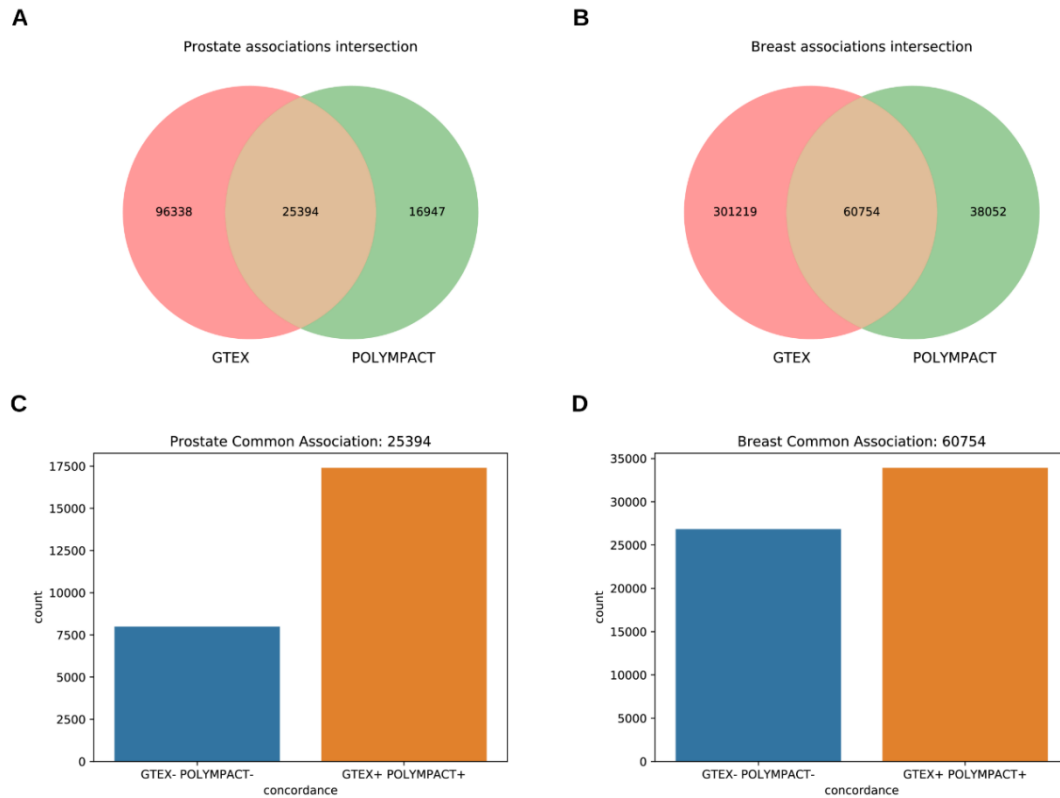
### Supplementary Figures



*Figure S3.1 A) SNP alteration of a TFBS consensus motif in position 8 annotated as change since the difference in score is greater than 10%. B) Small deletion of the red nucleotides starting from position 7 annotated as motif addition since reference score is negative and alternative score is positive. C) Small deletion of the red nucleotides starting from position 7 annotated as motif deletion since reference score is positive and alternative score is negative. D) Small insertion of the red nucleotides starting from position 6 annotated as motif addition because the motif starts inside the insertion region. E) Small insertion of the red nucleotides starting from position 8 annotated as deletion because the reference score is positive and the alternative score is negative.*




*Figure S3.2. Total number of genotype/transcript associations in various tissues stratified by the three different association models.*



*Figure S3.3. A) Venn diagram of Polympact and GTeX prostate tissue cis associations. B) Same as A) but considering breast tissue cis associations. C) Comparison of the direction of the associations in the prostate tissue data; GTeX- Polympact- represent associations with negative effect on the transcript levels in both GTeX and Polympact, while GTeX+ Polympact+ represent associations with positive effects. D) Same as C) but with breast data.*



**A**



**B**

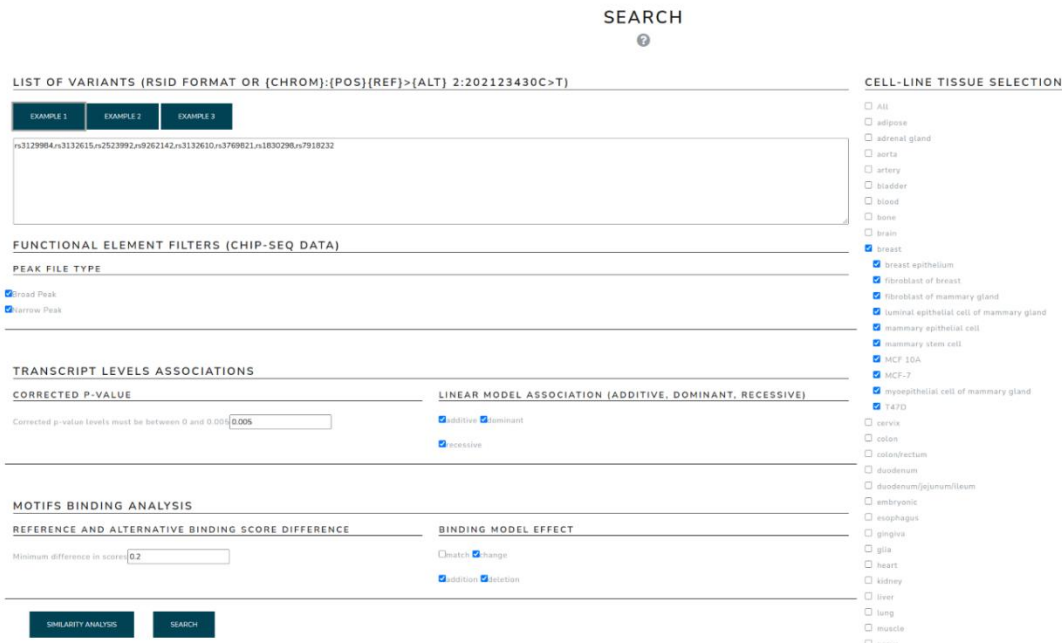


Figure S3.4. **A)** Quick search interface. The only required input are the variants provided in different formats and separated by commas. The search is performed using default parameters. **B)** Advanced search interface. The user can select specific tissues/cell lines, ChIP-seq peaks files types, transcript level association models and stringency and type of motif binding to refine the search.

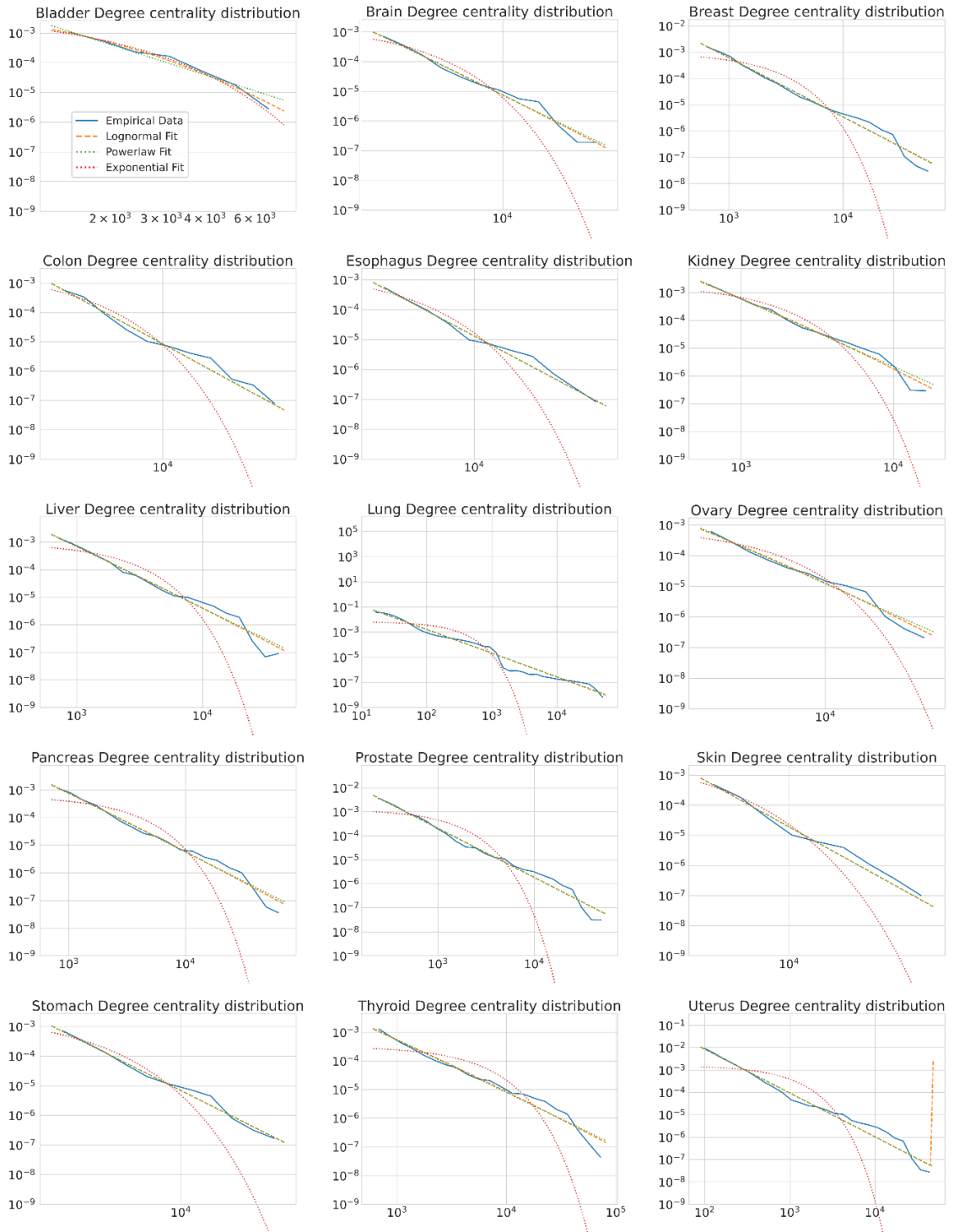


Figure S3.5. Degree centrality distributions of the variant-gene networks across all tissues analyzed in Polymapt.

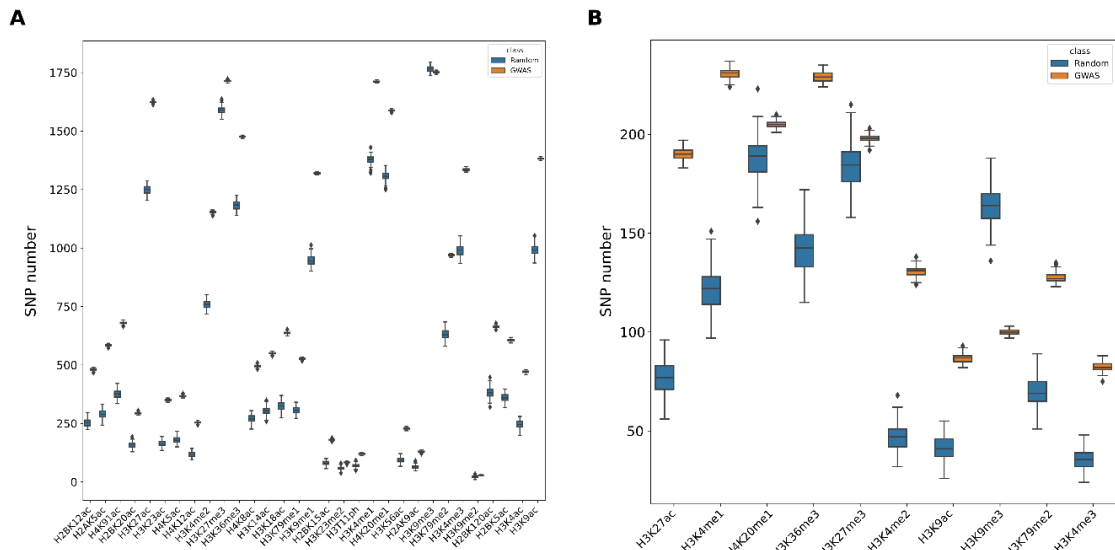


Figure S3.6. **A)** Histone marks enriched by GWAS cancer risk variants. **B)** Same as A) but considering GWAS breast cancer risk variants.

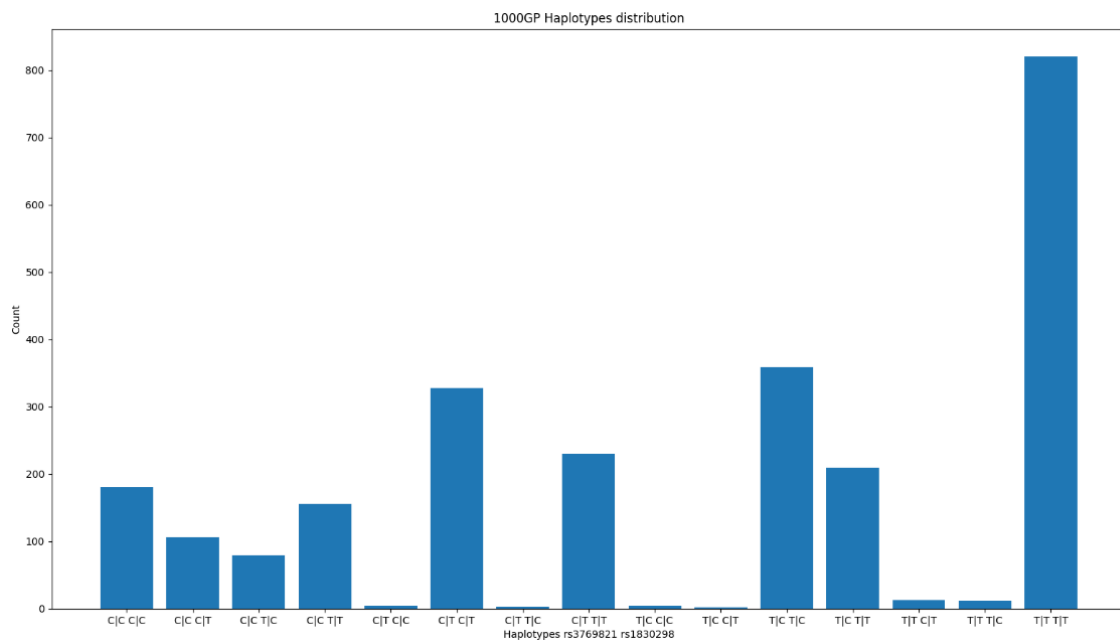


Figure S3.7. 1,000 Genomes Project genotypes for variants rs3769821 and rs1830298. Genotypes are phased. The first element in the pair refers to the two alleles of variant rs3769821 while the second element refers to the two alleles of variant rs1830298.

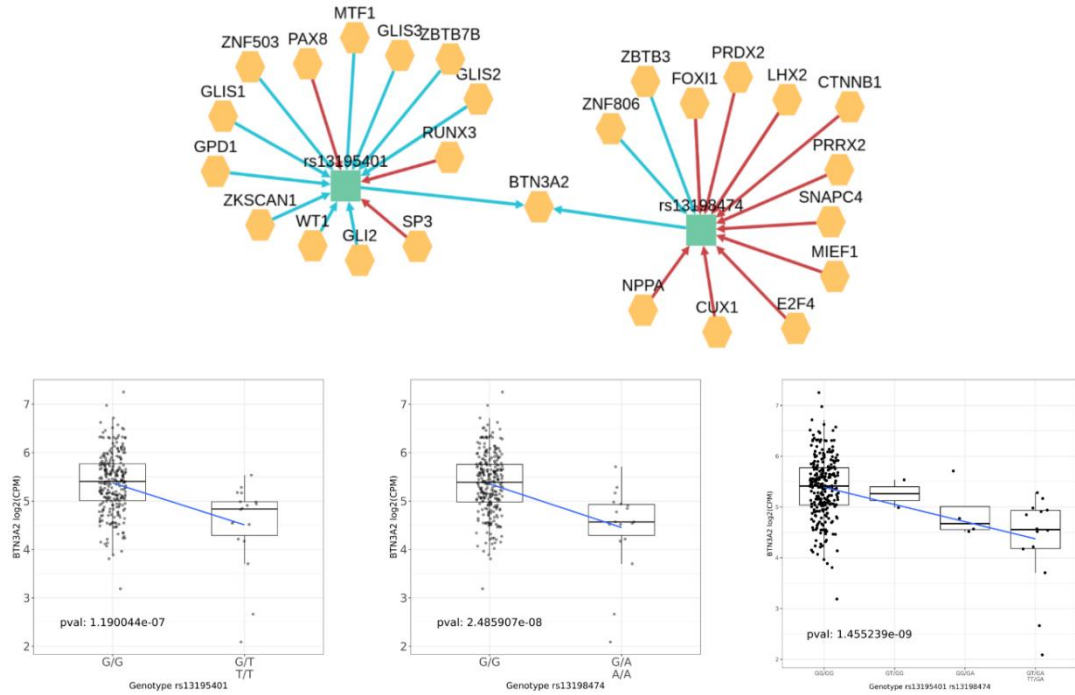


Figure S3.8. Variant-gene network (top) and effect of variants rs13195401 and rs13198474 on *BTN3A2* transcript in the breast tissue under the dominant model (bottom left and middle). Combined effect of the two variants (bottom, right) where the first boxplot contains samples with no dominant effect for both variants, the second has samples with a dominant effect for rs13195401 but no effect for rs13198474, the third has sample with a dominant effect for rs13198474 but no effect for rs13195401 and the fourth has samples with a dominant effect for both.

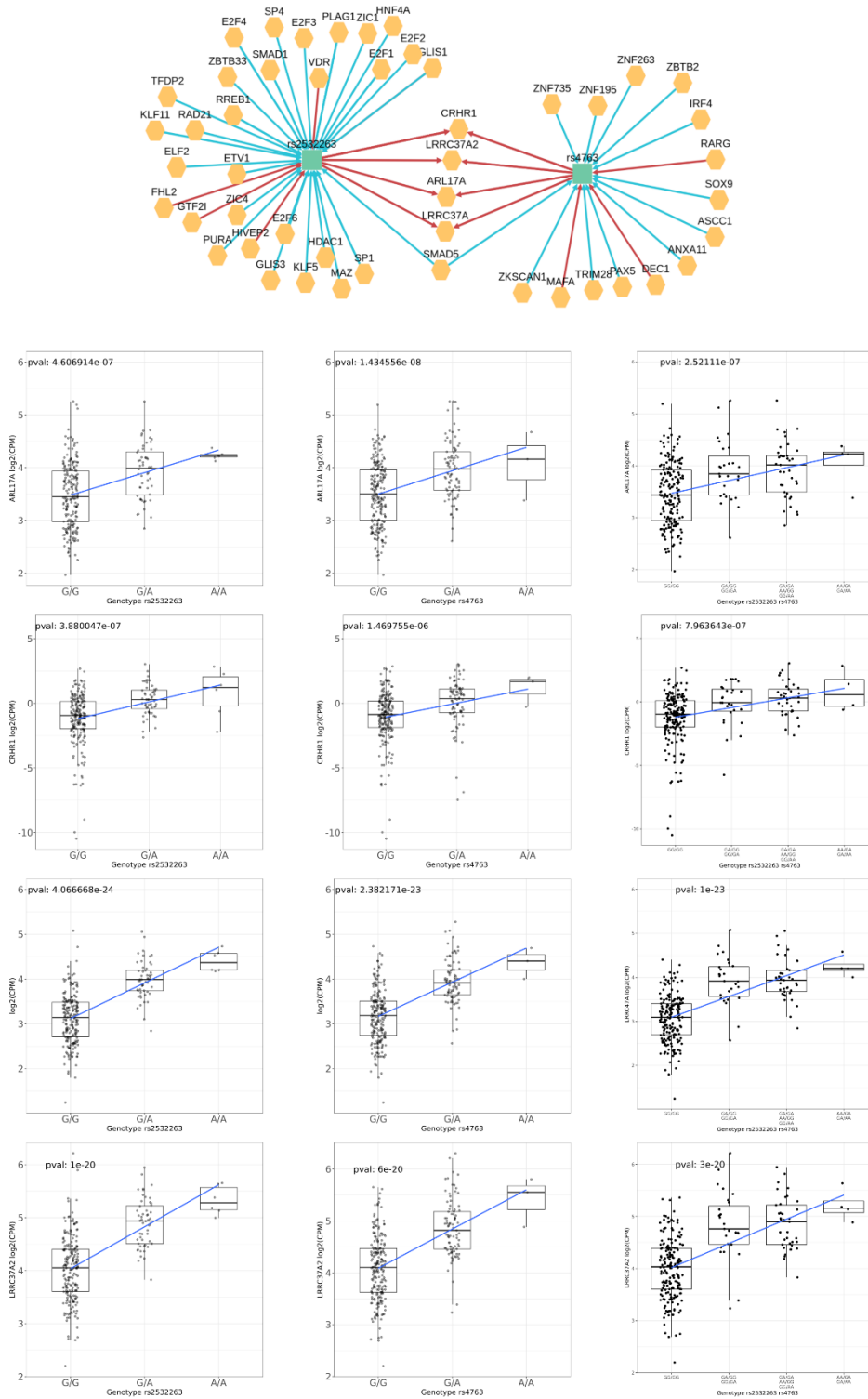
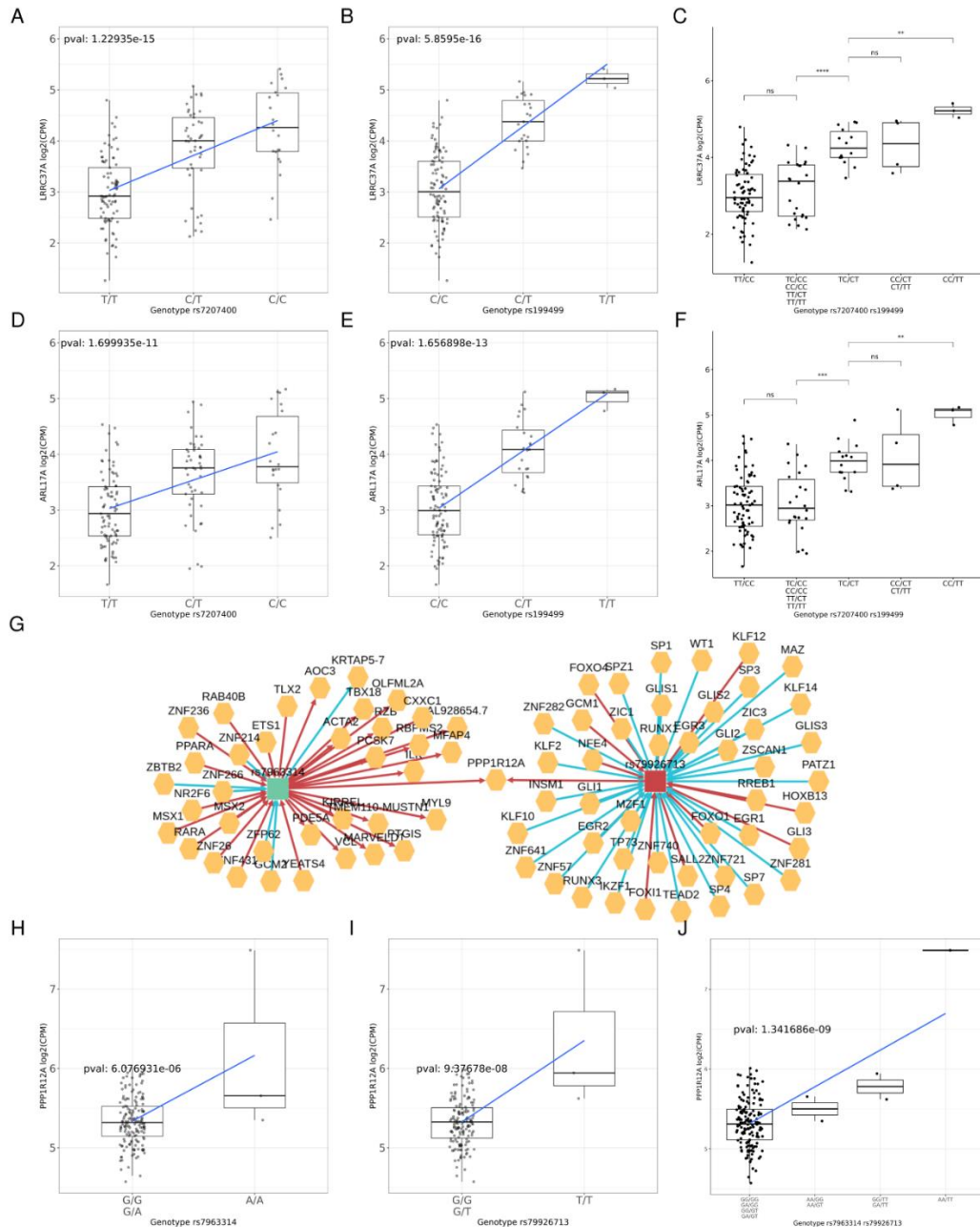
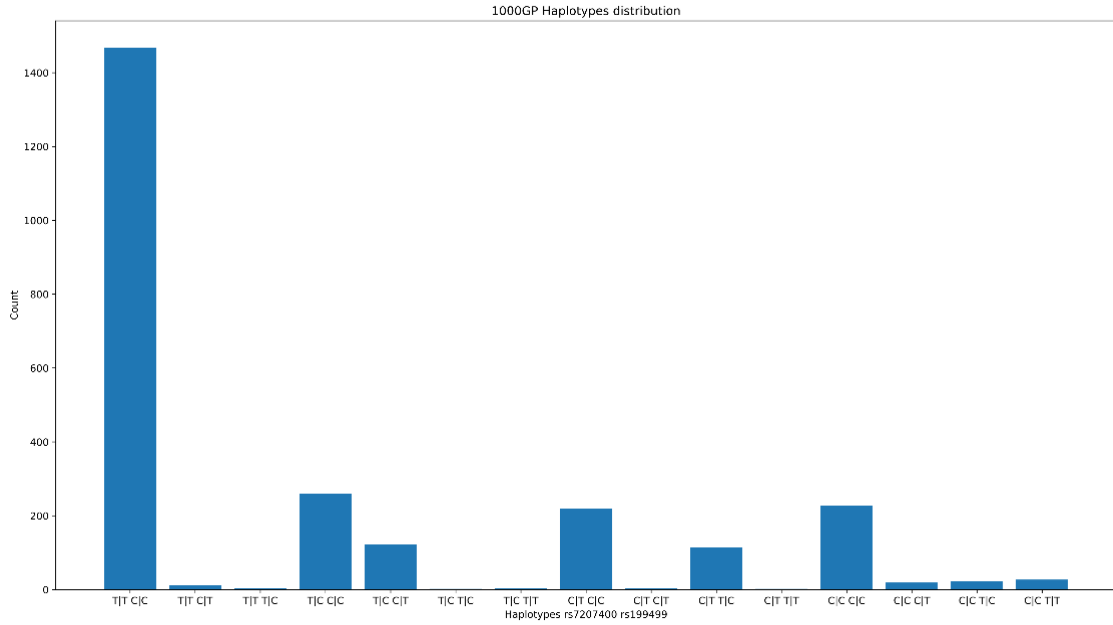


Figure S3.9. Variant-gene network of variants rs2532263 and rs4763 in breast tissue (top). Effects of rs2532263 and rs4763 on genes ARL17A, CRHR1, LRR37A and LRR37A2 under an additive model in breast tissue (bottom).



**Figure S3.10. A)** Effect of variant rs7207400 on the transcript level of LRRC37A under an additive model. **B)** Effect of variant rs199499 on the transcript level of LRRC37A under an additive model. **C)** Combined effect of the two variants rs7207400 (first pair of nucleotides in the label) and rs199499 (second pair) on LRRC37A. **D)** Effect of variant rs7207400 on the transcript level of ARL17A under an additive model. **E)** Effect of variant rs199499 on the transcript level of ARL17A under an additive model. **F)** Combined effect of the two variants rs7207400 (first pair of nucleotides in the label) and rs199499 (second pair) on ARL17A. **G)** Variant-gene network of variants rs7963314 and rs79926713 in the brain tissue. **H)** Effect of variant rs7963314 on transcript PPP1R12A under a recessive model. **I)** Effect of variant rs79926713 on transcript PPP1R12A under a recessive model. **J)** Combined effect of the two variants where the first boxplot contains samples with no recessive effect for both variants, the second has samples with a recessive effect for rs7963314 but no effect for rs79926713, the third has sample with a recessive effect for rs79926713 but no effect for rs7963314 and the fourth has samples with a recessive effect for both.



*Figure S3.11. 1,000 Genomes Project genotypes for variants rs7207400 and rs199499. Genotypes are phased. The first element in the pair refers to the two alleles of variant rs7207400 while the second element in the pair refers to the two alleles of variant rs199499.*

*Supplementary Tables*

Supplementary tables are available at:

[www.academic.oup.com/nar/article/50/3/1335/6513575#supplementary-data](http://www.academic.oup.com/nar/article/50/3/1335/6513575#supplementary-data)

# Chapter 4: Somatic mutations and rare variants in protein interfaces

## Introduction

Human genetic variants were firstly studied in Mendelian disorders, where a single variant is linked to a specific disease because it exerts a huge effect on a phenotype, an effect so large that can be selected against and tend to affect a small percentage of the population. In particular, highly penetrant variants associated to monogenic disorders tend to create stop codons or they affect the folding leading to a non-functional protein. Mendelian disorders have been fundamental to identify functions of genes but, recently, the focus of researchers have moved to polygenic complex traits usually caused by thousands of variants (182). Variants in complex traits are usually non-coding and their effect, mostly still unknown, are related to regulatory effects. Nevertheless, GWAS variants have been found recently among rare coding variants (12), starting to find some mechanistic effects and possibly drug targets.

Structures that are possibly altered by genetic variants are protein interfaces. Protein interfaces are residues of a protein where a physical interaction with other molecules happen. Protein interfaces are fundamental for the correct protein functions and their alteration can lead to disease while also offering an optimal target for drugs (183). Similarly, in cancer, non-synonymous somatic mutations have been observed to be more present on protein interfaces with respect to other parts of a protein (184).

Starting from these results we hypothesize that also non-synonymous rare germline variants can be localized in protein interfaces. BioBanks offer a tremendous opportunity to study rare variants given the required sample size to gain enough statistical power. In particular, the UKBioBank offers hundreds of thousands of samples with matching clinical and genomic data that allows researcher to explore complex hypothesis on germline variants.

In this chapter, developed during my stay at the Barcelona Supercomputing Center, we replicated the study on somatic interfaces reported in (184), using TCGA updated data,



and finally we analyzed the UKBioBank looking for enrichment of missense mutations in protein interfaces of cancer genes.

## Results

### Identification of somatic mutations in protein interfaces in TCGA

We replicated and extended the study reported in (184) where the authors analyzed the whole TCGA dataset to detect if the somatic mutations are randomly distributed across proteins or if they are located in functional regions. Starting from that results we replicated the enrichment analysis in protein interfaces using the new public TCGA mutations published in (185). (**Table 4.1**).

Cancer	#Samples	Missense Mutations	Genes Enriched	Interfaces Enriched
ACC	91	5588	0	0
BLCA	412	78906	11	51
BRCA	1087	60059	3	11
CESC	306	44278	4	9
CHOL	36	1868	1	7
COAD	445	124526	7	44
DLBC	50	3681	0	0
ESCA	183	19716	1	16
GBM	599	36364	3	21
HNSC	523	63002	9	47
KICH	65	1625	0	0
KIRC	518	14811	0	0
KIRP	286	16189	0	0
LAML	200	3742	1	6
LGG	514	22300	4	17
LIHC	374	28939	0	0
LUAD	575	127120	5	20
LUSC	490	103312	3	15
MESO	87	1803	0	0
OV	600	26422	2	15
PAAD	184	16552	2	12
PCPG	178	1467	1	8

PRAD	495	17619	3	15
READ	159	33372	5	27
SARC	257	13749	1	1
SKCM	470	275344	8	27
STAD	440	114951	6	20
TGCT	149	1658	2	3
THCA	503	5915	2	18
THYM	124	1923	1	7
UCEC	543	412839	6	26
UCS	57	5409	4	24
UVM	80	1057	2	7
PANCAN	11080	1704921	35	130

*Table 4.1 TCGA datasets samples, number of mutations, genes and interfaces enriched by dataset.*

The dataset showing the highest enrichment in both genes and interfaces in TCGA is the pan-cancer (including all tumor types) dataset having the largest number of samples and mutations and thus the greatest statistical power. The smallest datasets in terms of samples and mutations did not show enrichment apart from UCS and UVM with 4 and 2 genes enriched for mutations, respectively. Surprisingly, large datasets like KIRC and KIRP did not show any enrichment in interfaces.

In TCGA we found enriched many interface types (**Figure 4.1**). In particular, the pan-cancer dataset shows the greatest number of enriched genes and interfaces mutated while some cancers, like UVM, show only a small number of genes enriched. TP53 is the most common enriched gene across the datasets and usually its protein, nucleic and ligands interfaces are enriched for mutations. Also, the RAS gene family is recurrently mutated in their ligand and protein interfaces across most tissues. Some tissues show specific enrichments in mutations like SPOP in prostate cancer or GNA11/GNAQ in Uveal Melanoma which are genes that are recurrently mutated in those cancers.

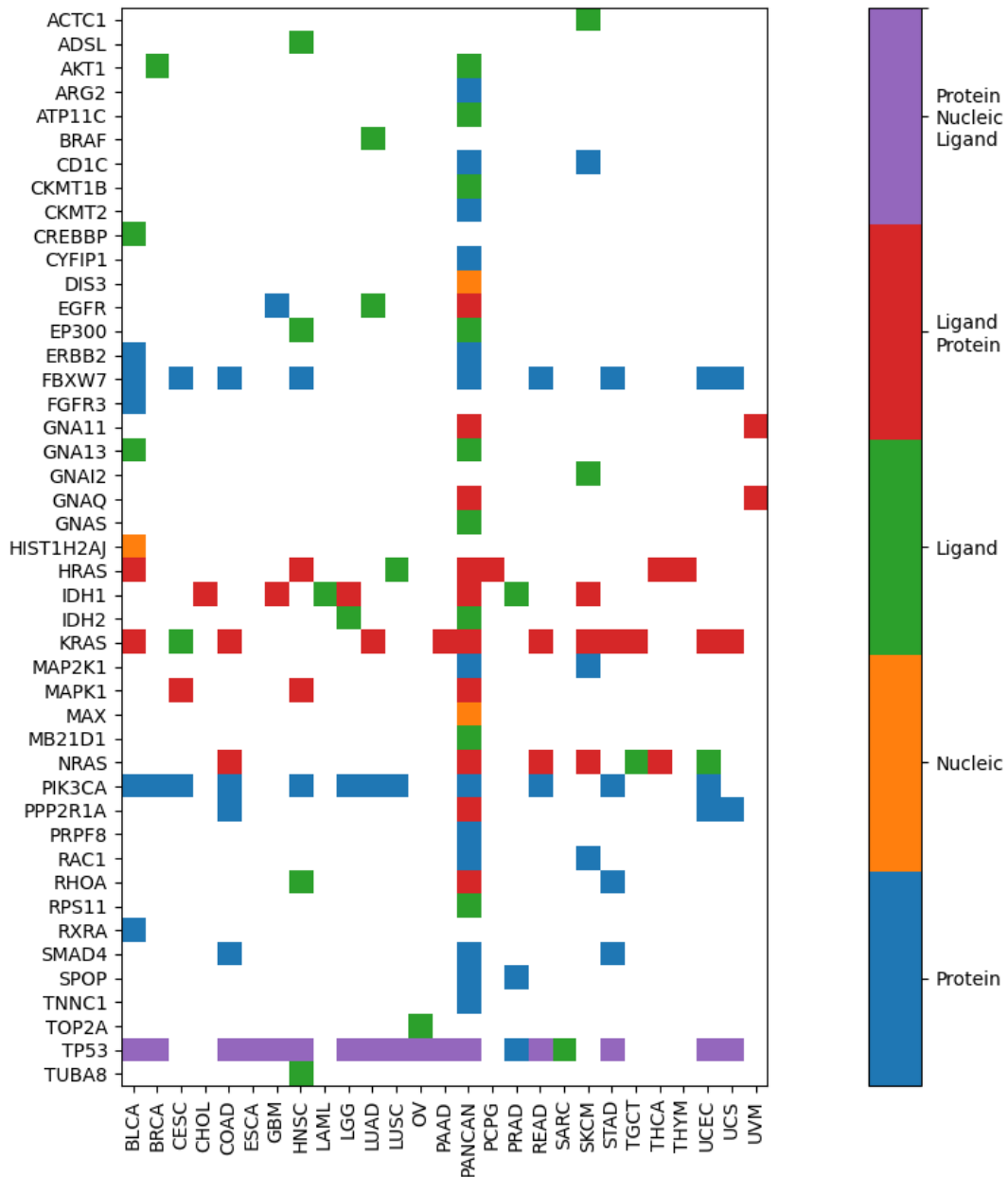


Figure 4.1 TCGA interfaces enrichment across the different cancers types. Colors represent if a protein interface, a nucleic acid interface or a ligand interface of a gene are enriched for mutations in the tissue.

## cBioPortal enrichment

We then extended the analysis to the cBioPortal datasets by selecting tissue specific cancers Whole-Exome Sequencing studies (**Table 4.2**).

Study	Cancer	#Samples	Missense Mutations	Genes Enriched	Interfaces Enriched
all_phase2_target_2018_pub	ALL	145	580	6	25
aml_ohsu_2018	AML	570	6943	13	71
aml_target_2018_pub	AML	148	447	2	20
ampca_bcm_2016	AMPCA	152	14679	4	20
bcc_unige_2016	BCC	292	118002	4	13
brca_broad	BRCA	103	2821	1	3
brca_igr_2015	BRCA	211	12166	3	25
brca_mbcproject_wagle_2017	BRCA	237	13391	4	23
brca_sanger	BRCA	100	4123	1	1
brca_smc_2018	BRCA	185	5273	2	4
ccrcc_utokyo_2013	CCRCC	106	3940	0	0
chol_icgc_2017	CHOL	393	3402	6	32
cll_broad_2015	CLL	535	7232	6	29
cll_iuopa_2015	CLL	506	5474	0	0
clll_icgc_2011	CLLSLL	105	836	0	0
coad_cptac_2019	COAD	106	38306	3	6
coadread_dfc_2016	COAD	619	159037	7	49
difg_glass_2019	DIFG	444	77376	6	25
dlbcl_dfc_2018	DLBCL	135	12018	2	2
dlbcl_duke_2017	DLBCL	954	5315	9	28
es_iocurie_2014	ES	106	629	0	0
esca_broad	ESCA	146	15361	2	16
hcc_inserm_fr_2015	HCC	240	15369	0	0
hccihch_pku_2019	HCC	171	10297	3	11
ihch_smmu_2014	IHCH	103	4947	6	20
lcll_broad_2013	LCLL	157	1788	1	1
lihc_amc_prv	LIHC	231	16242	0	0
luad_broad	LUAD	181	38010	3	12
luad_oncosg_2020	LUAD	302	24214	3	9

mbl_dkfz_2017	MBL	384	1398	6	13
mel_dfcj_2019	MEL	144	69645	4	18
metastatic_solid_tumors_mich_2017	MIXED	499	60197	7	40
mixed_allen_2018	MIXED	249	91127	5	22
mixed_pipseq_2017	MIXED	99	18333	10	30
mm_broad	MM	204	6592	4	21
mpcproject_broad_2021	MPCPROJECT	82	6584	1	6
mpn_cimr_2013	MPN	146	859	2	7
nbl_broad_2013	NBL	227	3097	0	0
nbl_target_2018_pub	NBL	117	163	0	0
nccrcc_genentech_2014	NCCRCC	138	3370	0	0
nepc_wcm_2016	NEPC	114	6101	4	12
paad_qcmg_uq_2016	PAAD	377	12260	2	11
paad_utsu_2015	PAAD	109	5669	2	5
pediatric_dkfz_2017	PEDIATRIC	706	9253	7	46
pptc_2019	PPTC	240	34940	5	15
prad_broad	PRAD	112	3467	1	7
prad_fhcr	PRAD	141	11590	28	77
prad_p1000	PRAD	1011	47224	5	28
prad_su2c_2015	PRAD	150	13202	3	10
prad_su2c_2019	PRAD	421	29885	3	15
prostate_dkfz_2018	PRAD	308	7176	3	13
sclc_ucologne_2015	SCLC	110	21520	1	1
skcm_broad	SKCM	121	47380	2	12
skcm_dfcj_2015	SKCM	110	44828	1	10
stad_oncosg_2018	STAD	147	18175	2	7
stad_pfizer_uhongkong	STAD	100	11969	1	4
wt_target_2018_pub	WT	99	554	0	0

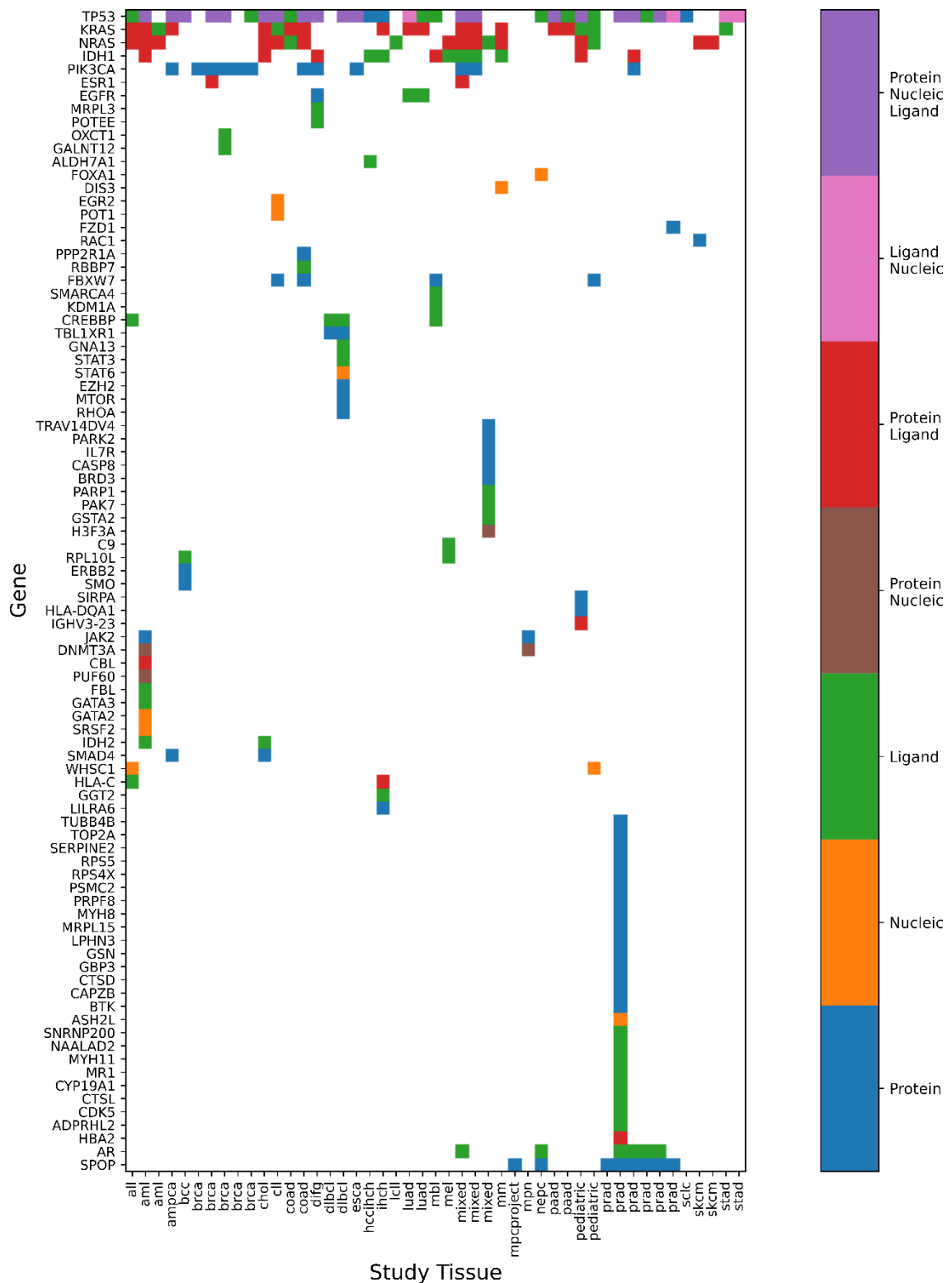
*Table 4.2 cBioPortal datasets samples, number of mutations, genes and interfaces enriched by dataset.*

The analysis allowed us to discover more tissue specific associations (**Figure 4.2**). Out of 57 studies we found enrichments in 47 of them. In particular, TP53 interfaces are recurrently mutated in almost all the datasets together with KRAS and NRAS, similarly to the TCGA dataset. Other frequently mutated genes are IDH1 and PIK3CA. In particular, PIK3CA is mostly enriched in protein interfaces in breast cancer datasets. A similar pattern can be observed for AR and SPOP that are enriched in most prostate adenocarcinoma studies (AR in mixed primary/metastatic and only metastatic cancer datasets). Other interesting patterns are related to the gene EGFR recurrently mutated in Lung Adenocarcinoma and in the single dataset of Diffuse Glioma. Other common enriched mutations are CREBBP and TBL1XR1 in Diffuse Large B Cell Lymphoma. Also, WHSC1 is recurrently mutated in Acute Lymphoblastic Leukemia, typically a pediatric cancer, where its DNA interface is the only one enriched in the pediatric pan-cancer study. Examples of interfaces enriched for somatic mutations are reported in **Table 4.3**.

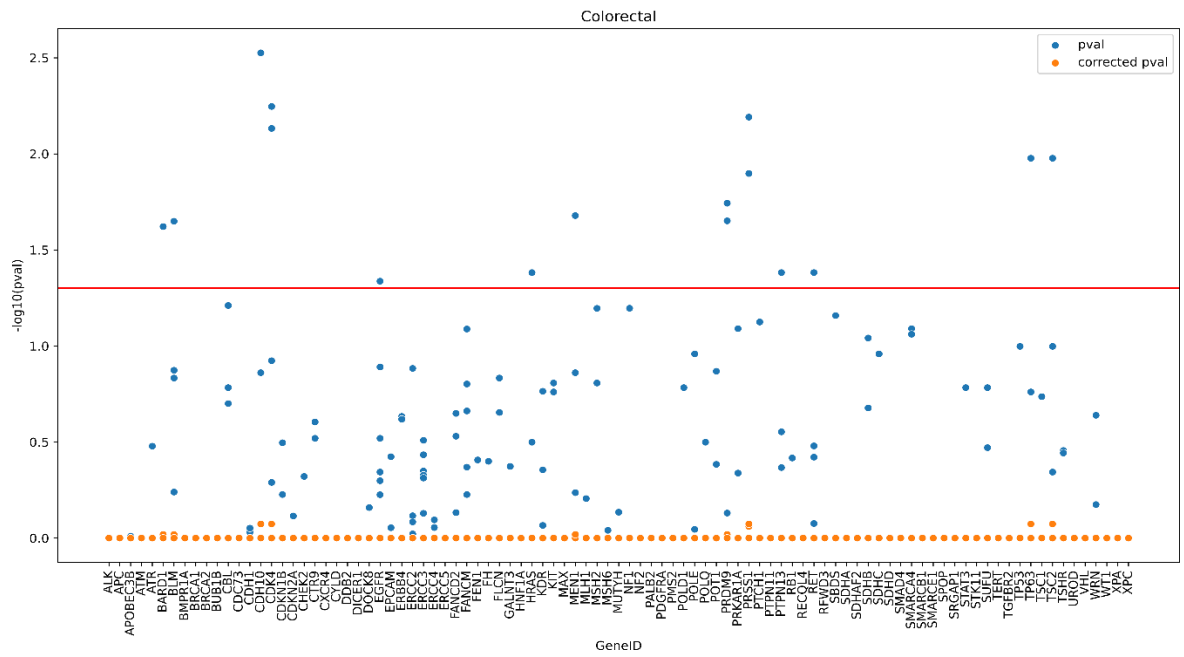
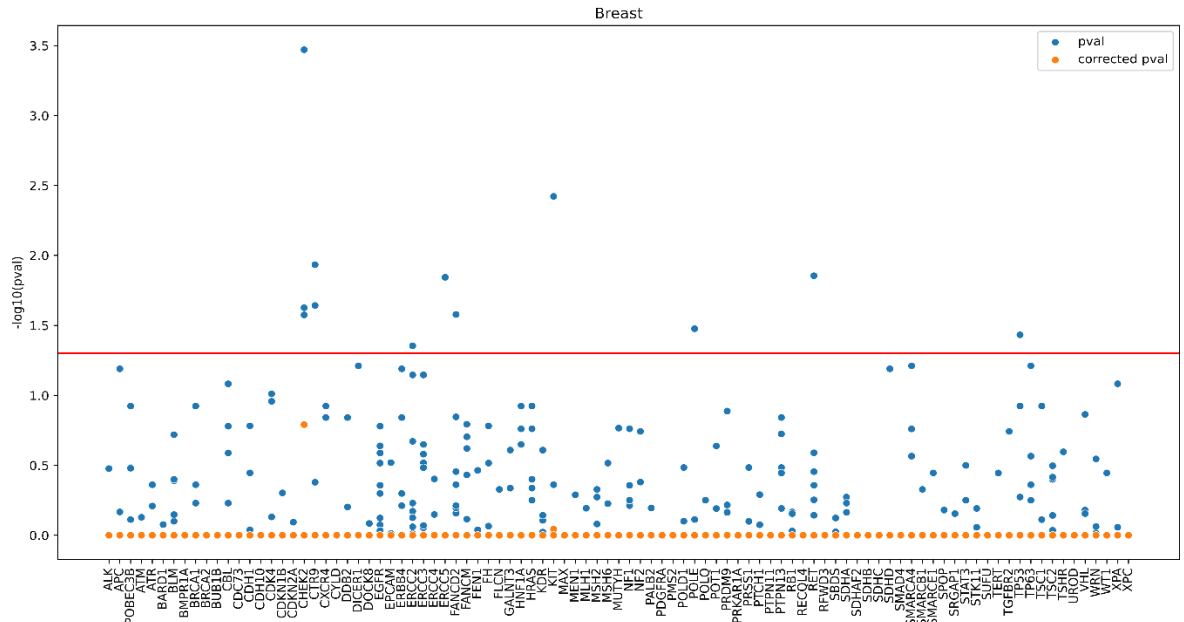
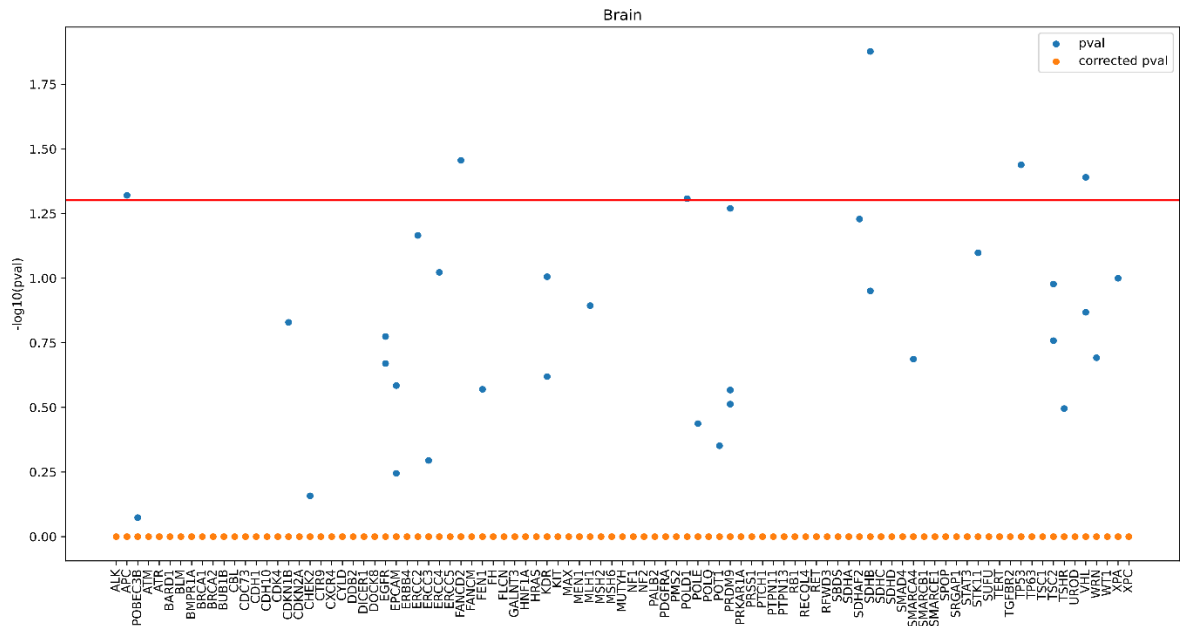
### Rare variants in cancer genes interface enrichment

Provided the enrichment of cancer mutations we found, we started to hypothesize that germline variants can have a similar effect on cancer genes.

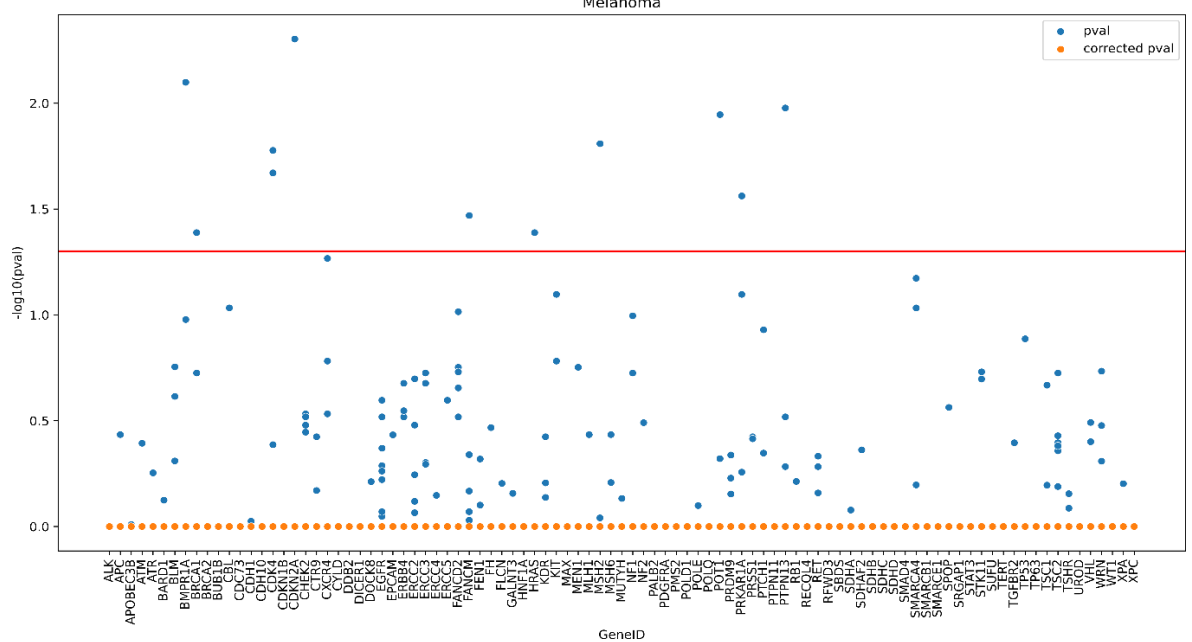
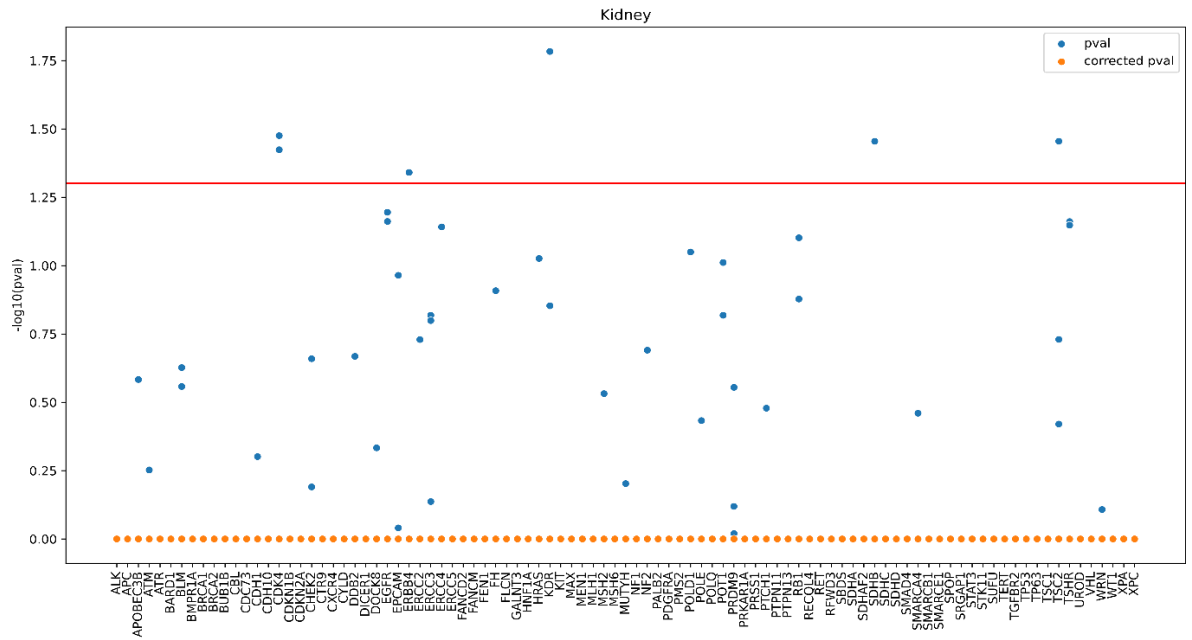
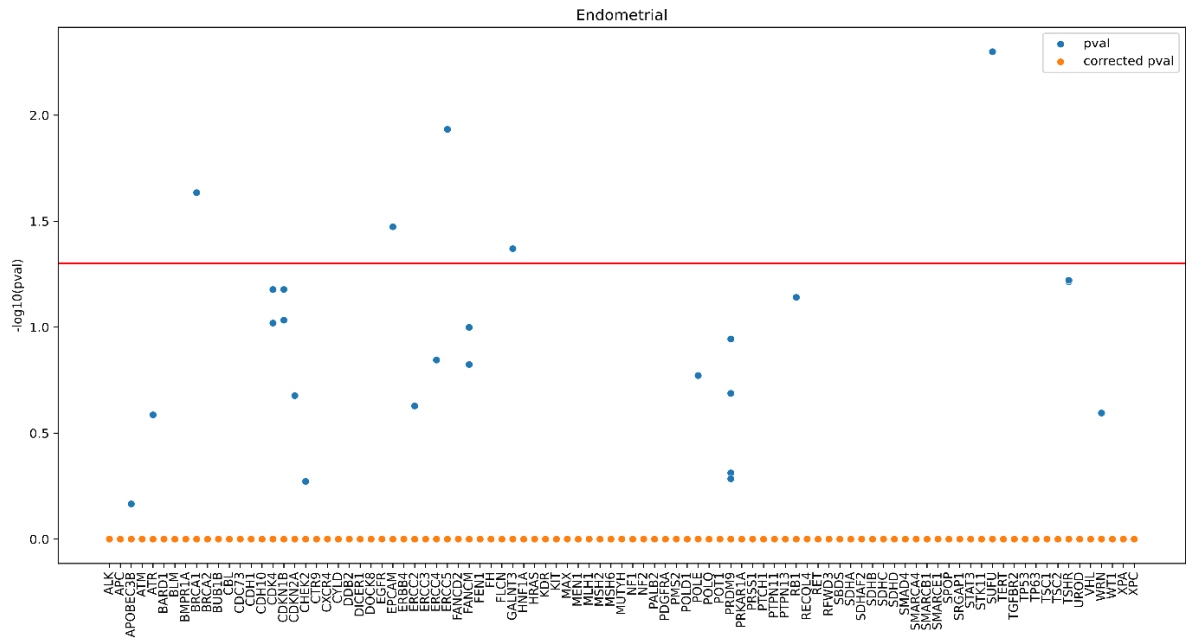
The UKBioBank project offers a unique opportunity to explore variant-phenotype associations on rare variants given the huge sample sizes. We retrieved 200,643 Whole-Exome Sequencing genotype data with associated clinical data from the UKBioBank and we analyzed if rare germline variants are more present in protein interfaces with respect to other positions. We analyzed 8 tissues: Brain, Breast, Colorectal, Endometrial, Kidney, Melanoma, Prostate and Thyroid (**Figure 4.3**). We also performed a pan-cancer analysis (**Figure 4.4**).

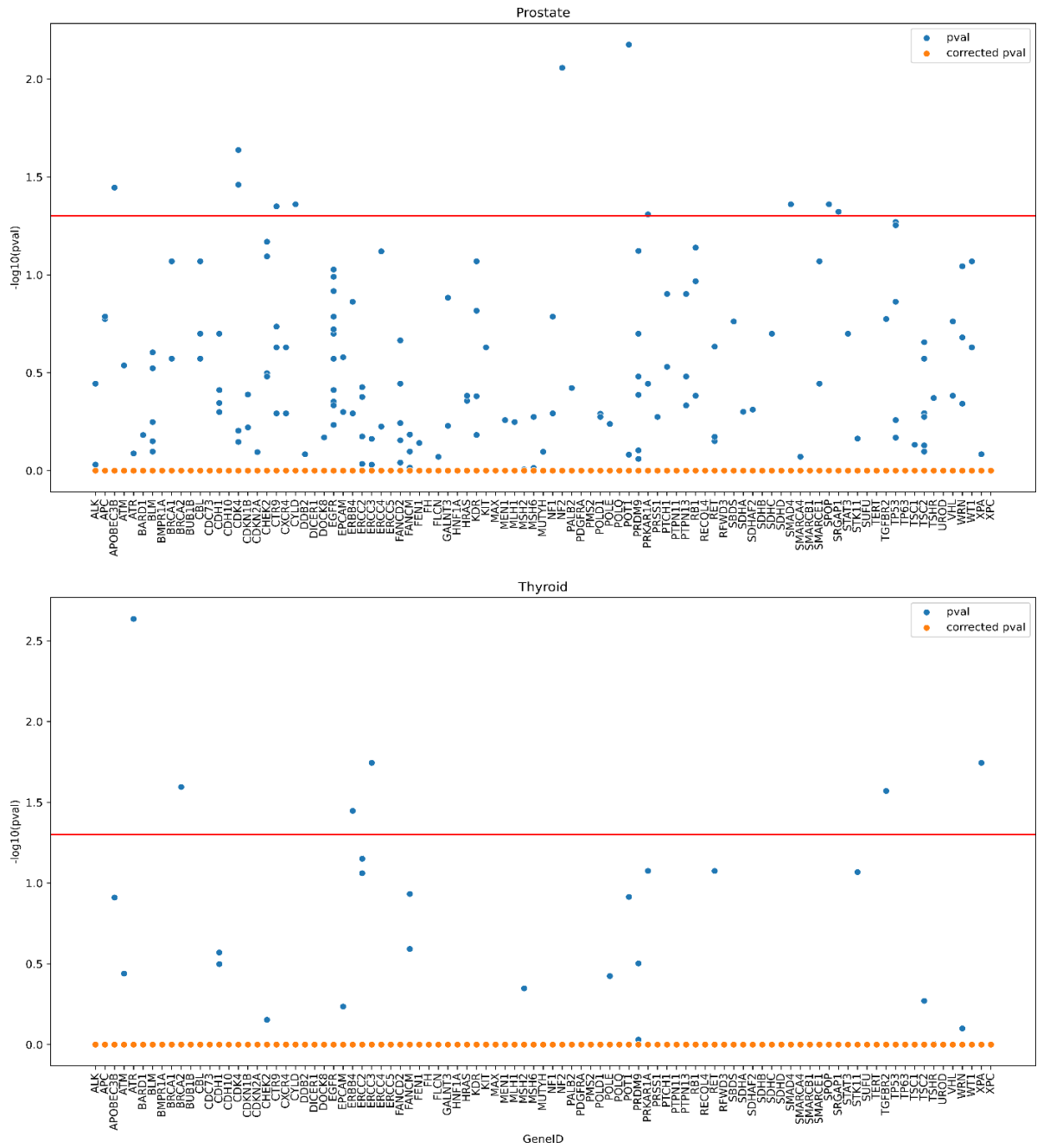


**Figure 4.2** *Enrichment of protein interfaces across cBioPortal studies. Genes are clustered using seriation clustering while studies are grouped by tissue. Colors represent if a protein interface, a nucleic acid interface or a ligand interface of a gene are enriched for mutations in the tissue*









**Figure 4.3 Enrichments of rare germline variants in protein interfaces of cancer genes across the Brain, Breast, Colorectal, Endometrial, Kidney, Melanoma, Prostate and Thyroid tissues. The red line shows a significance threshold of 0.05.**



Gene	Interface Residues	Protein Domain	Cancers
<i>SPOP</i>	70/76/77/80/83/87/102/115/116/117/119/123/129/130/131/ /132/133/134/135/136/137/138/141	MATH Domain	Prostate Cancer
<i>AR</i>	702/705/706/708/709/710/712/742/743/746/747/750/753/ 765/781/788/874/877/878/881/892/896/900	Nuclear Receptor Ligand Binding	Prostate Cancer
<i>GNA11</i>	187/188/189/189/191/204/208/209/210/211/212/214/215/ 216/218/219/220/221/263	GTP-Alpha	Uveal Melanoma
<i>GNAQ</i>	77/78/81/85/92/117/119/184/185/186/187/188/189/190/ 192/209/211/212/214/215/218/240/241/242/243	GTP-Alpha	Uveal Melanoma
<i>PIK3CA</i>	5/6/11/23/25/26/27/28/29/30/31/57/58/60/61/71/72/73/74/ 75/77/78/79/95/96/98/100/343/345/346/347/349/357/364/ 365/367/368/369/409/410/412/421/422/447/448/449/450/ 452/453/454/455/467/469/510/511/512/513/542/543/544/ 545/546/549/573/678/843/1017/1029/	Multiple	Breast Cancer
<i>EGFR</i>	718/719/720/726/743/744/745/766/775/776/777/788/789/790/ 791/792/793/794/796/797/844/853/854/855/856/858/997/1001	Kinase Domain	Lung Cancer
<i>EGFR</i>	110/217/218/219/220/227/228/229/233/234/254/263/270/ 273/274/275/276/277/286/287/288/289/299/302/304/306/ 307/308/309/310	Extracellular domain	Gliomas
<i>CREBBP</i>	1410/1431/1432/1433/1434/1435/1436/1443/1446/1447/ 1450/1471/1472/1473/1474/1475/1476/1480/1482/1487/ 1491/1492/1493/1494/1495/1498/1499/1502/1503/1542	HAT Domain	DLBCL
<i>TBL1XR1</i>	171/173/229/245/270/312/313/351/353/354/369/395/ 420/446/462/487/	Multiple	DLBCL
<i>WHSC1</i>	998/1099/1124/1152	SET	Pediatric
<i>CDK4</i>	41/42/45/46/47/48/49/50/51/53/54/55/57/58/59/61/ 62/64/65/66/75/76/77/79/87/89	Protein kinase	Germline Pancancer
<i>TP53</i>	94/95/96/166/167/170/171/172/174/175/176/177/ 180/210/211/212/213/244/245/249	DNA Binding	Germline Pancancer/ Somatic Pancancer

*Table 4.3 Examples of interfaces recurrently mutated in different cancer types or with enrichments in rare germline variants*

## Material and methods

### Protein Interfaces

Protein interfaces have been computed using 3DMapper (<https://www.github.com/vicruiser/3Dmapper>) and were already available when I joined the Barcelona Supercomputing Center for my stay.

3Dmapper is a novel tool that maps variants or positions to protein structures using an annotated vcf/maf file and protein structures in PDB format. 3Dmapper creates a structural database by aligning a target proteome against all the PDB files provided by the user using BLAST. This step is necessary to identify the relative positions between the protein sequences and each PDB file. Also, thanks to the sequence alignment, it allows for to detection of proteins that do not have an available structure but share sequence similarity. Finally, for each PDB with at least one BLAST hit, inter-protein interfaces are computed. For each pair of chains in a PDB file, a residue is considered part of an interface if their heavy atoms are closer than 5Å using the Euclidian distance. Ligands that are considered experimental artifacts are not considered.

3DMapper also provides an efficient and computational scalable tool that allows mapping the position of a variant on a protein structure.

TCGA and cBioPortal interfaces have been obtained by aligning the PDB chain sequence to the sequence from Uniprot (release 2021-04) while the UKBioBank interfaces have been aligned against the Ensembl sequence (release 104).

### TCGA dataset

The TCGA mutation data have been retrieved from (185). We retrieved data for about 33 tumors and 11,080 patients for a total of 3,600,963 mutations. Mutations were filtered keeping only missense mutations for a total of 1,704,921 missense mutations.

## cBioPortal dataset

cBioPortal mutational datasets were retrieved from (186) and were manually curated and filtered to include only Whole-Exome sequencing experiments with at least 50 samples. We retrieved 57 studies for a total of 14,348 patients and 1,204,776 missense mutations.

## Enrichment in protein interfaces and multiple test correction

We started from the null hypothesis that the mutations inside a protein are randomly distributed. Then for each interface we performed a binomial test:

$$\binom{n}{k} p^k (1-p)^{n-k}$$

where  $n$  is the total number of mutations,  $k$  are the mutations happening inside an interface and,  $p$  is estimated by the ratio between the length in number of residues of the interface and the length of the protein.

For each tumor we corrected resulting p-values using a Benjamini-Hochberg correction and we selected the interfaces having a corrected p-value smaller than 0.01.

## Interface filtering for germline analysis

We retrieved 1,031 cancer genes with germline signal from COSMIC (187) and we merged them with genes retrieved from (188) obtaining 1,080 unique transcripts. Of those, we selected the canonical interface using BioMart, obtaining 858 interfaces across 95 proteins. Then, for each pair of interfaces belonging to the same protein, we computed the overlapping index and if the overlap was greater than 0.5, we kept the longest interface in a greedy manner until no pair of interfaces has an overlap index greater than 0.5, obtaining 471 non-overlapping interfaces.

## UKBioBank dataset

The UKBioBank Whole-Exome Sequencing dataset was retrieved from the UKBioBank(189), obtaining 200,643 whole-exomes with matching clinical data. We analyzed 8 tumor types: Brain, Breast, Colorectal, Endometrial, Kidney, Melanoma,

Prostate and Thyroid. Also, an analysis across all cancers was performed integrating all cancer types available in the UKBioBank. Variants inside the 95 genes were extracted from the UKBioBank variant files using PLINK (152) and, finally, variants were annotated using VEP (190). We obtained a total of 36,474 rare missense variants.

## Proportion tests

For each interface we counted the number of patients with cancer and at least one variant in the interface, patients without cancer and at least one variant in the interface, patients with cancer and no variants in the interface and finally patients without cancer nor mutation. We obtained the following contingency table:

Cancer Patient and Mutation in interface	Cancer Patient and no mutations in interface
Non cancer and Mutation in interface	No cancer and no mutations in interface

We performed a one tailed Fisher Exact Test and finally we corrected the p-values obtained using Benjamini-Hochberg method considering each interface in a tested tissue as a separate hypothesis.

## Discussion

Coding variants were the first variants identified as causal in monogenic diseases and evidence that coding variants can influence the risk of disease was available even before the creation of the human reference genome. During the years many coding variants have been identified, like variants in the BRCA1 gene (191) in breast and ovarian cancers and variants in the APOE gene in Alzheimer's disease (192). Usually, coding variants have larger effects on phenotypes and risk of diseases with respect to non-coding variants identified by Genome-Wide Associations Studies. Rare coding variants are still understudied since they require an enormous sample size to be correctly characterized and only recently, thanks to biobanks, the firsts GWAS on rare coding variants was performed (12). However, even with strong associations and a clear effect on a gene, clear mechanistic links between the variants and the phenotypes is still missing. Missense variants can be indeed very difficult to characterize since they can alter the 3D structure of a protein, impacting protein folding or protein binding and making the prediction extremely complicate.

Studies on somatic mutations, showed that they are not distributed at random in proteins but they tend to affect protein interfaces (184). The first part of this chapter was dedicated to replicate the study on a more recent TCGA dataset version and on additional cancers available from cBioPortal. The study replicated most of the results, in particular we confirmed that most of the mutations that are recurrent in cancer are not happening at random, but they tend to be localized in protein interfaces. This is particularly evident in TP53 and the RAS family. Interestingly, some genes recurrently mutated in specific cancer types have enrichment in interfaces in their specific tissues, like SPOP in prostate cancer. SPOP is one of the most recurrent mutated genes in prostate cancer and studies found that it's MATH domain is the most mutated one (193). Also, GNA11 and GNAQ in Uveal Melanoma show hotspots of mutations related to a GTP binding domain (194). Mutational hotspots for SPOP, GNA11 and GNAQ are highlighted in our results.

In the cBioPortal study we found that PIK3CA protein interfaces are recurrently mutated in breast cancer. This gene is recurrently mutated, has been extensively studied



molecularly and its mutation hotspot in positions 542 and 545 in its helical domain of the catalytic subunit has been predicted to be pathogenic (195) and we show that both hotspots are located in a protein interfaces.

Similar to the TCGA dataset in prostate cancer, SPOP shows recurrent mutations and, in most datasets, also AR interfaces are recurrently mutated in more advanced cancers (196) strengthening our results.

The Epidermal Growth Factor Receptor (*EGFR*) is a recurrently mutated gene in many tumors (197). We found it enriched in non-smoker lung adenocarcinoma and its mutations were found primarily located in the tyrosine kinase domain (198). We found it is also enriched in gliomas, in particular in its extracellular domain (199). Interestingly, the two tumors are mutated in different domains and interfaces, suggesting that *EGFR* may activate different oncogenic mechanisms in different tumors. As seen in **Table 4.3** we found interfaces in both domain indicating that, probably, those mutations are altering important bindings in *EGFR* but in different domains.

Other recurrently mutated interfaces belong to CREBBP and TBL1XR1 proteins in the Diffuse Large B Cell Lymphoma (DLBCL). DLBCL is a rare lymphoma that cause the B lymphocyte to grow abnormally. CREBBP is recurrently mutated in DLBCL, and a mutational hotspot was found in an interface in the HAT domain. This hotspot has already been described but its effects are still unclear nevertheless CREBBP mutations are believed to be a main driver of DLBCL (200). TBL1XR1 is also a major driver of DLBCL and it has been shown that its missense mutations change the immune system into producing more B cells (201).

We found enriched for mutations an interface located in the SET domain of the gene WHSC1 in the Acute Lymphoblastic Leukemia, a common pediatric tumor, and in a mixed pediatric cancer dataset. This protein mutation has been shown to alter the chromatin methylation by enhancing the methyltransferase activity and the recurrent mutation of residue 1099 (located in our interface **Table 4.3**) is believed to be a driver of Acute Lymphoblastic Leukemia (202).

In the first part of this chapter, we showed that cancer mutations are not randomly distributed across proteins and, in fact, many recurrent mutated proteins in specific tissues shows an altered interface. These results can help in identifying specific altered molecular mechanisms in cancers and therapeutic approaches.

In the second part of this chapter, we explored the rare variant landscape in the UKBioBank. Here, we found that, after multiple test correction, only two partially overlapping interfaces show an enrichment in the pan-cancer dataset while none resulted significant in the single tissue studies. The enriched interface is related to the binding between CDK4 and CCND3, two genes involved in the cell cycle progression. Taking a less stringent threshold of nominal p-values  $< 0.05$  in the pancancer dataset, we found an interface of TP53 that is both enriched for the presence of rare inherited variants and somatic mutations in the TCGA pancancer analysis. This result suggests that a common mechanism, that can be acquired through somatic mutations or inherited, is commonly altered in TP53 in multiple cancers. However, a bigger sample size is required to better detect other possible shared mechanism between somatic mutations and inherited variants or to detect entirely novel alterations induced by coding germline polymorphisms.

Our results suggest that the study of protein interfaces can be a viable solution to prioritize genes and interface to study and, with new protein folding prediction technology like AlphaFold (203), it maybe will be possible to describe the change in structure and to make predictions about the altered molecular mechanisms.

## Discussion, conclusion and future works

The study of human germline variants allowed to discover the genetic cause of many rare Mendelian disorders and it is now starting to reveal the complex effects of common variants on common traits. Despite the progress made many questions are still open. In particular, GWAS studies are still unable to explain the estimated heritability of most traits and the biological links between GWAS hits and molecular mechanisms are still mostly unknown. Since most of GWAS hits are located in non-coding regions, many mechanisms have been proposed to explain the effects of those variants to phenotype, such as eQTLs, modifications of chromatin loops and allele-specific expression phenomena.

Also, the missing heritability problem is still one of the main open questions, even with constantly growing studies' sample sizes. Many hypotheses have been formulated to explain the missing heritability: from variant-variant interactions to rare and ultra-rare coding variants, effect of copy numbers and more technical reasons like the frequent exclusion of the X chromosome from associations studies.

In this thesis we explored effects and interactions of human germline variants while developing new tools and bioinformatics resources.

The first challenge when working with germline variants is the process of variant calling. Variant calling is the first step in many genomic data analyses and many subsequent results depends on it. Many methods and tools have been developed like the famous GATK (204), SNVer (205) or VarScan (206). Those tools provide very precise variant calling and variant discovery on NGS dataset however their usefulness is reduced by the amount of time required when calling variants of several samples like in huge genomic cohorts or biobanks. To improve the execution time of known variants genotyping on large cohorts in Chapter 1 we introduced PaCBAM, a new tool for pileup computation and variant genotyping. PaCBAM improves the running time and memory usage of pileup computation and variant genotyping on whole-exome and targeted sequencing data. The tool substantially improves the execution time for known SNPs variant calling in exomes and targeted sequencing. PaCBAM is not suited to detect variants in whole genome sequencing data and its current version is not able to characterize INDELS.

PaCBAM is a useful resource that can be used to characterize NGS data from large-scale sequencing projects like TCGA (207).

After genotyping, variants data can be exploited in many ways by trying to detect some of their possible molecular effects or by finding statistical associations to specific phenotypes.

We started by exploring molecular effects that can be affected by germline variants. In particular, we investigated the translational regulation that can be induced by variants that can induce the allelic specific expression phenomena where an allele is more expressed with respect to the other. This phenomenon has been seen associated to many diseases (208) like cancer (209), schizophrenia (210), and Parkinson's Disease (210). In Chapter 2 we introduced the novel concept of tranSNP, a class of SNPs that alter mRNA translation potential. Impact on translational potential was associated with prognostic effects in cancer, showing that UTR variants can be linked to cancer progression or evolution and can be helpful to stratify patients based on clinical endpoints. Our findings are compatible with the hypothesis that allele-specific expression is a possible molecular effect of variants on phenotypes and in fact it can alter diseases progression. However, given the complexity of the human haplotypes in terms of linkage disequilibrium blocks the causal variants are still elusive and may require further investigation to determine the altered molecular effects and the causal link to a disease.

After investigating more direct molecular effects of variants, we started to study more elusive effects in form of statistical associations with phenotypes. In particular, we explored the variant-variant interaction hypothesis trying to investigate if multiple variants can have a synergistic effect on phenotypes. This hypothesis has been proved to be extremely difficult to test given the enormous sample size required, the computational resources needed, and the number of tests and relative corrections required to test every possible combination of variant (211). Several tools have been implemented to test epistatic interactions. The first technique introduced was multifactor dimensionality reduction (212) where the authors implemented a method to project the information of multiple genomics loci into a one dimensional space. From multifactor dimensionality reduction many other methods evolved based on several techniques like TEAM (213) which is based on minimum spanning trees, techniques based on neural networks (214)

or the very recent Epi-MEIF based on random forests (215). All those approaches are tackling the interaction problem from a statistical point of view but none of them are considering the effects that variants can have on traits. To approach the problem from a functional point of view in Chapter 3 we introduced Polypact, a new web platform that allows the exploration of putative interactions between germline variants. Polypact characterizes the putative effects on transcript levels and transcription factor binding motifs of over 18 million variants. Polypact, also, introduces new network models that allowed us to explore and find novel putative interactions. In particular, we found functional relations among pairs of variants in GWAS, suggesting that the effect of variants on gene expression is mediated by many independent or cooperative interactions. Those effects are in support of the hypothesis that complex variant-variant interactions can affect complex traits and they can be a source of a part of the missing heritability. However, further investigations are needed to better understand how complex epistatic effects can affect traits and diseases.

Finally, in this thesis we analyzed another possible source of missing heritability: rare coding variants. The study of rare coding variants has been neglected in the past years mainly due to the sample size required to have enough statistical power to perform associations studies (216). Now, thanks to biobanks, we were able to explore the effects of rare germline variants and somatic mutations by aggregating them based on their effect on protein interfaces.

In Chapter 4 we explored the effects of rare coding germline variants on protein interfaces of cancer genes in the UKBioBank dataset. Even if the associations did not reach significance after multiple hypothesis correction, we believe our results can be used to prioritize the study of specific genes and interfaces that act as drivers in cancer. We believe that, thanks to the expansion of the biobanks, a part of the missing heritability can probably be explained by the effects of rare coding variants.

Most of the work presented in this thesis is under evolution. We are currently extending Polypact to include protein-protein interactions data, extending our notion of functional relation not only among variants and single genes but also on complex functional modules. We are also extending the models presented in Chapter 2 to identify

and validate additional instances of allele-specific expression across different cancer cell lines.

It is clear that entering in the post-GWAS era, we should start to shift our studies from associations to functional validation, since linking molecular mechanisms to polymorphisms can help in the diagnosis, prognosis and treatment of several diseases. However, finding mechanistic links have been proved to be tougher than expected. Most of the GWAS variants identified are non-coding and their biological effects are difficult to detect. eQTLs studies started to clarify the regulatory effects of variants located in non-coding part of the genome, however the mechanistic biological links are still mostly missing. Also, for a complex trait, GWAS may identify many variants affecting many genes, so a mechanistic link is probably going to involve complex interaction across gene regulatory networks. Currently, GWAS and eQTL studies are analyzing each variant as independent since testing for interactions is currently unfeasible due to the sample size and statistical testing requirements. However, this approach does not allow to detect interactions that can be fundamental in shaping a particular trait. In this thesis we defined and implemented novel functional approaches to study possible functional interactions between multiple variants and genes. These approaches allowed us to limit our hypothesis space, permitting us to detect and explore some putative variant-variant interactions.

We believe that this thesis introduced novel notions and ideas to explore mechanistic effects of human germline variants and potential functional relations and interactions among them. This work produced also technical tools and bioinformatics resources that are freely available for the scientific community.

## Bibliography

1. Kerem B, Rommens JM, Buchanan JA, Markiewicz D, Cox TK, Chakravarti A, et al. Identification of the cystic fibrosis gene: genetic analysis. *Science*. 1989 Sep 8;245(4922):1073–80.
2. Kremer B, Goldberg P, Andrew SE, Theilmann J, Telenius H, Zeisler J, et al. A worldwide study of the Huntington's disease mutation. The sensitivity and specificity of measuring CAG repeats. *N Engl J Med*. 1994 May 19;330(20):1401–6.
3. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. *Nature*. 2001 Feb 15;409(6822):860–921.
4. Haines JL, Hauser MA, Schmidt S, Scott WK, Olson LM, Gallins P, et al. Complement factor H variant increases the risk of age-related macular degeneration. *Science*. 2005 Apr 15;308(5720):419–21.
5. Jiang L, Zheng Z, Fang H, Yang J. A generalized linear mixed model association tool for biobank-scale data. *Nat Genet*. 2021 Nov;53(11):1616–21.
6. Wightman DP, Jansen IE, Savage JE, Shadrin AA, Bahrami S, Holland D, et al. A genome-wide association study with 1,126,563 individuals identifies new risk loci for Alzheimer's disease. *Nat Genet*. 2021 Sep;53(9):1276–82.
7. Sayaman RW, Saad M, Thorsson V, Hu D, Hendrickx W, Roelands J, et al. Germline genetic contribution to the immune landscape of cancer. *Immunity*. 2021 Feb 9;54(2):367-386.e8.
8. Carter H, Marty R, Hofree M, Gross AM, Jensen J, Fisch KM, et al. Interaction Landscape of Inherited Polymorphisms with Somatic Events in Cancer. *Cancer Discov*. 2017 Apr;7(4):410–23.
9. Young AI. Solving the missing heritability problem. *PLoS Genet*. 2019 Jun;15(6):e1008222.
10. Guindo-Martínez M, Amela R, Bonàs-Guarch S, Puiggròs M, Salvoró C, Miguel-Escalada I, et al. The impact of non-additive genetic associations on age-related complex diseases. *Nat Commun*. 2021 Apr 23;12(1):2436.
11. Liu DJ, Peloso GM, Zhan X, Holmen OL, Zawistowski M, Feng S, et al. Meta-analysis of gene-level tests for rare variant association. *Nat Genet*. 2014 Feb;46(2):200–4.
12. Sun BB, Kurki MI, Foley CN, Mechakra A, Chen CY, Marshall E, et al. Genetic associations of protein-coding variants in human disease. *Nature*. 2022 Mar;603(7899):95–102.

13. Sherman RM, Forman J, Antonescu V, Puiu D, Daya M, Rafaels N, et al. Assembly of a pan-genome from deep sequencing of 910 humans of African descent. *Nat Genet.* 2019 Jan;51(1):30–5.
14. Wellcome Trust Case Control Consortium, Craddock N, Hurles ME, Cardin N, Pearson RD, Plagnol V, et al. Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. *Nature.* 2010 Apr 1;464(7289):713–20.
15. Li YR, Glessner JT, Coe BP, Li J, Mohebnasab M, Chang X, et al. Rare copy number variants in over 100,000 European ancestry subjects reveal multiple disease associations. *Nat Commun.* 2020 Jan 14;11(1):255.
16. Auwerx C, Lepamets M, Sadler MC, Patxot M, Stojanov M, Baud D, et al. The individual and global impact of copy-number variants on complex human traits. *Am J Hum Genet.* 2022 Apr 7;109(4):647–68.
17. Hauser MT, Aufsatz W, Jonak C, Luschnig C. Transgenerational epigenetic inheritance in plants. *Biochim Biophys Acta.* 2011 Aug;1809(8):459–68.
18. Slatkin M. Epigenetic inheritance and the missing heritability problem. *Genetics.* 2009 Jul;182(3):845–50.
19. Bateson W, Mendel G. Mendel’s principles of heredity: a defence, with a translation of Mendel’s original papers on hybridisation [Internet]. 2010 [cited 2022 May 20]. Available from: <http://ezproxy.st-andrews.ac.uk/login?url=https://doi.org/10.1017/CBO9780511694462>
20. Fisher RA. XV.—The Correlation between Relatives on the Supposition of Mendelian Inheritance. *Trans R Soc Edinb.* 1919;52(2):399–433.
21. Phillips PC. Epistasis—the essential role of gene interactions in the structure and evolution of genetic systems. *Nat Rev Genet.* 2008 Nov;9(11):855–67.
22. Schaub MA, Boyle AP, Kundaje A, Batzoglou S, Snyder M. Linking disease associations with regulatory information in the human genome. *Genome Res.* 2012 Sep;22(9):1748–59.
23. Lyon MF. X-chromosome inactivation and human genetic disease. *Acta Paediatr Suppl.* 2002;91(439):107–12.
24. de la Chapelle A. Genetic predisposition to human disease: allele-specific expression and low-penetrance regulatory loci. *Oncogene.* 2009 Sep 24;28(38):3345–8.
25. Torkamani A, Wineinger NE, Topol EJ. The personal and clinical utility of polygenic risk scores. *Nature Reviews Genetics.* 2018 Sep 1;19(9):581–90.
26. Xu Y, Vuckovic D, Ritchie SC, Akbari P, Jiang T, Grealey J, et al. Machine learning optimized polygenic scores for blood cell traits identify sex-specific trajectories and genetic correlations with disease. *Cell Genom.* 2022 Jan 12;2(1):None.



27. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bembien LA, et al. Genome sequencing in microfabricated high-density picolitre reactors. *Nature*. 2005 Sep 15;437(7057):376–80.
28. Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, et al. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*. 2008 Nov 6;456(7218):53–9.
29. Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, et al. Real-time DNA sequencing from single polymerase molecules. *Science*. 2009 Jan 2;323(5910):133–8.
30. Clarke J, Wu HC, Jayasinghe L, Patel A, Reid S, Bayley H. Continuous base identification for single-molecule nanopore DNA sequencing. *Nat Nanotechnol*. 2009 Apr;4(4):265–70.
31. Koren S, Schatz MC, Walenz BP, Martin J, Howard JT, Ganapathy G, et al. Hybrid error correction and de novo assembly of single-molecule sequencing reads. *Nat Biotechnol*. 2012 Jul 1;30(7):693–700.
32. Nurk S, Koren S, Rhie A, Rautiainen M, Bzikadze AV, Mikheenko A, et al. The complete sequence of a human genome. *Science*. 2022 Apr;376(6588):44–53.
33. Wang DG, Fan JB, Siao CJ, Berno A, Young P, Sapolsky R, et al. Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science*. 1998 May 15;280(5366):1077–82.
34. Howie BN, Donnelly P, Marchini J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet*. 2009 Jun;5(6):e1000529.
35. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*. 2011 May;43(5):491–8.
36. Romanel A, Lago S, Prandi D, Sboner A, Demichelis F. ASEQ: fast allele-specific studies from next-generation sequencing data. *BMC Medical Genomics* [Internet]. 2015 Dec [cited 2018 May 25];8(1). Available from: <http://bmcmmedgenomics.biomedcentral.com/articles/10.1186/s12920-015-0084-2>
37. Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*. 2011 Nov 1;27(21):2987–93.
38. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. 2009 Aug 15;25(16):2078–9.
39. Korb J, Abyzov A, Mu XJ, Carriero N, Cayting P, Zhang Z, et al. PEMer: a computational framework with simulation-based error models for inferring genomic structural variants from massive paired-end sequencing data. *Genome Biol*. 2009 Feb 23;10(2):R23.

40. Poplin R, Ruano-Rubio V, DePristo MA, Fennell TJ, Carneiro MO, Van der Auwera GA, et al. Scaling accurate genetic variant discovery to tens of thousands of samples [Internet]. *Genomics*; 2017 Nov [cited 2022 May 31]. Available from: <http://biorxiv.org/lookup/doi/10.1101/201178>
41. Sahraeian SME, Liu R, Lau B, Podesta K, Mohiyuddin M, Lam HYK. Deep convolutional neural networks for accurate somatic mutation detection. *Nat Commun*. 2019 Mar 4;10(1):1041.
42. Benner C, Spencer CCA, Havulinna AS, Salomaa V, Ripatti S, Pirinen M. FINEMAP: efficient variable selection using summary data from genome-wide association studies. *Bioinformatics*. 2016 May 15;32(10):1493–501.
43. Hormozdiari F, van de Bunt M, Segrè AV, Li X, Joo JWJ, Bilow M, et al. Colocalization of GWAS and eQTL Signals Detects Target Genes. *Am J Hum Genet*. 2016 Dec 1;99(6):1245–60.
44. GTEx Consortium. The Genotype-Tissue Expression (GTEx) project. *Nat Genet*. 2013 Jun;45(6):580–5.
45. Võsa U, Claringbould A, Westra HJ, Bonder MJ, Deelen P, Zeng B, et al. Large-scale cis- and trans-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. *Nat Genet*. 2021 Sep;53(9):1300–10.
46. Liu X, Li YI, Pritchard JK. Trans Effects on Gene Expression Can Drive Omnigenic Inheritance. *Cell*. 2019 May 2;177(4):1022-1034.e6.
47. Luo Y, Hitz BC, Gabdank I, Hilton JA, Kagda MS, Lam B, et al. New developments on the Encyclopedia of DNA Elements (ENCODE) data portal. *Nucleic Acids Res*. 2020 Jan 8;48(D1):D882–9.
48. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*. 2010 Mar 15;26(6):841–2.
49. Pedersen BS, Quinlan AR. Mosdepth: quick coverage calculation for genomes and exomes. Hancock J, editor. *Bioinformatics*. 2018 Mar 1;34(5):867–8.
50. Tarasov A, Vilella AJ, Cuppen E, Nijman IJ, Prins P. Sambamba: fast processing of NGS alignment formats. *Bioinformatics*. 2015 Jun 15;31(12):2032–4.
51. Faust GG, Hall IM. SAMBLASTER: fast duplicate marking and structural variant read extraction. *Bioinformatics*. 2014 Sep 1;30(17):2503–5.
52. Romanel A, Lago S, Prandi D, Sboner A, Demichelis F. ASEQ: fast allele-specific studies from next-generation sequencing data. *BMC Med Genomics*. 2015 Mar 1;8:9.
53. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*. 2009 Jul 15;25(14):1754–60.

54. Bond GL, Hirshfield KM, Kirchhoff T, Alexe G, Bond EE, Robins H, et al. MDM2 SNP309 accelerates tumor formation in a gender-specific and hormone-dependent manner. *Cancer Res.* 2006;66(10):5104–10.
55. Tomso DJ, Inga A, Menendez D, Pittman GS, Campbell MR, Storicci F, et al. Functionally distinct polymorphic sequences in the human genome that are targets for p53 transactivation. *Proc Natl Acad Sci U S A.* 2005;102(18):6431–6.
56. Menendez D, Snipe J, Marzec J, Innes CL, Polack FP, Caballero MT, et al. P53-responsive TLR8 SNP enhances human innate immune response to respiratory syncytial virus. *Journal of Clinical Investigation.* 2019 Nov 1;129(11):4875–84.
57. Menendez D, Inga A, Snipe J, Krysiak O, Schonfelder G, Resnick MA. A single-nucleotide polymorphism in a half-binding site creates p53 and estrogen receptor control of vascular endothelial growth factor receptor 1. *Mol Cell Biol.* 2007;27(7):2590–600.
58. Bu H, Narisu N, Schlick B, Rainer J, Manke T, Schäfer G, et al. Putative Prostate Cancer Risk SNP in an Androgen Receptor-Binding Site of the Melanophilin Gene Illustrates Enrichment of Risk SNPs in Androgen Receptor Target Sites. *Hum Mutat.* 2016 Jan;37(1):52–64.
59. Romanel A, Garritano S, Stringa B, Blattner M, Dalfovo D, Chakravarty D, et al. Inherited determinants of early recurrent somatic mutations in prostate cancer. *Nat Commun.* 2017 29;8(1):48.
60. Bailey SD, Desai K, Kron KJ, Mazrooei P, Sinnott-Armstrong NA, Treloar AE, et al. Noncoding somatic and inherited single-nucleotide variants converge to promote ESR1 expression in breast cancer. *Nat Genet.* 2016;48(10):1260–6.
61. Dunning AM, Michailidou K, Kuchenbaecker KB, Thompson D, French JD, Beesley J, et al. Breast cancer risk variants at 6q25 display different phenotype associations and regulate ESR1, RMND1 and CCDC170. *Nat Genet.* 2016 Apr;48(4):374–86.
62. Pomerantz MM, Ahmadiyeh N, Jia L, Herman P, Verzi MP, Doddapaneni H, et al. The 8q24 cancer risk variant rs6983267 shows long-range interaction with MYC in colorectal cancer. *Nat Genet.* 2009 Aug;41(8):882–4.
63. Takatsuno Y, Mimori K, Yamamoto K, Sato T, Niida A, Inoue H, et al. The rs6983267 SNP is associated with MYC transcription efficiency, which promotes progression and worsens prognosis of colorectal cancer. *Ann Surg Oncol.* 2013 Apr;20(4):1395–402.
64. Landi D, Gemignani F, Landi S. Role of variations within microRNA-binding sites in cancer. *Mutagenesis.* 2012 Mar;27(2):205–10.
65. Rozowsky J, Abyzov A, Wang J, Alves P, Raha D, Harmanci A, et al. AlleleSeq: Analysis of allele-specific expression and binding in a network framework. *Molecular Systems Biology.* 2011;7:522.

66. Wei Y, Li X, Wang Q fei, Ji H. IASeq: Integrative analysis of allele-specificity of protein-DNA interactions in multiple ChIP-seq datasets. *BMC Genomics*. 2012;13:681.
67. Skelly DA, Johansson M, Madeoy J, Wakefield J, Akey JM. A powerful and flexible statistical framework for testing hypotheses of allele-specific gene expression from RNA-seq data. *Genome Research*. 2011;21(10):1728–37.
68. Pandey RV, Franssen SU, Futschik A, Schlötterer C. Allelic imbalance metre (Allim), a new tool for measuring allele-specific gene expression with RNA-seq data. *Molecular Ecology Resources*. 2013;13(4):740–5.
69. Mayba O, Gilbert HN, Liu J, Haverty PM, Jhunjhunwala S, Jiang Z, et al. MBASED: Allele-specific expression detection in cancer tissues and cell lines. *Genome Biology*. 2014;15(8):405.
70. Romanel A, Lago S, Prandi D, Sboner A, Demichelis F. ASEQ: Fast allele-specific studies from next-generation sequencing data. *BMC Medical Genomics*. 2015 Mar 1;8(1).
71. Romanel A. Allele-specific expression analysis in cancer using next-generation sequencing data. In: *Methods in Molecular Biology*. 2019. p. 125–37.
72. Przytycki PF, Singh M. Differential Allele-Specific Expression Uncovers Breast Cancer Genes Dysregulated by Cis Noncoding Mutations. *Cell Systems*. 2020;10(2):193–203.
73. Andreotti V, Bisio A, Bressac-de Paillerets B, Harland M, Cabaret O, Newton-Bishop J, et al. The CDKN2A/p16(INK) (4a) 5'UTR sequence and translational regulation: impact of novel variants predisposing to melanoma. *Pigment Cell Melanoma Res*. 2016;29(2):210–21.
74. Zaccara S, Tebaldi T, Pederiva C, Ciribilli Y, Bisio A, Inga A. p53-directed translational control can shape and expand the universe of p53 target genes. *Cell Death Differ*. 2014;21(10):1522–34.
75. Rizzotto D, Zaccara S, Rossi A, Galbraith MD, Andrysik Z, Pandey A, et al. Nutlin-Induced Apoptosis is Specified by a Translation Program Regulated by PCBP2 and DHX30. *Cell Rep*. 2020;30:1–15.
76. Rossi M, Bucci G, Rizzotto D, Bordo D, Marzi MJ, Puppo M, et al. LncRNA EPR controls epithelial proliferation by coordinating Cdkn1a transcription and mRNA decay response to TGF- $\beta$ . *Nature Communications*. 2019;10(1):1969.
77. Ha SA, Seung MS, Yong JL, Kim S, Hyun KK, Namkoong H, et al. HCCRBP-1 directly interacting with HCCR-1 induces tumorigenesis through P53 stabilization. *International Journal of Cancer*. 2008;
78. Cho GW, Shin SM, Kim HK, Ha SA, Kim S, Yoon JH, et al. HCCR-1, a novel oncogene, encodes a mitochondrial outer membrane protein and suppresses the UVC-induced apoptosis. *BMC Cell Biology*. 2007 Nov 28;8.

79. Kreis NN, Louwen F, Yuan J. The multifaceted p21 (Cip1/Waf1/CDKN1A) in cell differentiation, migration and cancer therapy. *Cancers*. 2019 Sep 1;11(9).
80. El-Deiry WS. p21(WAF1) mediates cell-cycle inhibition, relevant to cancer suppression and therapy. Vol. 76, *Cancer Research*. American Association for Cancer Research Inc.; 2016. p. 5189–91.
81. Wang W, Furneaux H, Cheng H, Caldwell MC, Hutter D, Liu Y, et al. HuR regulates p21 mRNA stabilization by UV light. *Mol Cell Biol*. 2000/01/11 ed. 2000;20(3):760–9.
82. Koboldt DC, Fulton RS, McLellan MD, Schmidt H, Kalicki-Veizer J, McMichael JF, et al. Comprehensive molecular portraits of human breast tumours. *Nature*. 2012;490(7418):61–70.
83. Liu J, Lichtenberg T, Hoadley KA, Poisson LM, Lazar AJ, Cherniack AD, et al. An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality Survival Outcome Analytics. *Cell*. 2018;173(2):400–16.
84. Cook KB, Kazan H, Zuberi K, Morris Q, Hughes TR. RBPDB: A database of RNA-binding specificities. *Nucleic Acids Research*. 2011;39:D301-8.
85. Schug J. Using TESS to predict transcription factor binding sites in DNA sequence. *Current Protocols in Bioinformatics*. 2008. p. unit 2.6.
86. Mazan-Mamczarz K, Galbán S, López de Silanes I, Martindale JL, Atasoy U, Keene JD, et al. RNA-binding protein HuR enhances p53 translation in response to ultraviolet light irradiation. *Proc Natl Acad Sci U S A*. 2003 Jul 8;100(14):8354–9.
87. Yasuda K, Zhang H, Loiselle D, Haystead T, Macara IG, Mili S. The RNA-binding protein Fus directs translation of localized mRNAs in APC-RNP granules. *J Cell Biol*. 2013 Dec 9;203(5):737–46.
88. Peixeiro I, Inácio Â, Barbosa C, Silva AL, Liebhaber SA, Romão L. Interaction of PABPC1 with the translation initiation complex is critical to the NMD resistance of AUG-proximal nonsense mutations. *Nucleic Acids Res*. 2012 Feb;40(3):1160–73.
89. Tanguay RL, Gallie DR. Translational efficiency is regulated by the length of the 3' untranslated region. *Mol Cell Biol*. 1996 Jan;16(1):146–56.
90. Behm-Ansmant I, Gatfield D, Rehwinkel J, Hilgers V, Izaurralde E. A conserved role for cytoplasmic poly(A)-binding protein 1 (PABPC1) in nonsense-mediated mRNA decay. *EMBO J*. 2007 Mar 21;26(6):1591–601.
91. Andrysik Z, Galbraith MD, Guarnieri AL, Zaccara S, Sullivan KD, Pandey A, et al. Identification of a core TP53 transcriptional program with highly distributed tumor suppressive activity. *Genome Res*. 2017;27(10):1645–57.

92. Tovar C, Rosinski J, Filipovic Z, Higgins B, Kolinsky K, Hilton H, et al. Small-molecule MDM2 antagonists reveal aberrant p53 signaling in cancer: implications for therapy. *Proc Natl Acad Sci U S A*. 2006;103(6):1888–93.
93. Nassiri I, Inga A, Meškytė EM, Alessandrini F, Ciribilli Y, Priami C. Regulatory Crosstalk of Doxorubicin, Estradiol and TNF $\alpha$  Combined Treatment in Breast Cancer-derived Cell Lines. *Scientific Reports*. 2019;9(1):15172.
94. Lion M, Bisio A, Tebaldi T, De Sanctis V, Menendez D, Resnick MA, et al. Interaction between p53 and estradiol pathways in transcriptional responses to chemotherapeutics. *Cell Cycle*. 2013/03/23 ed. 2013;12(8):1211–24.
95. Buniello A, MacArthur JAL, Cerezo M, Harris LW, Hayhurst J, Malangone C, et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res*. 2019 Jan 8;47(D1):D1005–12.
96. Stracquadanio G, Wang X, Wallace MD, Grawenda AM, Zhang P, Hewitt J, et al. The importance of p53 pathway genetics in inherited and somatic cancer genomes. *Nat Rev Cancer*. 2016 Apr;16(4):251–65.
97. Zhang P, Kitchen-Smith I, Xiong L, Stracquadanio G, Brown K, Richter PH, et al. Germline and Somatic Genetic Variants in the p53 Pathway Interact to Affect Cancer Risk, Progression, and Drug Response. *Cancer Res*. 2021 Apr 1;81(7):1667–80.
98. Griesemer D, Xue JR, Reilly SK, Ulirsch JC, Kukreja K, Davis JR, et al. Genome-wide functional screen of 3'UTR variants uncovers causal variants for human disease and evolution. *Cell*. 2021 Sep 30;184(20):5247-5260.e19.
99. Sicari D, Fantuz M, Bellazzo A, Valentino E, Apollonio M, Pontisso I, et al. Mutant p53 improves cancer cells' resistance to endoplasmic reticulum stress by sustaining activation of the UPR regulator ATF6. *Oncogene*. 2019 Aug 22;38(34):6184–95.
100. Guan BJ, van Hoef V, Jobava R, Elroy-Stein O, Valasek LS, Cargnello M, et al. A Unique ISR Program Determines Cellular Responses to Chronic Stress. *Molecular Cell*. 2017 Dec 7;68(5):885-900.e6.
101. Hassan N, Rutsch N, Györffy B, Espinoza-Sánchez NA, Götte M. SETD3 acts as a prognostic marker in breast cancer patients and modulates the viability and invasion of breast cancer cells. *Scientific Reports*. 2020 Dec 1;10(1).
102. Abaev-Schneiderman E, Admoni-Elisha L, Levy D. SETD3 is a positive regulator of DNA-damage-induced apoptosis. *Cell Death and Disease*. 2019 Feb 1;10(2).
103. Norberg E, Lako A, Chen PH, Stanley IA, Zhou F, Ficarro SB, et al. Differential contribution of the mitochondrial translation pathway to the survival of diffuse large B-cell lymphoma subsets. *Cell Death and Differentiation*. 2017 Feb 1;24(2):251–62.

104. Annunziato S, de Ruiter JR, Henneman L, Brambillasca CS, Lutz C, Vaillant F, et al. Comparative oncogenomics identifies combinations of driver genes and drug targets in BRCA1-mutated breast cancer. *Nature Communications*. 2019;10(1):397.
105. Horvath A, Pakala SB, Mudvari P, Reddy SDN, Ohshiro K, Casimiro S, et al. Novel insights into breast cancer genetic variance through RNA sequencing. *Scientific Reports*. 2013;3:2256.
106. Niwa T, Saito H, Imajoh-ohmi S, Kaminishi M, Seto Y, Miki Y, et al. BRCA2 interacts with the cytoskeletal linker protein plectin to form a complex controlling centrosome localization. *Cancer Science*. 2009;100(11):2115–25.
107. Reinhold WC, Sunshine M, Liu H, Varma S, Kohn KW, Morris J, et al. CellMiner: A web-based suite of genomic and pharmacologic tools to explore transcript and drug patterns in the NCI-60 cell line set. *Cancer Research*. 2012;
108. Obenchain V, Lawrence M, Carey V, Gogarten S, Shannon P, Morgan M. VariantAnnotation: a Bioconductor package for exploration and annotation of genetic variants. *Bioinformatics*. 2014 Jul 15;30(14):2076–8.
109. Yates A, Beal K, Keenan S, McLaren W, Pignatelli M, Ritchie GRS, et al. The Ensembl REST API: Ensembl Data for Any Language. *Bioinformatics*. 2015;31(1):143–5.
110. Delaneau O, Howie B, Cox AJ, Zagury JF, Marchini J. Haplotype estimation using sequencing reads. *Am J Hum Genet*. 2013 Oct 3;93(4):687–96.
111. Therneau TM, Grambsch PM. Modeling Survival Data: Extending the Cox Model [Internet]. New York, NY: Springer New York; 2000 [cited 2020 Nov 20]. (Dietz K, Gail M, Krickeberg K, Samet J, Tsiatis A, editors. *Statistics for Biology and Health*). Available from: <http://link.springer.com/10.1007/978-1-4757-3294-8>
112. Dassi E, Greco V, Sidarovich V, Zuccotti P, Arseni N, Scaruffi P, et al. Translational compensation of genomic instability in neuroblastoma. *Sci Rep*. 2015;5:14364.
113. Provenzani A, Fronza R, Loreni F, Pascale A, Amadio M, Quattrone A. Global alterations in mRNA polysomal recruitment in a cell model of colorectal cancer progression to metastasis. *Carcinogenesis*. 2006;27(7):1323–33.
114. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. TopHat2: Accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biology*. 2013;14(4):R36.
115. Trapnell C, Hendrickson DG, Sauvageau M, Goff L, Rinn JL, Pachter L. Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nat Biotechnol*. 2013 Jan;31(1):46–53.
116. Yu G, Wang LG, Han Y, He QY. clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS*. 2012 May;16(5):284–7.

117. Kanehisa M, Furumichi M, Sato Y, Ishiguro-Watanabe M, Tanabe M. KEGG: integrating viruses and cellular organisms. *Nucleic Acids Res.* 2020 Oct 30;
118. Zhou Y, Zhou B, Pache L, Chang M, Khodabakhshi AH, Tanaseichuk O, et al. Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. *Nat Commun.* 2019 Apr 3;10(1):1523.
119. Keene JD, Komisarow JM, Friedersdorf MB. RIP-Chip: the isolation and identification of mRNAs, microRNAs and protein components of ribonucleoprotein complexes from cell extracts. *Nat Protoc.* 2006;1(1):302–7.
120. Uroda T, Anastasakou E, Rossi A, Teulon JM, Pellequer JL, Annibale P, et al. Conserved Pseudoknots in lncRNA MEG3 Are Essential for Stimulation of the p53 Pathway. *Molecular Cell.* 2019 Sep;75(5):982-995.e9.
121. Tam V, Patel N, Turcotte M, Bossé Y, Paré G, Meyre D. Benefits and limitations of genome-wide association studies. *Nat Rev Genet.* 2019 Aug;20(8):467–84.
122. Nica AC, Dermitzakis ET. Expression quantitative trait loci: present and future. *Philos Trans R Soc Lond B Biol Sci.* 2013;368(1620):20120362.
123. Libioulle C, Louis E, Hansoul S, Sandor C, Farnir F, Franchimont D, et al. Novel Crohn disease locus identified by genome-wide association maps to a gene desert on 5p13.1 and modulates expression of PTGER4. *PLoS Genet.* 2007 Apr 20;3(4):e58.
124. Carter H, Marty R, Hofree M, Gross AM, Jensen J, Fisch KM, et al. Interaction Landscape of Inherited Polymorphisms with Somatic Events in Cancer. *Cancer Discov.* 2017 Apr;7(4):410–23.
125. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, et al. Finding the missing heritability of complex diseases. *Nature.* 2009 Oct 8;461(7265):747–53.
126. Li P, Guo M, Wang C, Liu X, Zou Q. An overview of SNP interactions in genome-wide association studies. *Brief Funct Genomics.* 2015 Mar;14(2):143–55.
127. Kumar S, Ambrosini G, Bucher P. SNP2TFBS - a database of regulatory SNPs affecting predicted transcription factor binding site affinity. *Nucleic Acids Res.* 2017 Jan 4;45(D1):D139–44.
128. Wang J, Dai X, Berry LD, Cogan JD, Liu Q, Shyr Y. HACER: an atlas of human active enhancers to interpret regulatory variants. *Nucleic Acids Res.* 2019 Jan 8;47(D1):D106–12.
129. Boyle AP, Hong EL, Hariharan M, Cheng Y, Schaub MA, Kasowski M, et al. Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res.* 2012 Sep;22(9):1790–7.



130. Ward LD, Kellis M. HaploReg v4: systematic mining of putative causal variants, cell types, regulators and target genes for human complex traits and disease. *Nucleic Acids Res.* 2016 Jan 4;44(D1):D877-881.
131. Pan Q, Liu YJ, Bai XF, Han XL, Jiang Y, Ai B, et al. VARAdb: a comprehensive variation annotation database for human. *Nucleic Acids Res.* 2021 Jan 8;49(D1):D1431-44.
132. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, et al. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* 2001 Jan 1;29(1):308-11.
133. 1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, et al. A global reference for human genetic variation. *Nature.* 2015 Oct 1;526(7571):68-74.
134. Taliun D, Harris DN, Kessler MD, Carlson J, Szpiech ZA, Torres R, et al. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program [Internet]. *Genomics*; 2019 Mar [cited 2021 Feb 4]. Available from: <http://biorxiv.org/lookup/doi/10.1101/563866>
135. Davis CA, Hitz BC, Sloan CA, Chan ET, Davidson JM, Gabdank I, et al. The Encyclopedia of DNA elements (ENCODE): data portal update. *Nucleic Acids Res.* 2018 Jan 4;46(D1):D794-801.
136. Bernstein BE, Stamatoyannopoulos JA, Costello JF, Ren B, Milosavljevic A, Meissner A, et al. The NIH Roadmap Epigenomics Mapping Consortium. *Nat Biotechnol.* 2010 Oct;28(10):1045-8.
137. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics.* 2010 Mar 15;26(6):841-2.
138. Dalfovo D, Valentini S, Romanel A. Exploring functionally annotated transcriptional consensus regulatory elements with CONREL. *Database.* 2020 Jan 1;2020:baaa071.
139. Matys V, Kel-Margoulis OV, Fricke E, Liebich I, Land S, Barre-Dirrie A, et al. TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res.* 2006 Jan 1;34(Database issue):D108-110.
140. Kulakovskiy IV, Medvedeva YA, Schaefer U, Kasianov AS, Vorontsov IE, Bajic VB, et al. HOCOMOCO: a comprehensive collection of human transcription factor binding sites models. *Nucleic Acids Res.* 2013 Jan;41(Database issue):D195-202.
141. Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell.* 2010 May 28;38(4):576-89.
142. Khan A, Fornes O, Stigliani A, Gheorghe M, Castro-Mondragon JA, van der Lee R, et al. JASPAR 2018: update of the open-access database of transcription factor

- binding profiles and its web framework. *Nucleic Acids Res.* 2018 Jan 4;46(D1):D260–6.
143. Cook KB, Kazan H, Zuberi K, Morris Q, Hughes TR. RBPDB: a database of RNA-binding specificities. *Nucleic Acids Res.* 2011 Jan;39(Database issue):D301-308.
  144. Romanel A, Garritano S, Stringa B, Blattner M, Dalfovo D, Chakravarty D, et al. Inherited determinants of early recurrent somatic mutations in prostate cancer. *Nat Commun.* 2017 Jun 29;8(1):48.
  145. Schug J. Using TESS to predict transcription factor binding sites in DNA sequence. *Curr Protoc Bioinformatics.* 2008 Mar;Chapter 2:Unit 2.6.
  146. Cancer Genome Atlas Research Network, Weinstein JN, Collisson EA, Mills GB, Shaw KRM, Ozenberger BA, et al. The Cancer Genome Atlas Pan-Cancer analysis project. *Nat Genet.* 2013 Oct;45(10):1113–20.
  147. GTEx Consortium. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science.* 2015 May 8;348(6235):648–60.
  148. Collado-Torres L, Nellore A, Kammers K, Ellis SE, Taub MA, Hansen KD, et al. Reproducible RNA-seq analysis using recount2. *Nat Biotechnol.* 2017 Apr 11;35(4):319–21.
  149. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics.* 2010 Jan 1;26(1):139–40.
  150. Law CW, Chen Y, Shi W, Smyth GK. voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol.* 2014;15(2):R29.
  151. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics.* 2007 Jan;8(1):118–27.
  152. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet.* 2007 Sep;81(3):559–75.
  153. Romanel A, Zhang T, Elemento O, Demichelis F. EthSEQ: ethnicity annotation from whole exome sequencing data. *Bioinformatics.* 2017 Aug 1;33(15):2402–4.
  154. Carrot-Zhang J, Chambwe N, Damrauer JS, Knijnenburg TA, Robertson AG, Yau C, et al. Comprehensive Analysis of Genetic Ancestry and Its Molecular Correlates in Cancer. *Cancer Cell.* 2020 May 11;37(5):639-654.e6.
  155. Staudt CL, Sazonovs A, Meyerhenke H. NetworKit: A tool suite for large-scale complex network analysis. *Net Sci.* 2016 Dec;4(4):508–30.

156. Lancichinetti A, Fortunato S. Community detection algorithms: a comparative analysis. *Phys Rev E Stat Nonlin Soft Matter Phys.* 2009 Nov;80(5 Pt 2):056117.
157. Yates A, Beal K, Keenan S, McLaren W, Pignatelli M, Ritchie GRS, et al. The Ensembl REST API: Ensembl Data for Any Language. *Bioinformatics.* 2015 Jan 1;31(1):143–5.
158. Valentini S, Marchioretto C, Bisio A, Rossi A, Zaccara S, Romanel A, et al. TransSNPs: A class of functional SNPs affecting mRNA translation potential revealed by fraction-based allelic imbalance. *iScience.* 2021 Dec 17;24(12):103531.
159. Parca L, Truglio M, Biagini T, Castellana S, Petrizzelli F, Capocéfalo D, et al. Pyntacle: a parallel computing-enabled framework for large-scale network biology analysis. *Gigascience.* 2020 Oct 21;9(10):giaa115.
160. Chen H, Kichaev G, Bien SA, MacDonald JW, Wang L, Bammler TK, et al. Genetic associations of breast and prostate cancer are enriched for regulatory elements identified in disease-related tissues. *Hum Genet.* 2019 Oct;138(10):1091–104.
161. Hazelett DJ, Rhie SK, Gaddis M, Yan C, Lakeland DL, Coetzee SG, et al. Comprehensive functional annotation of 77 prostate cancer risk loci. *PLoS Genet.* 2014 Jan;10(1):e1004102.
162. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature.* 2012 Sep 6;489(7414):57–74.
163. Parikh P, Palazzo JP, Rose LJ, Daskalakis C, Weigel RJ. GATA-3 expression as a predictor of hormone response in breast cancer. *J Am Coll Surg.* 2005 May;200(5):705–10.
164. Wu Y, Alvarez M, Slamon DJ, Koeffler P, Vadgama JV. Caspase 8 and maspin are downregulated in breast cancer cells due to CpG site promoter methylation. *BMC Cancer.* 2010 Feb 4;10:32.
165. Aghababazadeh M, Dorraki N, Javan FA, Fattahi AS, Gharib M, Pasdar A. Downregulation of Caspase 8 in a group of Iranian breast cancer patients - A pilot study. *J Egypt Natl Canc Inst.* 2017 Dec;29(4):191–5.
166. Olsson M, Zhivotovsky B. Caspases and cancer. *Cell Death Differ.* 2011 Sep;18(9):1441–9.
167. Krelin Y, Zhang L, Kang TB, Appel E, Kovalenko A, Wallach D. Caspase-8 deficiency facilitates cellular transformation in vitro. *Cell Death Differ.* 2008 Sep;15(9):1350–5.
168. Lin WY, Camp NJ, Ghoussaini M, Beesley J, Michailidou K, Hopper JL, et al. Identification and characterization of novel associations in the CASP8/ALS2CR12 region on chromosome 2 with breast cancer risk. *Hum Mol Genet.* 2015 Jan 1;24(1):285–98.

169. He L, Fan C, Kapoor A, Ingram AJ, Rybak AP, Austin RC, et al.  $\alpha$ -Mannosidase 2C1 attenuates PTEN function in prostate cancer cells. *Nat Commun.* 2011 Sep;2(1):307.
170. Shu X, Long J, Cai Q, Kweon SS, Choi JY, Kubo M, et al. Identification of novel breast cancer susceptibility loci in meta-analyses conducted among Asian and European descendants. *Nat Commun.* 2020 Mar 5;11(1):1217.
171. Huang C, Jia Y, Yang S, Chen B, Sun H, Shen F, et al. Characterization of ZNF23, a KRAB-containing protein that is downregulated in human cancers and inhibits cell cycle progression. *Exp Cell Res.* 2007 Jan 15;313(2):254–63.
172. Sobocińska J, Molenda S, Machnik M, Oleksiewicz U. KRAB-ZFP Transcriptional Regulators Acting as Oncogenes and Tumor Suppressors: An Overview. *Int J Mol Sci.* 2021 Feb 23;22(4):2212.
173. Ballatore C, Lee VMY, Trojanowski JQ. Tau-mediated neurodegeneration in Alzheimer's disease and related disorders. *Nat Rev Neurosci.* 2007 Sep;8(9):663–72.
174. Strang KH, Golde TE, Giasson BI. MAPT mutations, tauopathy, and mechanisms of neurodegeneration. *Lab Invest.* 2019 Jul;99(7):912–28.
175. Zody MC, Jiang Z, Fung HC, Antonacci F, Hillier LW, Cardone MF, et al. Evolutionary toggling of the MAPT 17q21.31 inversion region. *Nat Genet.* 2008 Sep;40(9):1076–83.
176. Höglinger GU, Melhem NM, Dickson DW, Sleiman PMA, Wang LS, Klei L, et al. Identification of common variants influencing risk of the tauopathy progressive supranuclear palsy. *Nat Genet.* 2011 Jun 19;43(7):699–705.
177. Achim K, Peltopuro P, Lahti L, Tsai HH, Zachariah A, Astrand M, et al. The role of Tal2 and Tal1 in the differentiation of midbrain GABAergic neuron precursors. *Biol Open.* 2013;2(10):990–7.
178. Crux S, Herms J, Dorostkar MM. Tcf4 regulates dendritic spine density and morphology in the adult brain. *PLoS One.* 2018;13(6):e0199359.
179. Badhwar A, Brown R, Stanimirovic DB, Haqqani AS, Hamel E. Proteomic differences in brain vessels of Alzheimer's disease mice: Normalization by PPAR $\gamma$  agonist pioglitazone. *J Cereb Blood Flow Metab.* 2017 Mar;37(3):1120–36.
180. Levanon D, Bettoun D, Harris-Cerruti C, Woolf E, Negreanu V, Eilam R, et al. The Runx3 transcription factor regulates development and survival of TrkC dorsal root ganglia neurons. *EMBO J.* 2002 Jul 1;21(13):3454–63.
181. Farley JE, Burdett TC, Barria R, Neukomm LJ, Kenna KP, Landers JE, et al. Transcription factor Pebbled/RREB1 regulates injury-induced axon degeneration. *Proc Natl Acad Sci U S A.* 2018 Feb 6;115(6):1358–63.

182. Antonarakis SE, Beckmann JS. Mendelian disorders deserve more attention. *Nat Rev Genet.* 2006 Apr;7(4):277–82.
183. Wells JA, McClendon CL. Reaching for high-hanging fruit in drug discovery at protein-protein interfaces. *Nature.* 2007 Dec 13;450(7172):1001–9.
184. Porta-Pardo E, Garcia-Alonso L, Hrade T, Dopazo J, Godzik A. A Pan-Cancer Catalogue of Cancer Driver Protein Interaction Interfaces. *PLoS Comput Biol.* 2015 Oct;11(10):e1004518.
185. Ellrott K, Bailey MH, Saksena G, Covington KR, Kandath C, Stewart C, et al. Scalable Open Science Approach for Mutation Calling of Tumor Exomes Using Multiple Genomic Pipelines. *Cell Syst.* 2018 Mar 28;6(3):271-281.e7.
186. Gao J, Aksoy BA, Dogrusoz U, Dresdner G, Gross B, Sumer SO, et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci Signal.* 2013 Apr 2;6(269):pl1.
187. Tate JG, Bamford S, Jubb HC, Sondka Z, Beare DM, Bindal N, et al. COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Res.* 2019 Jan 8;47(D1):D941–7.
188. Huang KL, Mashl RJ, Wu Y, Ritter DI, Wang J, Oh C, et al. Pathogenic Germline Variants in 10,389 Adult Cancers. *Cell.* 2018 Apr 5;173(2):355-370.e14.
189. Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, et al. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med.* 2015 Mar;12(3):e1001779.
190. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, et al. The Ensembl Variant Effect Predictor. *Genome Biol.* 2016 Jun 6;17(1):122.
191. Miki Y, Swensen J, Shattuck-Eidens D, Futreal PA, Harshman K, Tavtigian S, et al. A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1. *Science.* 1994 Oct 7;266(5182):66–71.
192. Liu CC, Liu CC, Kanekiyo T, Xu H, Bu G. Apolipoprotein E and Alzheimer disease: risk, mechanisms and therapy. *Nat Rev Neurol.* 2013 Feb;9(2):106–18.
193. Barbieri CE, Baca SC, Lawrence MS, Demichelis F, Blattner M, Theurillat JP, et al. Exome sequencing identifies recurrent SPOP, FOXA1 and MED12 mutations in prostate cancer. *Nat Genet.* 2012 May 20;44(6):685–9.
194. Shoushtari AN, Carvajal RD. GNAQ and GNA11 mutations in uveal melanoma. *Melanoma Res.* 2014 Dec;24(6):525–34.
195. Thirumal Kumar D, George Priya Doss C. Role of E542 and E545 missense mutations of PIK3CA in breast cancer: a comparative computational approach. *J Biomol Struct Dyn.* 2017 Sep;35(12):2745–57.
196. Culig Z, Santer FR. Androgen receptor signaling in prostate cancer. *Cancer Metastasis Rev.* 2014 Sep;33(2–3):413–27.

197. Normanno N, De Luca A, Bianco C, Strizzi L, Mancino M, Maiello MR, et al. Epidermal growth factor receptor (EGFR) signaling in cancer. *Gene*. 2006 Jan 17;366(1):2–16.
198. Sonobe M, Manabe T, Wada H, Tanaka F. Mutations in the epidermal growth factor receptor gene are linked to smoking-independent, lung adenocarcinoma. *Br J Cancer*. 2005 Aug 8;93(3):355–63.
199. Binder ZA, Thorne AH, Bakas S, Wileyto EP, Bilello M, Akbari H, et al. Epidermal Growth Factor Receptor Extracellular Domain Mutations in Glioblastoma Present Opportunities for Clinical Imaging and Therapeutic Development. *Cancer Cell*. 2018 Jul 9;34(1):163-177.e7.
200. Pasqualucci L, Dominguez-Sola D, Chiarenza A, Fabbri G, Grunn A, Trifonov V, et al. Inactivating mutations of acetyltransferase genes in B-cell lymphoma. *Nature*. 2011 Mar 10;471(7337):189–95.
201. Venturutti L, Teater M, Zhai A, Chadburn A, Babiker L, Kim D, et al. TBL1XR1 Mutations Drive Extranodal Lymphoma by Inducing a Pro-tumorigenic Memory Fate. *Cell*. 2020 Jul 23;182(2):297-316.e27.
202. Oyer JA, Huang X, Zheng Y, Shim J, Ezponda T, Carpenter Z, et al. Point mutation E1099K in MMSET/NSD2 enhances its methyltransferase activity and leads to altered global chromatin methylation in lymphoid malignancies. *Leukemia*. 2014 Jan;28(1):198–201.
203. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. *Nature*. 2021 Aug;596(7873):583–9.
204. Poplin R, Ruano-Rubio V, DePristo MA, Fennell TJ, Carneiro MO, Van der Auwera GA, et al. Scaling accurate genetic variant discovery to tens of thousands of samples [Internet]. *Genomics*; 2017 Nov [cited 2022 Sep 28]. Available from: <http://biorxiv.org/lookup/doi/10.1101/201178>
205. Wei Z, Wang W, Hu P, Lyon GJ, Hakonarson H. SNVer: a statistical tool for variant calling in analysis of pooled or individual next-generation sequencing data. *Nucleic Acids Research*. 2011 Oct;39(19):e132–e132.
206. Koboldt DC, Chen K, Wylie T, Larson DE, McLellan MD, Mardis ER, et al. VarScan: variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics*. 2009 Sep 1;25(17):2283–5.
207. Ciani Y, Fedrizzi T, Prandi D, Lorenzin F, Locallo A, Gasperini P, et al. Allele-specific genomic data elucidate the role of somatic gain and copy-number neutral loss of heterozygosity in cancer. *Cell Syst*. 2022 Feb 16;13(2):183-193.e7.
208. de la Chapelle A. Genetic predisposition to human disease: allele-specific expression and low-penetrance regulatory loci. *Oncogene*. 2009 Sep 24;28(38):3345–8.

209. Przytycki PF, Singh M. Differential Allele-Specific Expression Uncovers Breast Cancer Genes Dysregulated by Cis Noncoding Mutations. *Cell Systems*. 2020 Feb;10(2):193-203.e4.
210. Guella I, Sequeira A, Rollins B, Morgan L, Myers RM, Watson SJ, et al. Evidence of allelic imbalance in the schizophrenia susceptibility gene ZNF804A in human dorsolateral prefrontal cortex. *Schizophrenia Research*. 2014 Jan;152(1):111–6.
211. Chattopadhyay A, Lu TP. Gene-gene interaction: the curse of dimensionality. *Ann Transl Med*. 2019 Dec;7(24):813–813.
212. Ritchie MD, Hahn LW, Roodi N, Bailey LR, Dupont WD, Parl FF, et al. Multifactor-Dimensionality Reduction Reveals High-Order Interactions among Estrogen-Metabolism Genes in Sporadic Breast Cancer. *The American Journal of Human Genetics*. 2001 Jul;69(1):138–47.
213. Zhang X, Huang S, Zou F, Wang W. TEAM: efficient two-locus epistasis tests in human genome-wide association study. *Bioinformatics*. 2010 Jun 15;26(12):i217–27.
214. Motsinger-Reif AA, Fanelli TJ, Davis AC, Ritchie MD. Power of grammatical evolution neural networks to detect gene-gene interactions in the presence of error. *BMC Res Notes*. 2008 Dec;1(1):65.
215. Saha S, Perrin L, Röder L, Brun C, Spinelli L. Epi-MEIF: detecting higher order epistatic interactions for complex traits using mixed effect conditional inference forests. *Nucleic Acids Research*. 2022 Sep 15;gkac715.
216. Auer PL, Lettre G. Rare variant association studies: considerations, challenges and opportunities. *Genome Med*. 2015;7(1):16.