PhD Program in Biomolecular Sciences

Department of Cellular, Computational
and Integrative Biology – CIBIO
36th Cycle

Exploration of the interaction landscape between functional SNPs and somatic aberrations in cancer

Tutor

Alessandro ROMANEL

Advisor

Alberto INGA

Ph.D. Thesis of

Davide DALFOVO

Academic Year 2022 – 2023

Declaration

I (Davide Dalfovo) confirm that this is my own work and the use of all material from other sources has been properly and fully acknowledged.

Davide Dalfovo

Contents

Abstract

Cancer is a complex disease shaped by a heterogeneous landscape of inherited genetic variants and acquired somatic aberrations. Although specific patterns of somatic aberrations within key pathways are recognized as hallmarks of many cancers, and mounting evidence suggests a significant interplay between germline and somatic variants, the intricate relationship between germline predisposition and the disruption of these pathways remains poorly understood. Here, I present an integrative approach using multi-omics data to functionally characterize germline variants and explore the heterogeneous landscape of somatic mutations, with the aim of establish mechanistic links between functional variants and the disruption of cancer-related biological processes.

To enable the identification of functional variants, I initially performed a comprehensive characterization of functionally annotated transcriptional regulatory elements, establishing a hierarchy of 'consensus' elements across multiple levels of abstraction. This analysis generated a vast collection of consensus promoters, enhancers, and active enhancers, spanning 198 cell lines and 38 tissue types, with aggregate data providing global consensus definitions for each element type. Additionally, 'total binding affinity' method was employed, integrating 1000 Genomes Project genotype data and thousands of transcription factor binding motifs, to further characterize and functionally annotate these regulatory elements. The results generated from this analysis can be interactively explored and visualized through the CONREL web application.

To allow effective annotation of individual's ancestry, I developed and successfully employed an improved version of EthSEQ (version 3), an R package that provides a rapid and reliable pipeline for ancestry annotation. Accurate stratification of individual ancestry is essential for correctly interpreting the impact of genomic variations in associations studies. EthSEQ version 3 was successfully utilized to determine the genetic ancestry of over 500 pediatric patients diagnosed with 11 different tumor types, enabling further investigation into the genetic landscape of patients confidently identified as of European ancestry.

To further investigate into the interplay between germline and somatic variants, I conducted genome-wide association studies across 33 cancer types characterized by The Cancer Genome Atlas, using binary traits defined by somatic aberration profiles in ten oncogenic signaling pathways. Functional links between associated variants and somatic profiles were

investigated through cis-eQTL data to identify regulatory interactions with pathway-related genes. Additionally, using GWAS summary statistics I employed polygenic scores to examine the contribution of germline genetic variation to somatic molecular profiles, tumor subtypes, and clinical outcomes such as patient survival and tumor aggressiveness. Polygenic scores were validated using external data from PCAWG and CCLE datasets.

Lastly, to explore the heterogeneity of somatic mutational profiles, I employed a network-based approach to propagate somatic alterations through a molecular interaction network, aiming to reveal novel patterns of somatic alteration with potential significance in cancer. I then conducted a series of GWAS analyses, utilizing traits defined by combinations of these propagated somatic scores across genes involved in well-defined DNA repair pathways.

Overall, I demonstrate that germline genetics can describe patients' genetic liability to develop specific cancer molecular and clinical profiles. Understanding the functional roles of genetic variants can provide valuable insights into the biological mechanisms underlying a disease or trait.

List of abbreviations

AUC	Area under the curve
ChIP	Chromatin immunoprecipitation
CRE	Consensus regulatory element
eQTL	Expression quantitative trait locus
FDR	False discovery rate
GWAS	Genome-wide association study
LD	Linkage disequilibrium
MAF	Minor allele frequency
NBS^2	Network-Based Supervised Stratification
NGS	Next-generation sequencing
OS	Overall survival
PCA	Principal component analysis
PFI	Progression-free interval
PFM	Positional frequency matrix
PPI	Protein-protein interaction
pPSS	Pan-cancer polygenic somatic score
RWR	Random walk with restart
SNP	Single nucleotide polymorphism
TBA	Total binding affinity
TF	Transcription factor
TS	Targeted sequencing
TSS	Transcription start site
UMG	Upward mobility genes
WES	Whole-exome sequencing

Dataset:

CCLE	Cancer Cell Line Encyclopedia
CONREL	CONsensus Regulatory ELement
GTEx	Genotype-Tissue Expression
PCAWG	ICGC Pan-Cancer Analysis of Whole Genomes
PCNet	Parsimonious Composite Network
TCGA	The Cancer Genome Atlas

Introduction

Background

Cancers are complex diseases[1] driven by a combination of inherited genetic variants and somatic mutations that are accumulated during tumor progression, frequently disrupting crucial biological processes[2].

Over the past decades, advancement in genomic technologies have enabled the comprehensive characterization of disease-related alterations, leading to a deeper understanding of commonly dysregulated processes, and oncogenic pathways. The number of reported germline variants associated with cancer has grown considerably, thanks to various strategies[3]. Genome-wide linkage analysis, a method for tracking disease-related genetic markers in families with a strong history of cancer, has been particularly successful. More recently, genome-wide association studies (GWAS) have pinpointed hundreds of common and rare low-effect risk germline variants across multiple cancer types[4,5]. Through large-scale genomic analyses[6], rare germline variants have been identified linked to functional predisposition in 8% of adult cancer cases. The authors identified several germline variants with distinct associations. Variants within oncogenes tended to correlate with high gene expression, while variants within tumor suppressor genes were linked to low expression and loss of heterozygosity.

On the other side, a large number of somatic aberrations in tumor pathways are now used as hallmarks in many well-known forms of cancer[7]. However, the specific genes and pathways altered in cancer vary greatly across tumor types and individual patients. Some genes are recurrently altered and well-established as cancer drivers, while others are rarely or never mutated[8,9]. In[10] the authors investigate the patterns of alterations within established cancer pathways, comparing and contrasting these patterns across 33 different cancer types. Several crucial signaling pathways are observed frequently disrupted by genetic alterations in cancer. However, the alteration frequency within these pathways varies.

Advancements in next-generation sequencing have made large-scale genetic analysis both feasible and affordable. The availability of extensive sequencing data from both healthy individuals and cancer patients now allows for the identification of genetic factors that contribute to cancer susceptibility by examining both germline and somatic variations.

Lately, growing evidence supports an interplay between germline and somatic variants, demonstrating how inherited genetic predispositions can shape the somatic mutational landscape of tumors. In[11] the authors uncovers germline variants that have a direct impact on tumor evolution, either by promoting mutations in specific cancer genes or influencing the tissue of origin for tumor development. In our recent work[12], we provide evidence that germline genetics can shape the aberrant behavior of specific pathways, uncovering functional associations between SNPs and the biological alteration of oncogenic signaling pathways. In addition, very recently in[13] the authors investigated the impacts of germline cancer gene eQTLs on somatic mutations in a collection of cancer genes among >12,000 patients across 11 cancer types, demonstrating that germline variants regulate the expression of cancer genes and associate both with local and global somatic mutations' rates. Despite these studies, the functional links between germline variants and the somatic events of oncogenic pathways, and their impact on cancer genesis and progression remains largely unexplored.

Thesis aim

This thesis aims to fill the gap between individuals' genetic background and somatic events. Although, a substantial number of somatic aberrations in oncogenic signaling pathways have been observed to be linked with many well-known forms of cancer, the interaction landscape of germline variants and aberrant signaling pathways is still largely unknown.

To elucidate the impact of genetic variations within biological systems, it is important to correctly identify functional variants and understand their potential effects on specific biological pathways. Moreover, accurate identification of individuals in genetic studies is crucial for interpreting results and ensuring the correct attribution of variants to observed phenotypes.

In the first part of the thesis, I started identifying regulatory elements and their interactions with transcription. I implemented CONREL, a web resource to explore functionally annotated transcriptional regulatory elements across different cell lines and tissue types. Regulatory elements were constructed using a consensus approach, integrating patterns of various histone modifications derived from ChIP-seq experiments. Consensus regulatory elements were generated by aggregating ChIP-seq data at multiple levels of abstraction, resulting in a comprehensive collection of CREs across 198 cell lines and 38 tissue types, including global

consensus elements derived from the combined data. CONREL provides collections of TFs that show enriched TBAs across common alleles at different significance thresholds and can hence be used to elucidate regulatory mechanisms at specific regions in only a fraction of individuals.

Furthermore, I delved into identification and stratification of individual's ancestries for the correct interpretation of genetic and genomic profiling. I developed a new version of EthSEQ, a tool that provides a fast and automated computational workflow to annotate ancestry information from next-generation sequencing (NGS) data.

Given the growth of data generated by large-scale projects, optimizing software performance has become necessary. To address scalability challenges with large datasets, I developed a new version of EthSEQ optimized for efficient processing of extensive genetic data while ensuring compatibility with the latest VCF format. Critical steps in the workflow, which were previously bottlenecks in terms of memory usage and runtime, have been reimplemented in C++. This language is renowned for its speed and memory efficiency, particularly when compared to R, significantly enhancing the scalability and overall performance of EthSEQ. Moreover, a protocol paper has been published to describe detailed steps to perform ancestry analysis to a broader audience using different input file formats[14].

The main part of the thesis regards the identification, functional annotation, and characterization of inherited variants. Specifically, I performed a collection of genome-wide association studies (GWAS) across 33 cancer types characterized by TCGA and considering binary traits defined using a large collection of somatic aberration profiles across ten well-known oncogenic signaling pathways. I investigated functional links between associated variants and somatic profiles exploring cis-eQTL data to identify cis-regulatory interactions with genes directly within the pathways, or genes co-expressed and functionally close to genes within the pathways. I then leveraged polygenic scores approach to explore the contribution of germline genetic variation to somatic molecular profiles, tumor subtypes, and clinical outcomes including patient survival and tumor aggressiveness. Polygenic scores were validated using external data from PCAWG and CCLE datasets.

Finally, I investigated the heterogeneity of somatic mutational profiles aggregating tumor mutations in the context of molecular networks. In detail, I performed a network-based approach to propagate somatic alterations in cancer through a molecular interaction network to uncover low-rate mutated genes or new somatic alteration patterns that could

play an important role in cancer. Finally, I conducted a collection of GWAS analysis considering traits defined by combinations of these propagated somatic scores across well-defined DNA repair pathways genes.

Chapter 1. Exploring regulatory elements and TF:DNA interactions

Introduction

Cis-regulatory elements are regions of non-coding DNA that regulate transcription of neighboring genes. Promoters initiate gene transcription near the transcription start site (TSS) of a gene and consist of short sequences. Enhancer, on the other hand, influence gene transcription from various genomic positions relative to the gene(s) and can be of varying length.

Transcriptional regulation is a critical biological process that orchestrates gene activity and regulates the conversion of DNA to RNA (transcript). This process is finely tuned and involves physical interactions among multiple transcription factors (TFs) with core promoter elements and through distal enhancer elements. Understanding these interactions is crucial for deciphering gene regulatory networks. Various genomic factors, including sequence specificity and histone structure, influence how TFs bind to their target genes. The development of recent next-generation sequencing techniques has enabled detailed characterization of these genomic features. Recently, genome-wide chromatin annotations, based on histone modification patterns, have enabled the identification of potential regulatory elements across diverse human cell types[15–18]. Based on specific combination of different histone modification patterns it is possible to define distinct regulatory elements: trimethylation of H3 lysine 4 (H3K4me3) at promoters/transcription start sites, monomethylation of H3 lysine 4 (H3K4me1) at enhancers, and acetylation of H3 lysine 27 (H3K27ac) at active regulatory elements.

The Encyclopedia of DNA Elements (ENCODE)[19] and the NIH Roadmap Epigenomics Program[20] were established to identify all human genome functional elements. Both studies have performed a variety of assays to identify functional elements. Regulatory elements are mostly investigated though chromatin immunoprecipitation, followed by sequencing (ChIP-seq) experiments, uniformly curated, processed and validated, and publicly accessible though the ENCODE website (www.encodeproject.org). Several resources enable the investigation of regulatory elements, such as promoters and/or enhancers. This has been achieved through histone marker ChIP-seq experiments[21–23] or by analyzing their global

accumulation and integration[19,24]. Other resources utilize TF ChIP-seq data to explore potential interactions between transcription factors and DNA[25,26].

TFs are a class of proteins that play a vital role in gene regulation by binding to specific DNA sequences at enhancer or promoter regions. TF binding sites are short and usually degenerated sequences. The human genome encodes thousands of different TFs, which exhibit marked selectivity in their DNA binding and demonstrate a preference for specific sequences that can be over 1000-fold higher compared to others. This remarkable specificity allows a single TF to regulate distinct genes in different cell types, highlighting the dynamic nature of gene regulatory networks within an organism. A model summarizing the preferred DNA-binding sequences of a TF, is often represented by a positional frequency matrix (PFM). This matrix captures the nucleotide frequency distribution at each position within the TF binding site. Scores derived from PFM quantify the similarity between a DNA sequence and the TF's binding motif. While most of the methods to date predict TF:DNA interactions when these scores exceeds a predetermined threshold, recent advancements propose an alternative cutoff-independent methods for TF binding prediction[27,28]. Among them, an effective method considers the total binding affinity (TBA) of a sequence[29,30], which evaluate the entire sequence incorporating both high- and low-affinity binding sites, leading to a more accurate prediction of TF binding.

Here, I implemented CONREL (CONsensus Regulatory ELement), a web application for exploring regulatory elements across the human genome[31]. Employing a 'consensus' approach, I have implemented a workflow to build regulatory elements and provide annotations of TFs with enriched TBAs. By integrating data from multiple experiments, tissue types and cell lines, CONREL characterizes regulatory elements conserved across various conditions. Specifically, I combined ENCODE peak regions data across sample replicates and multiple experiments. For each cell line, 'consensus regions' for regulatory elements (CREs) are computed integrating TSS data. Then, the consensus regions are combined across similar tissues and across all cell lines to create a comprehensive map. Finally, I characterized all tissue and global regions by identifying all TFs showing enriched TBA and by determining the fraction of common alleles among 1000 Genomes Project individuals and Mouse Genomes Project strains that support TFs TBA enrichment in human and mouse respectively.

Initially, I implemented CONREL using the GRCh37 version of the human genome assembly. Since then, I have expanded CONREL to include the last GRCh38 human genome assembly.

Most recently, I have supervised and contributed to the expansion of CONREL to include a mouse model organism. The web application now facilitates the exploration of annotated CREs derived from both the human and mouse genomes.

CONREL offers a unique resource to explore regulatory elements and their functional properties across different genomic loci, genes, cell lines, and tissue types, filling a gap in the comprehensive landscape of TF TBAs across the human and mouse genomes.

Overview of data analysis workflow

I implemented and performed the computational workflow depicted in Figure 1.1 to identify robust annotated consensus regions for transcriptional regulatory elements in the human and mouse genomes.

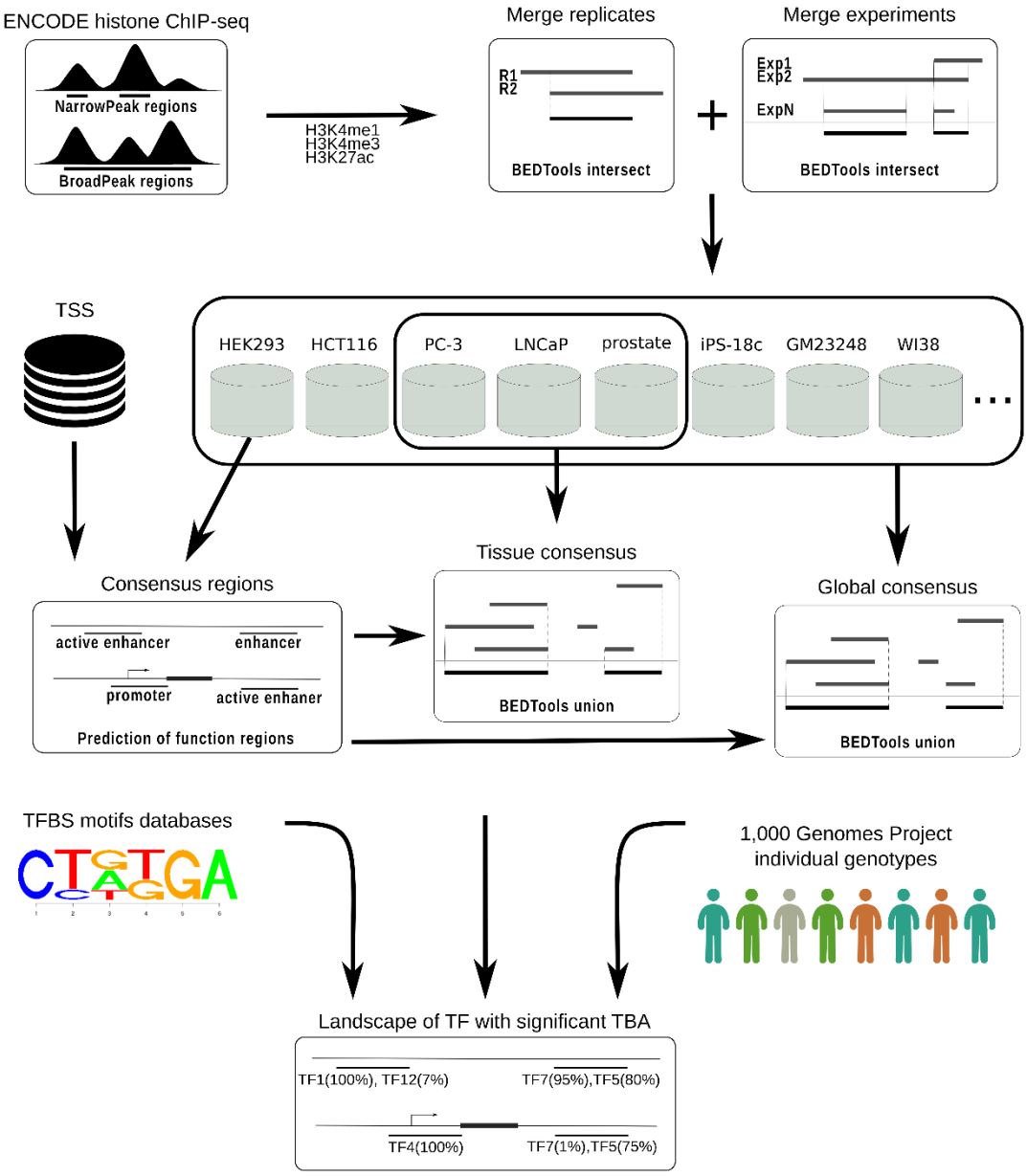


Figure 1.1 CONREL workflow for the identification of consensus regulatory regions (CRE) and transcription binding affinity annotations (TBA).

Results

Landscape of human transcriptional CREs

CONREL provides a vast collection of global and specific CREs for 38 tissue types and 198 different cell lines. This collection is based on over 1,000 ChIP-seq experiments from the ENCODE project. ENCODE provides peak data in two distinct formats: narrow and broad peaks, each computed using the peak calling tools with different thresholds. Notably, for certain experiments, both narrow and broad peak data are available, while for others, only one of the two formats is present. To ensure a comprehensive analysis, I consider both narrow and broad peak data separately.

Table 1.1 summarizes the global number of CREs identified for promoters, enhancers, and active enhancers, along with the corresponding percentage of the genome spanned by these regions for both narrow and broad peak data. Figure 1.2 expands upon these statistics by providing distribution plots for tissue- and cell line-specific CREs regarding their length distributions.

Table 1.1 Comparison of CREs coverage. Number of global CREs identified by CONREL using narrow and broad peak data, along with the percentage of the human genome covered by these CREs. It also includes a comparison with data from the ENCODE and RoadMap collections.

	Promoters		**Enhancers**		**Active enhancers**	
	No. of regions	**%**	**No. of regions**	**%**	**No. of regions**	**%**
Global narrowPeak	25 512	0.80	716 249	30.63	290 424	15.92
Global broadPeak	28 307	0.96	303 125	42.10	115 720	22.62
ENCODE	70 292	NA	399 124	NA	NA	NA
RoadMap	81 232	1.44	NA	NA	2 328 936	12.64

Global CRE promoters encompass roughly 1% of the human genome. In contrast, tissue- and cell line-specific CRE promoters exhibit greater variability, spanning a range of 0.27% to 0.67%. Interestingly, global CRE enhancers and active enhancers cover a more substantial

portion of the genome, ranging from 30% to 40% and 15% to 20%, respectively. However, tissue- and cell line-specific CREs for these elements demonstrate significant variability, with consensuses encompassing as little as 0.005% to a maximum of 15% of the genome.

A direct comparison of CREs derived from narrow and broad peak data (Figure 1.3) reveals a high degree of similarity for global CREs across all regulatory element types. Conversely, tissue-specific CREs display good concordance only for promoters, with substantial divergence observed for enhancers and active enhancers. This disparity likely reflects the limited and variable number of experiments available for specific tissue types within both narrow and broad peak datasets, where some tissues may have just a single experiment represented. Interestingly, I observed a significant correlation between the number of experiments and the degree of similarity considering both enhancers and active enhancers (correlation=0.63, p-value=5.56e-04, and correlation=0.51, p-value=8.31e-03, respectively). As examples, among all the tissue-specific CREs and regulatory element types, glia tissue for enhancer displays the lowest similarity between narrow and broad peak data. In this specific case, data considering narrow peak were available from two different glial cell lines (i.e. mid-neurogenesis radial glial cells and radial glial cell), with two independent experiments and

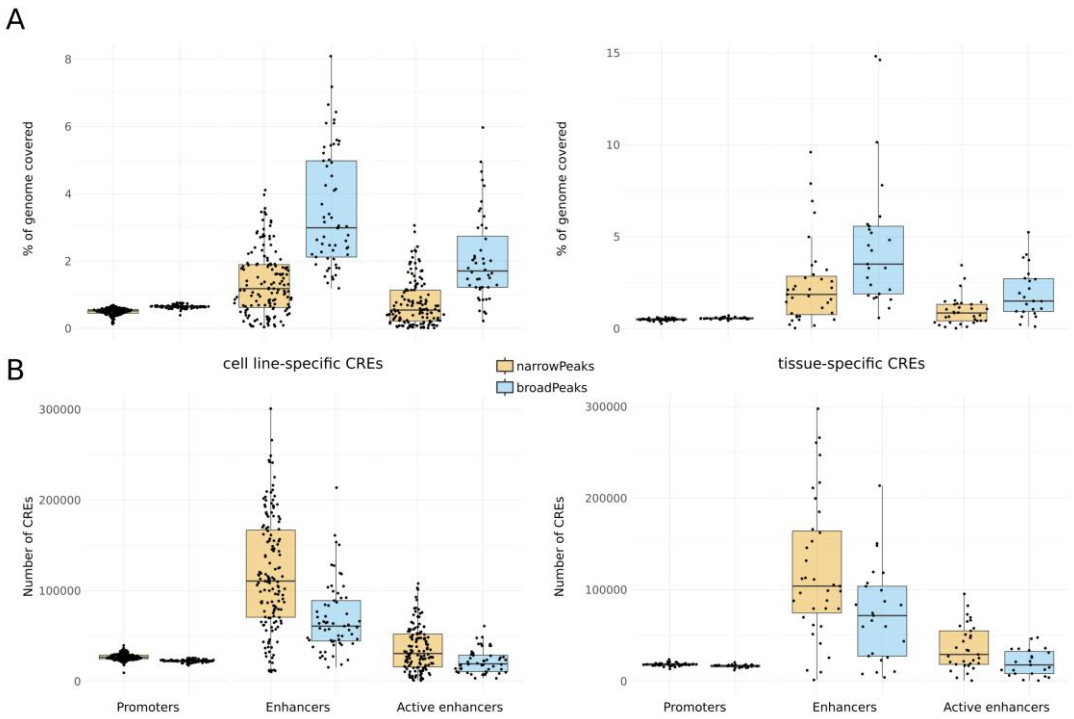


Figure 1.2 Comparison of CREs coverage for cell lines and tissues. Distribution of fractions of the human genome covered (A) and number of CREs (B) for CONREL cell lines (left) and tissue (right) consensus regions computed using both narrow and broad peak data.

two repeat measurements for each experiment. However, considering the broad peak, data were available only from astrocyte cell line, with three experiments but no repeat measurements. In contrast, muscle tissue-specific CREs exhibit some of the greatest similarity between narrow and broad peak data among promoters, enhancers, and active enhancers (with 88%, 32%, and 23% similarity, respectively). Actually, narrow peak data were available for 10 distinct cell lines originating from muscle tissue. Interestingly, broad peak data was also available for 5 cell lines, all these 5 were also included and available from narrow peak data, which may explain the higher similarity in muscle tissue observed compared to glia tissue.

Table 1.1 additionally integrates global regulatory element annotations derived from both ENCODE and RoadMap projects. Notably, while the number of globally annotated regions exhibits some variation across these three resources, the overall percentage of the genome covered by these regions remains comparable. Of note, this analysis is the only among the three annotations to offer data at both tissue and cell line level. This expansion facilitates a more comprehensive analysis at different abstraction levels of biological complexity.

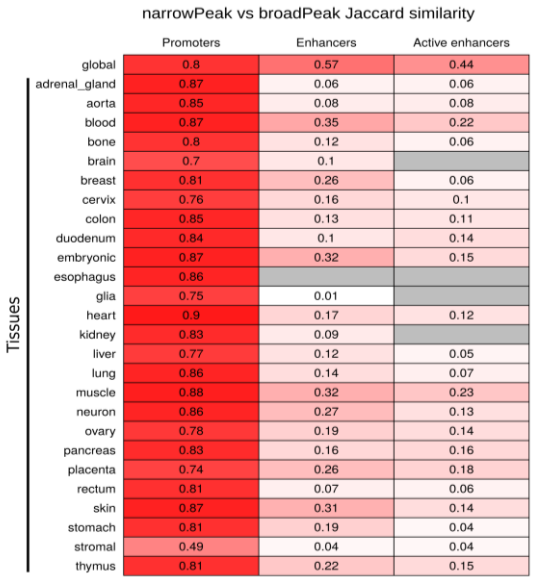


Figure 1.3 Jaccard similarity of CREs. Similarity comparison between narrow and broad peak data. Intensity in the red color of red areas represent higher overlap, while gray areas indicate no data for comparison.

Comparison with other regulatory elements resources

Given the absence of a definitive benchmark to validate our CREs, I decided to compare our global annotations with regulatory elements identified by other established resources. Specifically, for promoter annotations, I compared both narrow and broad global CREs to SCREEN[32], Ensembl[21], and GeneHancer[24]. For enhancer and active enhancer annotations, the comparison included EnhancerAtlas[22], DENdb[23], SCREEN, Ensembl, and GeneHancer. All regulatory region collections were converted into a uniform BED format. When necessary, coordinates were transformed to the human genome assembly GRCh37 using the UCSC Genome Browser's liftOver tool and chain file.

I employed an asymmetric pairwise comparison to calculate two distinct coefficients for each resource pair: 1) the percentage of regions from one resource that overlap with regions from the other, and 2) the ratio between the portion covered by one resource and the portion of the genome covered by both resources.

Figure 1.4A shows the pairwise comparison results, revealing an average promoter overlap of approximately 75% across all resource comparisons (excluding Ensembl). This indicates a generally good level of agreement among the promoter annotations provided by most resources. Additionally, genome coverage analysis, detailed in Table 1.2, reflects the observed concordance while accounting for the inherent differences in the size of the genome covered by each annotation (approximately 1% for CONREL, 2% for GeneHancer and Ensembl, and 0.3% for SCREEN).

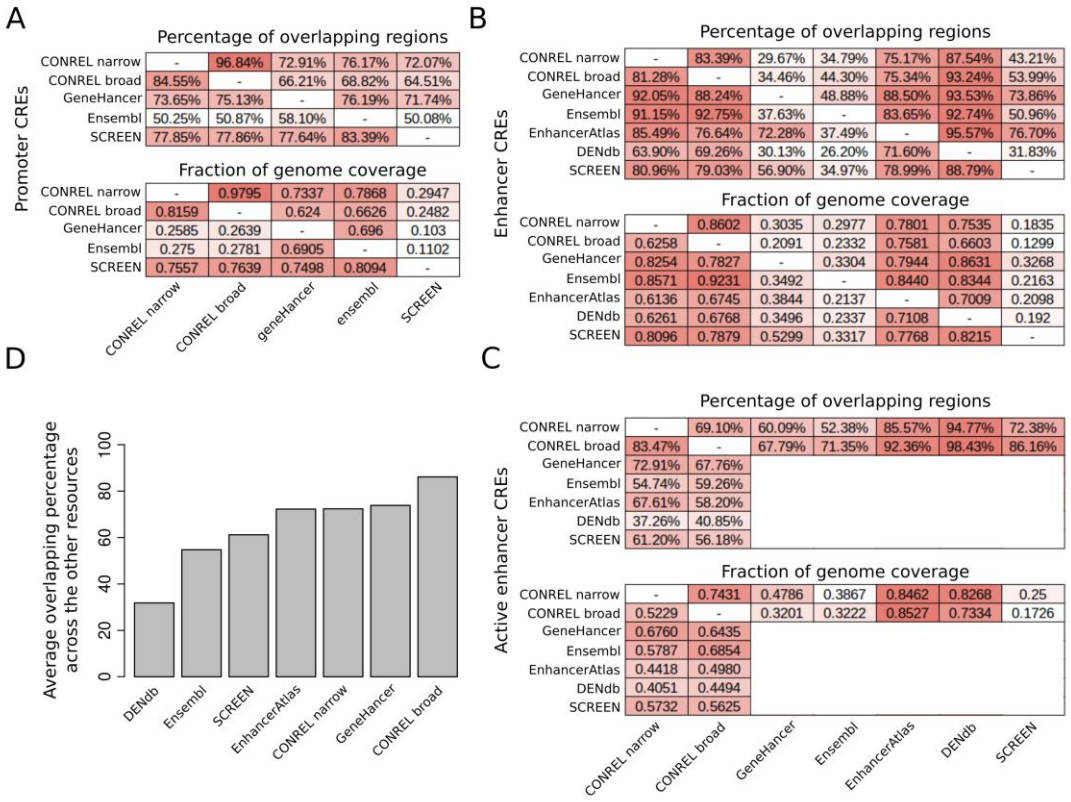


Figure 1.4 Comparison of CONREL CREs with other regulatory elements resources. A comparative analysis of Consensus Regulatory Elements (CREs) derived from the CONREL database with those from established regulatory element resources. A-C) Pairwise comparisons between all resources for promoter (A), enhancer (B), and active enhancer (C), both in terms of the percentage of promoters shared and the genomic coverage captured. The top matrix quantifies the percentage of promoters in a given resource (row) that overlap with regions in another (column), while the bottom matrix indicates the proportion of genomic space covered by regions in one resource that is also encompassed by those in another. (D) Average percentage of CONREL active enhancer and GeneHancer, Ensembl, EnhancerAtlas, DENdb and SCREEN enhancers that have overlapping with each of the other resources.

It is remarkable that, as far as we know, CONREL stands out as the sole resource offering promoter annotations at three distinct resolution levels (global, tissue, and cell line). Furthermore, CONREL differentiates between annotations derived from narrow and broad peak data. While Ensembl provides global annotations and annotations for individual experiments, both SCREEN and GeneHancer solely offer global annotations. The comparison of enhancer annotations reveals a generally good level of concordance between CONREL and other resources, evident in both the percentage of overlapping regions and the shared fraction of genome coverage (Figure 1.4B,C).

Table 1.2 Human **genome coverage of all resources considered in the comparison analysis.** Total coverage of the genome, corresponding fraction, and number of consensus regulatory elements for CONREL global consensus regions computed using both narrow and broad peak data and all other data collections.

	Genome coverage	Genome coverage fraction	Number of regions
Promoter			
CONREL narrow (global)	25,488,643	0.80%	25,512
CONREL broad (global)	30,599,944	0.96%	28,307
GeneHancer	72,447,469	2.26%	23,725
Ensembl	72,858,670	2.28%	35,035
SCREEN	9,941,504	0.31%	34,734
Enhancer			
CONREL narrow (global)	980,080,993	30.63%	716,249
CONREL broad (global)	1,347,251,757	42.10%	303,125
geneHancer	360,359,358	11.26%	246,906
ensembl	340,176,755	10.63%	273,175
enhancerAtlas	1,818,995,370	56.84%	2,464,777
DENdb	1,383,043,500	43.22%	3,506,396
SCREEN	222,228,613	6.94%	808,157
Active enhancer			
CONREL narrow (global)	509,458,088	15.92%	290,424
CONREL broad (global)	723,939,328	22.62%	115,720

Overall, a higher degree of heterogeneity is observed among the different enhancer resources. While the genome coverage of CONREL enhancers (~30%) and active enhancers (~20%) is more conservative compared to EnhancerAtlas (~55%) and DENdb (~45%) (Table 1.2), enhancers from SCREEN, GeneHancer, and Ensembl exhibit the most conservative coverage, encompassing roughly 10% of the genome. Although conservative annotations might mitigate the presence of artifacts, the overlap between SCREEN, GeneHancer, and Ensembl is not optimal. This suggests a potential divergence in how these resources functionally characterize specific genomic regions.

It is important to note that, as illustrated in Figure 1.4D, CONREL active enhancers display the highest average representation across all other resources. Additionally, CONREL remains the only resource offering enhancer annotations at three resolution levels (global, tissue, and cell line) and differentiating between annotations derived from narrow and broad peak data.

To facilitate the exploration of relationships between our CREs and regulatory elements identified by other resources, I integrated annotations into CONREL web application, highlighting all identified overlaps within consensus regions. This allows users browsing global CREs through the web application to readily identify which other resources support the specific regulatory elements.

Global and allele-specific distribution of transcription binding affinities across human CREs

Table 1.3 summarizes for each CRE type across various p-value cutoffs, the average number of transcription factors with enriched total binding affinity per CRE, alongside the percentage of regions exhibiting enriched TFs.

Employing broad peak data with the most stringent statistical threshold, we observed enriched TBAs in approximately 95% of promoters, 85% of enhancers, and 95% of active enhancers. Conversely, utilizing narrow peak data yielded lower percentages, with enriched TBAs detected in roughly 80% of promoters, 60% of enhancers, and 70% of active enhancers. Utilizing a more relaxed statistical approach resulted in enriched TBAs identified within all CREs.

Table 1.3 Enriched TFs through CREs. Mean number of TFs with enriched TBAs at promoter, enhancer, and active enhancer CREs at different significance cutoff, and percentage of CREs with at least one enriched TF TBAs.

	TBA significance p-value cutoff	Promoters		Enhancers		Active enhancers	
		Mean number of TF	CREs %	Mean number of TF	CREs %	Mean number of TF	CREs %
Narrow peak	1e-02	281	100	256	100	286	100
	1e-03	125	99.9	94	99.8	116	99.8
	1e-04	76	93.7	53	84.2	70	86.5
	1e-05	51	83.9	34	59	46	68.5
Broad peak	1e-02	302	100	431	100	522	100
	1e-03	140	100	231	99.9	299	100
	1e-04	86	97.8	164	93.3	218	98
	1e-05	56	94.7	123	84.4	166	95.6

By characterizing common CRE alleles using 1000 Genomes Project genotype data, we were able to identify TF TBAs that exhibited enrichment or depletion in only a subset of alleles. This finding suggests the potential existence of allele-specific regulatory mechanisms. For instance, analyzing global CREs revealed that roughly 1% and 4% of promoter and active enhancer regions, respectively, displayed TF TBAs enriched in less than 10% of common alleles from the 1000 Genomes Project when employing the most stringent significance cutoff. The full distribution of TF TBA enrichment scores across CRE promoters and active enhancers is presented in Figure 1.5.

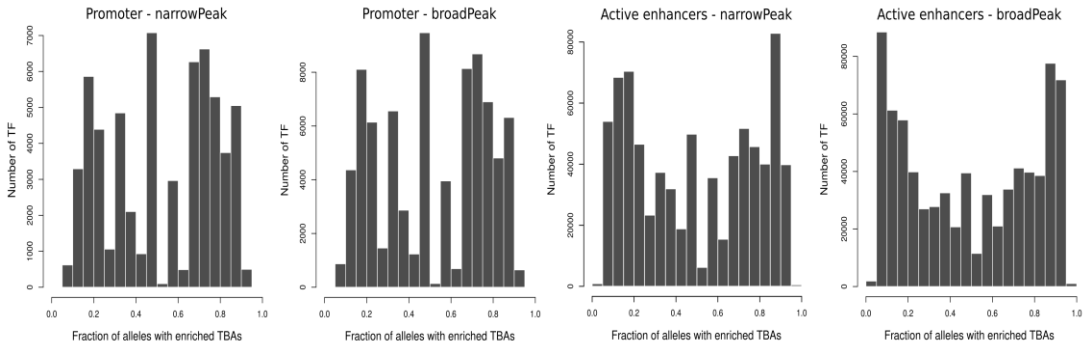


Figure 1.5 Allele-specific TF TBAs enrichment. Distribution of TFs exhibiting a specific enrichment fraction for common alleles from the 1000 Genomes Project across CRE promoters and active enhancers.

Comparative analysis of TBA annotations and TF-target regulatory relationships resources

We investigated how TBA annotations capture transcriptional regulatory networks. We retrieve a list of manually curated TF-target relations from the TRRUST database[33], focusing on those involving TFs in our data. We analyzed CONREL global promoter and active enhancer regions, looking for the closest protein-coding genes nearby CREs enriching a TF predicted by TBA with the strictest criteria (p-value cutoff 1e-05) and present in TRRUST. Our promoter TBA annotations (Figure 1.6A) explained about 15% of TRRUST relationships, increasing to 35.5% when including active enhancers (Figure 1.6B,C).

While CONREL TBA annotations identified many more relationships (Table 1.3) than TRRUST (around 7300), they still captured a statistically significant portion of TRRUST data. Specifically, we shuffled regulators and targets in TRRUST randomly 1000 times and compared the overlap with our results. We observed a statistically significant enrichment for both promoters (p-values < 0.001 for both broad and narrow data) and active enhancers (p-value = 0.001 and p-value = 0.012 for narrow and broad peak data respectively).

While TRRUST and CONREL rely on distinct input data, the results suggest CONREL has the potential to analyze the structure of transcriptional regulatory networks.

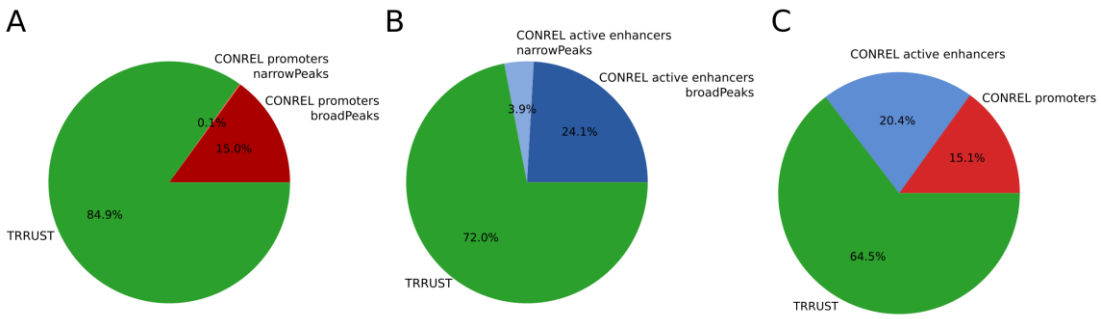


Figure 1.6 TRRUST transcriptional regulatory relationships captured by CONREL. The cumulative fraction of TRRUST relationship captured by CONREL-derived promoters (A) and active enhancers (B), considering both broadPeak and narrowPeak-derived CREs. (C) The combined contribution of promoters and active enhancers capture by CONREL.

Web-interface implementation and usage example in human

I implemented CONREL, a web application to easily explore CREs and their annotations about TF:DNA interactions. CONREL is developed in R (v3.6.1) and the Shiny package (v1.3.2) running on a Shiny server (v1.5.12.933). The user interface is accessed through a web browser. Several R packages are utilized for various functionalities: '*shinyDashBoardPlus*' for

interface design, '*TnT*' for genome browser generation, '*biovizBase*' and '*GenomicFeatures*' for genomic data utilities, and '*EnsDb.Hsapiens.v75*', '*EnsDb.Hsapiens.v86*' for providing genomic annotations for human reference genomes GRCh37 and GRCh38, respectively.

For deployment, CONREL utilizes a virtual server with 4GB RAM, 40GB disk space, and 2 CPUs running Ubuntu 16.04 LTS Linux. The application is containerized within a Singularity image, which is available for download alongside configuration scripts to enable local server execution. The source code for the web interface can be found on GitHub at https://github.com/cibiobcg/CONREL.

The user interface, accessed through a web browser, facilitates the exploration and analysis of regulatory elements within a genomic context. Users need to define the gene name or a region of interest (Figure 1.7A) and then select at least three mandatory inputs from the

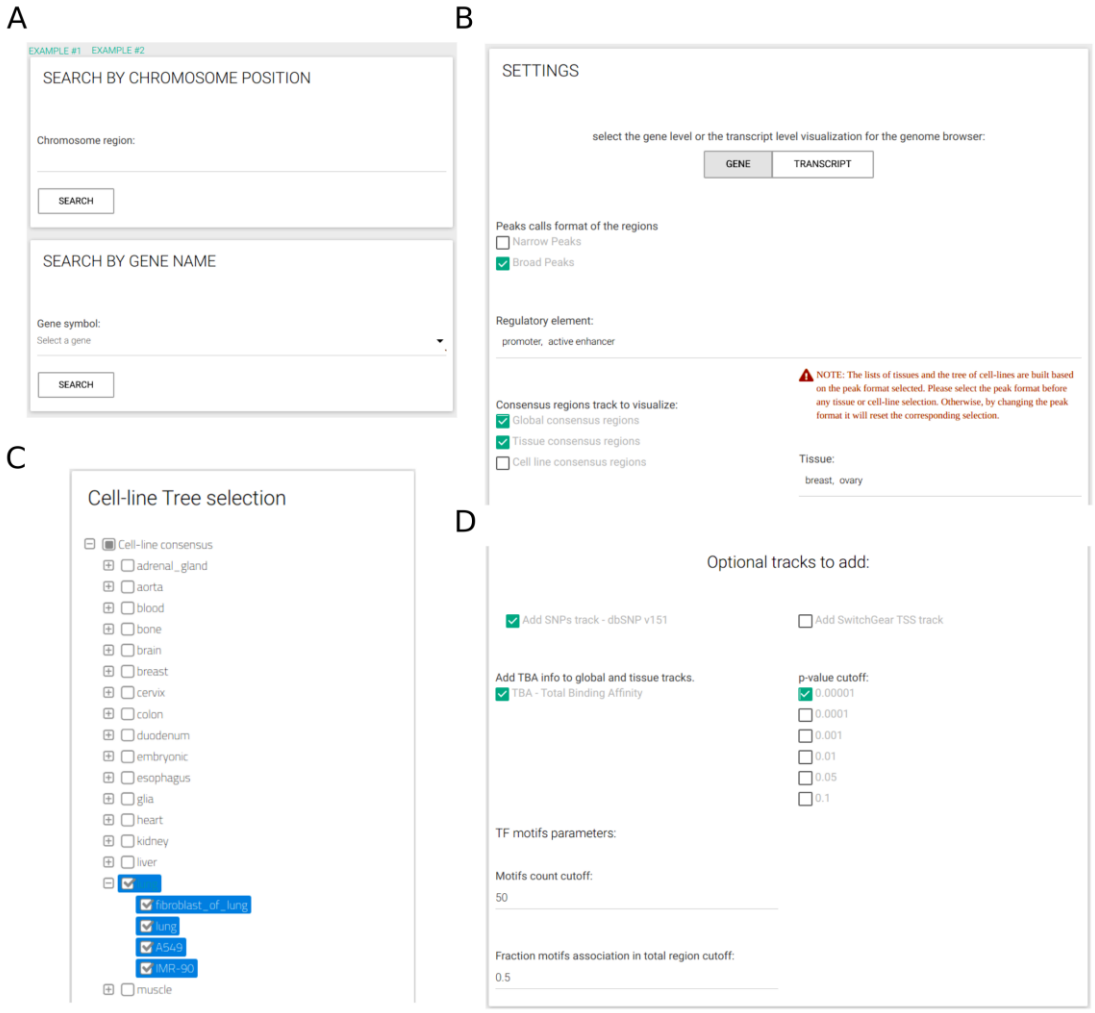


Figure 1.7 CONREL interface. (A) Search tab allows to select a specific genomic region or a gene name. Input tabs allow for the selection of source peak data and types of CRE to be displayed (B) using, when needed, also a cell line selection tree when activated (C) and TBA statistical filters (D).

available options (Figure 1.7B): (i) narrow or broad peak format for ChIP-seq data, (ii) at least one type of regulatory element (e.g., promoter, enhancer), and (iii) at least one CRE. For both tissue and cell line CREs, users need to select at least one CRE out of all available CREs. A selection tree displaying all cell lines categorized by tissue of origin is used to facilitate the selection of specific cell lines of interest (Figure 1.7C).

Additional tracks and parameters can be selected, including for example TBA significance threshold (default: 1e-05) and two filters for transcription factor position weight matrices (PFMs) used in the analysis (Figure 1.7D). These PFM filters allow users to exclude potentially low-confidence motifs by setting a minimum number of sequences defining a PFM (default: 50) and a maximum fraction of CREs an enriched PFM can be associated with (default: 0.50).

Upon selection, a genome browser tab is displayed (Figure 1.8). This browser allows users to navigate the surrounding genomic region (±1 Mbp) and visualize various features, including genes, transcripts, and consensus regions. Selecting a specific CRE within the browser, the bottom panels display detailed information: genomic coordinates, strand, the number of experiments used to build the consensus, and all associated transcription factor TBA

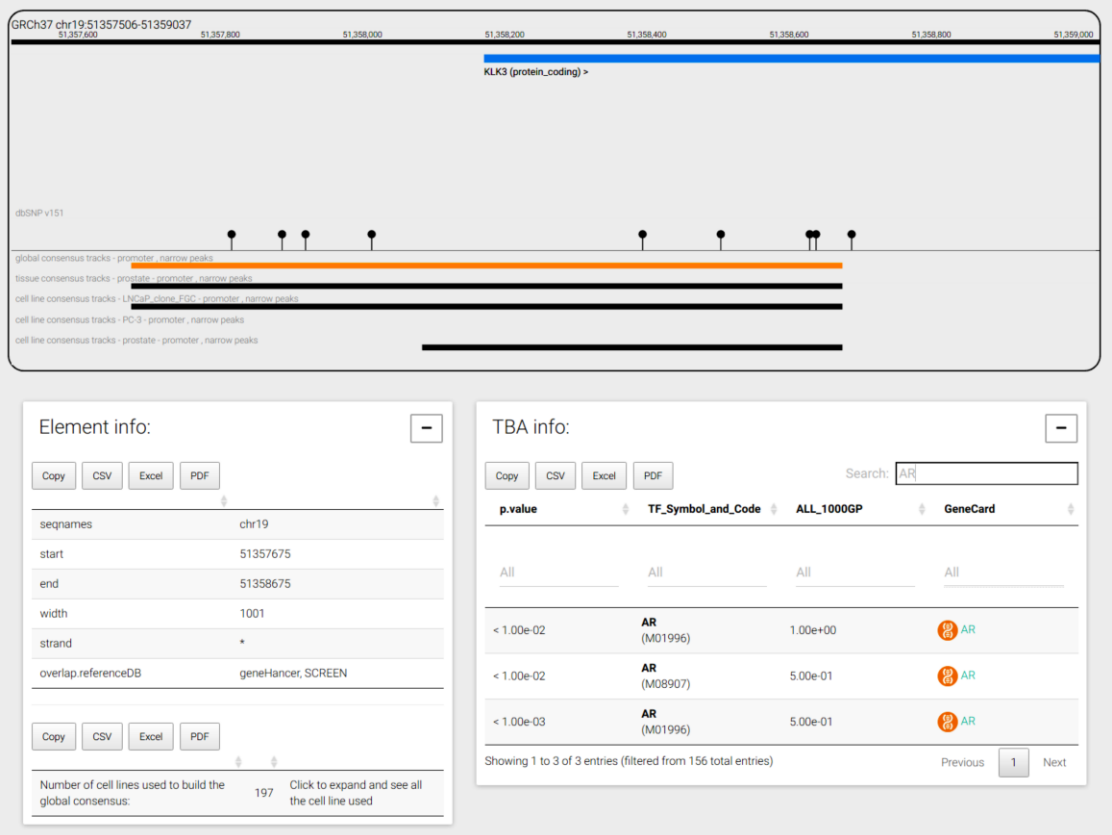


Figure 1.8 CONREL navigation webpage. The genome browser allows users to navigate specific regions or individual genes (top). Additional panels offer a deeper overview into CRE and TBA information (bottom).

enrichments. Figure 1.8 exemplifies this functionality by highlighting TBA enrichments for androgen receptor (*AR*) PFMs below a significance of 0.01 for the promoter region of *KLK3* gene, also known as Prostate-Specific Antigen (*PSA*). Interestingly, only half of the common alleles from the 1000 Genomes Project exhibit significant *AR* enrichment, suggesting the influence of single-nucleotide polymorphisms (SNPs) on the PFM score within the promoter region of *KLK3*. Moreover, I deeply investigated promoter CREs near the start site of the *KLK3* gene. Figure 1.8 shows narrow peak for global data, prostate-specific tissue, and prostate cell lines. Interestingly, both the global and tissue-specific data show a consensus promoter region at the beginning of the *KLK3* gene. Additionally, the data from LNCaP and prostate cell lines align in terms of the promoter regions, while the PC3 cell line lacks this specific CRE. This is interesting because according to scientific literature[34], LNCaP cells express the PSA protein, whereas PC3 cells do not.

Lastly, users can generate the link to the DNA sequence of the displayed genomic window. The interface displays additional functionalities for copying or downloading the selected consensus region information or the TBA information using different file formats (CSV, Excel, or PDF).

Mouse data integration and usage example

We extended CONREL to include mouse data, following the same workflow used for human data. ChIP-seq data from ENCODE for 37 cell lines across 18 tissues were utilized. As shown in Table 1.4, promoters cover approximately 0.5% of the genome, while enhancer regions range from 11 to 24% for narrow- and broad-peak, respectively. Coverage is more heterogeneous across cell lines and tissues (data not shown), with percentages ranging from 0.2 to 10%.

Following the approach used for human CRE comparison, we decided to conduct a comparative analysis of our global annotations with regulatory elements identified by other established resources. For promoter annotations, both narrow and broad global CREs were compared to SCREEN, Ensembl, and EPDnew[35]. Regarding enhancer and active enhancer annotations, the comparison encompassed EnhancerAtlas, SCREEN, and Ensembl. It is important to note that EPDnew represents promoter-like elements as single genomic coordinates. This allows for the comparison of overlapping regions but precludes the calculation of the shared fraction of the genome covered by two resources.

As previously observed with human genome, CONREL is more conservative than enhancerAtlas, while SCREEN and Ensembl are the most conservative amongst all analyzed resources in defining consensus regions. The enhancerAtlas database demonstrates a remarkable abundance of putative enhancer regions, covering approximately 82% of the genome. Of note, this comprehensive resource integrates enhancer predictions derived from a collection of 241 different cell lines and tissues obtained using 12 distinct high-throughput experimental techniques.

Table 1.4 Mouse **genome coverage of all resources considered in the comparison analysis**. Total coverage of the genome, corresponding fraction, and number of consensus regulatory elements for CONREL global consensus regions computed using both narrow and broad peak data and all other data collections.

	Genome coverage	Genome coverage fraction	Number of regions
Promoter			
CONREL narrow (global)	13,423,798	0.49%	14,445
CONREL broad (global)	13,838,769	0.51%	13,794
EPDnew	NA	NA	25,111
Ensembl	52,478,880	1.94%	25,110
SCREEN	6,935,143	0.26%	23,271
Enhancer			
CONREL narrow (global)	314,139,915	11.63%	456,313
CONREL broad (global)	657,282,252	24.34%	187,884
SCREEN	72,377,655	2.68%	262,393
Ensembl	54,215,051	2.01%	69,963
enhancerAtlas	2,225,958,966	82.44%	520,179
Active enhancer			
CONREL narrow (global)	103,100,009	3.82%	95,056
CONREL broad (global)	214,572,344	7.95%	47,887

Figure 1.9 presents a pairwise comparison of CRE annotations across various resources, assessing both the percentage of overlapping regions and the shared fraction of the genome covered by two resources. The results demonstrate a generally high level of agreement among promoter annotations across most resources. Regarding enhancer annotations,

CONREL exhibits a good degree of concordance with other resources, as evidenced by the substantial percentage of overlapping regions. However, a relatively low shared fraction of genome coverage is observed for all resources when compared to both SCREEN and Ensembl datasets. Conversely, a high fraction of shared genome coverage is observed when compared to EnhancerAtlas. This discrepancy can be attributed to the varying total genome coverage of each annotation, as detailed in Table 1.4 (approximately 2% for SCREEN and Ensembl, and 82% for EnhancerAtlas). These findings highlight the importance of considering both the extent of overlap and the overall genomic context when evaluating the concordance of CRE annotations across different resources. Notably, both the absolute number of annotated regions and the overall proportion of the genome covered by these annotations varies across these three resources. This divergence is likely attributable to the high diversity in input data utilized to construct the consensus regions for each dataset. Importantly, our analysis offers data at both the tissue and cell line levels, enabling a more comprehensive assessment of regulatory element annotations across different biological

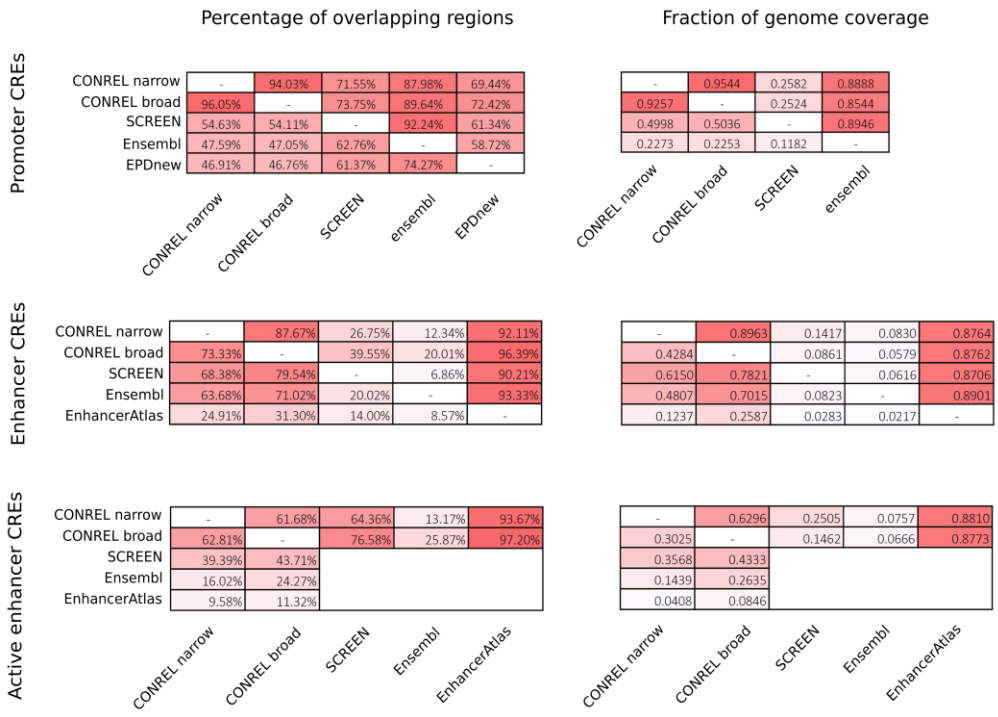


Figure 1.9 Comparison of CONREL CREs with other regulatory elements resources. A pairwise comparative analysis of all CRE types derived from the CONREL database with those from established regulatory element resources. The top matrix quantifies the percentage of promoters in a given resource (row) that overlap with regions in another (column), while the bottom matrix indicates the proportion of genomic space covered by regions in one resource that is also encompassed by those in another.

contexts and levels of complexity. This expanded scope facilitates a deeper understanding of the dynamic and context-specific nature of gene regulation.

We computed TBA scores for the mouse data using the same pipeline applied to human CREs. Following the integration of these results into our web application, we explored the data to analyze the presence of CREs across various cell lines and tissues of specific genes. As an illustrative example, we investigate the presence and characteristics of a promoter consensus region for the Hepatocyte Growth Factor (*Hgf*) gene (Figure 1.10). *Hgf* plays a crucial role in stimulating epithelial cell proliferation, motility, morphogenesis, and angiogenesis across diverse organ systems. Moreover, endogenous Hgf is essential for the self-repair mechanisms of injured tissues, including the liver, kidneys, and lungs[36]. Given the liver's high proportion (80%) of hepatocytes, *Hgf* gene likely plays a pivotal role in hepatic function. Furthermore, the widespread tissue expression of *Hgf* suggests the potential presence of promoter annotations within various tissues. As expected, our analysis revealed promoter annotations for *Hgf* gene in nearly all tissue specific CREs, apart from placental tissue (Figure 1.10A). This finding aligns with observations from cell lines (data not shown). Intriguingly, we detected the Hepatocyte Growth Factor activator (*Hgfac*) promoter annotation exclusively within liver specific CREs (Figure 1.10B). These results suggest a widespread distribution of inactive *Hgf*, consistent with existing literature, while potentially limiting its activation to the liver due to the liver specific co-expression of *Hgfac*.

Figure 1.10 CONREL navigation webpage for case example in mouse. The genome browser showing all available tissue specific CREs at promoter consensus regions of *Hgf* gene (A) and *Hgfac* gene (B).

Methods

CRE identification in human genome

ChIP-seq data from ENCODE, based on the GRCh37 assembly, was downloaded for cell lines with H3K4me1, H3K4me3, or H3K27ac histone markers peak data available. Data was obtained for both narrowPeak and broadPeak formats.

BroadPeak peaks were filtered based on a p-value threshold of less than 0.01, while no filters were applied to narrowPeak peaks as they all had p-values below 0.01. The peak files were converted into BrowserExtensibleData (BED) format files, representing each peak region with chromosome and genomic position information (BED3 format).

Peak regions derived from sample replicates were merged, preserving only overlapping regions. Afterward, merged peak regions from different experiments for the same cell line were combined, considering only regions overlapping in at least two experiments. Consensus regions for each cell line were computed based on available markers, defining promoters as regions occupied by H3K4me3 within 1 kb of a TSS, and enhancers as regions occupied by H3K4me1, depleted of H3K4me3, and at least 1 kb away from TSS. Active enhancers were each enhancer consensus region overlapping with H3K27ac peaks. TSS data were obtained from the UCSC Genome Browser, retaining only TSS with scores ≥ 10.

Consensus regions were also characterized at tissue and global levels by merging regions across cell lines from the same tissue or across all considered cell lines. The consensus was computed by considering regions overlapping in at least two cell lines and retaining the union of overlapping regions.

Due to limited availability of ChIP-seq experiments available in the ENCODE dataset and aligned to the GRCh38 human genome assembly at the time of implementation, I employed liftOver[37] methodology to translate all global, tissue and cell line specific CREs obtained using the GRCh37 genome assembly to their corresponding GRCh38 coordinates.

TBA scores at human CREs

To characterize the consensus regions, I performed an *ad hoc* computational strategy using the Total Binding Affinity (TBA) approach. This method quantifies the affinity of a DNA sequence for a TF described by a PFM with a single score (Equation 1).

Specifically, I computed TBA scores across all CREs using TF DNA-binding site motifs from public databases. TBA scores were computed both for the reference genome sequence and

for common alleles identified from the 1000 Genomes Project. Statistical significance of TBA scores was determined using a permutation approach and pre-computed thresholds based on a reference distribution of normalized TBA scores.

Formally, the TBA a_{rw} of a sequence r for a PFM w is given by:

$$a_{rw} = \sum_{i=1}^{L-l-1} max\left(\prod_{j=1}^{l} \frac{P(w_j, r_{i+j-1})}{P(b, r_{i+j-1})}, \prod_{j=1}^{l} \frac{P(w_{l-j+1}, r'_{i+j-1})}{P(b, r'_{i+j-1})}\right) \quad [1]$$

where l is the length of the PFM w, L is the length of the sequence r, r_i is the nucleotide at the position i of the sequence r on the plus strand, r'_i is the nucleotide in the same position but on the other strand, $P(w_j, r_i)$ is the probability to observe the given nucleotide $r_i$at the position j of the PFM w and $P(b, r_i)$ is the background probability to observe the same nucleotide r_i.

TBA method produces a single score considering binding sites of all possible affinities and weights them based on a physical model of TF:DNA interactions. Initially applied to study yeast transcriptional regulation, TBA has more recently been used to explore the evolution of cis-regulatory elements in humans and to detect and characterize Expression Quantitative Trait Loci (eQTLs).

To analyze TBA scores across all cis-regulatory elements (CREs), I collected 5424 unique TF DNA-binding site motifs in the form of PFM from public databases such as Jaspar[38], hPDI[39], SwissRegulon[40] and HOCOMOCO[41], and from TRANSFAC Professional[42].

I computed TBA scores for all TF PFMs across both tissue-specific and global CREs. For each combination of TF PFM and CRE, I computed TBA scores considering the CRE sequence described by both version of the human reference genomes (GRCh37and GRCh38), as well as TBA scores computed on all common alleles identified from individuals in the 1000 Genomes Project. Common alleles with an observed frequency >1% were retained for analysis.

To assess the statistical significance of a TBA score for a TF PFM at a specific CRE, I employed a permutation approach. Due to the extensive number of TBA scores computed across all global and tissue CREs (approximately 5.6e10), I implemented strategies to reduce computational costs of TBA significance calculation. TBA scores were normalized with respect to the corresponding CRE length, and significance was determined by comparing the

TBA value against a PFM-specific reference distribution of normalized TBA scores computed across 100,000 random genomic regions of varying lengths. Pre-computed TBA normalized score thresholds for different p-value cutoffs (ranging from 5e-02 to 1e-05) were utilized to determine TBA significance at various cutoffs, with the default cutoff in CONREL set at 1e-05 for stringent multiple hypothesis correction.

Mouse genome: CRE identification and TBA scores

Consensus regions and TBA scores were also characterized for mouse model. First, ChIP-seq data from ENCODE based mainly on MGSCv37 and GRCm38 assemblies were downloaded. All genomic coordinates based on MGSCv37 genome assembly were converted to their corresponding GRCm38 coordinates. Over 1,000 ChIP-seq replicates across various cell lines were downloaded, encompassing peak data for H3K4me1, H3K4me3, and/or H3K27ac histone markers in both broadPeak and narrowPeak formats. All peaks were filtered based on a p-value threshold of less than 0.01. The same method used for the human genome assembly was applied to the murine data to obtain consensus regulatory elements in mouse. To characterize these elements at both tissue-specific and global levels, consensus regions were merged across cell lines derived from the same tissue or across all cell lines considered in the analysis.

To analyze TBA scores across all mouse CREs, we collected 2159 unique TF DNA-binding site motifs in the form of PFM from public databases such as Jaspar[38], UniProbe[43], CIS-BP[44], Jolma et al.[45] and HOCOMOCO[41], and from TRANSFAC Professional[42]. Unlike the human 1000 Genomes Project, no equivalent comprehensive dataset exists for the laboratory mouse to facilitate the detailed characterization of genotype calls across diverse samples of mouse model. To address this, we employed an analysis of common alleles among common laboratory mouse strains to identify shared genetic variants. The Mouse Genomes Project[46] is an ongoing effort with the goal to comprehensively catalog genetic variants for common key mouse strains. The authors identified various small-scale genomic modifications, including single nucleotide polymorphisms and indels, relative to the C57BL/6J mouse reference genome. Genotype calls were retrieved for more than 78M SNPs and indels across 52 mouse samples for 36 distinct mouse strains. Then, we phased genotype calls using SHAPEIT2[47] to identify common alleles (frequency >1%). We computed TBA scores considering GRCm38 mouse reference genome assembly and common alleles identified

within the Mouse Genomes Project. TBA scores were computed for each PFM and CRE sequence across both tissue specific and global elements.

Discussion

In this chapter, I introduced CONREL, a web tool designed for exploring transcriptional cis-regulatory elements (CREs) and understanding TF:DNA interactions using TF total binding affinities (TBAs). Utilizing ENCODE ChIP-seq peak data, CONREL offers a comprehensive database of promoters, enhancers, and active enhancers, defined by combining histone markers H3K4me1, H3K4me3, and H3K27ac. While various resources exist for exploring ENCODE ChIP-seq data, CONREL stands out by aggregating experiments at different levels of abstraction, providing a unique collection of human CREs for 198 cell lines and 38 tissue types, mouse CREs for 37 cell lines and 18 tissue types, as well as global consensuses. I observed distinct similarities between narrow and broad peak CREs at tissue and cell-line levels, indicating the need for expanding input experiments to better characterize consensus regions while highlighting CONREL's effectiveness in integrating diverse CREs for deeper genomic exploration.

CONREL offers collections of TFs showing enriched TBAs at various significance thresholds for each regulatory element, aiding in the elucidation of regulatory mechanisms. Additionally, it provides information on TF TBA enrichment frequencies across common alleles in the 1000 Genomes Project, facilitating the identification of TFs regulating transcripts in specific individuals. Comparison with the TRRUST database suggests CONREL's utility in exploring transcriptional regulatory network structure and topology. Moreover, CONREL offers TF TBA enrichment frequencies information for 36 different mouse strains, providing identification of TFs regulating transcripts in specific mouse models, facilitating the identification of TFs regulating transcript within a limited fraction of mouse model strains.

Implemented as an R Shiny application, CONREL offers an intuitive interface for exploring all these data. This versatile resource provides comprehensive information for researchers interested in studying specific genomic regions or TFs, with all resources available for download. Future updates will focus on incorporating additional ChIP-seq experiments to reinforce CRE confidence and expand the range of supported transcriptional regulatory element types (e.g. poised enhancer, or silencer). Additionally, CONREL will be potentially

expanded integrating different animal models CREs. CONREL is freely accessible via web browser or through a downloadable singularity image, ensuring convenient usage for the wider scientific community.

Chapter 2. Analysis of Genetic Ancestry

Introduction

The advent of next-generation sequencing (NGS) has revolutionized the study of the genetic architecture of complex diseases, playing a key role in cancer research aiming to translate discoveries into clinical applications and personalized medicine efforts. In recent decades, genome-wide association studies (GWAS) have successfully identified thousands of common variants associated with human diseases and traits. However, these association variants often explain only a small fraction of the heritability and provide limited insights into the underlying functional mechanisms of disease. Consequently, many recent studies have shifted their focus to rare variants, which are more likely to exert direct functional effects on gene products. Due to the low frequency of these rare variants, large sample sizes and cost-effective sequencing approaches, such as whole-exome sequencing (WES) or targeted sequencing (TS), are favored approaches for exploring patient genomes. In this setting, a correct estimation of ancestry stratification of individuals is required to investigate results from GWAS studies and evaluate the importance of personal genomic variations[48]. Recent large-scale studies[49,50] have revealed a significant role of ancestry in influencing mutation rates, DNA methylation patterns, and mRNA expression levels. These findings emphasize the importance of considering ancestry information when investigating disease mechanisms and predicting responses to therapies. To address this, several model-based tools and tools based on Principal Component Analysis (PCA) have been realized and proposed so far[51–53]. Among them, EthSEQ[54] has been developed and used[49] for the rapid and automatic assignment of ancestry information to individuals based on their WES data.

The increasing availability and affordability of high-throughput genomic data have necessitated an upgrade of EthSEQ. Previous versions of EthSEQ forced the user to follow stringent input requirements, accepting VCF files only in a highly specific format. For instance, only positions with a single reference and alternative base and unphased genotype were permitted, and the genotype field was restricted to the only "GT" format. This limitation put challenges as most haplotype calling pipelines generate VCF files that may not hold to these constraints, limiting an easy and smooth integration of EthSEQ into existing pipelines designed for haplotype calling. The improved version (EthSEQ v3, Figure 2.1) aims to improve its capabilities in several key areas. First, it has been designed to automatically operate with diverse genome assemblies and reference populations, ensuring greater flexibility and

adaptability. Second, it provides pre-computed models for the most widely used WES kits. Last, it is now fully compatible with the standard Variant Call Format (VCF), a widely used format for storing genetic variation data, facilitating seamless integration with existing workflows. In addition, EthSEQ v3 exhibits significantly improved computational performance, enabling fast and efficient processing of large-scale genomic datasets.

EthSEQ is available as an R package, I have implemented and released the new version accompanied by comprehensive protocol paper detailing its features, the step-by-step procedures for performing ancestry analysis, and how interpret the results[14]. This protocol aims to make EthSeq v3 accessible to a broader audience, empowering researchers with a versatile and efficient tool for investigating population genetics and ancestry.

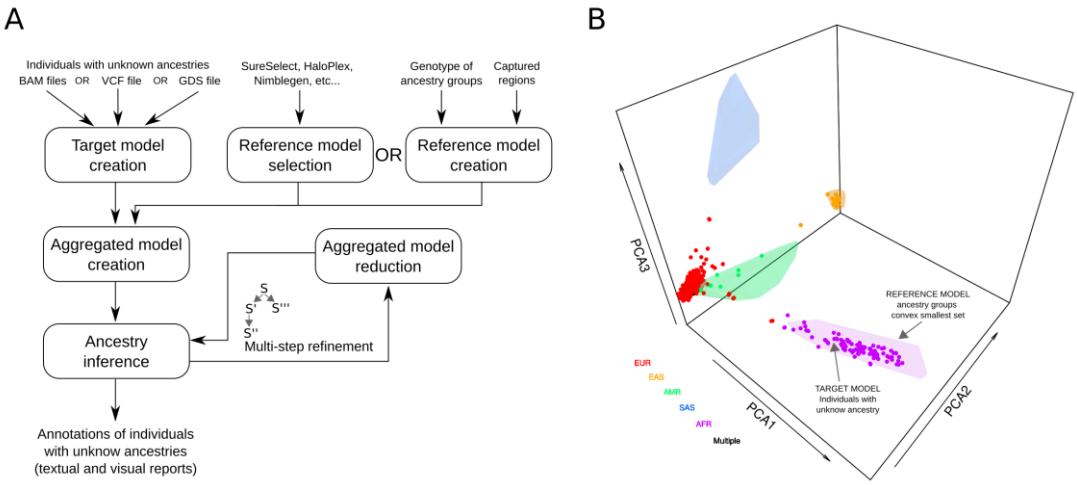


Figure 2.1 EthSEQ v3 analysis. (A) Schematic representation of the EthSEQ computational workflow. (B) Visual report example generated by EthSEQ, illustrating the three-dimensional PCA space. The smallest convex hulls delineate the ancestry groups within the reference model, while individual points represent the target model's individuals, color-coded according to their assigned reference ancestries.

Results

Performance analysis

To evaluate the performance of EthSEQ v2 and v3, I exploited the ICGC dataset and assessed both memory usage and computational time across various combinations of sample sizes and variant numbers. While EthSEQ v3 incorporates additional preprocessing steps within the software, it demonstrates comparable execution times to EthSEQ v2. As shown in Figure 2.2A, the marginal slowdown observed in EthSEQ v3 is in average around 20%. On the other side, the advent of large-scale cohorts showed an exponential increase in memory usage by

EthSEQ v2. As depicted in Figure 2.2B, ancestry analysis of the ICGC dataset, comprising 2,000 samples and 800,000 SNPs, consumed up to 80GB of memory. While feasible on high-end computing resources, this made EthSEQ v2 impractical for standard computers with limited resources, especially considering that standard computers with 128GB of RAM are not yet commonplace. In contrast, the new version, EthSEQ v3, demonstrates a remarkable reduction in memory consumption, requiring more than threefold less memory for the same analysis.

This streamlined approach eliminates the need for users to preprocess their data before conducting ancestry inference, enhancing overall user-friendliness and convenience. The optimization allows users to execute EthSEQ on standard computers with typical RAM capacities of 32GB, suffering only a minor increase in computational time.

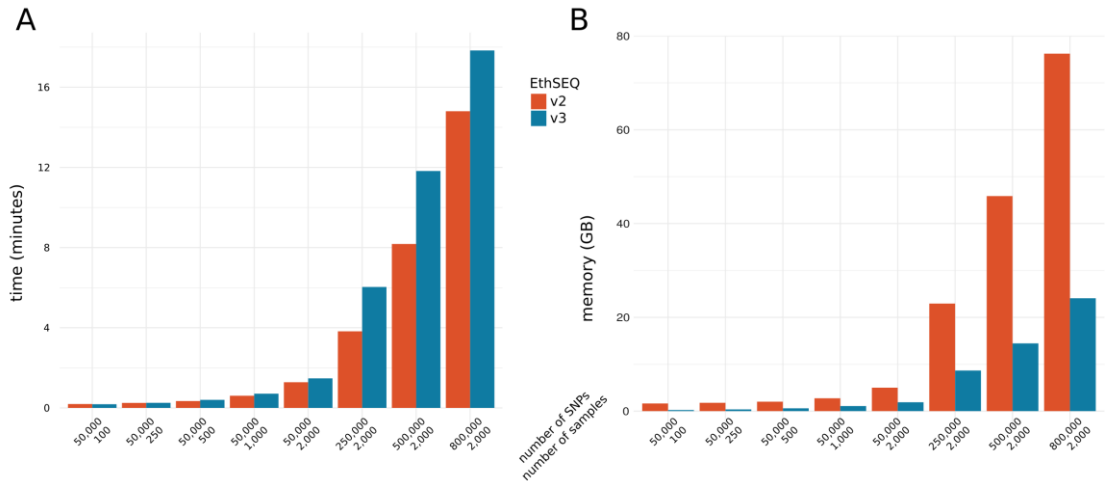


Figure 2.2 performance analysis. Execution time (A) and memory usage (B) comparison between EthSEQ v2 and v3 using different scale of target model.

Ancestry inference using EthSEQ

To evaluate the output consistency of the new EthSEQ version, I conducted a comparative analysis using both the original and upgraded versions of EthSEQ on the same target model of unknown ancestry. Specifically, since the original EthSEQ was implemented with pre-computed reference model built solely on the GRCh37 genome assembly, I extracted genotype calls from six individuals within the 1000 Genomes Project phase 3 dataset, not used to build any pre-computed reference models, and aligned against GRCh37 assembly. Genotype data for 121,012 SNPs captured by the Agilent SureSelect v2 regions were provided as input to EthSEQ in VCF format. The reference model was selected from the set

of pre-computed reference models, representing genotype data for 1000 Genomes Project individuals for SNPs overlapping exonic regions annotated by GENCODE.

Table 2.1 Comparative analysis of inferred ancestries. Ancestry analyses of six individuals from 1000 Genomes Project dataset. The first column presents the self-reported ancestry from the 1000 Genomes Project dataset. The results demonstrate a high degree of concordance between the self-reported ancestries and those inferred by both versions of EthSEQ.

		EthSEQ v2			EthSEQ v3		
ID	**self-reported**	**pop**	**type**	**contribution**	**pop**	**type**	**contribution**
HG00096	EUR	EUR	INSIDE		EUR	INSIDE	
HG00384	EUR	EUR	INSIDE		EUR	INSIDE	
HG01161	AMR	AMR	INSIDE		AMR	INSIDE	
HG02367	EAS	EAS	INSIDE		EAS	INSIDE	
NA18499	AFR	AFR	INSIDE		AFR	INSIDE	
HG03800	SAS	SAS	CLOSEST	SAS(85.09%) EUR(14.91%)	SAS	CLOSEST	SAS(84.41%) EUR(15.59%)

As showed in Table 2.1 and Figure 2.3A, the inferred ancestries for both EthSEQ versions exhibited a high degree of concordance. For 5 out of 6 samples, the inferred ancestry aligned with the self-reported ancestry from the 1000 Genomes Project. In the case of one individual (HG03800), both analyses positioned the individual outside any defined ancestry group. However, both versions consistently inferred the same major ancestry contribution, which was concordant with the self-reported ancestry. These results indicate a high degree of reproducibility and robustness between the original and upgraded EthSEQ versions.

Furthermore, an analysis of all 1000 Genomes Project individuals not used to build the pre-computed reference models demonstrated a high concordance (946 out of 954, 99.16%) between inferred ancestry and self-reported ancestry (Figure 2.3B). Notably, 107 individuals were positioned outside any defined ancestry group, suggesting admixed ancestry. Anyway, for 100 of them, the annotated inferred major ancestry contribution reflected the self-reported ancestry. To note, only one individual was assigned as EUR (European) ancestry, despite their self-reported ancestry being AMR (Ad Mixed American). The observed discrepancies between inferred and self-reported ancestries highlight the potential for complex or ambiguous ancestral origins in certain individuals, as well as the limitations of self-reported ancestry data. These inconsistencies underline the importance of utilizing

genetic-based methods for ancestry inference to complement and refine self-reported information, particularly in admixed populations or individuals with diverse genetic backgrounds.

Analysis of admixed populations

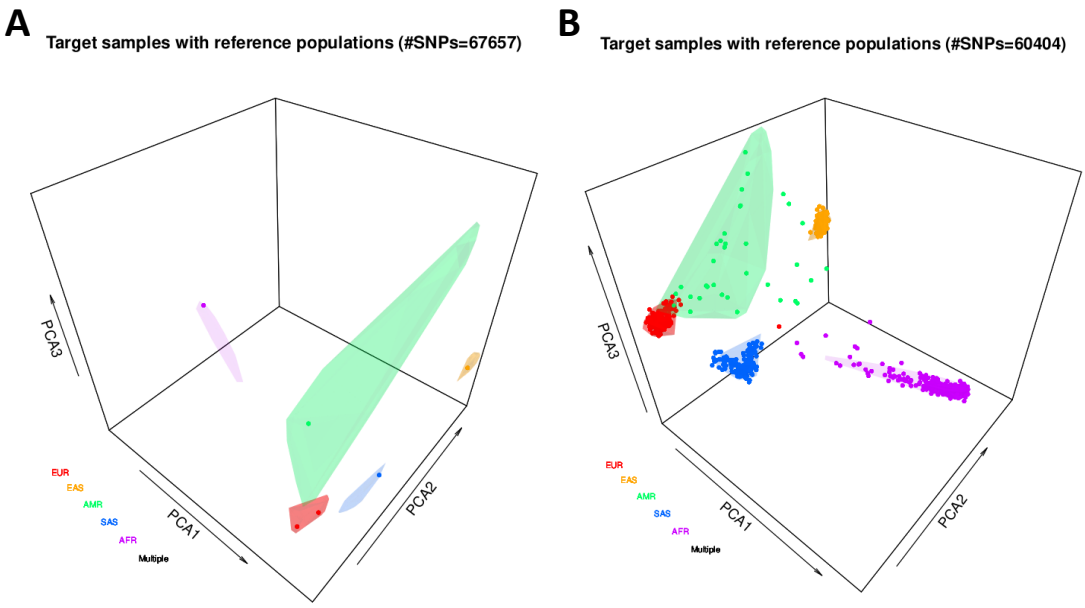


Figure 2.3 EthSEQ v3 analysis results of 1000 Genomes Project individuals. (A) Analysis performed on 6 individuals reported in Table 2.1. (B) Analysis performed on all 1000 Genomes Project individuals. The polygons represent the smallest convex sets identifying the ancestry groups described in the reference model.

To further evaluate the performance of EthSEQ v3 in ancestry inference, I explored the Human Genome Diversity Project (HGDP)[55] dataset, a collection of genetic profiles of more than 900 individuals across 55 indigenous populations. I inferred ancestry using EthSEQ v3 and the pre-computed reference model representing genotype data for 1000 Genomes Project individuals for SNPs overlapping exonic regions annotated by GENCODE.

The HGDP was proposed as a complement to the 1000 Genomes Project dataset with the aim to analyze interpopulation genetic variability. As expected, a majority (62%) of individuals were positioned outside any defined ancestry group in the analysis (Figure 2.4A). This observation highlights the diverse ancestral origins of individuals within the HGDP compared to 1000 Genomes Project dataset. In detail, the number of individuals for each self-reported major ancestry population within the HGDP, along with the corresponding inferred ancestry population assigned by EthSEQ v3, is showed in Table 2.2. The results highlight the good performance of EthSEQ v3 in accurately assigning ancestry to individuals from the diverse and underrepresented populations captured in the HGDP dataset. In

particular, populations reported by HGDP that lack direct correspondence with those in the 1000 Genomes Project (e.g., Central South Asia, Middle East, and Oceania) form distinct clusters in the PCA results (Figure 2.4B). For each individual, EthSEQ assigned ancestry based on the closest major population of the reference model from the 1000 Genomes Project.

Table 2.2 Summary of HGDP ancestry. Distribution of individuals across various combinations of self-reported and inferred ancestries. For 3 out of 4 self-reported populations in HGDP that are already represented in 1000 Genomes Project (e.g. Africa, America, and Europe), EthSEQ correctly inferred populations for all individuals. For East Asia individuals, demonstrated a high accuracy rate of 99%.

Ancestry		
self-reported	EthSEQ	Number of individuals
Africa	AFR	79
America	AMR	49
Central South Asia	SAS	133
Central South Asia	EUR	46
East Asia	EAS	183
East Asia	SAS	2
Europe	EUR	135
Middle East	EUR	149
Middle East	AFR	3
Oceania	SAS	14
Oceania	EAS	9

Of note, the ancestries assigned to individuals from these populations tend to be geographically close to the corresponding self-reported ancestries. Individuals self-reported as Middle East were predominantly (98%) annotated as EUR, with the rest classified as AFR. This population includes individuals from four indigenous groups spanning territories in both Northern Africa and the Middle East, including Lebanon, Iraq, and Palestine. Central South Asian individuals were primarily (81%) annotated as SAS, with the remaining individuals classified as EUR. This population includes individuals from Pakistan, Iran, Punjab, and Afghanistan. For the Oceania population, approximately 61% were inferred as SAS, while the rest were classified as EAS. This group represents four indigenous inhabitants from Papua New Guinea. In the past, genetic studies have been extensively utilized to investigate ancient migratory pathways of human dispersal from Africa across Europe and Asia[56]. These results

align with these established theories for human migration and offer valuable insights into the ancestry of indigenous populations and their genetic proximity to major populations, aligning and integrating with recent findings that challenge the strict "out-of-Africa" model, suggesting a more complex pattern of ancient human dispersal. This demonstrated robust performance in inferring ancestry with reasonable accuracy, even for underrepresented populations. Furthermore, a more comprehensive annotation and reference model, incorporating a wider range of populations, could be in the future developed to achieve even more fine-grained and accurate ancestry inference.

Usage example

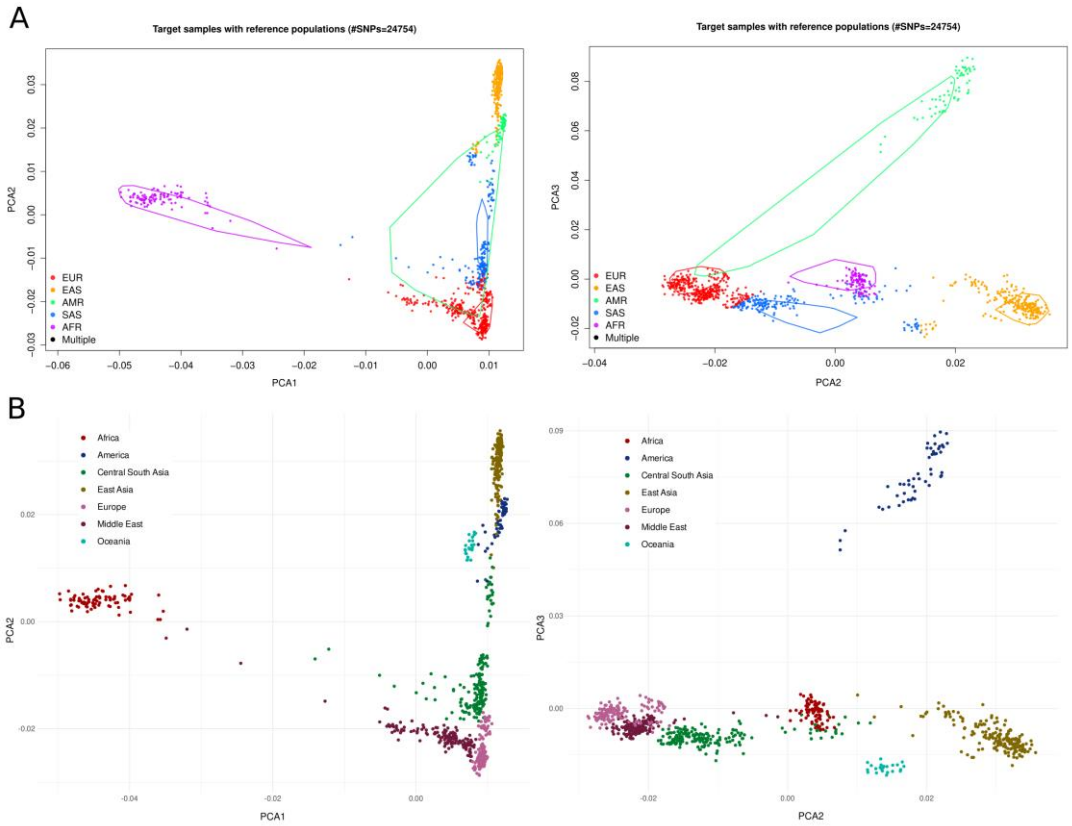


Figure 2.4 Inferred and self-reported ancestries of HGDP individuals. 2-dimensional principal component spaces representing inferred and self-reported ancestry. (A) Ethseq outputs of inferred ancestry for HGDP individuals, based on a pre-computed reference model constructed from variants overlapping exonic regions as reported by GENCODE. (B) principal component values for all HGDP individuals, annotated according to their self-reported ancestry.

The upgraded version of EthSEQ was successfully used in a collaborative project to perform ancestry analysis for over 550 patients across 11 cohorts with recurrent or refractory pediatric solid cancers[57]. The study focused on the high-confidence determination and

characterization of human leukocyte antigen (HLA) genotypes. To explore genotype inference accurately, we evaluated the ancestry distribution of the cohort considered. By analyzing normal whole-exome sequencing (WES) data, EthSEQ identified patients with European ancestry (Figure 2.5). Using an ancestry fraction threshold of ≥ 70% for population assignment, 455 out of 576 (79%) patients showed a predominant EUR ancestry, 30 (5.2%) were AFR, seven (1.2%) were SAS, and 80 (13.9%) patients with no ancestry fraction above the threshold were classified as admixed.

Subsequent analyses focused exclusively on EUR patients across nine tumor cohorts, each comprising at least 20 individuals. These cohorts were further analyzed to infer HLA haplotypes, homozygosity frequencies, and potential candidate allelic associations, providing valuable insights into the genetic landscape of these specific patient populations.

Of note, the characterization of HLA peptidome revealed an increased occurrence of certain variant alleles and haplotypes. Notably, the patient cohort in this study primarily originated from Europe, with a majority from France. In contrast, the reference allele and haplotype frequencies for European/Caucasian individuals were derived from the US population. This discrepancy may slightly affect frequency comparisons. This underscores the importance of

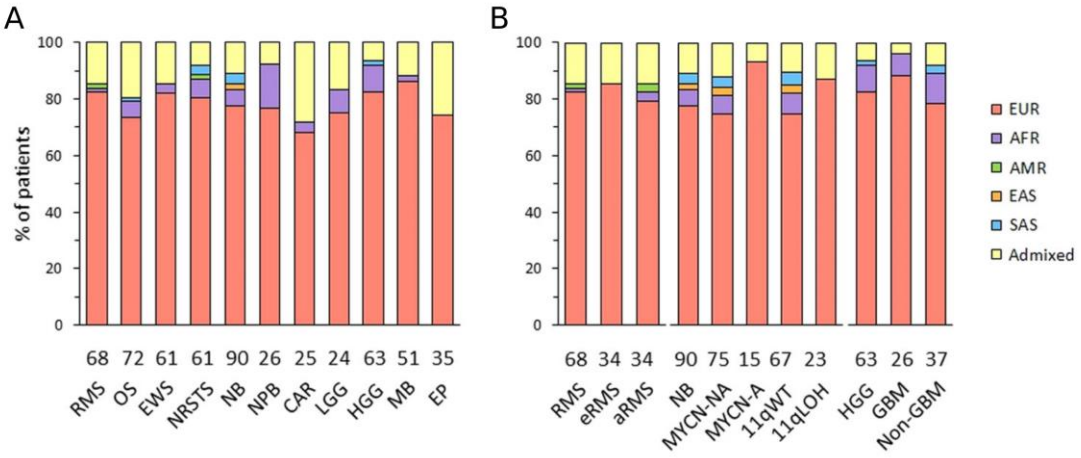


Figure 2.5 Genetic ancestry in patients with advanced pediatric solid cancers. Predominant genetic ancestry fractions (≥ 70%) of patients with specific tumor types (A) and subtypes (B), as determined using EthSEQ. Patients with no predominant genetic ancestry fraction were classified as admixed. The number of patients in each cohort is indicated at the bottom of the corresponding bar charts. RMS, rhabdomyosarcoma; eRMS, embryonal/fusion negative RMS; aRMS, alveolar/fusion positive RMS; OS, osteosarcoma; EWS, Ewing sarcoma; NRSTS, non-rhabdomyosarcoma soft-tissue sarcoma; NB, neuroblastoma; NPB, nephroblastoma; CAR, carcinoma; LGG, low-grade glioma; HGG, high-grade glioma; GBM, glioblastoma; MB, medulloblastoma; EP, ependymoma.

developing more comprehensive reference models that encompass a broader range of populations to enable more accurate and fine-grained ancestry inference.

Methods

EthSEQ is an R package that automates the annotation of individual ancestry from WES or TS data. It analyzes differential SNP genotype profiles, leveraging variants specific to the sequencing assay. As input, EthSEQ requires a set of individuals with unknown ancestry (the target model) and a set of individuals with known ancestry (the reference model). Both models are required by EthSEQ to be in GDS (CoreArray Genomic Data Structures) format[58]. EthSEQ accommodates diverse target model input file formats, automatically generating the appropriate GDS file when required. This feature enhances user-friendliness and streamlines the analysis process by eliminating the need for manual format conversions.

Reference model

Pre-computed reference models are available within EthSEQ. I implemented an automated pipeline (https://github.com/ddalfovo/ModelCreationGDS) to generate reference models compatible with EthSEQ analysis. This pipeline accepts two key inputs: target region files that define the specific genomic regions of interest, typically targeted by sequencing assays (e.g., whole-exome capture kits) and VCF genotype files containing genotype calls for individuals with defined ancestry. The pipeline is implemented using Snakemake (Figure 2.6) and run into a Singularity container for reproducibility. The tool is optimized for parallel processing, enabling the efficient generation of multiple reference models simultaneously.

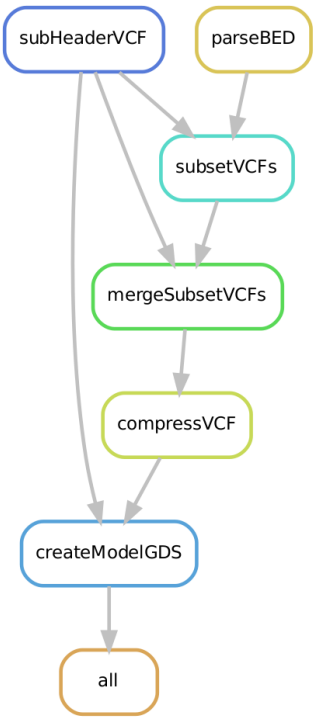


Figure 2.6 Schematic representation of the Snakemake workflow for generating pre-computed models.

Pre-computed reference models, covering a variety of populations, genome assemblies, and RNA-sequencing kits, are readily available for automatic download and utilization within EthSEQ. To retrieve the list of available reference models, EthSeq provides a dedicated function, facilitating user selection and customization of the analysis process. These models are constructed using 1000 Genomes Project data, focusing on major ethnic groups (AFR, AMR, EUR, EAS, and SAS), using both genome

assembly GRCh37 and GRCh38, and supporting multiple WES designs (such as Agilent SureSelect, Twist Bioscience, Roche MedExome, and Roche KAPA). Additionally, a generic reference model is constructed by considering the overlapping exonic regions of SNPs as annotated by GENCODE[59]. This comprehensive model allows for broader applicability and flexibility in ancestry analysis, accommodating a wider range of genetic data.

Alternatively, a custom reference model can be generated based on user-provided sets of genomic regions and combined genotype/ancestry data.

Target model

EthSEQ accommodates diverse input formats for creating the target model, using genotype data in VCF or GDS formats, as well as user-provided lists of control (non-tumor) sequencing Binary Alignment Map (BAM) files. This flexibility enhances the utility of EthSEQ across various genomic data sources. Specifically, VCF and GDS formats are directly incorporated by EthSEQ, aggregated with the reference model, and subsequently utilized for ancestry inference of the target individuals. In contrast, BAM files undergo an initial conversion into an intermediate VCF format. Utilizing the genotyping module of the ASEQ tool[60] with user-defined quality thresholds (default: depth ≥ 10X, mapping quality ≥ 20), EthSEQ determines the genotype calls for each individual at all available reference model SNPs. The merged genotype calls across all individuals are then employed for ancestry inference.

Ancestry inference

To estimate genetic ancestry, PCA is performed on the aggregated genotype data of both the target and reference models. Utilizing the space defined by the first two or three principal components (Figure 2.1B), the smallest convex sets that delineate each ancestry groups described in the reference model are constructed. Then, individuals within the target model (i.e., those for whom ancestry analysis is unknown) are annotated based on their proximity to these ancestry groups. Specifically, individuals located within an ancestry group (or intersecting more than one group) are assigned the corresponding ancestry. For individuals positioned outside all ancestry groups, the relative contribution of each group is calculated based on their distances from the group centroids, providing a fine-grained assessment of their ancestry composition.

To enhance the accuracy of annotations among ancestrally close groups within a study cohort, a multi-step inference procedure is implemented (Figure 2.1A). This procedure utilizes a hierarchical tree structure of ancestry group sets, defined by the user. EthSEQ

performs the ancestry annotation step reducing both the reference and target models including only individuals from ancestry subgroups. The global annotations of all individuals are then updated throughout the tree traversal, ensuring a more precise and granular assignment of ethnicity, particularly for individuals belonging to closely related ancestral groups.

EthSEQ version 3: improvements and tests

The R programming language is widely recognized for its statistical capabilities and user-friendly interface. It has also faced criticism for its memory management, potentially limiting its effectiveness in large-scale data analysis or memory-intensive tasks. The advent of NGS technologies and the increasing use of large-scale cohorts in research necessitate improved efficiency in computational tools. To address these challenges, I implemented an upgrade of EthSEQ to enable compatibility with recent WES kits, provide a comprehensive protocol for a broader range of users, and offers improved performance, making it feasible to run effectively on standard computers with limited hardware resources compared to high-end computing resources.

The original EthSEQ version implemented totally in R bottlenecked when dealing with high-dimensional datasets, particularly struggling with memory constraints, and exhibiting suboptimal computational speed. I implemented a new function to replace the preprocessing steps that convert target model from VCF to GDS format using C++, known for its efficiency and control over system resources. This new function leveraged lower-level memory manipulation and optimized algorithms, leading to a substantial reduction in memory consumption. At the same time, this function introduces additional steps for manipulating the VCF file, specifically extracting and converting genotype data to ensure compatibility with the reference model. While this allows for more complex transformations of the input data, the overall preprocessing time remains comparable to the original R implementation.

To assess the computational performance of EthSEQ, I conducted ancestry inference on all samples from the ICGC dataset, evaluating both execution time and memory usage. The ICGC dataset aggregated whole-genome sequencing (WGS) data from over 2,000 cancer patients of diverse ancestries, encompassing 38 different tumor types. EthSEQ's pre-computed reference models are specifically designed for some common commercial WES kits, an analysis of WGS data necessitates a distinct model. To perform ancestry analysis on WGS

data, a reference model containing roughly 1M germline variants has been previously generated and used to infer ancestry for ICGC patients. Then, a VCF file containing genotype information for roughly 1 million germline variants, overlapping with those in the reference model, across nearly 2,000 ICGC individuals was used for this analysis.

To evaluate the computational performance of EthSEQ (both execution time and memory usage) across varying data scales, ancestry inference was conducted on random subsets of the ICGC dataset generated at different thresholds for both the number of samples and the number of SNPs. To further evaluate the performance of the upgraded EthSEQ, I conducted ancestry inference on the 1000 Genomes Project dataset. Individuals used to construct the pre-computed reference models were excluded from this analysis, and genotype data were down-sampled to 1 million SNPs. A total of 954 individuals from 26 populations were included, considering major ancestry group annotations (EUR, AFR, AMR, EAS, and SAS). Additionally, I explored the inferred ancestries for the HGDP dataset, considering genotype data for over 800 individuals across 7 major ancestry group annotations (Africa, America, Central South Asia, East Asia, Europe, Middle East, Oceania). The HGDP genotype data were similarly down-sampled to 1 million SNPs. Of note, this project comprised 55 underrepresented human populations, aiming to record the genetic profiles of indigenous and isolated populations to understand the genetic frequencies, human evolution, and migration patterns.

Finally, to enhance usability and provide a benchmark for testing, EthSEQ now includes updated sample data, offering users a practical reference for their analyses. Additionally, a comprehensive protocol[14] has been developed, complete with detailed instructions and accompanying commands for running the tool. This protocol is designed to guide users through the necessary parameters and settings based on their specific input file types, ensuring a smoother and more efficient user analysis.

Discussion

Here, I presented an upgraded version of EthSEQ, a rapid, reliable, and user-friendly R package for annotating individual ancestry from WES and TS data. EthSEQ is versatile, capable of processing single or multi-sample datasets, and offers a wide array of pre-computed platform-specific reference models. It provides a streamlined approach for

generating ethnicity annotations directly from a list of BAM files, facilitating seamless integration into existing WES-based processing pipelines.

The improved version facilitates users to smoothly apply EthSEQ to any VCF file containing SNP genotype data generated by most of variant calling software, eliminating the need for data preprocessing. This automated procedure generates detailed information about each individual's inferred ancestry and includes an informative visual report. Additionally, a multi-step refinement procedure is available to enhance the accuracy of annotations for ancestrally close groups of individuals. Furthermore, I compared inferred ancestries derived from genotype data with self-reported ancestries in a dataset comprising admixed populations, offering valuable insights into the accuracy and potential limitations of ancestry inference methods. Moreover, a comprehensive and well-documented version of EthSEQ v3 is now available, highlighting its diverse features and making this powerful tool accessible to a wider audience of researchers. Finally, the new version of EthSEQ has been successfully used to infer ancestry across several pediatric tumors cohort to focus exclusively on patients of predominantly EUR ancestry. This effectively mitigate potential biases in genetic analyses that could arise from population stratification.

Chapter 3. Exploring associations between functional SNPs and somatic aberrations

In this chapter, I present my recent published article[12] investigating the intricate relationship between germline variants and somatic aberrations in cancer.

I first conducted a comprehensive collection of genome-wide association studies (GWAS) on a large cohort of samples across 33 cancer types. I identified 276 common single nucleotide polymorphisms (SNPs) by constructing phenotypic traits based on well-characterized oncogenic signaling pathways. Through linkage disequilibrium (LD) analysis, many LD-extended SNPs were found to reside within regulatory elements and to potentially alter the binding affinity of transcription factor binding motifs, including those of known oncogenes and tumor suppressor genes. Moreover, exploiting *cis*-eQTL and transcriptomic data from the Genotype-Tissue Expression (GTEx) project, I conducted a systematic investigation and identified 247 *cis*-eQTL links, involving 94 variants and 134 transcripts. Further analysis, incorporating an integrated protein-protein interaction (PPI) network, revealed that many *cis*-eQTL genes present in the PPI network were connected to genes implicated in cancer. These results show a potential link between cis-eQTL genes and genes involved in oncogenic pathways, mediated through cancer-related genes. Suggesting a potential effect of cancer genes on the dysregulation of genes within oncogenic pathways.

Taken together, these results support the hypothesis that functional links exist between functional germline variation and the dysregulation of key oncogenic pathways. The identification of this relationship provides additional support for the validity and biological relevance of the GWAS findings.

Next, I explored to what extent polygenic score theory, to elucidate the relationship between an individual's unique combination of germline alleles and their predisposition to specific patterns of somatic aberrations in cancer. A customized workflow was implemented to determine optimal cutoff parameters through a five-fold cross-validation approach and compute polygenic somatic scores (PSS). Statistical significance was assessed via permutation analysis, incorporating multiple hypothesis correction to control for false discovery rate (FDR). This rigorous analysis revealed 24 PSS exhibiting an FDR<0.25 across 9 oncogenic signaling pathways. The 24 identified PSSs were explored to demonstrate their

ability to stratify cancer patients based on prognostic outcomes, such as survival and aggressiveness, and by tumor subtype classifications. To ensure the robustness and generalizability of the findings, I performed a validation of the PSSs using independent pan-cancer datasets from ICGC and CCLE, as well as a cancer-specific independent dataset. This accurate validation process underscores the potential clinical applicability of PSSs in tailoring treatment strategies and predicting patient outcomes based on their unique genetic predispositions.

In conclusion, this article provides a deep exploration of the complex interplay between germline variants and somatic aberrations in cancer, integrating diverse biological data across multiple levels. Consistent with other research, these results highlight the substantial influence of germline variants on specific occurrence of somatic aberrations in key oncogenic pathways. Furthermore, polygenic scores have recently emerged as a promising tool for cancer risk prediction and are currently undergoing validation in various clinical settings, demonstrating that an individual's genetic background can influence the aberration of oncogenic processes.

Future large-scale studies that collect both germline and somatic omics data should continue to investigate the interplay between inherited genetic variation and acquired somatic mutations in cancer. The ultimate goal of such works is the identification of robust biomarkers that can accurately predict cancer risk and inform personalized prevention and treatment strategies.

Germline determinants of aberrant signaling pathways in cancer

Davide Dalfovo[1], Riccardo Scandino[1], Marta Paoli[1], Samuel Valentini[1], Alessandro Romanel[1,*]

[1] Department of Cellular, Computational and Integrative Biology (CIBIO), University of Trento, 38123 Trento (TN), Italy.

*Corresponding author

Journal: npj Precision Oncology

Publisher: Springer Nature

DOI: https://doi.org/10.1038/s41698-024-00546-5

Abstract

Cancer is a complex disease influenced by a heterogeneous landscape of both germline genetic variants and somatic aberrations. While there is growing evidence suggesting an interplay between germline and somatic variants, and a substantial number of somatic aberrations in specific pathways are now recognized as hallmarks in many well-known forms of cancer, the interaction landscape between germline variants and the aberration of those pathways in cancer remains largely unexplored. Utilizing over 8,500 human samples across 33 cancer types characterized by TCGA and considering binary traits defined using a large collection of somatic aberration profiles across ten well-known oncogenic signaling pathways, we conducted a series of GWAS and identified genome-wide and suggestive associations involving 276 SNPs. Among these, 94 SNPs revealed *cis*-eQTL links with cancer-related genes or with genes functionally correlated with the corresponding traits' oncogenic pathways. GWAS summary statistics for all tested traits were then used to construct a set of polygenic scores employing a customized computational strategy. Polygenic scores for 24 traits demonstrated significant performance and were validated using data from PCAWG and CCLE datasets. These scores showed prognostic value for clinical variables and exhibited significant effectiveness in classifying patients into specific cancer subtypes or stratifying

patients with cancer-specific aggressive phenotypes. Overall, we demonstrate that germline genetics can describe patients' genetic liability to develop specific cancer molecular and clinical profiles.

Introduction

Common germline variants in the form of Single Nucleotide Polymorphisms (SNPs) represent the main form of DNA polymorphism. In the last fifteen years, genome-wide association studies (GWAS) identified thousands of variants linked with susceptibility to different types of cancers[61–63]. However, most of these variants exhibited low relative risk, suggesting that they individually have a small effect on the heritability of cancer[64–66]. Polygenic scores hence emerged as an effective approach to integrate multiple small effects across hundreds or even thousands of variants summarizing in a single measure the patients' genetic liability to develop specific cancer types[67].

Cancer, however, is a complex disease[68] influenced by both germline variants and a heterogeneous landscape of somatic aberrations acquired during tumor formation and evolution which recurrently target core cellular pathways and processes[69]. A growing number of studies support the presence of intricate links between germline variants and somatic aberrations. For example, a pan-cancer study[70] exploiting genomic data for >5,000 tumors revealed hundreds of significant associations between germline variants and tumor formation in specific tissues or somatic aberration of specific cancer genes. Further, in[71] a network-based approach was developed to study interactions between multiple germline variants and acquired somatic events in breast cancer and in[72] we queried genomic data from more than 500 prostate cancer patients and found strong signal of association between a germline SNP and SPOP mutated prostate cancer molecular subtype. In addition, in[13] it was demonstrated that germline variants regulate the expression of cancer genes and associate both with local and global somatic mutations and in[73] it was recently demonstrated that polygenic background underlying common hematological traits influence the clonal selection of specific somatic mutations and the development of specific hematological cancer subtypes.

Overall, although there is an increasing evidence suggesting an interplay between germline and somatic variants and a large number of somatic aberrations in specific pathways are now

used as hallmarks in many well-known forms of cancer[74], an exhaustive exploration of the interaction landscape between germline variants and the aberration of these pathways in cancer is still largely missing.

Here we exploit data from The Cancer Genome Atlas (TCGA)[75], ICGC Pan-Cancer Analysis of Whole Genomes (PCAWG)[76] and Cancer Cell Line Encyclopedia (CCLE)[77,78] projects, together with other cancer specific studies, to integrate germline genotypes with somatic aberration profiles in a set of well characterized oncogenic signaling pathways to obtain a pan-cancer and cancer specific view of how common germline SNPs may contribute or predispose to the progression and evolution of tumors. We first identify and characterize an array of common SNPs that increase or decrease the predisposition of these somatic events patterns to occur and then exploit the theory of polygenic scores to explore to what extent germline genetics correlates with somatic molecular profiles, tumor subtypes and clinical variables such as patients' survival and tumor aggressiveness.

Results

SNP genotypes associate with somatic aberrations in oncogenic signaling pathways

To examine to what extent germline genetics primes aberrations in oncogenic signaling pathways we first conducted genome-wide association studies (GWAS) using >8,500 human samples across 33 cancer types characterized by TCGA and exploiting phenotypic traits built considering 10 oncogenic signaling pathways previously described and characterized in[10]; considered pathway include Cell Cycle, HIPPO, MYC, NOTCH, NRF2, PI3K, RTK RAS, TGF Beta, TP53 and WNT. Specifically, using TCGA SNP Affymetrix 6.0 array data, a collection of pan-cancer GWAS were performed by means of logistic regression considering the genotypes of 833,130 high quality SNPs across 8,682 TCGA high quality normal samples (patient's control samples, non-tumor) using additive, dominant and recessive models. Forty binary traits were tested, 10 of which considering for each oncogenic signaling pathway the presence/absence of a somatically altered gene (as described in[10] and here referred to as *somatic traits*, Figure 3.1A), and the remaining ones (here referred to as *somatic transcriptomic traits*, Supplementary Figure 3.1A) considering for each pathway the presence/absence of up-regulated genes (10 traits), down-regulated genes (10 traits) or generally deregulated genes (10 traits). The aberration frequencies of all traits across all

tumor types are reported in Supplementary Figure 3.2. All analyses were adjusted for age at diagnosis, sex and the first six components from a principal component analysis (Supplementary Figure 3.3). Genomic inflation (GI) was inspected (Supplementary Figure 3.4) and TP53 downregulation recessive trait (TP53 DOWN recessive) was removed due to an inflation >1.1. In addition, heterogeneity of associations across tumor types was determined and investigated.

We identified 6 genome-wide significant (p-value<4.2e-10) associations between 6 SNPs (1 intronic and 5 intergenic) and 5 traits (Figure 3.1B, Supplementary Table 3.1), no one reported in the GWAS catalog[79] or listed in[70]. We also identified additional 320 suggestive (p-value<1e-06) associations between 272 SNPs (3 exonic, 7 promoter, 2 3'UTR, 85 intronic and 175 intergenic) and 36 traits, 7 already reported in the GWAS catalog, one associated *with Core binding factor acute myeloid leukemia* and six associated to non-cancer traits (Figure 3.1B, Supplementary Figure 3.1B, Supplementary Table 3.1), and no one listed in[70]. Of these suggestive associations, 8 had a p-value <1e-08 and 71 a p-value <1e-07. Overall, the majority of associations were trait specific, with 39 SNPs associated to at least two traits. We found both risk and protective alleles with associations, especially those derived from dominant and recessive models, often exhibiting high/low ORs. In particular, recessive models applied in the association of low frequency variants and low case/control ratios resulted in significant though unstable results (high ORs and large CIs), demanding for careful interpretation of effect sizes. Of all 326 associations, about 97% demonstrated zero to moderate heterogeneity across tumor types (64% of associations with $I^2 = 0$, 21% with $0 < I^2 < 0.25$ and 13% with $0.25 \leq I^2 < 0.5$) while of the remaining ones only 1 had $I^2 \geq 0.75$. All 9 associations with $I^2 \geq 0.5$ were recessive, suggesting that the variable sample size of the different tumor type datasets (from 36 in the CHOL and DLBC datasets to 953 in the BRCA dataset) was probably the major contributor[80] for the high heterogeneity of those associations. Of note, the global Minor Allele Frequency (MAF) distribution of genome-wide significant SNPs was not significantly different than the MAF distribution of suggestive SNPs (Supplementary Figure 3.5). Linkage disequilibrium (LD) analysis was performed to retrieve variants in strong LD (D'=1 and $R^2 \geq 0.8$) with associated SNPs, obtaining 1105 LD variants for 133 associated SNPs.

Using our resource CONREL[31] we found that 654 of the LD extended associated SNPs (59%) lie in enhancer elements conserved across 34 tissue types, 331 SNPs (30%) lie in active enhancer elements conserved across 33 tissue types and 15 SNPs lie in promoter regions

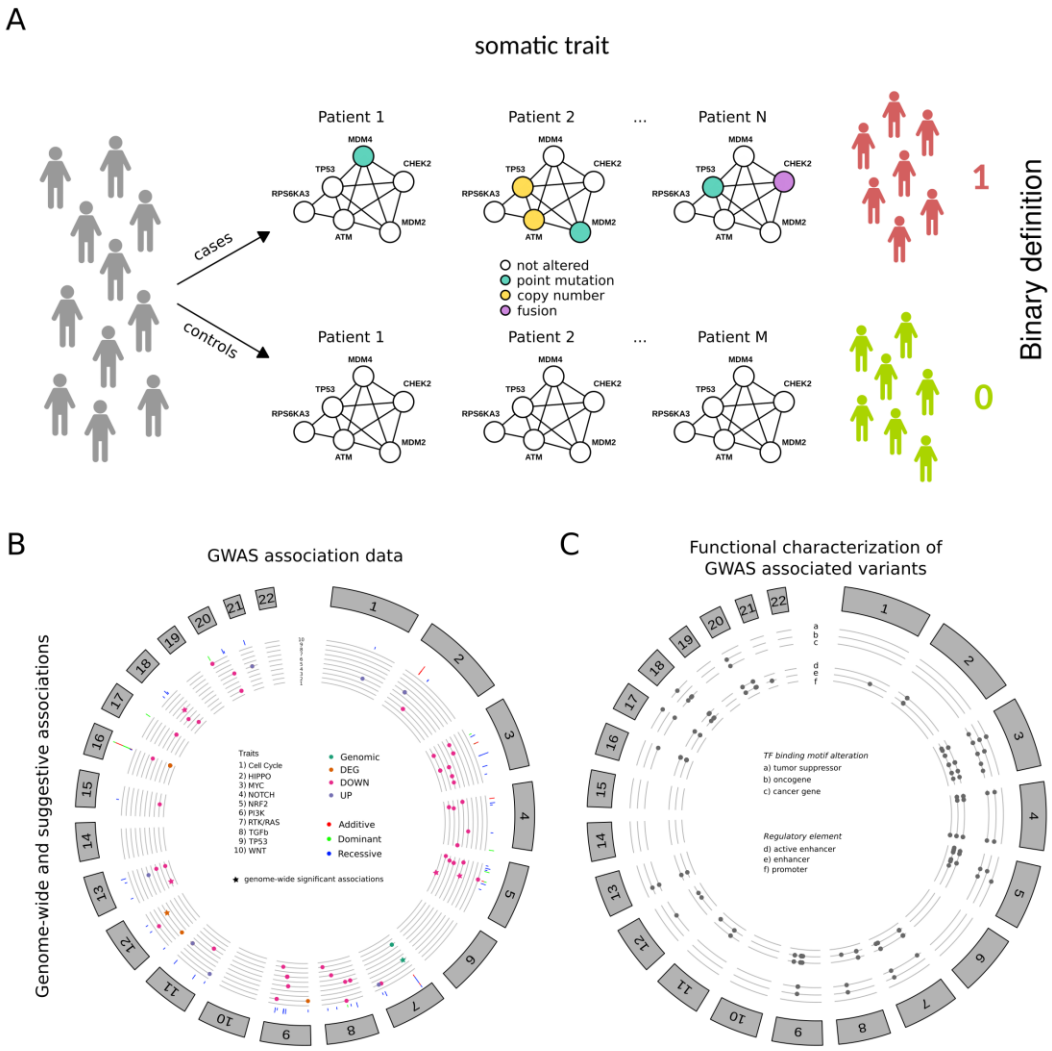


Figure 3.1 Somatic trait definition and GWAS analysis results. A) Cancer patients are stratified based on the presence of aberrant genes in specific oncogenic signaling pathways to build binary somatic traits. TP53 somatic trait construction is shown as example. B) Circular plots showing GWAS results for genome-wide significant associations (highlighted with the star symbol) and suggestive associations with p-value < 1e-07. The chromosomal positions (outer track) of the associations are shown for the forty traits in the inner track. The associations for different oncogenic pathways are reported on different rows and shown with different colors based on the trait's definition. In the middle track, the statistical models used for each association are shown in different colors. c Circular plots showing functional characterization of genome-wide significant associations (highlighted with the star symbol) and suggestive associations with p-value < 1e-07. The functional characterization is performed on LD extended associated variants. LD extended sets of associated variants are characterized for genomic overlaps with regulatory elements (inner track) and to cause a change in the transcription factor binding motifs of genes implicated in cancer (middle track). The chromosomal positions (outer track) are reported for the

(Figure 3.1C, Supplementary Figure 3.1C and Supplementary Table 3.2). Exploiting our resource Polympact[81] we found that 523 of the 678 functional SNPs we identified (77%) cause a putative absolute relative change >0.5 in the scores of 594 transcription factor binding motifs, of which 19 are oncogenes (including *MYC*, *JUN*, and *CTNNB1*), ten are tumor suppressor genes (including , *TP53, PTEN, BRCA1* and *CEBPA*) and more generally 90 (15%) are genes implicated in cancer (Figure 3.1C, Supplementary Figure 3.1C and Supplementary Table 3.2).

Overall, the data support the presence of wide association signal between functional germline SNPs and the occurrence of somatic aberrations in specific oncogenic signaling pathways.

Associated variants are functionally linked to oncogenic signaling pathways

To further explore GWAS results, we asked whether the observed associations could be due to downstream effects that SNPs may have on the transcription of genes linked to the activity of traits' oncogenic signaling pathways. We hence exploited *cis*-eQTL and transcriptomic data available from the Genotype-Tissue Expression (GTEx) project to search, among the 276 GWAS associated variants, for *cis* interactions with genes in the pathways, or *cis* interactions with genes co-expressed and functionally close to genes in the pathways.

Overall, we retrieved 247 *cis*-eQTL links (of which 123 identified across multiple GTEx tissues) involving 94 variants and 134 transcripts (Supplementary Table 3.3). Of these transcripts, 89 were protein coding genes with an associated gene symbol, while the remaining ones were mostly categorized as novel transcripts. Interestingly, although only three of these 89 *cis*-eQTL genes are known to be involved in cancer, when exploiting data from an integrated protein-protein interaction (PPI) network, 66% of the 74 *cis*-eQTL genes that are characterized in the PPI network were found connected to genes involved in cancer, of which 15 were connected to oncogenes and 16 were connected to tumor suppressor genes (Figure 3.2A). Further, of the 89 *cis*-eQTL genes 53 demonstrated significant transcript level correlations with oncogenic signaling pathway related genes, 25 of which exhibiting consistent significant correlations across multiple tissues (Supplementary Table 3.4). Of note, those co-expression signals span across several traits, with some oncogenic pathways exhibiting enriched signal in specific traits, like downregulation based somatic transcriptomic traits, which show the richest signal.

Overall, 50 SNPs were involved in *cis* interactions with genes that were observed co-expressed with members of the corresponding traits' oncogenic pathways, for a total of 1,802 putative links (Figure 3.2B and Supplementary Table 3.4). Interestingly, mean PPI distance among *cis*-eQTL genes and co-expressed genes was 2.94, a distance that was smaller (p-value<1e-03) when compared to the ones obtained from permuted gene sets. Of note, 205 putative links demonstrated a distance less than or equal to 2. Among those latter links, we may highlight variant rs2722888, a SNP we found associated to TP53 somatic trait (additive), which was observed with an effect size lower than 1 (Supplementary Table 3.1). This indicates that aberrations in TP53 pathway is less likely to occur when the alternative allele is present. Interestingly, variant rs2722888 alternative allele was linked to increased expression of *ELP3* gene in multiple GTEx tissues, which was positively correlated (correlations across tissues in the range 0.6-0.7) with *TP53* transcript level (Figure 3.2C, Supplementary Figure 3.6A and Supplementary Table 3.4) with PPI interaction data supporting a close link (PPI distance 2) between the two proteins. We can hence speculate that patients carrying rs2722888 SNP may constitutively have higher expression of *TP53* gene, likely protecting cells from the accumulation of somatic aberrations in the TP53 signaling pathway and hence supporting the observed GWAS association.

Another interesting example is variant rs12686004, which was found additively associated to Cell Cycle downregulation trait with an OR of 3.4 (Supplementary Table 3.1), indicating a strong enrichment of variant's alternative allele in patients with downregulation of genes part of the Cell Cycle pathway. Variant rs12686004 alternative allele was linked to increased expression of ABCA1 gene, which was negatively correlated (-0.7) with RB1 transcript level (Supplementary Figure 3.6B and Supplementary Table 3.4) and closely linked (PPI distance 2) to it. Interestingly, *RB1* is a tumor suppressor gene and is dysfunctional in many major cancers[82]. Hence, we can hypothesize that patients carrying rs12686004 SNP may constitutively have lower expression of *RB1* gene, likely enhancing the cancerous phenotype of cells that accumulate a somatic deregulation of Cell Cycle genes.

Further, we may highlight variant rs436898, associated with NRF2 downregulation trait (NRF2 DOWN recessive). The SNP was found linked to increased expression of *TMEM30A* gene in multiple GTEx tissues, which was in turn negatively correlated to KEAP1 gene expression (correlations across tissues in the range 0.53-0.58) and closely PPI connected to it (Supplementary Figure 3.6CD and Supplementary Table 3.4). Based on these observations,

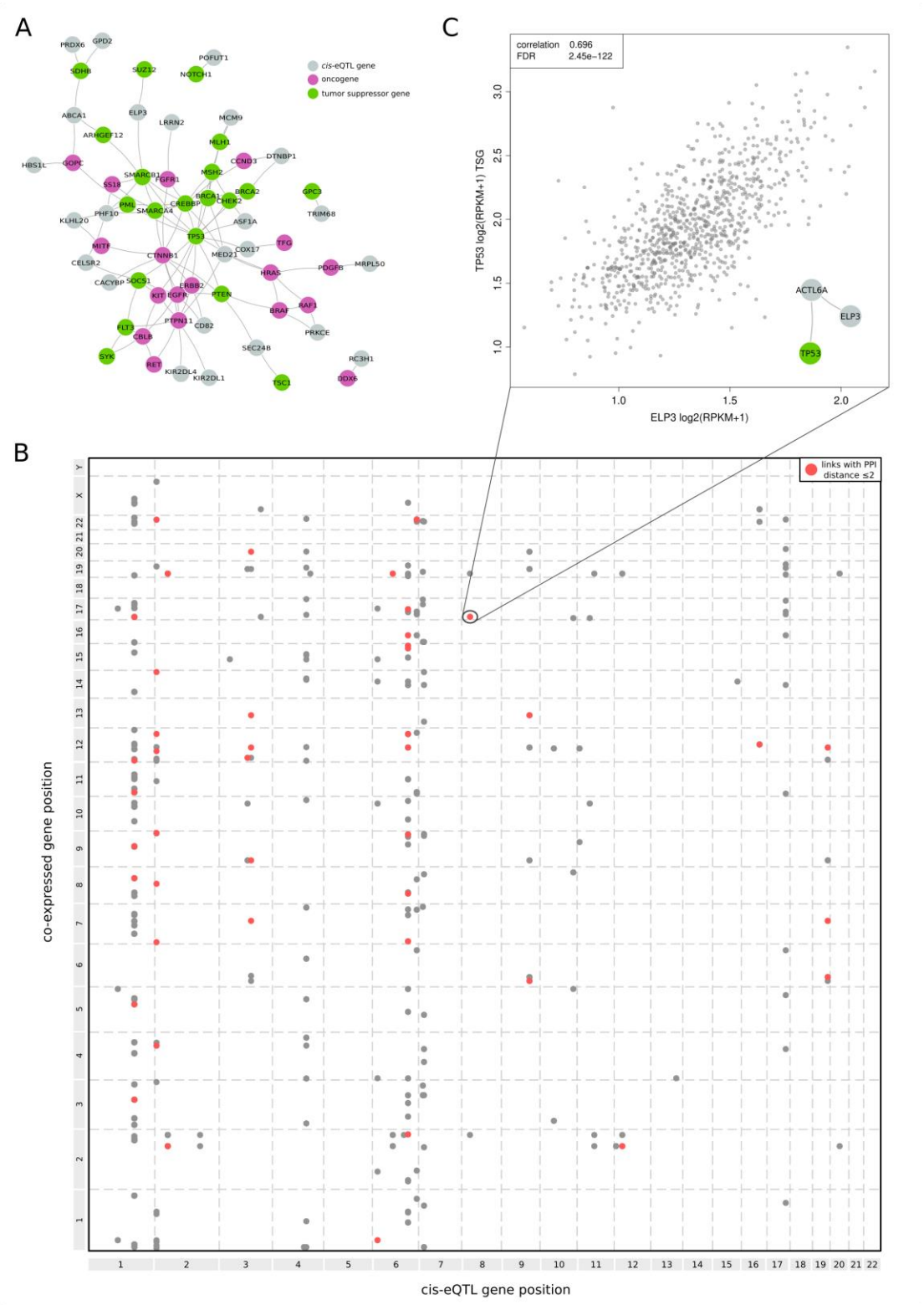


Figure 3.2 cis-eQTL and co-expression analyses. A) PPI network showing cis-eQTL genes that were found connected to cancer-related genes. B) Grid visualization highlighting coordinates of cis-eQTL genes in one dimension and coordinates of co-expressed genes in the other dimension. Points in red represent links between genes with PPI interaction data supporting a close link (PPI distance ≤2) between the two proteins. C) An example representing variant rs2722888 alternative allele (associated with TP53 somatic trait) linked to increased expression of ELP3 gene in Whole Blood tissue, which was positively correlated with TP53 transcript level with PPI interaction data supporting a close link (PPI distance 2) between the two proteins.

GWAS association of rs436898 variant can be supported by the observation that patients carrying the SNP may have reduced expression of *KEAP1*, which combined with somatic downregulation of other NRF2 pathway genes likely exposes cells to a cancerous phenotype characterized by an increased induction of *NRF2*.

Taken together, these results support the hypothesis that functional links between GWAS associated variants, the corresponding traits' oncogenic signaling pathways and cancer genes exist, further strengthening the validity of our GWAS results.

Polygenic Somatic Scores

Provided the strong and broad association signal we identified in the TCGA dataset and the putative functional links we observed, we then explored to what extent polygenic scores can capture the relationship between the unique combination of alleles in a cancer patient and its likelihood to present aberrations in specific oncogenic signaling pathways. A new class of polygenic scores, referred to as *Polygenic Somatic Scores (PSS)*, were computed in the TCGA dataset for all considered traits across additive, recessive and dominant models using a five-fold cross-validation approach. Given a trait, the computational strategy we developed first identifies the best p-value cutoff to build the PSS across different LD clumps, then determines the PSS performances in terms of AUC across the different LD clumps, selecting the best performing one, and finally determines its statistical significance using permutation analysis and multiple hypotheses correction.

Overall, we observed 24 PSS showing an FDR<0.25 across 9 oncogenic signaling pathways and different association models (Supplementary Table 3.5). Among the obtained PSS, NRF2 downregulation traits (NRF2 DOWN) presented consistent high AUC values across the different association models with an AUC of 0.75 for the additive model and 0.72 for the recessive model. Of note, the baseline distributions built on NRF2 transcriptomic traits show a high variance due to the low ratio between cases and controls patients (0.3% for NRF2 DOWN and 1.6% for NRF2 UP). The other somatic traits, including traits for Cell Cycle, TP53, MYC, PI3K and RTK RAS oncogenic pathways were observed with AUC values ranging from 0.53 to 0.61 and with an observed AUC greater than all the corresponding baseline distribution values (Figure 3.3A). As shown in Figure 3.3B, quantile plots obtained from PSS calculated using the identified LD-clump and p-value thresholds but exploiting the entire TCGA dataset clearly demonstrate how high PSS predominantly identify patients with altered

oncogenic pathways. As shown in Supplementary Figure 3.7, no specific tumor type is segregated by our PSS.

The 24 PSS with FDR<0.25 (Figure 3.3A), denoted as *pan-cancer PSS (pPSS)*, were retained for further analyses.

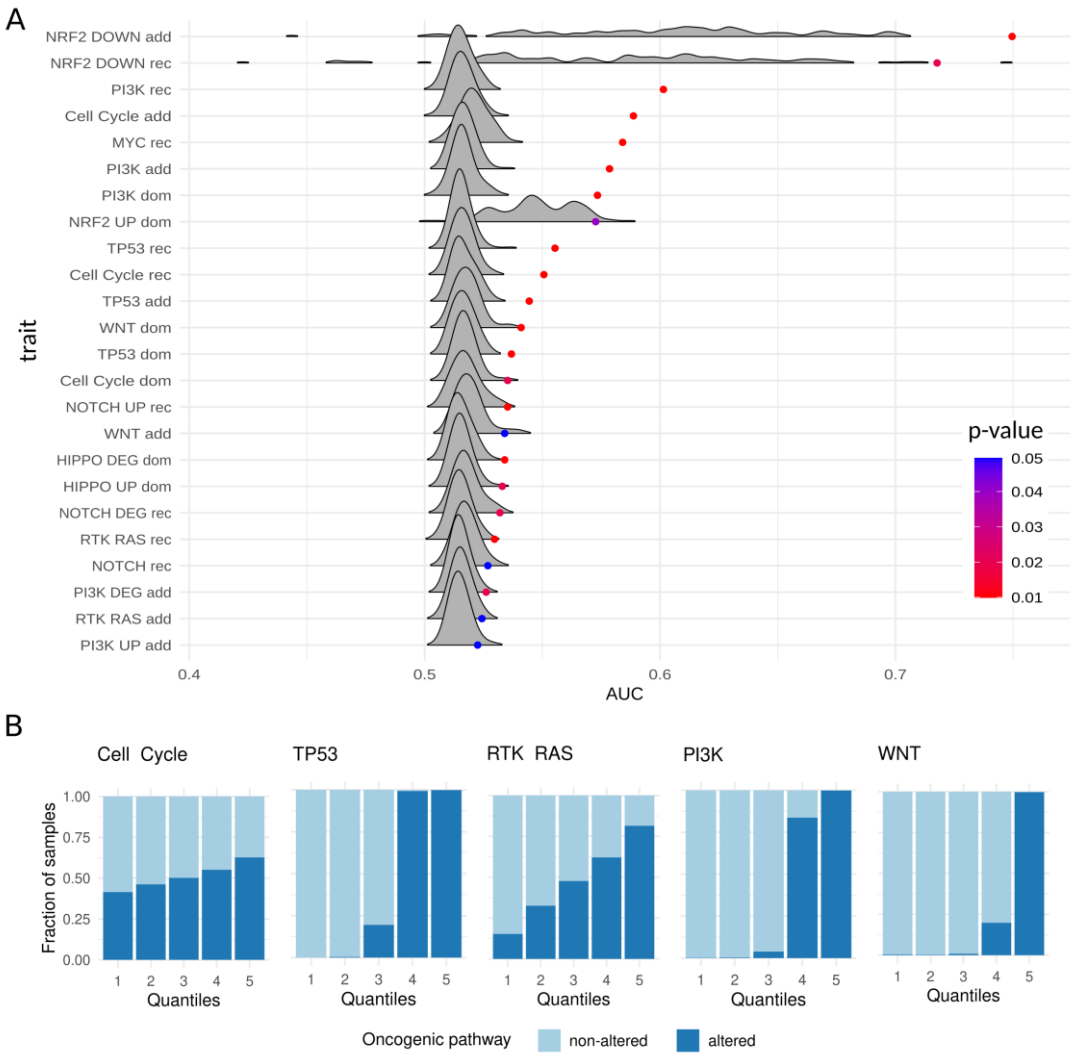


Figure 3.3 Polygenic somatic score (PSS) analysis. A) Ridgeline plot of all PSS with a FDR smaller than 0.25, ordered by AUC value, showing the distribution of AUC values generated from random permutations and the observed AUC values (dots) colored by the corresponding p-value. B) Quantile plots with 5 quantiles of increasing PSS for all the somatic traits with significant FDR using the additive model showing the fraction of samples with altered and non-altered phenotypes.

PSS associate with patient's clinical endpoints

To determine the effectiveness of pPSS, we first explored to what extent they can reproduce the prognostic value of somatic (transcriptomic) traits. Tumor types were analyzed separately and Overall Survival (OS) and Progression-Free Interval (PFI) data for TCGA patients was retrieved from[83]. Patients were stratified based on both traits' oncogenic pathways aberration status and pPSS quantiles (considering the median values) and tumor type specific analyses were performed using a Cox proportional hazards regression model considering age, sex, and principal components as covariates. Also in this case, models'

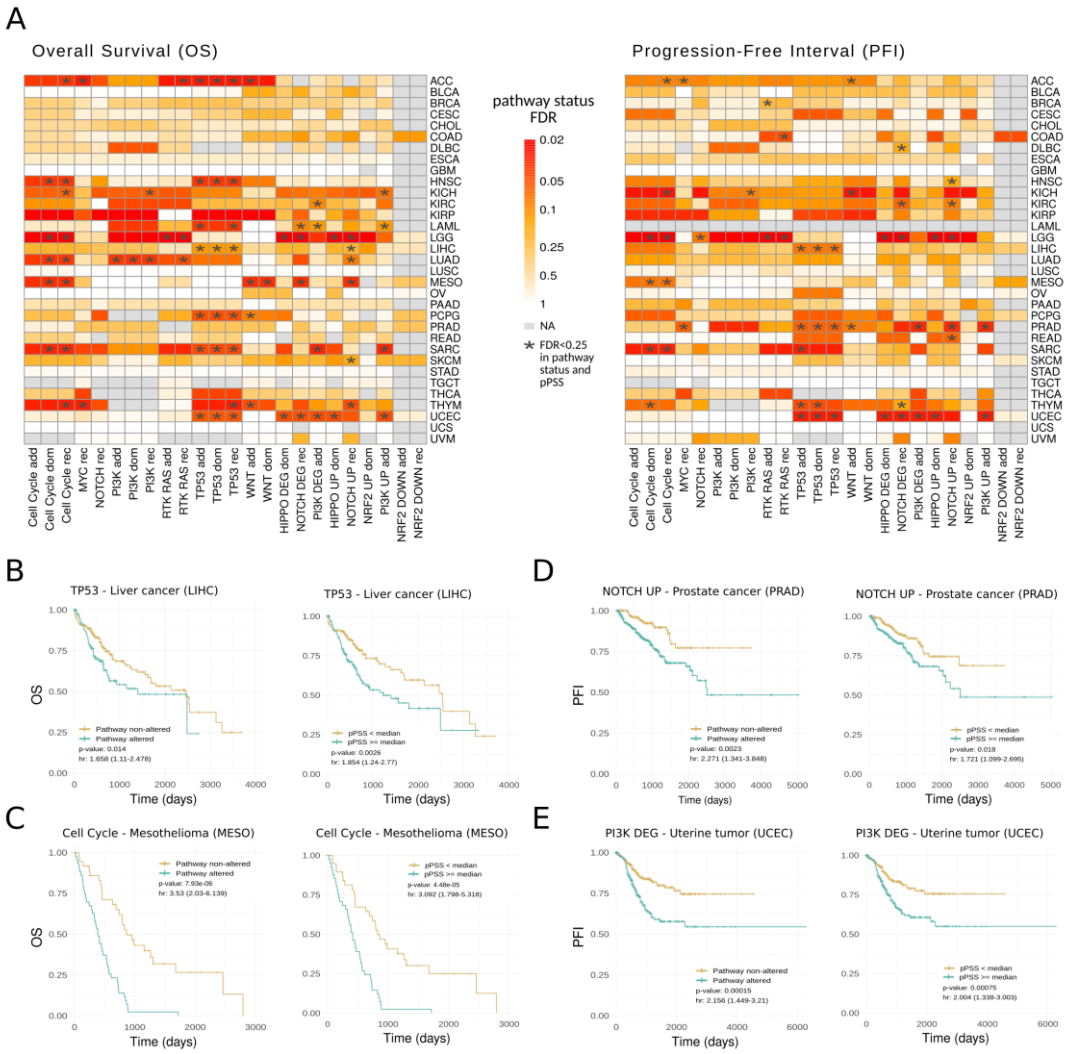


Figure 3.4 Clinical endpoints analysis. A) Tile plots recapitulating the traits survival analysis results. Results are divided based on PFI and OS events. For each trait's oncogenic pathway aberrations status and tumor type, corrected (FDR) empirical p-values computed comparing the observed AUC with the corresponding AUC baseline reference distribution are reported. Combinations of trait and tumor type were both trait's pathways aberration status and pPSS survival analyses resulted statistically significant (FDR < 0.25) are highlighted with an asterisk. B–E) Kaplan–Meier curves showing significant survival analyses for specific examples in both trait's pathway aberration status (left) and pPSS (right).

performances (AUC) were computed using a five-fold cross validation approach and were then tested for statistical significance against reference baseline distributions generated using permutation analyses, finally correcting for multiple hypotheses. Overall, we observed 87 significant (FDR<0.25) traits showing also a significant (FDR<0.25) pPPS (70 from OS analysis, 46 from PFI analysis) across 19 tumor types (Figure 3.4A, Supplementary Table 3.6). pPSS reproduced traits' OS and PFI prognostic value across different tumor types, with Cell Cycle and TP53 somatic traits showing significant OS associations across 8 tumor types and significant PFI associations across 6 and 5 different tumor types, respectively. As examples, TP53 pathway aberrations status and pPSS (TP53 additive trait) showed a strong OS prognostic value in LIHC tumors (Figure 3.4B), Cell Cycle pathway aberrations status and pPSS (Cell Cycle dominant trait) demonstrated OS prognostic value in MESO tumor (Figure 3.4C), NOTCH UP pathway aberrations status and pPSS (NOTCH UP recessive trait) demonstrated PFI prognostic value in PRAD (Figure 3.4D) and PI3K DEG pathway aberrations status and pPSS (PI3K DEG additive trait) showed significant PFI prognostic value in UCEC tumors (Figure 3.4e).

Overall, our data demonstrate that pPSS can be potentially used to stratify patients with poor survival or treatment response.

PSS and tumor subtypes

We then asked to what extent pPSS can be used to identify tumor specific subtypes. For each tumor type we tested the presence of a significant deviation in the distribution of pPSS across different tumor subtypes. Interestingly, we identified several tumor types were pPSS demonstrated strong shifts across specific subtypes (Figure 3.5). Examples are UCEC CN_HIGH subtype (Figure 3.5A), ESCA CIN subtype (Figure 3.5B), TGCT non-seminoma and seminoma subtypes (Figure 3.5C), STAD CIN subtype (Figure 3.5D), LGG IDHmut codel subtype (Figure 3.5E), BRCA Basal and Her2 subtypes (Figure 3.5F). Of note, several pPSS demonstrated significant shifts across subtypes of multiple tumor types.

To explore further this relationship, we built logistic regression models and by comparing observed AUC against AUC baseline distributions obtained from permutation analysis, we identified 22 pPSS across the subtypes of 7 tumor types with statistically significant (FDR<0.25) classification performances (Figure 3.5G, Supplementary Table 3.7). Additionally, in most of those cases an extended logistic regression model integrating all significant subtype-specific pPSS achieved same or better performances in classifying tumor subtypes

(Supplementary Table 3.8). In particular, integrated models for subtypes UCEC *CN_HIGH*, TGCT *non seminoma* and TGCT *seminoma* achieved much better classification performances with respect to models built with single pPSS. Instead, integrated models for subtypes BRCA *Basal,* BRCA *Her2*, STAD *CIN*, STAD *GS*, ESCA *CIN* and ESCA *ESCC* exhibited classification performances that were comparable to the single most significant pPSS. Of note, the majority of the subtype-specific pPSS were non transcriptomic and combinations of Cell

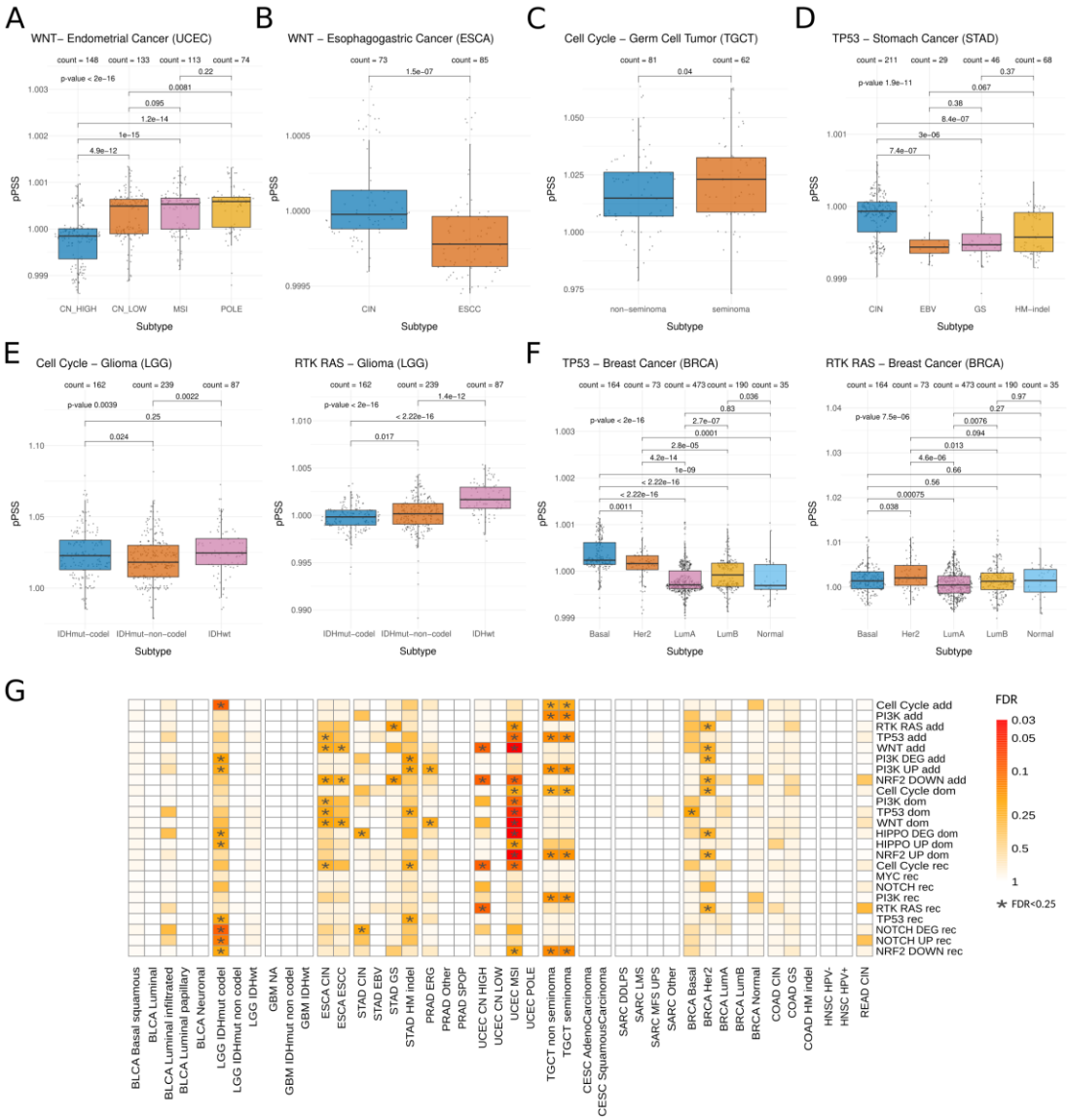


Figure 3.5 pPSS and tumor subtypes. A–F) Boxplots showing the distributions of the pPSS values across different tumor subtypes. pPSS in each cancer subtype are compared using Kruskal–Wallis test and pPSS for each cancer subtypes pair are compared using Wilcoxon-test. G) Tile plot recapitulating the tumor subtype analysis results. For each pPSS and tumor subtype, FDR values of empirical p-values computed comparing the observed AUC with the corresponding baseline reference distribution are reported. The combinations of pPSS and tumor subtype statistically significant (FDR < 0.25) are highlighted with '*'.

Cycle, NRF2 DOWN, PI3K, TP53 and WNT pPSS were observed as particularly effective in identifying specific tumor subtypes.

Overall, our results demonstrate that pPSS can be used across several tumor types to stratify patients based on specific tumor subtypes.

Validation of PSS in an independent pan-cancer dataset

We next tested the effectiveness of our 15 non transcriptomic pPSS using data from the ICGC PCAWG project[76], a large collection of cancer and matched normal whole-genomes from patients spanning over 40 tumor types. Although the differences in PCAWG and TCGA projects data collection limit our ability to test and validate pPSS in PCAWG patients, we exploited PCAWG germline and somatic processed data to test the presence of statistically significant shifts in the distribution of pPSS among PCAWG patients with somatic trait specific aberrations.

In detail, by exploiting GWAS summary statistics trained in the TCGA dataset, PCAWG germline genotype calls were used to calculate the 15 pPSS of interest across 1,823 PCAWG patients. Somatic trait specific aberrations for each patient were determined considering (separately or in combination) reported somatic point mutations, homozygous deletions and amplifications data identified within the corresponding oncogenic signaling pathways. For 5 of the 15 tested pPSS (33%) we found a statistically significant (FDR<0.25) increase of pPSS distribution in PCAWG patients harboring somatic trait specific aberrations (Supplementary Table 3.9). For example, patients harboring point mutations in RTK RAS signaling pathway genes showed increased RTK RAS pPSS values (Figure 3.6A, left) and patients harboring homozygous deletions or point mutations in WNT signaling pathway genes showed increased WNT pPSS value (Figure 3.6B, left).

Overall, the predictive power of pPSS in identifying patients' genetic liability to develop specific cancer molecular profiles was validated in an independent pan-cancer dataset.

Validation of PSS in cancer cell line data

The 5 pPSS showing significant associations in the ICGC dataset were further tested for confirmation using data from the Cancer Cell Line Encyclopedia CCLE[77,78], a large collection of SNP array and omics data for cancer cell lines. Also in this case by exploiting GWAS summary statistics trained in the TCGA dataset, CCLE germline genotype calls were used to calculate the 5 pPSS of interest across 995 CCLE cell lines. Somatic trait specific aberrations

for each cell line sample were determined considering (separately or in combination) reported somatic point mutations, homozygous deletions and amplifications data identified within the corresponding oncogenic signaling pathways. For 2 of the 5 tested pPSS (40%) we found a statistically significant increase (p-value<0.05) of pPSS distribution in CCLE samples harboring somatic trait specific aberrations (Supplementary Table 3.10). We found, for example, that patients harboring homozygous deletions in the RTK RAS showed increased RTK RAS pPSS values (Figure 3.6A, right) and that patients harboring point mutations in WNT signaling pathway showed increased WNT pPSS values (Figure 3.6B, right).

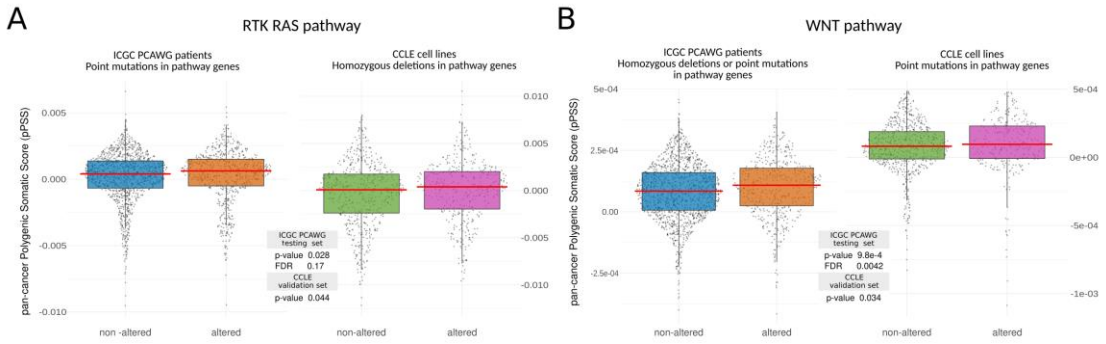


Figure 3.6 pPSS validation using data for ICGC PCAWG and CCLE. Boxplots showing statistically significant shift of pPSS distributions in patients harboring specific aberrations in somatic traits. Specific examples for RTK RAS (A) and WNT (B) pathways significant in ICGC PCAWG dataset (left) and confirmed in the CCLE dataset (right) are reported. Wilcoxon-test was performed (two-tail statistic with FDR correction for ICGC PCAWG and one-tail statistic for further confirmation in CCLE) and reported in the figure.

Validation of PSS in an independent cancer specific dataset

We finally evaluated our pPSS in the Tyrol cohort[84,85], a prostate cancer (PCa) dataset including 1,036 control samples and 837 cancer samples, of which 280 (of 492 with ERG gene status annotation) are annotated as PCa samples collected from patients overexpressing the ERG gene due to a TMPRSS2-ERG fusion (i.e. ERG subtype patients). Considering the effective ERG subtype classification performances that we observed in the TCGA PCa dataset (PRAD) for 5 pPSS, we tested to what extent this result could be validated in the Tyrol cohort. Exploiting GWAS summary statistics trained in the TCGA dataset, the 5 pPSS were calculated for all 837 cancer samples in the Tyrol dataset exploiting the available Tyrol genotype data. Two of the five pPSS (40%) also validated in the Tyrol cohort (Figure 3.7A), and one demonstrated a similar (though not significant) trend. Notably, a logistic

regression model built using the two validated pPPS demonstrated in the Tyrol cohort statistically significant performances (p-value = 0.033) in ERG subtype classification.

The Tyrol cohort provides also clinical information about patients' Gleason Score (GS), a grading system representing one of the best independent predictor of prostate cancer clinical outcome[86]. Of the 19 pPSS that in the discovery TCGA dataset demonstrated a significant association with moderate/high grade prostate cancer patients (i.e., patients with GS equal to 4+3 or greater than 7, respectively), four (21%) also validated in the Tyrol cohort (Figure 3.7B) and one other demonstrated a similar (though not significant) trend.

Overall, the predictive power of pPSS was further validated in an independent cancer specific dataset and we additionally demonstrated that pPSS could be effective in stratifying patients with more aggressive cancer phenotypes.

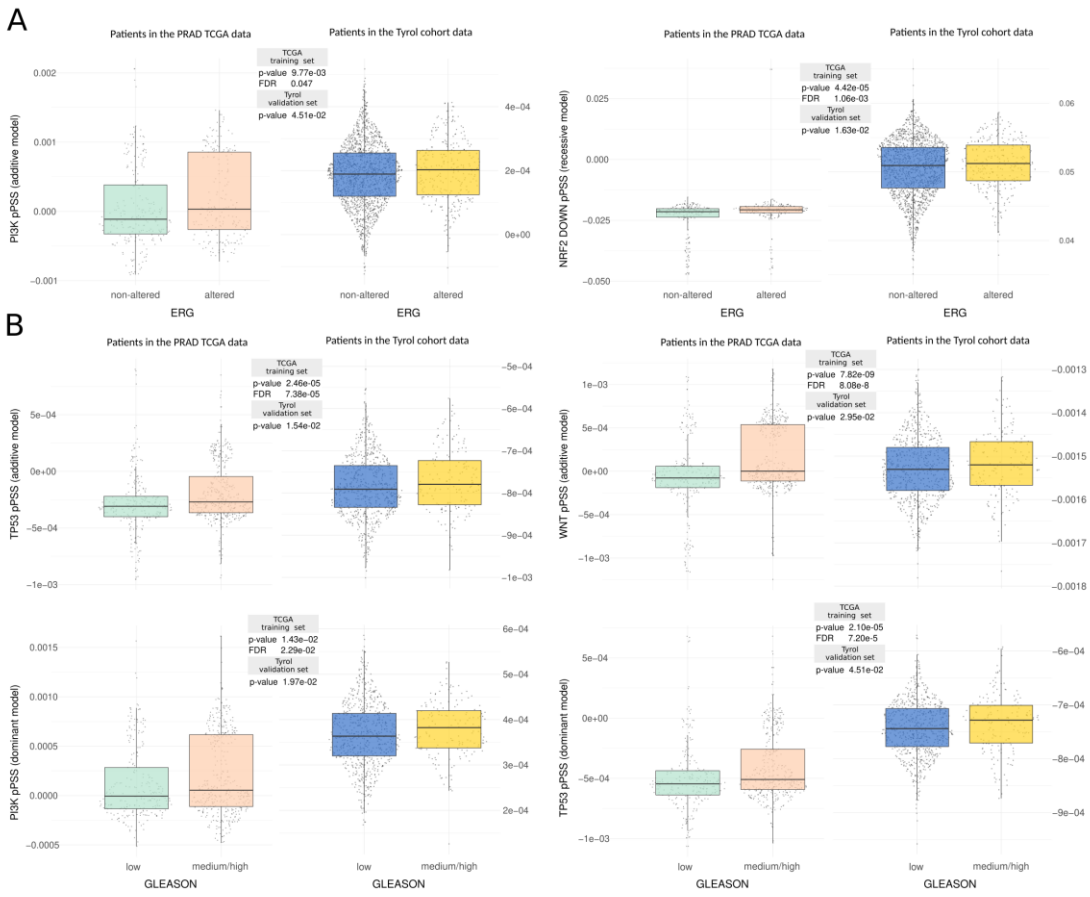


Figure 3.7 pPSS validation in a prostate cancer dataset. Boxplots showing statistically significant shift of pPSS distribution for ERG subtype (A) and in patients with moderate/high Gleason Score (GS) (B) in both TCGA dataset (left) and their confirmation in the Tyrol dataset (right). Kruskal–Wallis rank test sum was performed (two-tail test with FDR correction for TCGA and one-tail test for confirmation in the Tyrol dataset).

Discussion

Over the past 15 years, despite numerous common SNPs have been linked by GWAS studies to the susceptibility of developing different cancer types, most of the identified associations demonstrated modest albeit significant effects. GWAS studies have been usually designed to measure the increased risk that individuals have in developing a specific cancer type. However, in the last ten years, cancer genomes studies based on next generation sequencing data have unveiled how cancer is heterogeneous, characterized by the presence of multiple molecular subtypes and recurrently targeting signaling pathways and biological processes that are now recognized as hallmarks across many well-known forms of cancer.

This motived a deeper exploration of germline-somatic interactions, leading to a clear evidence that genetic background can influence the somatic evolution of tumors[13,70–73,87–90]. Here, we dug further into the exploration of this germline and somatic interplay, using a GWAS-based approach with additive and non-additive[91,92] models and exploiting the availability of matched germline genotypes and somatic phenotypes from large scale projects like TCGA, ICGC PCAWG and CCLE. The datasets utilized in our analyses are multi-ancestry, with European ancestry being the dominant population. Although we employed logistic regression combined with principal component analysis instead of more advanced models, extensive evidence has demonstrated the effectiveness of our approach, particularly in the context of case-control studies[93–97]. Further, other recent GWAS studies successfully used logistic regression with PCA correction on TCGA data[87,98].

Overall, we found evidence that germline genetics can influence the aberration of specific oncogenic signaling pathways, highlighting hence how individuals' genetic background may contribute to the activity and stability of fundamental biological processes that are recurrently disrupted in cancer. A large fraction of the SNPs we found associated in our GWAS were indeed known *cis*-eQTLs of genes closely connected to oncogenes, tumor suppressor genes or cancer related genes. In addition, we identified functional links between specific GWAS associated SNPs and the corresponding oncogenic pathways traits, exploring for some of them putative biological interpretations that are in-line with scientific knowledge and literature. As an example, we highlighted a SNP associated with NRF2 signaling pathway deregulation that is linked in *cis* to genes that are co-expressed with genes in the pathway across multiple tissues. Of note, the alternative allele of the SNP was indicative of a

transcriptional signature associated with downregulation of KEAP1/CUL3/RBX1 complex, which acts as regulator of NRF2 levels in various cancers[99,100].

The ability to analyze and integrate different matched omics data enabled us not only to identify and functionally characterize putative links between specific SNPs genotypes and the aberration of specific oncogenic signaling pathways, but also to exploit the theory of polygenic scores to investigate patients' genetic liability to develop specific molecular profiles or particularly aggressive forms of cancer. While polygenic scores have been recently proven valuable in cancer risk prediction with multiple areas where they can have strong clinical utility, recent reports demonstrate that they can preferentially predict patients belonging to certain tumor subtypes or carrying specific somatic aberrations[101], highlighting hence the importance to better understand their association with molecular and clinical variables. In line with this, our study demonstrates that individuals' genetic background may influence the aberration of oncogenic processes in a way that is orthogonal with respect to the tumor type but important for specific tumor subtypes or to cancers that are particularly aggressive.

Our results are also in line with[70], were the authors identified polymorphisms associated to specific tumor types or specific cancer driver gene alterations. While in both cases a genome-wide association approach was exploited to study germline-somatic links, our approach is substantially different. Indeed, we performed a pan-cancer analysis that explores germline-somatic links at the level of pathway and in particular we investigated the polygenic nature of those links. Although, and as expected, we had no specific overlap with polymorphism reported in[70], the two studies can be considered complementary, since by exploring different dimensions of germline-somatic links they both converge to the same conclusion that germline variants have a significant influence on specific somatic changes in tumors.

While the specific germline-somatic interactions we identified and reported may be used to generate testable hypothesis about mechanistic processes related to cancer genesis and progression, an important question would be to what extent our PSS could be useful in a clinical setting. Although the PPS we have studied demonstrated AUC below 0.8 (which represent a well-recognized threshold of high predictive power), some of our pan cancer PSS were able to stratify patients based on OS and PFI in an extremely effective and cancer specific manner. In addition, classification models built from our PSS demonstrated effective

in identifying tumor subtypes and tumors with more aggressive phenotypes both in the discovery but also in external pan-cancer and cancer-specific datasets.

This study has several limitations, including the relatively small size of the TCGA dataset, the absence of an independent validation dataset with specular data characteristics and the limited clinical utility that our OS and PFI results could have given that TCGA was not designed for clinical outcome studies. We, however, envision that our approach could be exploited and refined to intercept cancer patients with a genetic background that could more likely make their cancer evolve and progress towards specific molecular and clinical trajectories (Figure 3.8).

We want to underline that due to the subtle links that can relate tumor types and pathway aberration profiles, no explicit inclusion of the tumor type in the association model was considered in the current study. Indeed, while it has been established that genetics influences tumor type formation[70], the extent at which it can act as a collider or mediator variable with respect to pathway aberration profiles is not easily definable and further

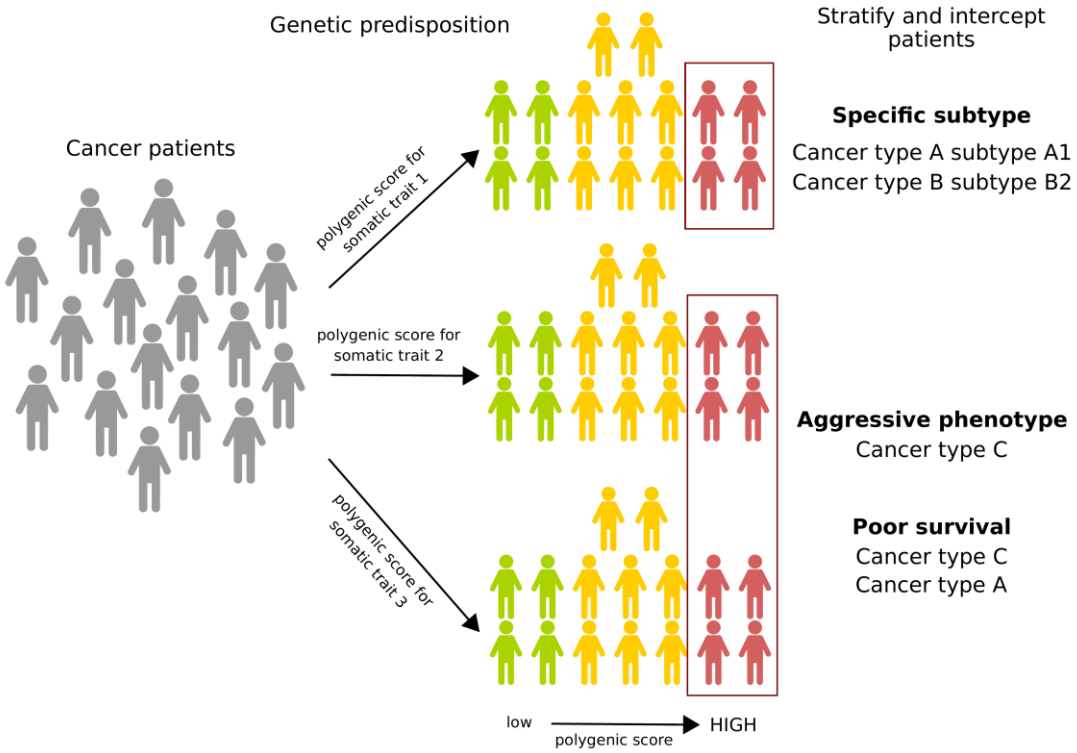


Figure 3.8 Polygenic scores model to describe patients' genetic liability to develop specific cancer profiles. Cancer patients are stratified based on multiple polygenic scores built from somatic phenotypic traits. Somatic traits represent patients' predisposition to carry somatic aberrations in specific oncogenic signaling pathways. Single polygenic scores or combination of polygenic scores can identify patients with more aggressive phenotypes, specific tumor subtypes or patients with poorer survival.

investigations are required. Furthermore, an increased number of recessive associations, primarily involving downregulation traits with slightly elevated GIs, were observed. While an increased GI may suggest a polygenic trait[102], the instability of OR estimations observed across these traits made characterizing most of them challenging in our polygenic analyses. This necessitates future efforts to delve deeper into their characterization and their role in cancer predisposition and evolution.

In addition, while in this study we focused on a set of phenotypic traits derived from the aberration profiles of specific signaling pathways, more advanced methods could be explored to define somatic traits, were cancer specific disruption of specific biological processes could be identified by combining germline and somatic tumor omics data together with network data (e.g. gene networks, protein-protein interaction network)[103].

Future large-scale studies collecting both germline and somatic omics data should continue to explore links between germline genetics and somatic variants with the ultimate goal of identifying cancer risk biomarkers.

Methods

Landscape of inherited SNPs in cancer patients

Genotype calls generated from Affymetrix SNP Array 6.0 intensities of normal (non-tumor) samples were retrieved from the TCGA legacy archive (portal.gdc.cancer.gov/legacy-archive). Each SNP was there annotated with an allele count (0 = AA, 1 = AB, 2 = BB, −1 = missing) and a confidence score between 0 and 1. Genotype calls with a score larger than 0.1 (corresponding to an error rate of >10%) were set to missing and the data was reformatted with PLINK v2[104]. Only autosomal SNPs were considered. Hardy-Weinberg equilibrium (HWE) was calculated across European individuals, selected based on the ancestry calls previously defined in[49], and reported in Supplementary Table 3.11. Samples with SNP call rates <0.9 were discarded. Multi-allelic SNPs and SNPs with call rates <0.9, minor allele frequencies <0.01, or HWE test p-values <1e-06 were discarded resulting in 842,108 SNPs across 10,755 TCGA samples. Considering that batch effects associated with groups of samples processed together (plate effects) can lead to a bias in the estimation of variants allele frequencies[105], we then searched for the presence of variants displaying strong link with plate. In details, analysis of plates was performed stratifying samples by population (considering AFR, EUR,

AMR, EAS, SAS major populations as annotated by EthSEQ[14,54] in[49], Supplementary Table 3.11) and, for each population, comparing all samples of a particular plate with all other plate's samples pooled together. Each variant was tested for the enrichment of genotypes in specific plates (across 275 plates) performing Fisher exact test considering allelic, dominant, and recessive models. We discarded all the SNPs demonstrating a strong plate association (p-value<1e-08) in at least one population and one statistical model, retaining however variants associated with 4 or more plates. In addition, we searched for variants showing links with specific tumor types using a procedure that is similar to the one used for plate association analysis. All the variants displaying a strong association (p-value<1e-08) in at least one population and one statistical model with exactly one tumor type were excluded. Overall, genotype calls of 833,130 SNPs across 10,755 TCGA samples were finally considered. Principal Component Analysis (PCA) was performed on the final data using the *smartpca* function implemented in the EIGENSOFT tool[106] and the first 6 components were extracted.

GWAS traits definition

A set of phenotypic binary traits were defined based on the somatic aberration profiles corresponding to 10 oncogenic signaling pathways characterized in[10] using TCGA data. The considered oncogenic pathways include Cell Cycle, HIPPO, MYC, NOTCH, NRF2, PI3K, RTK RAS, TGF Beta, TP53 and WNT (Supplementary Table 3.12). A set of phenotypic binary traits (referred to as *somatic traits*) were defined based on the somatic aberration profiles described in[10], one for each oncogenic pathway considered. Figure 3.1A shows an example, based on TP53 pathway, of how a somatic trait is built. An additional set of phenotypic binary traits (referred to as *somatic transcriptomic traits*) were defined based on the expression deregulation profile of the list of genes defined in[10] for each oncogenic pathway (Supplementary Table 3.12). Specifically, mRNA expression z-scores (RNA Seq V2 RSEM) were retrieved from The cBioPortal for Cancer Genomics[107,108] for each patient and an oncogenic pathway was considered up-regulated, down-regulated, or generally deregulated if at least two genes in the pathway had, respectively, an expression z-score >2, <-2 or not in the range [-2,2]. Supplementary Figure 3.1A provides an example of how a somatic transcriptomic trait is built, with the TP53 pathway serving as an example. Overall, we defined 10 *somatic traits* and 30 *somatic transcriptomic traits.*

GWAS association analysis

GWAS analyses were performed for each considered trait within the TCGA dataset. Associations of SNPs and traits were performed with PLINK v2 using logistic regression with firth-fallback parameter active, indicating that firth regression is used when logistic regression fails. The analyses were performed using age at diagnosis, sex and the first 6 principal components previously calculated as covariates. Of note, the selection of the number of principal components (PCs) was based on the observation that the first six were sufficient to capture all TCGA populations and subpopulations described in[49]. PCs 1-3 captured the major population structure, while PCs 4-6 captured Asian and European substructures (Supplementary Figure 3.3). In addition, considering that in our scenario the assumption that the likelihood of a patient to have an oncogenic pathway altered is proportional to the number of alternative alleles may not be sufficient to explain the complex genetic architecture of cancer, all three allelic, dominant, and recessive models were investigated. Overall, 8,860 patients with phenotype and covariate data available were used in the analyses. Associations were calculated against the minor allele. Family structure in the analysis was controlled excluding 178 samples representing potential 3rd degree relatives using a scaled KING kinship coefficient of 0.0422 (--king-cutoff parameter was used while running the analyses). We extracted all associations that achieved a genome-wide statistical significance threshold of p-value<4.2e-10 (Bonferroni correction, adjusted also for the number of traits and models tested, i.e., 5e-08/120), but also suggestive associations considering a weaker threshold of p-value<1e-06. The latter threshold was chosen, similar to[87], based on the observation that our analyses were conducted across correlated traits (Supplementary Figure 3.8), involving hundreds of thousands of SNPs (some of which in linkage disequilibrium), and encompassing both additive and non-additive dependent models. Associations flagged by PLINK as UNFINISHED were excluded from reported results. Cross-cancer heterogeneity of the resulting associated variants was determined calculating the I^2 index. In detail, the set of significant associations were tested again in each tumor type separately. The analyses were performed with PLINK as described before. GWAS summary statistics were combined via meta-analysis across tumor types using PLINK. Associations flagged by PLINK as UNFINISHED were not considered in the meta-analyses. Heterogeneity values I^2 were extracted and collected.

Functional characterization of associated variants

For each GWAS (both genome-wide and suggestive) associated SNP, we identified all SNPs in strong linkage disequilibrium (LD) with them within a genomic window of 250kb centered around the SNP. LD data was retrieved from the ENSEMBL database. Strong LD was defined as R^2>0.8 and D`=1. This extended list of associated SNPs and LD SNPs was then queried for genomic overlaps with regulatory elements, cancer genes, oncogenes, or tumor suppressor genes, and their disruptive effect on transcription factor binding motifs. Oncogenes (OGs, N=82), tumor suppressor genes (TSGs, N=63) and more generally cancer related genes (N=920) were characterized using a comprehensive list we compiled from literature. Regulatory elements for promoters, enhancers and active enhancers were retrieved using our resource CONREL[31], while the impact of SNPs on putative transcription factor DNA binding motifs was retrieved from our resource Polympact[81], which characterizes the impact of >18 million common SNPs across >5,000 DNA motifs. SNPs were classified as disruptive when causing an absolute relative change of motifs' score >0.5.

Integrated protein-protein interaction network

A reference protein-protein interaction (PPI) network was built by merging information of five databases: BioGRID release 3.5.173[109]; HPRD release 9 20100413[110]; IntAct release 20150120[111]; BioPlex 3 release 20190502[112]; STRING release v11.0[113]. Interactions between nodes that represent human proteins and experimentally validated were retained. Predicted data, such as evolutionary analysis, gene expression data, and metabolic associations, were excluded. Interactions from STRING and IntAct databases were filtered considering only interactions with reported confidence scores higher than 700 and 0.6 respectively. Interactions from BioGRID, HPRD and BioPlex were all included because manually curated. After the removal of duplicated edges, the resulting network contains 245,787 interactions and 16,514 unique human proteins.

Cis-eQTL and co-expression analyses

GTEx v8 RNAseq count matrices were downloaded from recount3 database[114]. For each tissue, logarithm (two based) transformed RPKM+1 of each gene was calculated using R *recount* and *recount3* packages and quantile normalized using R *limma* package. A total of 16,805 RNA-seq samples across 42 tissues were used in the analysis. *cis*-eQTL data for GWAS SNPs (both genome-wide and suggestive) were retrieved from GTEx data portal (gtexportal.org). SNP/gene *cis*-eQTL links were stratified by tissue and for each tissue *cis*-

eQTL genes in that tissue were collected and tested for co-expression against all other protein coding genes expressed in the same tissue, using Pearson correlation and correcting p-values with FDR method. Only correlation values smaller than -0.50 or greater than 0.50 and with FDR<0.05 were considered significant.

Polygenic somatic scores construction

For each considered trait, a set of polygenic scores were computed using a five-fold cross-validation approach and exploiting the TCGA dataset. TCGA samples were randomly partitioned into five equal-sized disjoint subsets. For each fold, a partition was retained as validation set while the others were aggregated and used as training set. A set of GWAS runs was performed in the training sets as previously described. Specifically, logistic regression was used, considering allelic, dominant, and recessive models, and using age at diagnosis, sex, and the first 6 principal components as covariates. The generated GWAS summary statistics were then used in the validation set to build polygenic scores, referred to as polygenic somatic scores (PSS). PSS were calculated as the average number of minor alleles weighted by the allele's effect size using PRSice-2[115]. As shown in[62,63], using a more liberal but optimized p-value threshold instead of a genome-wide significant threshold, improves performance of polygenic scores prediction. Hence, a computational workflow was designed to build effective traits' PSS and test their performances and statistical significance. As described in Supplementary Figure 3.9, for each trait we first used PRSice-2 to determine the best p-value threshold (testing p-values ranging from 1e-08 to 1 and using a 1e-08 step) across different LD clumps (using R^2 of 0.2, 0.4, 0.6, 0.8 and 1). In particular, to determine the optimal p-value threshold for each clump, we averaged the p-value thresholds at the highest pseudo-R^2, when significant (p-value < 0.05), that we obtained across the five folds. Then, we used PRSice-2 again to generate for each LD clump a trait's score using the corresponding best p-value threshold and calculating its representative AUC performance score, which was obtained averaging the AUC values obtained across the five folds (R *pROC* package was used to compute the AUCs). This to finally select the best performing combination of p-value threshold and LD clump that was used to generate the trait's PSS. Further, to better characterize the statistical significance of PSS performances, we implemented an additional analysis step that is based on permutation analysis. In detail, for each of the 120 PSS (40 traits across 3 association models), 100 random PSS were generated by randomly shuffling trait's labels and for each of them performances in terms of AUC values were computed using the same computational workflow described before, producing a PSS's

specific AUC baseline reference distribution. Then, for each PSS the observed AUC value and the corresponding AUC baseline reference distribution were used to compute an empirical p-value. Specifically, each empirical p-value was computed as (r+1)/(n+1), where n is the size of the reference distribution and r is the number of AUC values in the reference distribution that are greater or equal to the observed AUC. P-values were finally corrected for multiple hypothesis testing using FDR method. A set of *pan-cancer PSS (pPSS)* was finally defined only considering PSS with an FDR<0.25.

Survival analysis

TCGA survival data was retrieved from[83]. Overall survival (OS) and Progression-Free Interval (PFI) data were used. Survival analysis was performed to examine to what extent clinical endpoints correlate with both the somatic (transcriptomic) traits and pPSS within individual tumor types. Also in this case, a five-fold cross-validation approach was applied. Analysis was performed using the R *survival* package. For the analysis based on somatic (transcriptomic) traits, patients were stratified based on traits definitions. For pPSS analysis, patients were grouped and tested on the median value of each selected pPSS. In detail, for each fold analysis, a Cox proportional hazards regression model was computed in the training set and then used in the validation set to compute the performance (AUC) which evaluates the ability of the model to discriminate patients with altered pathways or the patients with a higher pPSS. Also in this case, the performances of our survival models were compared against AUC baseline reference distributions generated by permutation analyses. Empirical p-values were computed as described previously. For both analyses, OS and PFI associations were corrected for multiple hypotheses separately and for each tumor type. OS and PFI associations with an FDR<0.25 for both somatic (transcriptomic) traits and pPSS analyses were highlighted.

Analysis of tumor subtypes

TCGA cancer subtypes were collected from[49]. A total of 5,148 samples were annotated with molecular subtypes for the following tumor types: BLCA, BRCA, CESC, COAD, ESCA, GBM, HNSC, LGG, READ, SARC, STAD, TGCT and UCEC. The molecular subtypes of TCGA prostate cancer (PRAD) dataset were retrieved from[72]. Only TCGA patients included in our polygenic scores computations were retained and then tumor subtypes with less than 20 patients were discarded. A total of 4,818 patients, representing 13 tumor types spanning more than 40 different tumor subtypes, were used in the analysis. For each tumor type, we tested the presence of significant deviation in the distribution of pPSS across different tumor subtypes

applying a five-fold cross-validation approach as described previously. In detail, for each combination of tumor subtype and pPSS, statistical significance was determined building a logistic regression model in the training set testing all samples of a particular tumor subtype against all other tumor samples of that tumor type. Then, the performance (AUC) of the model was computed in the validation set. Also in this case, the performances of our models were compared against AUC baseline reference distributions generated by permutation analyses. An empirical p-value for each combination of pPSS and tumor subtype was calculated as described previously. For each tumor subtype, associations were corrected for multiple hypotheses. Given the non-standard u-shape distribution of p-values that we observed for some combinations, associations were here corrected using the robust FDR method described in[118]. Only FDR<0.25 were considered significant. For each tumor subtype, significant pPSS were integrated using a logistic regression model to test their predictive power in identifying tumor subtypes.

Validation using PCAWG data

Data for somatic point mutations, somatic copy number aberrations, together with matched common SNPs genotype calls and relevant clinical information were obtained from the ICGC PCAWG project[76] for 1,823 patients. Based on available samples annotations, samples that are both in TCGA and ICGC projects were not considered in the analysis. Genotyping files (VCF format) representing a total of 67,207,291 germline variants were downloaded from the ICGC Data Portal (dcc.icgc.org). INDELS and SNPs not in the TCGA genotype dataset were excluded. A total of 830,168 variants were retrieved and used to build pPSS exploiting the weights previously trained in the TCGA dataset. Specifically, scores were calculated with PRSice-2 using TCGA GWAS summary statistics filtered based on PSS TCGA specific optimal p-value thresholds and LD clump cutoffs. Somatic point mutations and somatic copy number aberrations were downloaded for each patient and used to collect somatic trait specific genomic aberrations. Specifically, for each gene in a somatic trait defined by an oncogenic signaling pathway, we retrieved non-synonymous point mutations, homozygous deletions, and amplifications. We considered only the somatic copy number aberrations consistent with the role of the gene (deep deletion of TSGs and amplification of OGs, as defined above). Somatic alterations data representing the presence of gene aberration were integrated and summarized across patients. Due to the differences between data in TCGA and ICGC PCAWG projects, aberrations were not aggregated but kept separated. Binary somatic trait specific aberration profiles were defined for each patient considering separately or in different

combinations the three types of somatic aberrations. Distributions of pPSS in the different groups were compared using Wilcoxon test statistics (two-tail) and p-values were corrected for multiple hypotheses. Only results with FDR<0.25 were considered significant.

Validation using CCLE data

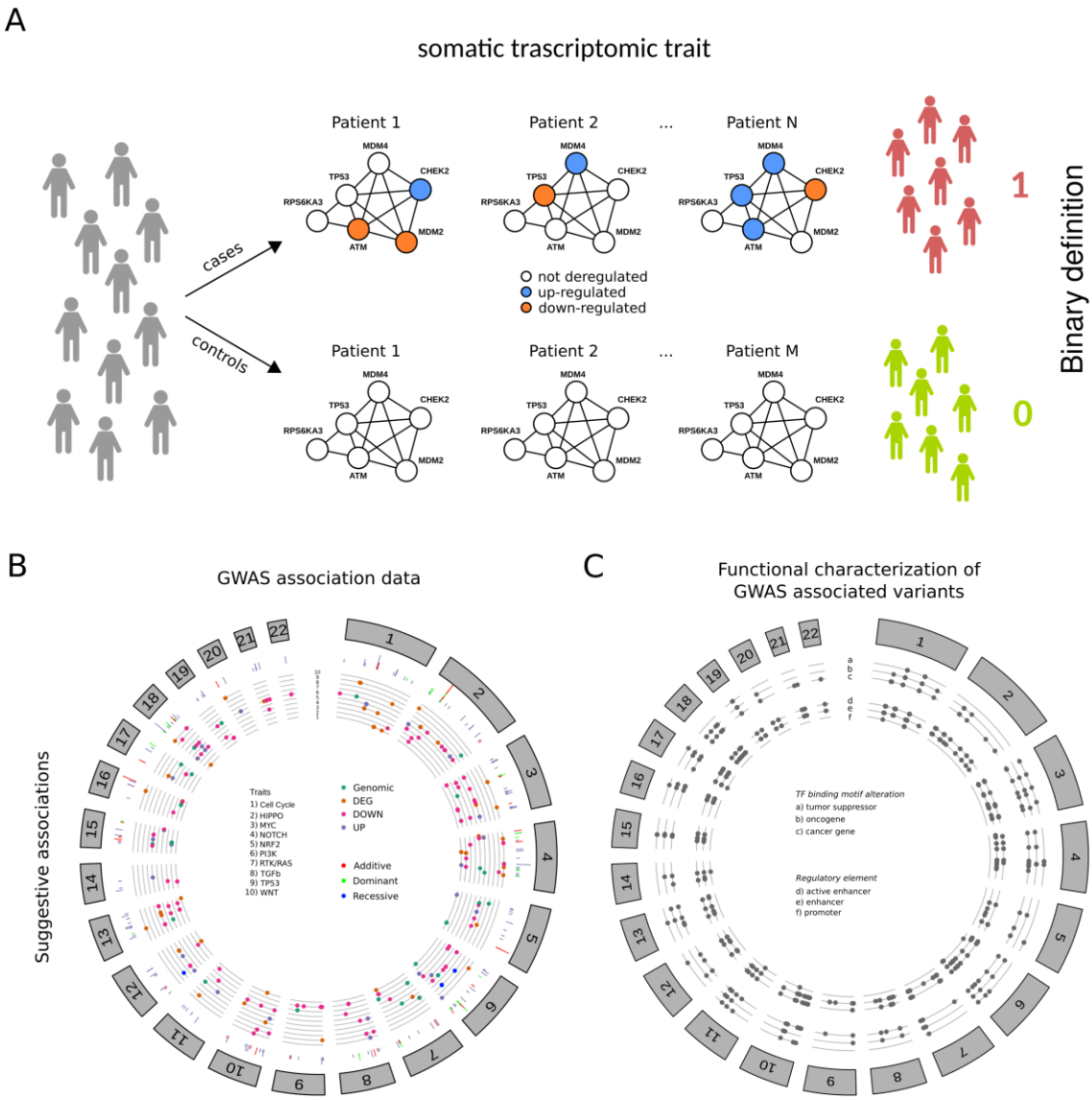
Data for somatic point mutations, somatic copy number aberrations, together with matched SNP Affymetrix 6.0 array Birdseed calls were obtained from the CCLE data portal for 995 cell lines[77,78]. Each SNP was there annotated with an allele count (0 = AA, 1 = AB, 2 = BB, −1 = missing) and a confidence score between 0 and 1. Genotype calls with a score larger than 0.1 were set to missing and the data were reformatted with PLINK v2[104]. A total of 868,261 variants were retrieved and used to build pPSS exploiting the weights previously trained in the TCGA dataset. As for ICGC, scores were calculated with PRSice-2 using TCGA GWAS summary statistics filtered based on PSS TCGA specific optimal p-value thresholds and LD clump cutoffs. Somatic point mutations and somatic copy number aberrations were downloaded for each cell line and used to collect somatic trait specific genomic aberrations. Data was processed as described in the previous section. Only pPSS resulting significant in the ICGC validation were tested for confirmation in CCLE data using a Wilcoxon test statistic (one-tail) with 0.05 p-value cutoff.

Validation using Tyrol cohort data

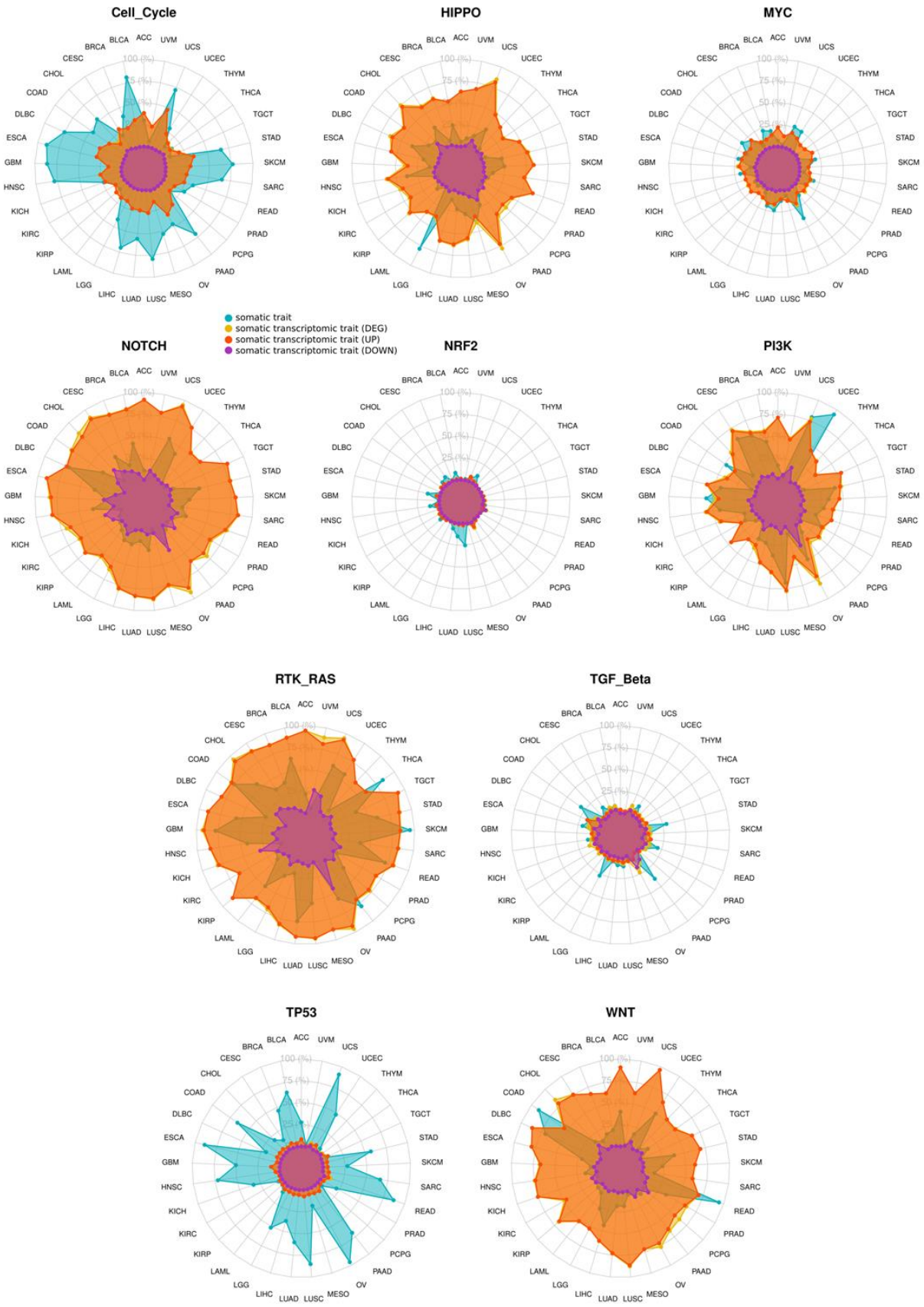
SNP genotype calls (Affymetrix SNP Array 6.0) data and clinical information for 1,903 individuals from the Tyrol Early Prostate Cancer Detection Program cohort were retrieved from[84,85]. The data include genotype calls for 1,036 healthy control individuals and 867 prostate cancer (PCa) patients. Of these, 492 had annotation for ERG status with 280 patients (57%) annotated as positive for the TMPRSS2-ERG fusion (ERG subtype patients). In addition, 159 patients were annotated as having a moderate/high Gleason Score (GS) of 4+3 (N=54) or >7 (N=105). A total of 871,856 SNPs were retrieved and used to build pPSS exploiting the weights previously trained in the TCGA dataset. Also in this case, scores were calculated with PRSice-2 using TCGA GWAS summary statistics filtered based on PSS TCGA specific optimal p-value thresholds and LD clump cutoffs. Only pPSS resulting significant (FDR<0.25) in the TCGA PRAD subset were tested for confirmation in the Tyrol dataset. Distributions of PSS were compared using Wilcoxon test statistic (one-tail) to identify PCa ERG subtype patients and patients with high GS with 0.05 p-value cutoff. Significant pPSS were integrated using a logistic regression model to test their predictive power in identifying ERG positive patients.

Supplementary Material

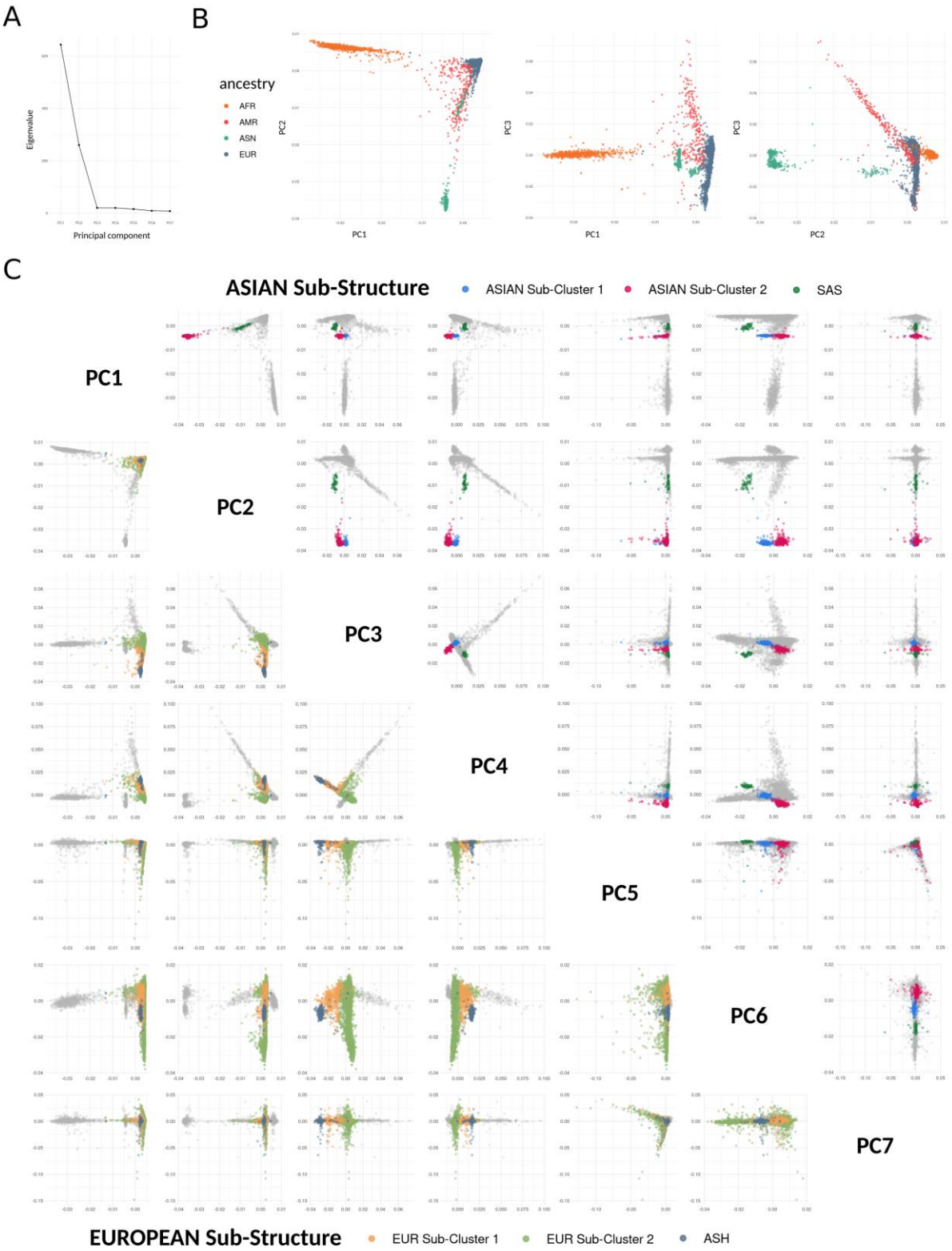
Supplementary Figures



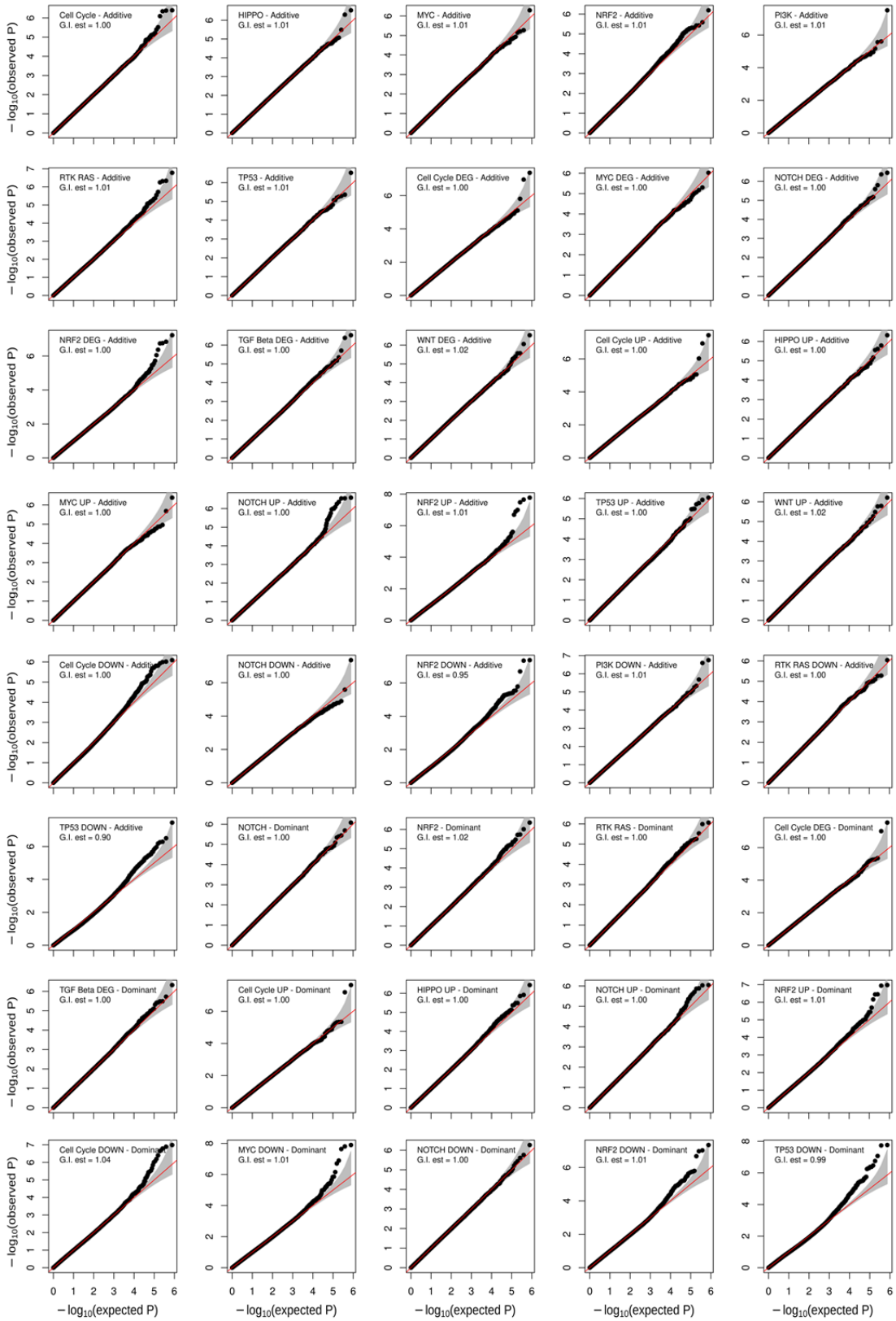
Supplementary Figure 3.1 Traits construction and GWAS results. A) Cancer patients are stratified based on the presence of transcriptomic alterations of genes in specific oncogenic signaling pathways to construct somatic transcriptomic binary traits. TP53 somatic transcriptomic trait is shown as an example of how the genes deregulation are used to build the trait. B) Circular plots showing GWAS results for suggestive significant associations with p-value in the range [1e-07,1e-06). The chromosomal positions (outer track) of the associations are shown for the forty traits in the inner track. The associations for different oncogenic pathways are reported on different rows and shown with different colors based on the trait's definition. In the middle track, the statistical models used for each association are shown in different colors. C) Circular plots showing functional characterization of suggestive associations with p-value in the range [1e-07,1e-06). The functional characterization is performed on LD extended associated variants. LD extended sets of associated variants are characterized for genomic overlaps with regulatory elements (inner track) and to cause a change in the transcription factor binding motifs of genes implicated in cancer (middle track). The chromosomal positions (outer track) are reported for the corresponding variant from the GWAS analyses.



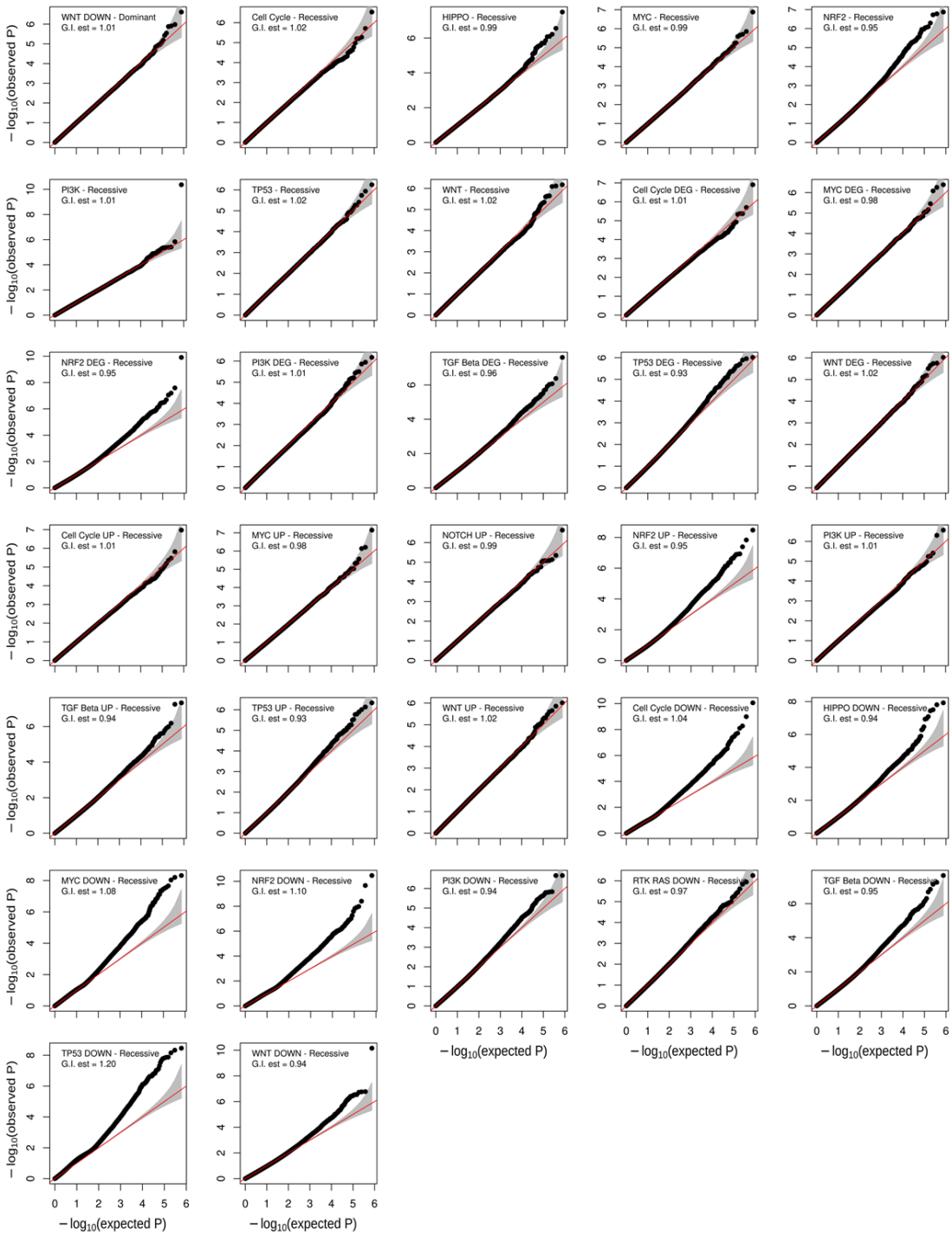
Supplementary Figure 3.2 Traits alteration frequencies. Radar plots showing the fraction of altered samples per trait across all tumor types.



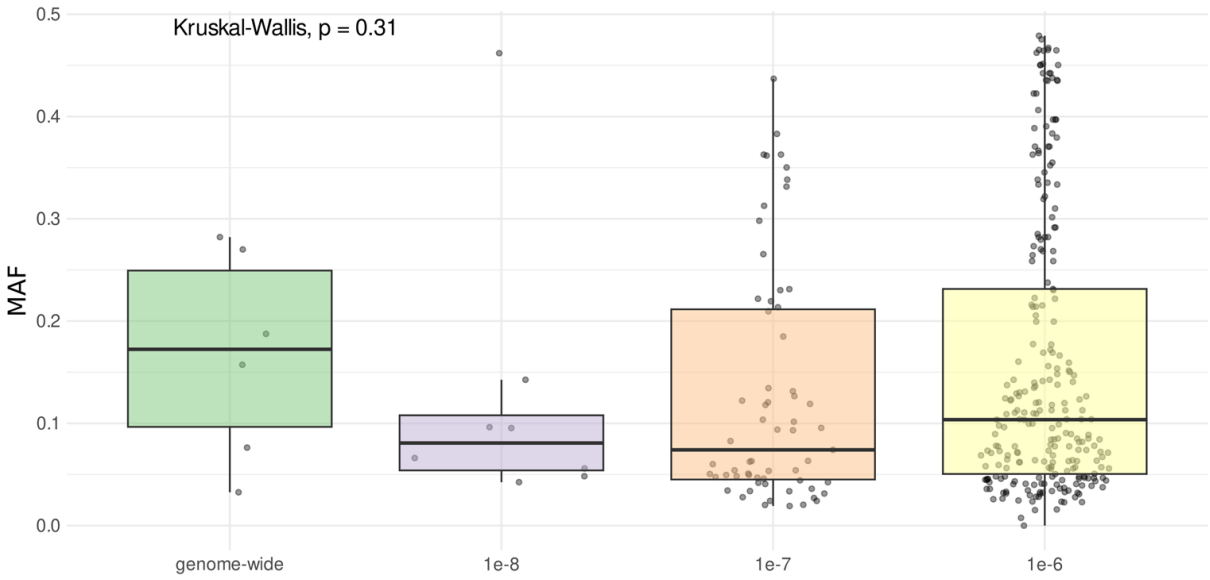
Supplementary Figure 3.3 Principal Component Analysis (PCA) and TCGA population structure. A) Scree plot of the first seven principal components (PCs); B) Major populations are captured by the first three PCs; C) Asian and European sub-populations are captured by the first six PCs. Annotations of populations and subpopulations are derived from (Carrot-Zhang et al., 2020).



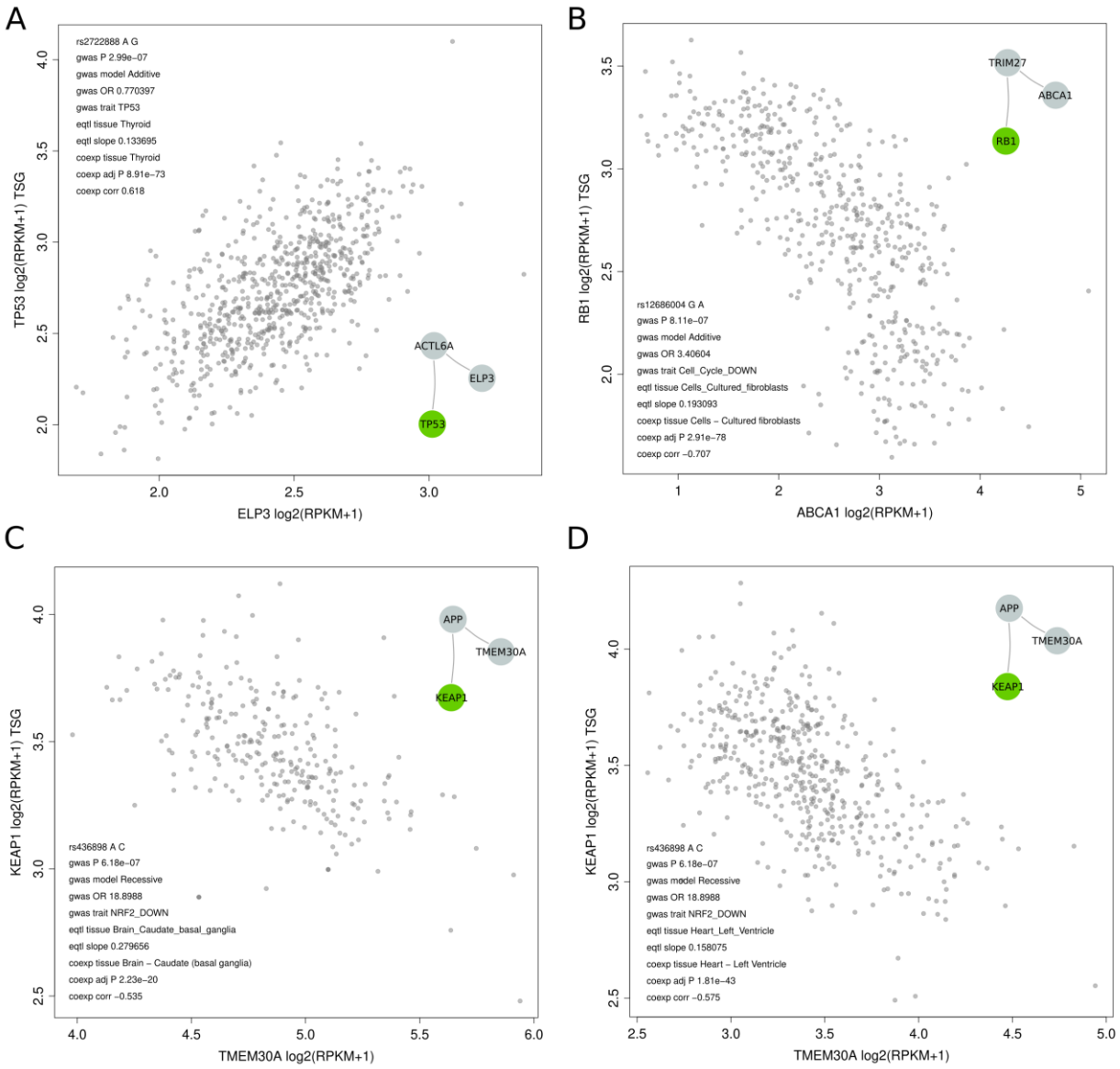
Supplementary Figure 3.4 Quantile-quantile (QQ) plots and genomic inflation (G.I.) estimates for GWAS with traits showing significantly associated SNPs. Red lines represent the expected distributions, the 95% confidence interval is shaded in gray.



Supplementary Figure 3.4 (see legend in previous page)

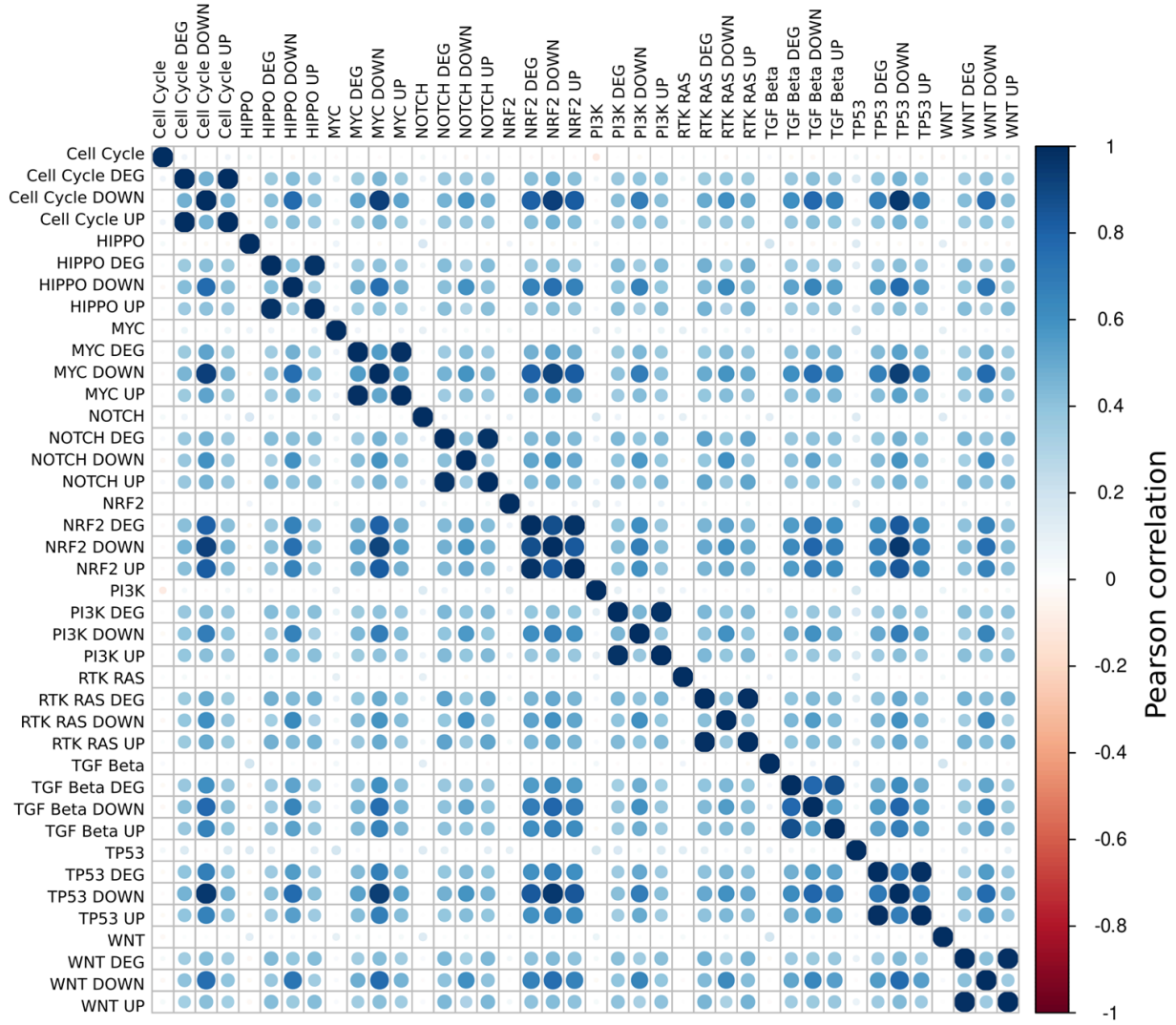


Supplementary Figure 3.5 Boxplots showing the distributions of the Minor Allele Frequencies (MAFs) of genome-wide and suggestive (<1e-8, <1e-7, <1e-6 p-value thresholds) associated SNPs.

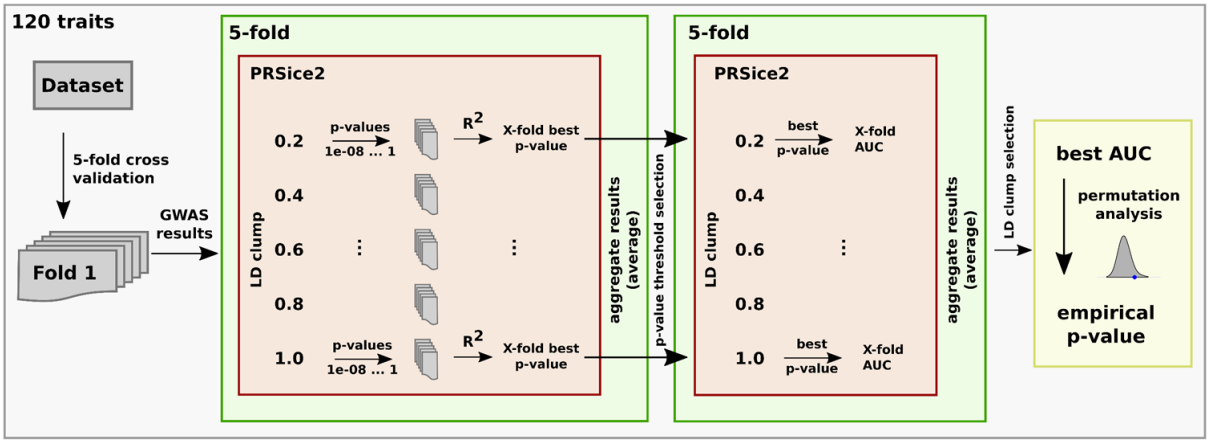


Supplementary Figure 3.6 Examples of cis interactions with genes that are co-expressed with genes in the oncogenic pathways of the corresponding traits. These putative links are supported by a close link (PPI distance 2) between the two proteins. A) shows the co-expression between cis-eQTL ELP3 gene of the variant rs2722888 (found associated with TP53 somatic trait) and TP53 gene. B) shows the co-expression between cis-eQTL ABCA1 gene of the variant rs12686004 (found associated with Cell cycle somatic transcriptomic DOWN trait) and RB1 gene. C-D) shows the co-expression between cis-eQTL TMEM30A gene of the variant rs436898 (found associated with NRF2 somatic transcriptomic DOWN trait) and KEAP1 gene in two different tissues.

Supplementary Figure 3.7 Principal Component Analysis (PCA) of the 24 significant PSSs. The first two principal components are represented, points are colored by tumor type.



Supplementary Figure 3.8 Traits correlation analysis. Heatmap showing the correlations between all trait' pairs.



Supplementary Figure 3.9 PSS computational workflow. The workflow explains how PSSs are built and how their statistical significance is computed. For each trait, a 5-fold cross-validation is used to compute GWAS statistics. The best p-value thresholds across different LD clumps and averaged across the five folds are computed using PRSice-2. Then, the AUC performance scores are computed for each LD clump at the corresponding best p-value threshold. The best performing combination of p-value threshold and LD clump are used to generate each trait's PSS. Finally, a permutation approach is used to compute empirical p-values for each trait comparing each observed AUC value and the corresponding AUC baseline reference distribution.

Supplementary Tables

Supplementary tables (named Supplementary Data) are available at:

https://www.nature.com/articles/s41698-024-00546-5#Sec26

Chapter 4. Propagated mutational scores in DNA repair pathways and variant associations

In the previous chapter, I employed a binary classification for gene mutations based on the presence or absence of somatic alterations within specific oncogenic pathway genes.

To further this research, I participated in a collaborative project at the Laboratory of Computational Cancer Genomics, at the University of California San Diego, under the supervision of Prof. Hannah Carter. The focus of this collaboration was to integrate somatic mutational profiles with gene networks.

Rationale

The hypothesis that uniquely recurrent mutations in a few driver genes account for malignant transformation is now recognized as overly simplistic. In the previous chapter, I analyzed aggregated somatic alterations in specific pathway genes, showing that germline genetics can influence the dysregulation of oncogenic signaling pathways. However, all cancers harbor numerous rarely recurrent mutations in unique combinations across hundreds of potentially cancer-relevant genes. This demands novel approaches that integrates germline, somatic, and molecular interaction data to assess the functional significance of these mutations, define somatic traits that capture cancer-specific disruptions of biological processes, and prioritize them for further investigation.

To address this challenge, I used a network-based method to explore somatic mutational profiles in a cohort of breast cancer patients. Additionally, I extended the analysis to cover somatic alterations in DNA damage repair (DDR) pathways, which were not examined in the previous chapter, to gain a more comprehensive understanding of the mutational landscape of breast cancer.

Introduction

In the last years, several techniques have been proposed and implemented to identify disease genes integrating somatic mutation data with network. While simpler network analysis approaches, such as predicting all neighboring genes[119] or calculating shortest

paths[120], offer a straightforward starting point for identifying phenotype-associated genes, they often not come up to expectations. These methods are prone to false predictions due to irrelevant interactions and fail to capture relevant genes that are not directly connected to the regulated ones, even though they might be strongly linked through multiple long-distance interactions. To address these limitations, global network-similarity approaches have emerged as a more powerful alternative, outperforming local distance measures. These studies focus on a method that considers the entire network structure. Network propagation[103] leverages the idea that genes sharing a phenotype tend to interact closely. By spreading the signal across the network, enabling the identification of altered pathways in a specific condition, offering a more comprehensive understanding of the underlying molecular mechanisms.

Network propagation describes multiple techniques discovered in numerous fields that follow the same underlying strategy[121–124]. Among these, a popular approach to interpret and aggregate somatic mutations heterogeneity is network propagation[103] using a random walk[125] model to diffuse information about gene mutations through network interactions. Network propagation works by integrating each gene's alteration with those of its neighboring genes within the network, taking into account all potential pathways between genes. Iteratively, the alteration information is spread to the neighbors of the corresponding node. This propagation process continues until the propagated scores converge to a steady state on the network.

The network propagation of somatic scores has been used for identifying cancer-related genes and pathways[126]. This approach leverages the concept of "guilt-by-association", assuming that genes mutated in cancer are likely to be functionally related and play a role in cancer development. Network propagation of somatic scores can identify novel cancer genes that may not be detectable from individual gene-level analyses, providing new insights into the molecular mechanisms of cancer.

Breast cancer is the most common cancer among women worldwide. BC is a complex and heterogeneous disease with various molecular subtypes and clinical outcomes. The genomic landscape of breast cancer is characterized by a complex interplay of germline mutations and somatic alterations that impact DNA repair pathways. Inherited mutations in *BRCA1* and *BRCA2* account for a significant proportion of hereditary breast cancers, and their identification has enabled targeted screening and prevention strategies[127]. Furthermore,

somatic mutations in genes such as *PIK3CA*, *TP53*, and *ERBB2* are common in breast cancer and can guide treatment decisions[128].

In this chapter, I explored the intricate relationship between germline variants and network-based propagated mutational scores within a set of well-defined DNA damage repair (DDR) pathways, focusing specifically on breast cancer. Initially, I investigated the effectiveness of propagated mutational scores in prioritizing rarely to moderately mutated genes implicated in cancer, revealing their potential utility in identifying novel cancer-related genes. Then, I identified and characterized common genomic loci that correlate with patterns of propagated mutational profiles across DDR pathways. This analysis aimed to elucidate how germline variants functionally correlate with the dysregulation of corresponding pathway genes and reveal the genetic mechanisms of DDR pathway disruption in breast cancer.

Results

Propagated mutational scores

To investigate the extent to which somatic mutation profiles propagate across breast cancer patients, I performed the Network-Based Supervised Stratification (NBS^2) [129] algorithm on the parsimonious composite network (PCNet)[130]. Specifically, I performed a three-fold cross-validation approach, utilizing 486 samples from the training set to optimize hyperparameters. The optimal hyperparameter values were determined through a grid search strategy, where each hyperparameter was evaluated across a range of values while the remaining two were held constant. The classification performance (AUC) was used to guide the selection of optimal hyperparameters. Overall, the best classification performance was achieved with $\alpha=0.5$, $\lambda=0.01$, and $\beta=2e\text{-}05$ (Figure 4.1).

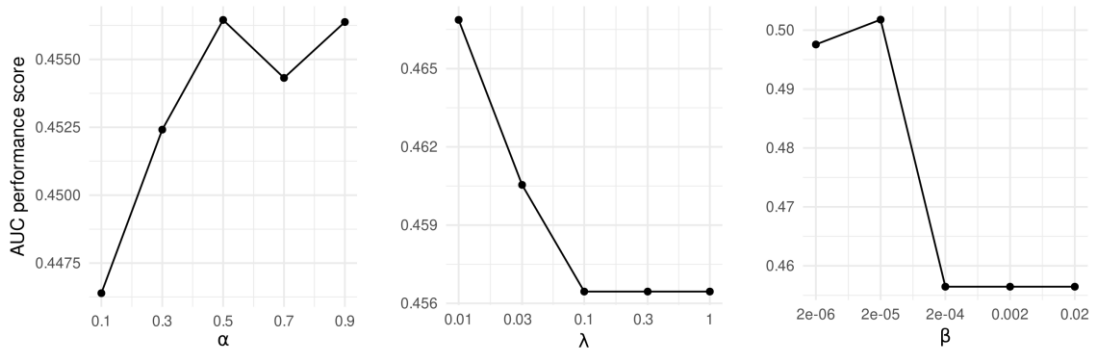


Figure 4.1 Hyperparameter optimization. The performance of NBS^2 to classify tumor subtypes with respect to different choices of hyperparameters.

Then, I used the complete training set for building the final classifier and the remaining set of 245 tumors for validation. Upon convergence, the model was applied to the entire breast cancer dataset to derive the propagated mutational profiles of all patients.

For each patient, I aggregated the propagated mutational scores for all genes belonging to nine DDR pathways previously described[131] including BER, DR, FA, HDR, MMR, NER, NHEJ, NP, and TLS. Interestingly, across all pathways, there were no statistically significant differences in the distributions of the propagated mutational scores between patients harboring at least one mutation within the genes of a given pathway and those without any observed mutations in these genes (Figure 4.2). This observation suggests that the network propagation of mutational signals may reveal underlying pathway dysregulation even in the absence of direct mutations within the pathway's genes.

Moreover, by aggregating the propagated mutational scores across all breast cancer patients, I obtained for each gene a score that represents its network proximity to all genes with mutations. Using these scores, I computed two gene rankings: one based on the non-propagated (i.e., raw mutation frequency) profiles and another based on the propagated mutational profiles. I performed the Wilcoxon-Mann-Whitney rank sum test to assess the significance of propagation-based rankings by measuring the enrichment of known

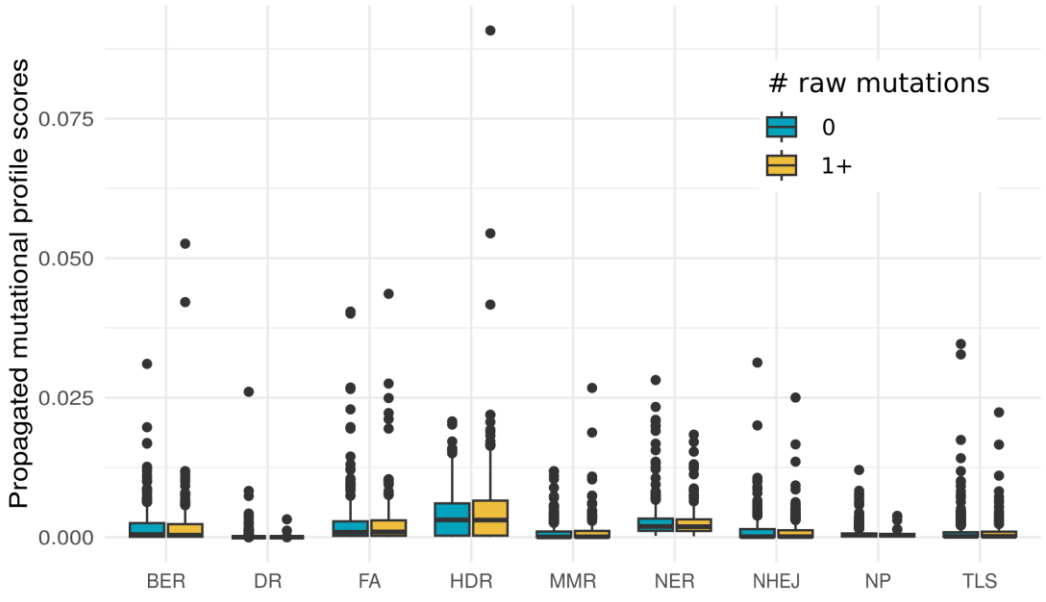


Figure 4.2 DDR propagated mutational scores. Distributions of aggregated propagated mutational scores for nine DDR pathways in breast cancer patients. Each patient's score is calculated as the sum of the propagated mutational scores of all genes within the corresponding pathway.

oncogenes (OG) and tumor suppressor genes (TGS) towards higher ranks before and after propagation. Notably, no significant enrichment (p-value=0.17) was observed in the non-propagated ranking. However, a highly statistical significance enrichment (p-value=2.01e-15) was observed after propagation. These results support the validity of the method to identify functional importance genes such as oncogenes and/or tumor suppressor genes. To further explore the broader relevance of this approach, I extended the analysis by incorporating a list of cancer in addition to the OGs and TSGs. As expected, considering that the most mutated genes in a tumor cohort are typically genes implicated in cancer, I observed a significant enrichment in both rankings, before and after propagation (p-values 3.43e-19 and 2.08e-47, respectively).

Identification of UMGs

I then calculated the difference in rank for each gene before and after propagation. Genes that move up in the rank order post-propagation were listed as upward mobility genes (UMGs). This method effectively filters out frequently mutated genes, including well-known cancer drivers, that occupy high ranks before propagation and therefore cannot meet the upward mobility threshold. I reported a total of 267 UMGs for breast cancer, of which 64 genes (24%) with established implications in cancer development and progression, including 13 oncogenes (such as JUN, KRAS, and PPARG), and eight tumor suppressor genes (including CEBPA, CREB1, and NOTCH1). Among the remaining UMGs not directly annotated as cancer-related genes, I exploited data from the PCNet network. I found that nearly all genes (198 out of 201) of those present in the PCNet network were connected to genes implicated in cancer.

Overall, the identification of UMGs reveals both known and novel genes potentially implicated in cancer, demonstrating that network propagation of mutational somatic profiles in combination with UMG approach can estimate the functional importance of genes potentially implicated in cancer development and evolution.

Variants associate with propagated mutational profiles in DDR pathways

I conducted genome-wide association studies (GWAS) using 731 breast cancer patients considering nine DDR pathways. These GWAS analyses employed linear regression, considering the genotypes of 8,560,450 imputed high-quality variants. Analyses, performed using PLINK v2[104], were adjusted for age at diagnosis, sequencing plate and the first three components from principal component analysis. Genomic inflation (GI) was inspected to

identify potential population structure and other technical artefacts in the data, no bias was found among all the GWAS results (average GI=1.01) (Figure 4.3, displays example results for TLS pathway). I identified 6,272 genome-wide significant (p-value<5e-08) variants across 1433 independent loci across the nine DDR pathways.

MAGMA gene-set enrichment analysis identified only six significant gene sets (Bonferroni adjusted p-value < 0.05) across all nine DDR pathways. Notably, one of these significant gene sets (GINESTIER_BREAST_CANCER_ZNF217_AMPLIFIED_DN) is associated with the mismatch repair (MMR) pathway and is directly relevant to breast cancer. This gene set has been shown[132] to be associated with down-regulation in non-metastatic breast cancer tumors exhibiting amplification in the 20q13 region, involving *ZNF217* locus only.

I finally investigated the potential impact of associated variants on the transcription of genes linked to DDR pathways. Functional mapping of variants to genes, based on eQTL information from breast tissue, identified 165 *cis*-eQTL links involving 154 variants, that mapped on 25 genomic risk loci, and 31 protein coding genes. Interestingly, while only one of these 31 *cis*-eQTL genes are known to be involved in cancer (*NNT*), a network analysis using PCNet revealed that 29 (94%) of these genes were connected to known cancer-related genes. Specifically, 21 *cis*-eQTL genes were connected to at least one oncogene, while 16 were connected to at least one tumor suppressor gene. Further, of the 31 *cis*-eQTL genes, six demonstrated significant transcript level correlations with DDR pathway related genes (Table 4.1).

Overall, 16 variants across six genomic loci were involved in cis interactions with genes that were observed co-expressed with members of the corresponding DDR pathways, for a total of 44 putative links. Interestingly, mean molecular network interactions distance among *cis*-

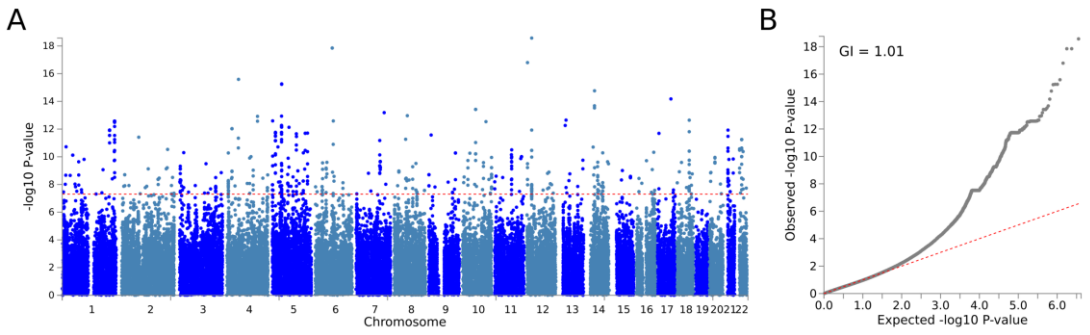


Figure 4.3 GWAS association results performed by FUMA for TLS pathway. (A) Manhattan plot displaying genome-wide associations. The red line represents genome-wide significance (5e−08). (B) Quantile-quantile (QQ) plot and genomic inflation (GI) estimates for GWAS results. Red line represents the expected

eQTL genes and co-expressed genes was 2.02, significantly smaller (p-value<1e-03) than that observed in permuted gene sets. Of note, one genomic locus demonstrated a direct connection between *cis*-eQTL and co-expressed genes within PCnet network. Specifically, this locus was found associated with TLS pathway and was observed to decrease the expression of *IBTK* gene in breast tissue, which in turn was positively correlated with *REV3L* transcript level, link that is supported by a direct interaction in the molecular network. This suggests that patients carrying alternative alleles at this locus may constitutively exhibit lower *REV3L* expression, potentially leading to TLS pathway dysregulation. While the role of *REV3L* in cancer is still under investigation, multiple studies have reported that *REV3L* down-regulation or depletion contribute to genomic instability during neoplastic transformation and progression[133], leading to the accumulation of double-strand breaks[134].

These results provide evidence supporting the existence of functional links between GWAS associated variants, the corresponding DDR pathways, and cancer-related genes, thereby strengthening the validity of the GWAS results and highlighting the potential impact of germline variation on DNA repair processes and cancer susceptibility.

Table 4.1 Cis-interactions with DDR pathways. Significant co-expression between eQTL genes and genes in the GWAS variant associated DDR pathway

Pathway	Genomic Locus	cis-eQTL gene	co-expressed gene	network distance
BER	14q23.1	C14orf39	OGG1	3
			NEIL1	3
HDR	19p13.2	ZNF266	NSMCE4A	2
			EME2	3
TLS	6q14.1	IBTK	REV3L	1
			UBE2A	2
			POLM	2
			UBE2N	2
	5q13.2	FCHO2	WDR48	2
			POLK	2
			SHPRH	2
	3p26.2	LRRN1	REV3L	2
			RAD18	2
			UBE2A	2
			UBE2V2	2
	5q22.2	EPB41L4A	REV3L	2

Methods

Somatic mutational profiles

Somatic mutation and copy number alteration data were collected and integrated from TCGA for 982 breast cancer patients. To control for population stratification, only patients identified of European ancestry[49] were considered. Briefly, a gene was classified as altered for each patient if it had a non-silent somatic mutation or fell within a CNA region. To maintain biological relevance, only CNAs consistent with the role of the gene (i.e., amplification of oncogenes and deep deletion of tumor suppressor genes) were retained. For each patient, somatic mutational profile is represented as a binary (1, 0) profile of gene alterations, in which a '1' indicates a gene for which mutation(s) has occurred in the tumor relative to germ line. Breast cancer subtypes annotations were collected from[49]. A total of 731 patients with somatic alterations in 18,684 genes were considered in the analysis.

PPI network and interaction features

I downloaded Parsimonious Composite Network (PCNet) via the NDEX browser (www.ndexbio.org/), a resource detailing molecular interactions among human genes. Within this network, nodes represent genes, and edges represent various types of functional relationships between genes, such as protein binding interactions, transcriptional regulation and signaling by phosphorylation. Molecular interaction was not preprocessed as the authors recommend it as a consensus network. I annotated each interaction with a set of edge features, including 76 distinct interaction features distributed across nine categories, derived from Pathway Commons (v11)[135] data and as completely explained in[129]. These features are designed to weigh the interactions between genes, guiding the direction of propagation to maximize the agreement among tumors of the same subtype.

Network propagation algorithm

I performed network propagation using NBS[2] algorithm to aggregate and amplify the effects of tumor mutations using knowledge of molecular interaction networks. The mutational profile for each patient independently is projected onto a human gene interaction network to learn the mutated subnetworks underlying tumor subtypes using a supervised approach (Figure 4.4).

Briefly, given the graph obtained from PCNet network, Random Walk with Restart (RWR) was conducted iteratively as follow:

$$P^{(t+1)} = (1-\alpha)P^{(t)} \cdot Q + \alpha P^{(0)}$$

Where $P^{(0)}$ is a tumor-by-gene binary matrix representing the mutational profile of each patient and Q is the degree-normalized adjacency matrix of the network graph. Adjacency matrix Q is directly learned from data, ensuring that the stratification of propagated mutation profiles resulting from the random walk closely aligns with the predefined tumor subtypes. The parameter α denotes the restart probability, governing the distance that mutation signal is allowed to propagate through the network. Upon convergence, when $P^{(t+1)} \approx P^{(t)}$, the propagated mutation profile matrix P represent a tumor-by-gene matrix where somatic alteration profiles have been 'smoothed' by the network. The score of each gene represents its network proximity to all genes with mutations. The cost function J is used to find optimal edges feature weights w to minimize the subtype classification error on the propagated mutational profiles P. The cost function is regularized using two hyperparameters λ and β to control respectively sparsity and non-linearity of the model (specific algorithm implementation is detailed in the NBS[2] publication methods section).

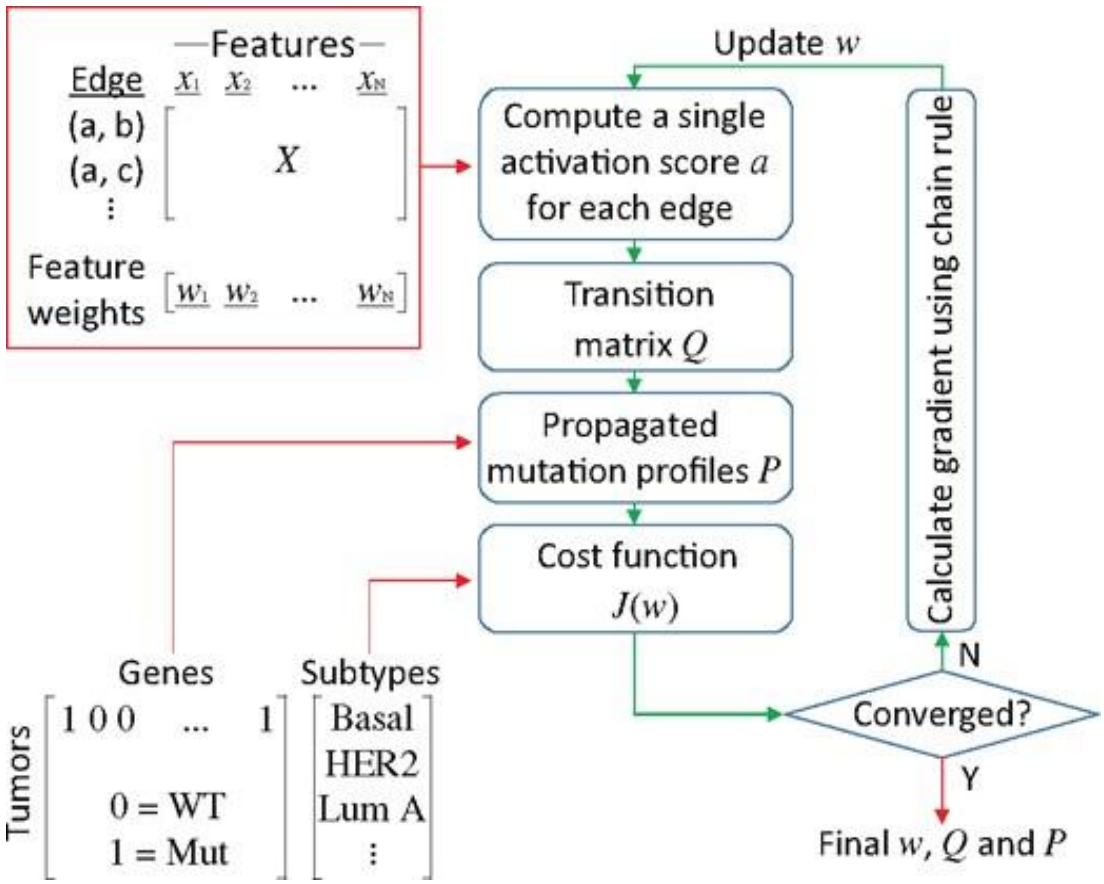


Figure 4.4 NBS[2] workflow (as published in[129] by permission of Oxford University Press). The final feature weights (w), transition matrix (Q), and propagated mutation profiles (P) are computed as described in the original NBS[2] publication.

Breast cancer propagated mutational scores

The somatic profile data for breast cancer patients was partitioned into training and validation sets (66% and 33% respectively). Within the training set a three-fold cross-validation approach was used to optimize the hyperparameters α, λ, and β. Specifically, the training set was randomly divided into three equal-sized, and non-overlapping subsets. The NBS[2] algorithm was applied to compute the AUC to assess the performance of the classification across various values of α, λ, and β performing a grid-search strategy. The optimal values for each hyperparameter were selected based on the highest AUC performance score achieved averaging the score across the three folds. The final classifier was built using the complete training set and the validation set to assess its performance for the classification of tumor subtypes on unseen data. To obtain the propagated mutational score for all breast cancer patients, the classifier is finally applied to the entire somatic profile dataset for further analysis.

Upward mobility genes identification

To extend the spectrum of cancer-relevant genes, I performed an integrative approach[130] to identify rarely mutated genes that show a significant rank improvement after mutational propagation. Specifically, the rank before and after propagation is calculated for each gene as the arithmetic average score across samples of the mutational profile P before or after propagation respectively. The mobility status of a gene is then calculated as the difference between initial and final rank scores. Finally, according with the authors' definition, genes classified as UMG were those that demonstrated a substantial improvement of at least $\beta \cdot |G|$ ranks (where β is equal to 0.25, previously determined by the authors and specific for breast cancer cohorts, and $|G|$ is the number of nodes of the network) and were ranked within the top 1,000 genes after the network propagation process. The gene ranking generated from the raw and propagated mutational profiles were both tested for enrichment of genes previously established as functionally important in various cancer types. Specifically, oncogenes (N=82), tumor suppressor genes (N=63), and a general set of cancer-related genes (N=920) were identified using a comprehensive list compiled from the scientific literature and used to evaluate whether the ranking scores of these genes were statistically higher than those of other genes.

GWAS associations with DNA damage repair pathways

I defined a set of continuous traits based on the mutational profiles after propagation corresponding to nine major DDR pathways: base excision repair (BER), nucleotide excision repair (NER), mismatch repair (MMR), the Fanconi anemia (FA) pathway, homology-dependent recombination (HR), non-homologous DNA end joining (NHEJ), direct damage reversal/repair (DR), translesion DNA synthesis (TLS), and nucleotide pool maintenance (NP). A total of 212 genes across all the DDR pathways were considered in the analysis. Specifically, a DNA damage repair (DDR) pathway score was calculated for each patient by summing the propagated somatic mutational scores from P of all genes belonging to that pathway.

Genotype calls derived from Affymetrix SNP Array 6.0 intensities of normal (non-tumor) samples from the Breast Invasive Carcinoma (BRCA) cohort were obtained from the TCGA legacy archive (portal.gdc.cancer.gov/legacy-archive). Genotype calls with an error rate greater than 10% were set to missing, and the data was reformatted using PLINK v2. Genotype calls with a call rate below 0.75 were removed. The haplotype structure was inferred with SHAPEIT v2[136]. To impute missing genotypes, IMPUTE v2.3.2[137] was performed, utilizing a reference panel constructed from the 1000 Genomes Project data. The imputed genotype calls were then intersected with imputed GTEx genotype data obtained from dbGaP (phs000424.p7.v2). Samples with an overall call rate less than 0.9 were excluded, and only variants with a minor allele frequency (MAF) of 1% or greater were retained.

I then performed a set of GWAS analyses for each considered DNA damage repair pathway within BRCA dataset. Associations of SNPs and traits were performed with PLINK v2 using linear regression with age at diagnosis, sequencing plate, and the first 3 principal components as covariates.

Functional, cis-eQTL and co-expression analyses

I then performed functional mapping and annotation of GWAS results with FUMA (v1.6.1), an integrated web-based platform[138]. Genomic risk loci were defined around significant variants (<5e-08); the genomic risk loci included all variants correlated (R2>0.6) with the most significant variant. Genome-wide gene association analysis was performed using MAGMA v.1.08[139], utilizing GWAS summary statistics. Additionally, MAGMA gene-set analysis was conducted on 17,023 gene sets from the MSigDB v2023.1Hs collection. Gene sets were considered significant if p-value<0.05 after Bonferroni correction for the number of tested gene sets. Finally, to establish links between associated variants and gene expression, eQTL

mapping was performed using FUMA for breast mammary tissue data from GTEx v.8. Each *cis*-eQTL gene identified in breast mammary tissue was tested for co-expression with all other protein-coding genes expressed in the same tissue, using Pearson correlation and correcting p-values with FDR method. Co-expressions were considered significant only if the correlation coefficient was smaller than –0.50 or greater than 0.50 and with FDR<0.05.

Discussion

In this chapter, I presented a deep exploration of the biological links between germline variants and their impact on the transcriptome of genes involved in DDR mechanisms. I first described a network propagation-based approach that is particularly effective in estimating the functional significance of rarely or moderately mutated genes in breast cancer. Moreover, I showed that upward mobility genes were enriched in cancer-related genes. This result underscores the importance of considering the broader mutational landscape, beyond high-frequency driver mutations, to understand the complex molecular mechanisms underlying oncogenesis. In combination with known driver genes, these UMGs contribute to a more comprehensive understanding of breast cancer mechanisms.

Using propagated mutational score profiles, I dug further into the exploration of germline and somatic interplay through a GWAS-based approach. This analysis revealed evidence that germline genetics can influence the mutational pattern of specific DDR pathways, highlighting the potential impact of individual genetic backgrounds on the activity and stability of fundamental biological processes that are frequently dysregulated in cancer. Notably, a substantial proportion of the identified GWAS-associated variants were known *cis*-eQTLs of genes closely connected to oncogenes, tumor suppressor genes, or other cancer-related genes. Furthermore, I identified functional links between specific associated variants and their corresponding DDR pathway's genes expression. The integration and analysis of diverse matched omics data enabled the identification and functional characterization of putative links between specific germline variants and the dysregulation of specific DDR pathways. This integrative approach highlights the power of multi-omics analyses in uncovering the complex genetic underpinnings of cancer.

Moving forward, it is important for future large-scale studies to continue exploring the intricate links between germline genetics and somatic aberrations. This method is broadly applicable to any cohort of cancer patients with the ultimate goal of identify robust cancer

risk biomarkers in both pan-cancer and cancer-specific contexts, ultimately advancing our understanding of cancer evolution and informing personalized prevention and treatment strategies.

Conclusion and future directions

In the context of genetic research, GWAS studies have emerged as a powerful and widely used methodology for detecting associations between phenotypes and genetic variants. Through a significant increase in published GWAS results, the utility of this method in advancing the understanding of complex disease genetics has become increasingly evident.

While GWAS has revolutionized the field, a critical challenge persists in the identification of causal variants from these results. Several studies have explored approaches to address these challenges, focusing on both data pre-processing and post-GWAS analyses.

Data pre-processing prior to GWAS analysis plays an important role in mitigating potential biases and improving the accuracy of results. Factors such as non-random sampling and population stratification can introduce biases and confound the identification of true genetic associations. At the same time, post-GWAS analyses support the accurate identification of causal variants and elucidate their potential mechanisms of action. The integration of advanced pre-processing techniques and comprehensive post-GWAS analyses enhance the understanding of genetics of complex diseases, contributing to more personalized approaches in medicine and healthcare.

In this thesis, I have explored the intricate relationship between inherited genetic variation and somatic events in adult cancer.

First, to elucidate the impact of genetic variations within biological systems, I developed CONREL, a web-based tool for exploring transcriptional cis-regulatory elements and understanding TF:DNA interactions using total binding affinities. This tool represents a significant advancement in our ability to identify functional variants and comprehend their potential effects on specific biological pathways. Furthermore, I explored and improved EthSEQ, a tool to define ancestry structure within individuals. This analysis underscores the critical importance of considering ancestry information in investigating disease mechanisms and prediction of therapy responses.

In the main part of the thesis, I implemented a GWAS-based approach to explore how germline genetics can influence the aberration of specific oncogenic signaling pathways. A comprehensive post-GWAS integrative analysis has revealed that germline variants can significantly impact the somatic evolution of tumors. Notably, a large fraction of the associated SNPs was known cis-eQTLs of genes closely connected to oncogenes, tumor

suppressor genes, or cancer-related genes. Moreover, integrating diverse matched omics data, I identified functional links between specific GWAS-associated SNPs and the dysregulation of oncogenic pathways. Extending upon this approach, I exploited the concept of polygenic scores to investigate patients' genetic liability to develop specific molecular profiles or particularly aggressive forms of cancer. This analysis demonstrated that an individual's genetic background may influence the dysregulation of biological oncogenic processes of specific tumor subtypes or particularly aggressive cancers.

Looking ahead, to identify cancer risk biomarkers, it is important for future large-scale studies to further investigate the complex links between germline genetics and somatic aberrations. In the last part of the thesis, I examined the method of network propagation. This approach has been used to rank genes and amplify weak associations of genes with phenotypes, offering a more comprehensive understanding of cancer mechanisms in specific pathway potentially implicated in cancer. The application of network propagation in combination with diverse matched omics data elucidated the power of multi-omics analyses in unraveling the intricate genetic landscape of cancer. This integrative approach expanded the list of potentially relevant genes in cancer, highlighting how genetic and molecular factors interact to influence cancer development and progression.

The methodologies developed in this thesis have broad applicability across various cohorts of cancer patients, with the ultimate goal of identifying robust cancer risk biomarkers in both pan-cancer and cancer-specific contexts. These advancements have the potential to significantly enhance the understanding of cancer evolution and inform personalized prevention and treatment strategies.

This approach can be extended, and more recent deep learning methodologies can be implemented to further enhance the integration and analysis of the diverse data utilized in this research. During the past years, the field of genetic research has dramatically changed with the introduction of deep learning methods. The use of deep learning techniques demonstrated superior performance in handling complex datasets and analytical tasks. The growing availability of combined high-dimension and multi-omics datasets enabled deep learning to unprecedented predictive performance in resolving intricate biological problems. Deep learning often yields better performance than traditional approaches due to its ability to scale with data size and model highly non-linear relationships. However, it is important to consider the limitations of these methods, particularly in terms of interpretability. The "black

box" nature of many deep learning models can make it challenging to understand the specific factors driving their predictions. In a clinical setting, the ability to explain and interpret results is crucial to driving decisions. To address this challenge, researchers are actively developing methods to enhance the interpretability of DL models in biological contexts.

In the near future, the integration of network propagation techniques, multi-omics analysis, and interpretable deep learning approaches will be central to explore the complex relationships between genetic background and dysregulation of oncogenic biological processes. This synergistic approach will potentially provide unprecedented understandings into cancer biology, underlying tumor initiation, progression, and treatment response.

Bibliography

1. Stratton, M. R., Campbell, P. J. & Futreal, P. A. The cancer genome. *Nature* **458**, 719–724 (2009).
2. Vogelstein, B. *et al.* Cancer Genome Landscapes. *Science* **339**, 1546–1558 (2013).
3. Rahman, N. Realizing the promise of cancer predisposition genes. *Nature* **505**, 302–308 (2014).
4. Yu, W. *et al.* GWAS Integrator: a bioinformatics tool to explore human genetic associations reported in published genome-wide association studies. *Eur J Hum Genet* **19**, 1095–1099 (2011).
5. Lu, C. *et al.* Patterns and functional implications of rare germline variants across 12 cancer types. *Nat Commun* **6**, 10086 (2015).
6. Huang, K. *et al.* Pathogenic Germline Variants in 10,389 Adult Cancers. *Cell* **173**, 355-370.e14 (2018).
7. Hanahan, D. & Weinberg, R. A. Hallmarks of Cancer: The Next Generation. *Cell* **144**, 646–674 (2011).
8. ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. Pan-cancer analysis of whole genomes. *Nature* **578**, 82–93 (2020).
9. Bailey, M. H. *et al.* Comprehensive Characterization of Cancer Driver Genes and Mutations. *Cell* **173**, 371-385.e18 (2018).
10. Sanchez-Vega, F. *et al.* Oncogenic Signaling Pathways in The Cancer Genome Atlas. *Cell* **173**, 321-337.e10 (2018).
11. Carter, H. *et al.* Interaction Landscape of Inherited Polymorphisms with Somatic Events in Cancer. *Cancer Discovery* **7**, 410–423 (2017).
12. Dalfovo, D., Scandino, R., Paoli, M., Valentini, S. & Romanel, A. Germline determinants of aberrant signaling pathways in cancer. *npj Precis. Onc.* **8**, 57 (2024).

13. Liu, Y., Gusev, A. & Kraft, P. Germline Cancer Gene Expression Quantitative Trait Loci Are Associated with Local and Global Tumor Mutations. *Cancer Research* **83**, 1191–1202 (2023).
14. Dalfovo, D. & Romanel, A. Analysis of Genetic Ancestry from NGS Data Using EthSEQ. *Current Protocols* **3**, (2023).
15. Rada-Iglesias, A. *et al.* A unique chromatin signature uncovers early developmental enhancers in humans. *Nature* **470**, 279–283 (2011).
16. Creyghton, M. P. *et al.* Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 21931–21936 (2010).
17. Heintzman, N. D. *et al.* Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature* **459**, 108–112 (2009).
18. Strahl, B. D. & Allis, C. D. The language of covalent histone modifications. *Nature* **403**, 41–45 (2000).
19. The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
20. Roadmap Epigenomics Consortium *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).
21. Zerbino, D. R., Wilder, S. P., Johnson, N., Juettemann, T. & Flicek, P. R. The Ensembl Regulatory Build. *Genome Biol* **16**, 56 (2015).
22. Gao, T. *et al.* EnhancerAtlas: a resource for enhancer annotation and analysis in 105 human cell/tissue types. *Bioinformatics* **32**, 3543–3551 (2016).
23. Ashoor, H., Kleftogiannis, D., Radovanovic, A. & Bajic, V. B. DENdb: database of integrated human enhancers. *Database* **2015**, bav085 (2015).
24. Fishilevich, S. *et al.* GeneHancer: genome-wide integration of enhancers and target genes in GeneCards. *Database* **2017**, (2017).

25. Griffon, A. *et al.* Integrative analysis of public ChIP-seq experiments reveals a complex multi-cell regulatory landscape. *Nucleic Acids Research* **43**, e27–e27 (2015).
26. Gheorghe, M. *et al.* A map of direct TF–DNA interactions in the human genome. *Nucleic Acids Research* **47**, e21–e21 (2019).
27. Tanay, A. Extensive low-affinity transcriptional interactions in the yeast genome. *Genome Res.* **16**, 962–972 (2006).
28. Thomas-Chollier, M. *et al.* Transcription factor binding predictions using TRAP for the analysis of ChIP-seq data and regulatory SNPs. *Nat Protoc* **6**, 1860–1869 (2011).
29. Foat, B. C., Morozov, A. V. & Bussemaker, H. J. Statistical mechanical modeling of genome-wide transcription factor occupancy data by MatrixREDUCE. *Bioinformatics* **22**, e141–e149 (2006).
30. Ward, L. D. & Bussemaker, H. J. Predicting functional transcription factor binding through alignment-free and affinity-based analysis of orthologous promoter sequences. *Bioinformatics* **24**, i165–i171 (2008).
31. Dalfovo, D., Valentini, S. & Romanel, A. Exploring functionally annotated transcriptional consensus regulatory elements with CONREL. *Database* **2020**, baaa071 (2020).
32. The ENCODE Project Consortium *et al.* Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* **583**, 699–710 (2020).
33. Han, H. *et al.* TRRUST v2: an expanded reference database of human and mouse transcriptional regulatory interactions. *Nucleic Acids Research* **46**, D380–D386 (2018).
34. Tai, S. *et al.* PC3 is a cell line characteristic of prostatic small cell carcinoma. *The Prostate* **71**, 1668–1679 (2011).
35. Dreos, R., Ambrosini, G., Groux, R., Cavin Périer, R. & Bucher, P. The eukaryotic promoter database in its 30th year: focus on non-vertebrate organisms. *Nucleic Acids Res* **45**, D51–D55 (2017).

36. Nakamura, T. & Mizuno, S. The discovery of Hepatocyte Growth Factor (HGF) and its significance for cell biology, life sciences and clinical medicine. *Proc. Jpn. Acad., Ser. B* **86**, 588–610 (2010).

37. Hinrichs, A. S. The UCSC Genome Browser Database: update 2006. *Nucleic Acids Research* **34**, D590–D598 (2006).

38. Khan, A. *et al.* JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework. *Nucleic Acids Research* **46**, D260–D266 (2018).

39. Xie, Z., Hu, S., Blackshaw, S., Zhu, H. & Qian, J. hPDI: a database of experimental human protein–DNA interactions. *Bioinformatics* **26**, 287–289 (2010).

40. Pachkov, M., Balwierz, P. J., Arnold, P., Ozonov, E. & Van Nimwegen, E. SwissRegulon, a database of genome-wide annotations of regulatory sites: recent updates. *Nucleic Acids Research* **41**, D214–D220 (2012).

41. Kulakovskiy, I. V. *et al.* HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis. *Nucleic Acids Research* **46**, D252–D259 (2018).

42. Matys, V. TRANSFAC(R) and its module TRANSCompel(R): transcriptional gene regulation in eukaryotes. *Nucleic Acids Research* **34**, D108–D110 (2006).

43. Hume, M. A., Barrera, L. A., Gisselbrecht, S. S. & Bulyk, M. L. UniPROBE, update 2015: new tools and content for the online database of protein-binding microarray data on protein–DNA interactions. *Nucleic Acids Research* **43**, D117–D122 (2015).

44. Weirauch, M. T. *et al.* Determination and Inference of Eukaryotic Transcription Factor Sequence Specificity. *Cell* **158**, 1431–1443 (2014).

45. Jolma, A. *et al.* DNA-Binding Specificities of Human Transcription Factors. *Cell* **152**, 327–339 (2013).

46. Keane, T. M. *et al.* Mouse genomic variation and its effect on phenotypes and gene regulation. *Nature* **477**, 289–294 (2011).
47. Delaneau, O., Marchini, J. & Zagury, J.-F. A linear complexity phasing method for thousands of genomes. *Nat Methods* **9**, 179–181 (2012).
48. Beltran, H. *et al.* Whole-Exome Sequencing of Metastatic Cancer and Biomarkers of Treatment Response. *JAMA Oncol* **1**, 466 (2015).
49. Carrot-Zhang, J. *et al.* Comprehensive Analysis of Genetic Ancestry and Its Molecular Correlates in Cancer. *Cancer Cell* **37**, 639-654.e6 (2020).
50. Yuan, J. *et al.* Integrated Analysis of Genetic Ancestry and Genomic Alterations across Cancers. *Cancer Cell* **34**, 549-560.e9 (2018).
51. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664 (2009).
52. Li, Y. *et al.* FastPop: a rapid principal component derived method to infer intercontinental ancestry using genetic data. *BMC Bioinformatics* **17**, 122 (2016).
53. Privé, F., Luu, K., Blum, M. G. B., McGrath, J. J. & Vilhjálmsson, B. J. Efficient toolkit implementing best practices for principal component analysis of population genetic data. *Bioinformatics* **36**, 4449–4457 (2020).
54. Romanel, A., Zhang, T., Elemento, O. & Demichelis, F. EthSEQ: ethnicity annotation from whole exome sequencing data. *Bioinformatics* **33**, 2402–2404 (2017).
55. Bergström, A. *et al.* Insights into human genetic variation and population history from 929 diverse genomes. *Science* **367**, eaay5012 (2020).
56. Bae, C. J., Douka, K. & Petraglia, M. D. On the origin of modern humans: Asian perspectives. *Science* **358**, eaai9067 (2017).

57. Lim, W. C. *et al.* Divergent HLA variations and heterogeneous expression but recurrent HLA loss-of- heterozygosity and common HLA-B and TAP transcriptional silencing across advanced pediatric solid cancers. *Front. Immunol.* **14**, 1265469 (2024).

58. Zheng, X. *et al.* SeqArray-a storage-efficient high-performance data format for WGS variant calls. *Bioinformatics* **33**, 2251–2257 (2017).

59. Frankish, A. *et al.* GENCODE 2021. *Nucleic Acids Research* **49**, D916–D923 (2021).

60. Romanel, A., Lago, S., Prandi, D., Sboner, A. & Demichelis, F. ASEQ: fast allele-specific studies from next-generation sequencing data. *BMC Med Genomics* **8**, 9 (2015).

61. Sud, A., Kinnersley, B. & Houlston, R. S. Genome-wide association studies of cancer: current insights and future perspectives. *Nat Rev Cancer* **17**, 692–704 (2017).

62. Hosking, F. J., Dobbins, S. E. & Houlston, R. S. Genome-wide association studies for detecting cancer susceptibility. *British Medical Bulletin* **97**, 27–46 (2011).

63. Galvan, A., Ioannidis, J. P. A. & Dragani, T. A. Beyond genome-wide association studies: genetic heterogeneity and individual predisposition to cancer. *Trends in Genetics* **26**, 132–141 (2010).

64. Manolio, T. A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461**, 747–753 (2009).

65. Chang, C. Q. *et al.* A systematic review of cancer GWAS and candidate gene meta-analyses reveals limited overlap but similar effect sizes. *Eur J Hum Genet* **22**, 402–408 (2014).

66. Varghese, J. S. & Easton, D. F. Genome-wide association studies in common cancers—what have we learnt? *Current Opinion in Genetics & Development* **20**, 201–209 (2010).

67. Torkamani, A., Wineinger, N. E. & Topol, E. J. The personal and clinical utility of polygenic risk scores. *Nat Rev Genet* **19**, 581–590 (2018).

68. Stratton, M. R., Campbell, P. J. & Futreal, P. A. The cancer genome. *Nature* **458**, 719–724 (2009).

69. Vogelstein, B. *et al.* Cancer Genome Landscapes. *Science* **339**, 1546–1558 (2013).

70. Carter, H. *et al.* Interaction Landscape of Inherited Polymorphisms with Somatic Events in Cancer. *Cancer Discov* **7**, 410–423 (2017).

71. Castro, M. A. A. *et al.* Regulators of genetic risk of breast cancer identified by integrative network analysis. *Nat Genet* **48**, 12–21 (2016).

72. Romanel, A. *et al.* Inherited determinants of early recurrent somatic mutations in prostate cancer. *Nat Commun* **8**, 48 (2017).

73. Guo, J. *et al.* Inherited polygenic effects on common hematological traits influence clonal selection on JAK2V617F and the development of myeloproliferative neoplasms. *Nat Genet* (2024) doi:10.1038/s41588-023-01638-x.

74. Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: the next generation. *Cell* **144**, 646–674 (2011).

75. The Cancer Genome Atlas Research Network *et al.* The Cancer Genome Atlas Pan-Cancer analysis project. *Nat Genet* **45**, 1113–1120 (2013).

76. ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. Pan-cancer analysis of whole genomes. *Nature* **578**, 82–93 (2020).

77. Barretina, J. *et al.* The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483**, 603–607 (2012).

78. Ghandi, M. *et al.* Next-generation characterization of the Cancer Cell Line Encyclopedia. *Nature* **569**, 503–508 (2019).

79. Buniello, A. *et al.* The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res* **47**, D1005–D1012 (2019).

80. Mittlböck, M. & Heinzl, H. A simulation study comparing properties of heterogeneity measures in meta-analyses. *Stat Med* **25**, 4321–4333 (2006).

81. Valentini, S. *et al.* Polympact: exploring functional relations among common human genetic variants. *Nucleic Acids Res* gkac024 (2022) doi:10.1093/nar/gkac024.

82. Murphree, A. L. & Benedict, W. F. Retinoblastoma: clues to human oncogenesis. *Science* **223**, 1028–1033 (1984).

83. Liu, J. *et al.* An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality Survival Outcome Analytics. *Cell* **173**, 400-416.e11 (2018).

84. Demichelis, F. *et al.* Identification of functionally active, low frequency copy number variants at 15q21.3 and 12q21.31 associated with prostate cancer risk. *Proceedings of the National Academy of Sciences* **109**, 6686–6691 (2012).

85. Schaefer, G. *et al.* Distinct ERG rearrangement prevalence in prostate cancer: higher frequency in young age and in low PSA prostate cancer. *Prostate Cancer Prostatic Dis* **16**, 132–138 (2013).

86. Gordetsky, J. & Epstein, J. Grading of prostatic adenocarcinoma: current state and prognostic implications. *Diagn Pathol* **11**, 25 (2016).

87. Sayaman, R. W. *et al.* Germline genetic contribution to the immune landscape of cancer. *Immunity* **54**, 367-386.e8 (2021).

88. Musa, J. & Grünewald, T. G. P. Interaction between somatic mutations and germline variants contributes to clinical heterogeneity in cancer. *Molecular & Cellular Oncology* **7**, 1682924 (2020).

89. Mamidi, T. K. K., Wu, J. & Hicks, C. Integrating germline and somatic variation information using genomic data for the discovery of biomarkers in prostate cancer. *BMC Cancer* **19**, 229 (2019).

90. Musa, J. *et al.* Cooperation of cancer drivers with regulatory germline variants shapes clinical outcomes. *Nat Commun* **10**, 4128 (2019).
91. Liu, H.-M. *et al.* Recessive/dominant model: Alternative choice in case-control-based genome-wide association studies. *PLoS ONE* **16**, e0254947 (2021).
92. Guindo-Martínez, M. *et al.* The impact of non-additive genetic associations on age-related complex diseases. *Nat Commun* **12**, 2436 (2021).
93. Liu, N., Zhao, H., Patki, A., Limdi, N. A. & Allison, D. B. Controlling Population Structure in Human Genetic Association Studies with Samples of Unrelated Individuals. *Stat Interface* **4**, 317–326 (2011).
94. Astle, W. & Balding, D. J. Population Structure and Cryptic Relatedness in Genetic Association Studies. *Statist. Sci.* **24**, (2009).
95. Yang, J., Zaitlen, N. A., Goddard, M. E., Visscher, P. M. & Price, A. L. Advantages and pitfalls in the application of mixed-model association methods. *Nat Genet* **46**, 100–106 (2014).
96. Sul, J. H. & Eskin, E. Mixed models can correct for population structure for genomic regions under selection. *Nat Rev Genet* **14**, 300–300 (2013).
97. Tucker, G., Price, A. L. & Berger, B. Improving the Power of GWAS and Avoiding Confounding from Population Stratification with PC-Select. *Genetics* **197**, 1045–1049 (2014).
98. TCGA Analysis Network *et al.* Ancestry-specific predisposing germline variants in cancer. *Genome Med* **12**, 51 (2020).
99. Martinez, V. D. *et al.* Disruption of KEAP1/CUL3/RBX1 E3-ubiquitin ligase complex components by multiple genetic mechanisms: Association with poor prognosis in head and neck cancer. *Head Neck* **37**, 727–734 (2015).

100. Martinez, V. D. *et al.* Unique pattern of component gene disruption in the NRF2 inhibitor KEAP1/CUL3/RBX1 E3-ubiquitin ligase complex in serous ovarian cancer. *Biomed Res Int* **2014**, 159459 (2014).

101. Porta-Pardo, E., Sayaman, R., Ziv, E. & Valencia, A. *The Landscape of Interactions between Cancer Polygenic Risk Scores and Somatic Alterations in Cancer Cells*. http://biorxiv.org/lookup/doi/10.1101/2020.09.28.316851 (2020)

102. Schizophrenia Working Group of the Psychiatric Genomics Consortium *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat Genet* **47**, 291–295 (2015).

103. Cowen, L., Ideker, T., Raphael, B. J. & Sharan, R. Network propagation: a universal amplifier of genetic associations. *Nat Rev Genet* **18**, 551–562 (2017).

104. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaSci* **4**, 7 (2015).

105. Pluzhnikov, A. *et al.* Spoiling the whole bunch: quality control aimed at preserving the integrity of high-throughput genotyping. *Am J Hum Genet* **87**, 123–128 (2010).

106. Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* **38**, 904–909 (2006).

107. Cerami, E. *et al.* The cBio Cancer Genomics Portal: An Open Platform for Exploring Multidimensional Cancer Genomics Data: Figure 1. *Cancer Discovery* **2**, 401–404 (2012).

108. Gao, J. *et al.* Integrative Analysis of Complex Cancer Genomics and Clinical Profiles Using the cBioPortal. *Science Signaling* **6**, pl1–pl1 (2013).

109. Oughtred, R. *et al.* The BioGRID interaction database: 2019 update. *Nucleic Acids Res* **47**, D529–D541 (2019).

110. Peri, S. *et al.* Human protein reference database as a discovery resource for proteomics. *Nucleic Acids Res* **32**, D497-501 (2004).

111. Orchard, S. *et al.* The MIntAct project--IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res* **42**, D358-363 (2014).

112. Huttlin, E. L. *et al.* Architecture of the human interactome defines protein communities and disease networks. *Nature* **545**, 505–509 (2017).

113. Szklarczyk, D. *et al.* STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res* **47**, D607–D613 (2019).

114. Wilks, C. *et al.* recount3: summaries and queries for large-scale RNA-seq expression and splicing. *Genome Biol* **22**, 323 (2021).

115. Choi, S. W., Mak, T. S.-H. & O'Reilly, P. F. Tutorial: a guide to performing polygenic risk score analyses. *Nat Protoc* **15**, 2759–2772 (2020).

116. Zhang, Y. D. *et al.* Assessment of polygenic architecture and risk prediction based on common variants across fourteen cancers. *Nat Commun* **11**, 3353 (2020).

117. Chatterjee, N., Shi, J. & García-Closas, M. Developing and evaluating polygenic risk prediction models for stratified disease prevention. *Nat Rev Genet* **17**, 392–406 (2016).

118. Pounds, S. & Cheng, C. Robust estimation of the false discovery rate. *Bioinformatics* **22**, 1979–1987 (2006).

119. Oti, M. Predicting disease genes using protein-protein interactions. *Journal of Medical Genetics* **43**, 691–698 (2006).

120. Franke, L. *et al.* Reconstruction of a Functional Human Gene Network, with an Application for Prioritizing Positional Candidate Genes. *The American Journal of Human Genetics* **78**, 1011–1025 (2006).

121. Page, Lawrence, Brin, Sergey, Motwani, Rajeev & Winograd, Terry. The PageRank Citation Ranking: Bringing Order to the Web. *Technical Report* Stanford InfoLab, (1999).

122. Krapivsky, P. L., Redner, S. & Ben-Naim, E. *A Kinetic View of Statistical Physics*. (Cambridge University Press, 2010). doi:10.1017/CBO9780511780516.

123. Noble, W. S., Kuang, R., Leslie, C. & Weston, J. Identifying remote protein homologs by network propagation. *The FEBS Journal* 272, 5119–5128 (2005).

124. Mitra, K., Carvunis, A.-R., Ramesh, S. K. & Ideker, T. Integrative approaches for finding modular structure in biological networks. *Nat Rev Genet* 14, 719–732 (2013).

125. Pearson, K. The Problem of the Random Walk. *Nature* 72, 342–342 (1905).

126. Hofree, M., Shen, J. P., Carter, H., Gross, A. & Ideker, T. Network-based stratification of tumor mutations. *Nat Methods* 10, 1108–1115 (2013).

127. Kuchenbaecker, K. B. *et al.* Risks of Breast, Ovarian, and Contralateral Breast Cancer for *BRCA1* and *BRCA2* Mutation Carriers. *JAMA* 317, 2402 (2017).

128. The Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature* 490, 61–70 (2012).

129. Zhang, W., Ma, J. & Ideker, T. Classifying tumors by supervised network propagation. *Bioinformatics* 34, i484–i493 (2018).

130. Mohsen, H. *et al.* Network propagation-based prioritization of long tail genes in 17 cancer types. *Genome Biol* 22, 287 (2021).

131. Knijnenburg, T. A. *et al.* Genomic and Molecular Landscape of DNA Damage Repair Deficiency across The Cancer Genome Atlas. *Cell Reports* 23, 239-254.e6 (2018).

132. Ginestier, C. *et al.* Prognosis and Gene Expression Profiling of 20q13-Amplified Breast Cancers. *Clinical Cancer Research* 12, 4533–4544 (2006).

133. Wittschieben, J. P., Reshmi, S. C., Gollin, S. M. & Wood, R. D. Loss of DNA Polymerase ζ Causes Chromosomal Instability in Mammalian Cells. *Cancer Research* **66**, 134–142 (2006).

134. Brondello, J.-M. *et al.* Novel evidences for a tumor suppressor role of Rev3, the catalytic subunit of Pol ζ. *Oncogene* **27**, 6093–6101 (2008).

135. Rodchenkov, I. *et al.* Pathway Commons 2019 Update: integration, analysis and exploration of pathway data. *Nucleic Acids Research* gkz946 (2019) doi:10.1093/nar/gkz946.

136. Delaneau, O., Howie, B., Cox, A. J., Zagury, J.-F. & Marchini, J. Haplotype Estimation Using Sequencing Reads. *The American Journal of Human Genetics* **93**, 687–696 (2013).

137. Howie, B. N., Donnelly, P. & Marchini, J. A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies. *PLoS Genet* **5**, e1000529 (2009).

138. Watanabe, K., Taskesen, E., Van Bochoven, A. & Posthuma, D. Functional mapping and annotation of genetic associations with FUMA. *Nat Commun* **8**, 1826 (2017).

139. De Leeuw, C. A., Mooij, J. M., Heskes, T. & Posthuma, D. MAGMA: Generalized Gene-Set Analysis of GWAS Data. *PLoS Comput Biol* **11**, e1004219 (2015).