PSL UNIVERSITÉ PARIS

Préparée à MINES ParisTech

THÈSE DE DOCTORAT
DE L'UNIVERSITÉ PSL

Network-guided genome-wide association studies

Soutenue par

Héctor GONZÁLEZ CLIMENTE

Le 1 Février 2020

École doctorale n°432

Sciences et Métiers de l'Ingénieur

Spécialité

Bio-informatique

Composition du jury :

Nadine ANDRIEU Mme., Institut Curie	*Examinatrice*
Kristel VAN STEEN Mme., Université de Liège	*Rapporteuse*
Antonio RAUSELL M., Imagine Institute	*Rapporteur*
Laura FURLONG Mme., Pompeu Fabra University	*Examinatrice*
Véronique STOVEN Mme., MINES ParisTech	*Directrice de thèse*
Chloé-Agathe AZENCOTT Mme., MINES ParisTech	*Co-encadrante de thèse*

Contents

Preface **1**

1 Context **3**
1.1 The common disease/common variant framework . . . 3
1.2 Epistasis . . . 4
1.3 Genome-wide association studies . . . 5
1.3.1 Challenges . . . 6
1.3.1.1 Low statistical power . . . 6
1.3.1.2 Choice of encoding . . . 7
1.3.1.3 Estimating individual risk . . . 7
1.3.1.4 Population structure . . . 8
1.3.1.5 Interpretability . . . 8
1.4 Genome-wide association interaction studies . . . 9
1.5 Diseases studied in this thesis . . . 10
1.5.1 Breast cancer . . . 10
1.5.1.1 The GENESIS dataset . . . 12
1.5.2 Inflammatory bowel disease . . . 12
1.5.2.1 The IIBDGC dataset . . . 13
1.6 Network view of complex diseases . . . 13
1.6.1 Networks in disease . . . 15
1.6.2 Network-guided approaches to disease study . . . 16
1.6.2.1 High-score subnetwork search . . . 16
1.6.2.2 Module detection . . . 16
1.6.2.3 Aggregation of networks . . . 16
1.7 Contributions . . . 17

2 Combining network-guided GWAS to discover susceptibility mechanisms for breast cancer **21**
2.1 Introduction . . . 21
2.2 Methods . . . 23
2.2.1 GENESIS . . . 23
2.2.2 Preprocessing and quality control . . . 23
2.2.3 High-score subnetwork search algorithms . . . 25
2.2.3.1 SNP and gene association . . . 25
2.2.3.2 Mathematical notation . . . 25
2.2.3.3 Methods used . . . 26
2.2.3.4 Gene-gene network . . . 30

2.2.3.5 SNP networks 30
2.2.3.6 Consensus network 30
2.2.4 Evaluation of methods 31
2.2.4.1 Classification accuracy of selected biomarkers 31
2.2.4.2 Biological relevance of the genes 31
2.2.5 Code availability 32
2.3 Results 32
2.3.1 A conventional GWAS shows that FGFR2 is strongly associated with familial breast cancer 32
2.3.2 Network methods successfully identify genes associated with breast cancer 34
2.3.3 heinz retrieves a small, highly informative set of biomarkers in a fast and stable fashion 41
2.3.4 No solution is perfect 41
2.3.5 Aggregating solutions provides insights into the biology of cancer 45
2.4 Discussion 49
Funding and acknowledgments 52

3 The *martini* R package 55
3.1 Introduction 55
3.2 Improvements over SConES 56
3.2.1 Additional measures of association 56
3.2.2 Hyperparameter optimization 56
3.2.2.1 Selection criterion 56
3.2.3 Network-based simulations 57
3.2.4 Interface, documentation and quality assurance 59
3.3 The `scones.nf` pipeline 60
3.4 Conclusions 60

4 Boosting interpretability and statistical power in epistasis detection by using prior biological knowledge 63
4.1 Introduction 63
4.2 Materials and methods 64
4.2.1 Dataset and initial quality control 64
4.2.2 Gene interaction detection procedure 64
4.2.2.1 Functional SNP pre-filtering 65
4.2.2.2 Post-filtering quality control 66
4.2.2.3 SNP-level epistasis detection and multiple test correction 67
4.2.2.4 From SNP-level to gene-level epistasis 67
4.3 Preliminary results 68

4.3.1 Chromatin contacts map more SNPs per gene than other mappings 68
4.3.2 The *physical* protocol does not recover any SNP interaction . . . 70
4.3.3 Gene-level network 70
4.3.4 Chromatin and Standard mappings partially replicate previous studies on IBD 73
4.3.5 The type I error of the protocol is controlled 73
4.4 Discussion 74
Acknowledgements 76

5 Epistasis networks **77**

6 Conclusions **79**

A Block HSIC Lasso: model-free biomarker detection for ultra-high dimensional data **83**

B The Functional Impact of Alternative Splicing in Cancer **93**

C Systematic Analysis of Splice-Site-Creating Mutations in Cancer **107**

D Susceptibility genes to breast cancer **125**
D.1 Homologous recombination repair 125
D.2 Replication fork stability 125
D.3 Transcription-replication collisions 126
D.4 Mismatch repair 126
D.5 DNA damage signaling, checkpoints and cell death 126

List of Tables

2.1 Summary statistics on the results of multiple network methods on the gene-gene interaction network. The first row contains the summary statistics on the whole network. 26
2.2 Summary statistics on the results of multiple network methods on the gene-gene interaction network. The first row contains the summary statistics on the whole network. 34
2.3 Summary statistics on the results of SConES on the three SNP-SNP interaction networks. The first row within each block contains the summary statistics on the whole network. 35

4.1 Properties of the different SNP-gene mappings and the filtered datasets. 70
4.2 Properties of the SNP networks from the different datasets. 70
4.3 Properties of the gene networks from the different datasets. 71

List of Figures

2.1 GENESIS shows no differential population structure between cases and controls. **(A,B,C,D)** Eight main principal components computed on the genotypes of GENESIS. Cases are colored in green, controls in orange. 24

2.2 Association in GENESIS. The red line represents the Bonferroni threshold. **(A)** SNP association, measured from the outcome of a 1df χ^2 allelic test. Significant SNPs that are within a coding gene, or within 50 kilobases of its boundaries, are annotated. The Bonferroni threshold is 2.54×10^{-7}. **(B)** Gene association, measured by P-value of VEGAS2v2 (Mishra and Macgregor 2015) using the 10% of SNPs with the lowest P-values. The Bonferroni threshold is 1.53×10^{-6}. 33

2.3 Overview of the subnetworks produced by the different network methods. **(dmGWAS, heinz, HotNet2, LEAN, and SigMod)** contain gene subnetworks; **(SConES GI)**, SNP subnetworks. 36

2.4 Manhattan plots showing the biomolecules selected by each method. In **(Consensus, dmGWAS, heinz, HotNet2, and SigMod)** datapoints are genes; in **(SConES GS, GM, and GI)**, SNPs. LEAN was excluded, as it did not select any gene. 37

2.5 Proportion of the selected genes by each of the methods on the GENESIS data that is a known breast cancer susceptibility gene (Section 2.2.4.2). Only genes present in the protein-protein interaction network were considered. LEAN is not displayed as it did not select any gene. The presented network methods recover a higher proportion of breast cancer susceptibility genes than of other genes, despite their lack of significance in GENESIS. 38

2.6 Proportion of the Bonferroni significant biomolecules (in either the GENESIS or the BCAC datasets) selected by each of the methods on the GENESIS data. **(Consensus, dmGWAS, heinz, HotNet2, and SigMod)** involve significant genes, only among those present in the protein-protein interaction network. **(SConES GS, GM and GI)** involve significant SNPs. LEAN is not displayed as it did not select any gene. The presented network methods recover a higher proportion of significant genes than of non-significant genes in both datasets, despite their lack of significance in GENESIS. 39

2.7 Genes on the consensus network. Breast cancer susceptibility genes are colored in pink; the rest are colored in grey. **(A)** Number of methods selecting every gene in the subnetwork. **(B)** VEGAS P-values of association of the genes, with regards to the number of methods that selected them. **(C)** Comparison of betweenness centrality of the genes in the consensus network and the other genes in the PPIN and not in the consensus network. To improve visualization, we removed outliers. **(D)** Relationship between the $\log_{10}$ of the betweenness centrality and the -$\log_{10}$ of the VEGAS P-value of the genes in the consensus network. The blue line represents a fitted generalized linear model. 40

2.8 Comparison of network-based GWAS methods on GENESIS. Each method was run 5 times of a random subset of the samples, and tested on the remaining samples (Section 2.2.4.1). **(A)** Number of SNPs selected by each method and number of SNPs on the active set used by the Lasso classifier. Points are the average over the 5 runs; lines represent the standard error of the mean. A grey diagonal line with slope 1 is added for comparison. For reference, the active set of Lasso using all the SNPs included, on average, 154 117.4 SNPs. **(B)** Sensitivity and specificity on test set of the L1-penalized logistic regression trained on the features selected by each of the methods. In addition, the performance of the classifier trained on all SNPs is displayed. Points are the average over the 5 runs; lines represent the standard error of the mean. **(C)** Pairwise Pearson's correlations of the solutions used by different methods. A Pearson's correlation of 1 means the two solutions are the same. A Pearson's correlation of 0 means that there is no SNP in common between the two solutions. **(D)** Runtime of the evaluated methods, by type of network used (gene or SNP). For gene network-based methods, inverted triangles represent the runtime of the algorithm itself, and circles the total time, which includes the algorithm themselves and the additional 119 980 seconds (1 day and 9.33 hours) which took VEGAS2v2 on average to compute the gene scores from SNP summary statistics. 42

2.9 Drawbacks confronted when using network guided methods. **(A)** dmGWAS solution subnetwork. Genes with a P-value < 0.1 are highlighted in red. **(B)** Centrality degree and -$\log_{10}$ of the VEGAS P-value for the nodes in SigMod solution subnetwork. **(C)** Genomic regions where either SConES GS, GM or GI select SNPs. 43

2.10 Pearson's correlation between the different solution subnetworks. **(A)** Correlation between selected SNPs. **(B)** Correlation between selected genes. In general, the solutions display a very low overlap. 44

2.11 Consensus subnetwork on GENESIS (Section 2.2.3.6). Each node is represented by a pie chart, which accounts the methods that selected it. The labeled genes have a VEGAS2v2 P-value < 0.001 and/or are known breast cancer susceptibility genes (colored in pink). 46

2.12 Consensus subnetwork on GENESIS (Section 2.2.3.6). **(A)** Each node is represented by a pie chart, which accounts the methods that selected it. The labeled genes have a VEGAS2v2 P-value < 0.001 and/or are known breast cancer susceptibility genes (colored in pink). This panel is equivalent to Figure 2.11. **(B)** The name of every gene is indicated. . 48

2.13 Biotypes of genes from the annotation that are not present in the HINT protein-protein interaction network. 51

2.14 Comparison of benchmark on high-throughput interactions to benchmark on both high-throughput and literature curated interactions. Grey lines represent no change between the benchmarks (1 for ratios, 0 for differences). **(A)** Ratios of the selected features between both benchmarks and of the active set. **(B)** Shifts in sensitivity and specificity. **(C)** Shift in Pearson's correlation between benchmarks. **(D)** Ratio between the runtimes of the benchmarks. For gene network-based methods, inverted triangles represent the ratio of runtimes of the algorithms themselves, and circles the total time, which includes the algorithm themselves and the additional 119 980 seconds (1 day and 9.33 hours) which took VEGAS2v2 on average to compute the gene scores from SNP summary statistics. In general, adding additional interactions slightly improves the stability of the solution, but increases the solution size, has mixed effects on the sensitivity and specificity, and impacts negatively the required runtime of the algorithms. 53

3.1 Allelic effect i as function of causal allele frequency p for different counts of causal allele in a patient ($x = 0, 1, 2$). 59

4.1 Overview of the gene-gene interaction detection procedure. The whole protocol is described in Section 4.2.2. 65

4.2 **(A)** Number of SNPs per gene for each of the three mappings described in Section 4.2.2.1. Outliers are not displayed to facilitate visualization. **(B)** Ranking of genes with most SNPs mapped using any of the mappings, colored by mapping. Only genes with more than 100 SNPs mapped to it are displayed. **(C,D,E)** Comparison between the rank of each gene according to the number of SNPs mapped to it using each mapping. 69

4.3 SNP-level epistasis networks for *Standard* (orange), *eQTL* (green), and *Chromatin* (violet) (Sections 4.2.2.1 and 4.2.2.3). The *physical* dataset is absent, as no SNP pairs were significant. 71
4.4 Relationship between the number of significant SNP pairs and of significant gene pairs. **(A)** Histogram of the number of significant gene pairs mapped to the same SNP pair. **(B)** Relationship between the total number of SNP pairs mapped to the same gene pair (y-axis), and the percentage of all significant SNP-pairs between all the SNP-pairs mapped to the same gene (x-axis). Data points are semi-transparent, so multiple points stacked result in a darker shade. 72
4.5 (ref:fig-gene-network-caption) 73

Preface

Ph.D. thesis

Supervised by Chloé-Agathe Azencott and Véronique Stoven.

This thesis was funded by the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie [666003].

Chapter 1

Context

1.1 The common disease/common variant framework

Complex diseases are those caused by a mixture of genetic, environmental and lifestyle factors. The object of study of this thesis are the methodologies to identify such genetic factors. This is of paramount importance for disease prevention, understanding the etiology of diseases, and providing better treatments.

The genetic architecture of a trait includes the variants that contribute to the risk, as well as their allelic frequencies, effect sizes, and their genetic mode of action (e.g. dominant or recessive). From this point of view, complex diseases are easier to understand in contrast with Mendelian traits. The latter are caused by a single locus with strong effects, and hence follow the Mendelian rules of inheritance. In essence, and barring considerations on reduced penetrance, whether an individual will develop a Mendelian disease or not depends exclusively on the two alleles at that particular locus, and its genetic mode of action. By contrast, the genetic architecture of complex is modeled by the *liability-threshold model*, an extension to binary traits of the infinitesimal model used to describe the genetics of continuous phenotypes like height. Under the infinitesimal model, a continuous trait is shaped by many Mendelian alleles, each of them with a small contribution to the trait (Barton, Etheridge, and Véber 2017). The liability-threshold model, and as in the infinitesimal model, computes a latent score for an individual based on the contribution of each of the alleles in the genetic architecture, plus the contribution of the environment. Then, if such score takes a value above a given threshold, disease will ensue.

Because the risk alleles have such small effect sizes, they are not under strong purifying selection. This allows them to be common (>1-5% of the population), unlike the mutations causing Mendelian diseases, which are rare, as they strongly decrease the fitness of the individual (Manolio et al. 2009). In consequence, the study of the genetics of complex diseases relies on the *common disease, common variant* hypothesis: common diseases are partly attributable to allelic variants present in more than 1-5% of the population, which cause, by themselves or in combinations, small increments in risk (1.1 - 1.5-fold). However, another consequence of this limited effect size is that only weak associations between causal variants and phenotypes can be expected. Again,

this notion is radically different from Mendelian diseases, where every carrier of the risk allele develops the disease under complete penetrance. In summary, the study of the genetics of complex diseases requires the identification of a large number of risk variants, among a humbling 88 millions known variants in humans (The 1000 Genomes Project Consortium et al. 2015).

The common source of genetic variation in humans are single base-pair changes in the DNA sequence, called single-nucleotide polymorphisms or SNPs (The 1000 Genomes Project Consortium et al. 2015). They usually involve two alleles, meaning that in a population there are two possible base-pairs for a genetic position. SNPs are characterized by their minor-allele frequency, that is, the frequency of the least common allele in the population. In this thesis, I focused my work on SNPs; I develop their involvement in the genetics of disease in Section 1.3. However, other forms of genetic variation exist and are relevant for human health. These are the structural variants, which involve variation in the structural and quantitative arrangement of the chromosomes (Spielmann, Lupiáñez, and Mundlos 2018). Copy number variants (CNVs), an instance of structural variants, consist on a repeated segment of the genome, where the specific number of repeats changes from person to person. A neurological disorder known as Huntington's disease ensues when a specific tri-nucleotide in the huntingtin gene is repeated more than 36 times (Macdonald 1993).

1.2 Epistasis

Epistasis is the phenomenon where the effect of one locus on the phenotype depends on the state of one or more additional loci. It has two variants: biological and statistical (Moore and Williams 2005). Biological epistasis refers to the physical interaction occurring between the loci, for instance via their protein products. It is possibly a consequence of the redundancy of biological mechanisms, which would require altering all redundant biological mechanisms to alter a biological function (Niel et al. 2015). Multiple cases of biological epistasis contributing to phenotypes in model organisms have been reported in the literature (**???**). Statistical epistasis, by contrast, is the observation that the association between one locus and the phenotype changes across the level of the other locus. In essence, statistical epistasis refers to the biological epistasis detectable at the population level. In this regard, links between epistasis and complex diseases like Alzheimer's disease (Combarros et al. 2009), inflammatory bowel disease (Cho et al. 1998) and hypertension (Kimura et al. 2012) have been found. As in this thesis I worked exclusively in the detection of the latter kind, for brevity's sake I will refer to it simply as *epistasis*.

Despite the links between epistasis and complex disease mentioned above, estimating the magnitude of its contribution to complex diseases in humans is hard (Gusareva and

Van Steen 2014). Nonetheless, (Zuk et al. 2012) proposed that incorrectly accounting for epistasis might be behind the so-called *missing heritability* of complex traits. Additionally, studies of traits model organisms suggest that epistasis plays a key role of their genetic architecture (**???**). This motivates further studying the involvement of epistasis in complex diseases.

1.3 Genome-wide association studies

Genome-wide association studies (GWAS) are experiments that explore large cohorts, systematically surveying both a high number of genetic variants, and phenotypes (Bush and Moore 2012). Their goal is to find associations between genotypes and the studied phenotype. Those associations might allow earlier diagnosis, choose a treatment appropriate for a patient's genetic background, and improve our understanding of the etiology of the disease. For that purpose, a classical GWAS setting involves a statistical test for association is conducted between each variant and the phenotype of interest. Often, that statistical test is a logistic regression, which allows using additional variables that might act as confounders. For instance, for a SNP i:

$$logit\,(p_i) = \alpha + \beta_i g_i + \gamma X \tag{1.1}$$

where g_i is the vector of genotypes at SNP i, β_i is the coefficient, and X and γ are respectively the matrix of covariates and the vector with the covariates' coefficients. A statistical test can be conducted on the value of β by transforming it into a Z-score. In order to evaluate the significance of the associated P-value, an appropriate threshold which accounts for multiple testing needs to be selected. Often, that P-value is chosen by setting the family-wise error rate to 0.05 i.e. 0.05/# SNPs (e.g. 10^{-7} if 500 000 SNPs are tested). Then, those genome-wide significant SNPs undergo a follow-up study on an independent cohort.

The 1000 Genome Projects catalogued 84.7 million SNPs across multiple human populations (The 1000 Genomes Project Consortium et al. 2015). However GWAS do not need to survey all of them to obtain a representative view of the genome. Instead, GWAS exploits the correlations between the variants as a consequence of the genetic history, a phenomenon known as linkage disequilibrium (LD). Thanks to LD, GWAS survey the whole genome in an inexpensive fashion using SNP arrays that need only contain only a small fraction of the known variants (Visscher et al. 2017). In consequence, the SNPs associated with a disease are likely not the causal ones, simply close and in LD with the causal variant.

Since the first GWAS in the late 2000s, more than 5 600 studies have shed light into

the genetics of complex traits, identifying more than 70 000 variant-trait associations (Buniello et al. 2019). From this community effort we took a few lessons about the architecture of complex trains. First, GWAS confirmed the infinitesimal model in all studied complex traits, whose variance can only be explained but by many loci with small effect sizes (Visscher et al. 2017). Such explanatory variants tend to be located in chromatin that is open and expressed in the tissues relevant for the disease (Boyle, Li, and Pritchard 2017). Also, even genes functionally implicated in the disease explain a small fraction of the trait variance (Boyle, Li, and Pritchard 2017). In fact GWAS has revealed widespread pleiotropy, as the same genomic regions are often found in association to multiple traits (Visscher et al. 2017). This has profound implications, showing how interrelated different biomolecules are. In fact, holistic models that add nuances to the infinitesimal model have been suggested, like the *omnigenic model* (Boyle, Li, and Pritchard 2017). The omnigenic model postulates that only a few core genes are directly implicated in the disease, and alterations on them have a strong effect. But alterations in many other, unrelated, genes can also lead to disease as they propagate through biological networks to affect the functionality of those core genes. A more complete view of biological networks and disease is available in Chapter 2.

1.3.1 Challenges

The discovery of the genetic basis of complex diseses is hindered by several intrincacies in both the GWAS setting and underlying biology. My thesis involved developing methodologies that tackle these challenges, which I present below.

1.3.1.1 Low statistical power

Due to the low effects sizes implied by the common disease / common variant hypothesis, GWAS would require very large sample sizes. Yet, for practical limitations, they have traditionally remained in the thousands, with the latest studies raising them up to hundreds of thousands. However, the number of biomarkers required to scan the whole genome is even larger, from hundreds of thousands to the millions. In consequence, GWAS is conducted in an ultradimensional setting, which altogether with the small effect sizes, leads to low statistical power (Button et al. 2013). Statistical power also takes a hit due to the partial correlations between the tests as multiple test correction procedures like Bonferroni consider the statistical tests independent, and hence overcorrect. Statistical power can be further reduced if the causal SNPs are in weak LD with the closest genotyped variants, for instance if they are rare (Visscher et al. 2017), or when the phenotype is heterogeneous (Visscher et al. 2017), as is common in complex diseases. This has several implications. First, it implies by definition a small chance of discovering effects that are true. Second, it raises the probability of a discovery to be a false positive. Third, when an underpowered study discovers a true effect, it is

likely that the effect size is overestimated. One practical consequence of this low power in GWAS studies, for example, is the difficulty to reproduce a GWAS result (Visscher, Hill, and Wray 2008).

As an illustration, Visscher conducted a classical GWAS inquiring about the heritability of height (Visscher, Hill, and Wray 2008). Three studies with a total sample size of 63k individuals (14k + 16k + 34k) identified 54 variants that are reliably associated with height. However, there was an alarmingly low overlap between the causal SNPs identified in the different studies. The reasons for this are multiple. First, it dealt with SNPs that have a very limited effect size which require sample sizes of the order of tens of thousands to reliably identify one causal SNP. Second, hundreds of thousands of association tests are conducted, which requires being very stringent with the significant threshold in order to minimize type I errors. Lastly, the number of SNPs in a population is between one and two orders of magnitude above the number of SNPs surveyed in a chip (Wray et al. 2013). Hence, it is likely that the association we measure between most of the markers and the phenotype is due to the linkage disequilibrium with the true causal SNP(s). This further dilutes the association that we intend to measure.

Another consideration regarding GWAS is that most studies have been carried out on European ancestry populations, which has a reduced variability in comparison to other human populations (Visscher, Hill, and Wray 2008). In fact, studies on non-European populations have yielded a big number of new, intriguing variants.

1.3.1.2 Choice of encoding

The most commonly used association tests in GWAS require making assumptions on the mode of action of the SNPs (dominance, codominance, etc.). For instance, the logistic regression presented in Equation (1.1) needs a single number that represents the two alleles of each individual. Converting the genotype into such a number is known as encoding a genotype. There are several such encodings, and its choice have implications on study: choosing an encoding which diverges from how the SNP acts will reduce the statistical power (Romagnoni et al. 2019). A common one is the additive encoding, which assumes that the minor allele is responsible for the phenotype, in linear way to the number of copies; hence, the major allele in homozygosis is represented by a 0, the heterozygous genotype by a 1, and the minor allele in homozygosis by a 2. In Appendix A we explore a feature selection algorithm which, when applied to GWAS, makes no assumptions on the mode of action of the SNP.

1.3.1.3 Estimating individual risk

Once the genetic architecture is fully understood, it will be possible to estimate the full genetic component of a patient. This involves moving moving from population-level

associations, to an individual assessment (Wray et al. 2013). Conventionally these are done through polygenic risk scores, which consist on a linear combination of genotypes from the associated loci in a patient, weighted by their effect sizes. However, their utility is far from clinical applications (Visscher et al. 2017). In this thesis, I apply machine learning algorithms for sample classification in Chapter 2 (Section 2.3.3) and in Appendix A. Also, it is worth stressing than even when the whole genetic architecture of a disease is uncovered, the R^2 of a linear predictor would be upper bounded by the heritability (h^2). A complete prediction will require fully understanding epistasis (Section 1.4), environmental, and the interaction between environment and genetics.

1.3.1.4 Population structure

As explained before, the GWAS exploits LD to avoid genotyping all known variants. However, these correlations between SNPs depend on the evolutionary history of each sample, and hence are population specific. Therefore, GWAS designs must account for samples with different ancestries or, in other words, capture the population structure in the data. Failure to do so might lead to overestimating allelic and genotypic frequencies, reducing statistical power and producing spurious associations (Wang, Cordell, and Van Steen 2018). Population structure can be captured by the principal components of the genotype matrix, and hence they are often used to account for it (Price et al. 2006). For instance, a logistic regression using the main principal components as covariates is often used to obtain measures of association at the SNP level correcting by potentially confounding population structure (Michailidou et al. 2015, 2017; D. Ellinghaus, Spain, et al. 2016).

1.3.1.5 Interpretability

By design, a genotyped SNP acts as a tag for the region in the genome in high LD with it. Hence, even if that tag SNP shows statistical association with a disease, fine-mapping studies are needed to pinpoint the specific SNP that is involved in the susceptibility to it. A frequent strategy is identifying the genes under the influence of that genomic region (Wang, Cordell, and Van Steen 2018), as those are considered the functional unit of inheritance. For instance, a gene might impact a gene by provoking an amino acid change in the protein product, altering its gene expression or its splicing. Yet, mapping genomic regions to the genes they might impact is not trivial. In the literature, we find three ways of of doing so. The first one is the physical mapping, which maps that region to the genes that whose genomic coordinates overlap with it. Often the gene boundaries are expanded by a fixed number of kilobases, as SNPs in promoters or nearby enhancers can affect gene expression (Segrè et al. 2010). Nonetheless, physical mapping can be ambiguous due to the overlap between genes (Section 2.3.5). The second SNP-gene mapping is through gene expression, when SNPs in the associated genomic region are

eQTLs of a gene. In this regard, the GTEx project (GTEx Consortium 2017) is a useful source of tissue-specific eQTL. Gene expression mapping is not exempt of the overlap problem that occurs in the physical mapping. Solutions which consider LD patterns and the association across the whole genomic region have been proposed (Liu et al. 2019). The third SNP-gene mapping is based on the 3D structure of the genome, which causes distant genomic regions to be close in space (Spielmann, Lupiáñez, and Mundlos 2018). In this situation, SNPs in a genomic region are associated to genes in the neighboring region.

1.4 Genome-wide association interaction studies

Genome-wide association interaction studies (GWAIS) share the experimental design with GWAS, but focus on the detection of epistasic associations. As opposed to GWAS, no standard GWAIS protocol exist yet , although some general recommendations have been issued (Gusareva and Van Steen 2014; Ritchie and Van Steen 2018).

Due to their similar experimental design, GWAIS and GWAS share a number of challenges (Chapter 1.3.1). Nonetheless, such problems are often aggravated in GWAIS. For instance, a larger number of tests implies a further reduction of the statistical power. Interpretation is also more complicated, as two or more genomic regions in LD with the tag SNPs need to be examined for the causal variants (Gusareva and Van Steen 2014). However it presents two additional considerations, related to the multiple genetic scenarios in which epistasis can occur. For instance (Li and Reich 2000) estimated that there are 50 different fully penetrant disease models involving two loci in epistasis.

First, in any epistasis related study, a number of arbitrary choices must be made, like the order of the explored interactions, whether to pre-filter the data according to function or detectable main effects, or the genetic encoding (see Section 1.3.1.2) (Romagnoni et al. 2019). For instance, fourth order epistasis involves four different loci jointly contributing to a phenotype. Nonetheless, as the number of interactions grows exponentially with the epistasis order, most methods and studies focus on second order epistasis. Second, epistasis introduces the challenge of quantifying and assessing the significance of an statistical interaction. Multiple strategies to detect epistasis have been proposed, from logistic regression with an interaction term to deep learning. (Niel et al. 2015) reviews the main families of epistasis detection strategies. In general, this diversity comes from tackling different aspects of the computational and statistical issues that arise from the large number of potential interactions. For instance, logistic regression with an interaction term is an exhaustive method with a parallelized implementation (Chang et al. 2015), but makes strong assumptions on the underlying relationship between the genotype and the phenotype. Hence, statistical power is com-

promised when that model is inaccurate. On the other hand, MDR (Moore et al. 2006) is model-free and exhaustive, but is limited to case-control phenotypes and the inability to compute P-values analytically makes it slower. In Chapter 5 we examine different epistasis detection methods.

1.5 Diseases studied in this thesis

The bulk of my work in this thesis revolved around two complex diseases: breast cancer and inflammatory bowel disease.

1.5.1 Breast cancer

Cancer is the name of a collection of related diseases. Specifically, all cancers undergo an uncontrolled proliferation of the patient cells, which spread into surrounding tissues. In a normal organism, cells grow and divide to maintain the tissue. As cells grow old, or accumulate too much damage, they undergo cell death and new cells will take their place. However, in cancer, this orderly process breaks down. Cells refuse to die when they get old, or accumulate damage. New cells are formed even if they are not needed. In consequence they form purpose-less growths called tumors.

This abnormal behavior occurs as consequence of the alteration of crucial genes. These alterations can be inherited from our parents, or acquired during our lifetime, due to replication errors or exposure to DNA-damaging substances. As with any other phenotypic trait, the likelihood of developing cancer will be determined by the interplay between our genetic background and the environment: genetic backgrounds may favor or hinder the acquisition of mutations, and so do environmental factors.

Breast cancer occurs when breast cells undergo this uncontrolled proliferation. In most of the cases they begin in the ducts that carry the milk to the nipple. However the tumor can originate in other tissues, mainly the milk-producing gland.

Breast cancer is the second most commonly diagnosed cancer among women, after non-melanoma skin cancer. It is also the second leading cause of cancer deaths after lung cancer. It is mostly a women's disease: only about 1% of the diagnosed cases are in men. Among the most important risk factors for breast cancer we can highlight age, family history, reproductive history, usage of oral contraceptives and exposure to radiation. Most breast cancers occur after age 50.

Breast cancer is a very heterogeneous disease: while all the tumors appear in the same organ, the tissue where they originate, the molecular mechanism involved, the response to therapy, etc. vastly differ. In general, clinical decisions are based on the expression of 3 molecular markers: the expression of the endocrine receptors for estrogen and

progesterone (ER and PgR, respectively) and the expression of the HER2 gene. The proteins these three genes code for are targets for chemotherapy. Based on the results, we distinguish three main breast cancer subtypes: hormone receptor positive, HER2 positive and triple negative.

- Hormone receptor positive: Hormone receptor positive tumors include the tumors expressing ER and/or PR, which respectively depend on estrogen and/or progesterone to grow. They happen mostly in postmenopausal women. HR+/HER2- also known as LuminalA are the majority of breast cancers (60-75%) and they present the best prognosis.
- HER2 positive: HER2+ tumors depend on the protein HER2/neu (human epidermal growth factor receptor 2) to proliferate, which they over-express. HR+/HER2+ (also known as LuminalB) constitutes 10% of the cases, while HR-/HER2 (also known as HER2-enriched) involves 5% of them. There are a couple of very effective drugs against it.
- Triple-negative: Triple-negative tumors lack the expression of all three of ER, PgR and HER2. These patients present a worse prognosis than the rest, due to the aggressiveness of the tumor and the lack of a clear molecular target. Still, the main treatment is chemotherapy.

In the mid-19th century a French medical doctor, Pierre Paul Broca, reported for the first time a case of familial breast cancer (Nielsen, Overeem Hansen, and Sørensen 2016). Indeed, his wife acquired breast cancer, as many women in her family had for, at least, 4 generations. Cases of familiar breast cancer usually occur in women younger than 50 years, and bilateral primary breast tumors are frequent. Epidemiological studies later quantified the relative risk conferred by a the presence of multiple breast cancers in the family at 2.7. Moreover, they exhibit a higher likelihood of acquiring triple-negative breast cancer.

It wasn't until the late 20th century that two genes involved in DNA repair, BRCA1 and BRCA2, were associated with hereditary breast and ovarian cancer (HBOC). Some mutations in these genes increase the risk of developing breast cancer, giving respectively a 57–65% or 45–55% risk of developing breast cancer by age 70 among women. For that reason, BRCA1 and BRCA2 mutations are rare in most populations (1 of 400).

HBOC follows an autosomal dominant inheritance pattern. While approximately 5–10% of all patients with breast cancer exhibit a monogenic predisposition to breast and ovarian cancer, only about 25% of them harbor BRCA1/2 mutations. An additional 23 genes have been associated with familial breast and/or ovarian cancer (Table D.5).

Nearly all known HBOC susceptibility genes encode tumor suppressors that participate

in genome stability pathways (homologous recombination repair, replication fork stability, transcription–replication collisions, mismatch repair, and DNA damage signaling, checkpoints and cell death; see Appendix D for more information).

1.5.1.1 The GENESIS dataset

The GENE Sisters (GENESIS) study was designed to investigate risk factors for familial breast cancer in the French population (Sinilnikova et al. 2016). Index cases are patients with infiltrating mammary or ductal adenocarcinoma, who had a sister with breast cancer, and who have been tested negative for BRCA1 and BRCA2 pathogenic variants. Controls are unaffected colleagues and/or friends of the cases, born around the year of birth of the corresponding case (± 3 years). We focused on the 2 577 samples of European ancestry, of which 1 279 are controls and 1 298 are cases. The genotyping was performed using the iCOGS array, a custom Illumina array designed to study genetic susceptibility of hormone-related cancers (Sakoda, Jorgenson, and Witte 2013). It contains 211 155 SNPs, including SNPs putatively associated with breast, ovarian, and prostate cancers, SNPs associated with survival after diagnosis, and SNPs associated to other cancer-related traits, as well as functional candidate variants in selected genes and pathways.

1.5.2 Inflammatory bowel disease

Inflammatory bowel disease (IBD) is a group of complex diseases that, as the name indicates, share a common theme of inflammation of the intestines. The two main subtypes are ulcerative colitis and Chron's disease. Clinically, these two share a lot of the symptoms, mainly intermittent abdominal pain and diarrhea (Liu and Stappenbeck 2016). However, they differ in the specific regions of the digestive tract that get affected, as well as the specific lesions. IBD's incidence worldwide has been continually growing, specially in newly industrialized countries, althouth after decades of growth it has stabilized in North America, Oceania and Europe (Ng et al. 2017). In these latter countries, the prevalence is slightly above 0.3%.

The genetic component of IBD was recognized more than a century ago (Ek, D'Amato, and Halfvarson 2014). However, it was not until the 2001 that the first gene, NOD2, was linked to IBD susceptibility. Ever since, hundreds loci have been associated to IBD as well, in positions related to immune system genes, both innate and adaptive (Loddo and Romano 2015; D. Ellinghaus, Spain, et al. 2016; Liu and Stappenbeck 2016). For instance NOD2, and other susceptibility genes like IL23R, and PTPN2, are related to cell signalling in immune cells. However, most of the loci associated to genes are rare variants ($< 0.5\%$) with large effect sizes. By contrast, several GWAS have identified very common SNPs (20-50%), with small effect sizes (odds ratio < 1.1)

(Liu and Stappenbeck 2016), but which do not encode any coding change (Jostins et al. 2012). This raises questions about the underlying biology (see Section 1.3.1.5). For a comprehensive view of the genetics of IBD, interested readers can read (Liu and Stappenbeck 2016) and (Loddo and Romano 2015).

1.5.2.1 The IIBDGC dataset

The International Inflammatory Bowel Disease Genetics Consortium carried out the largest case-control GWAS on ulcerative colitis and Chron's disease to date (Jostins et al. 2012). The dataset contain 66 280 samples, out of which 32 622 are cases and 33 658 are controls. The Immunochip SNP array was used for the genotyping (Cortes and Brown 2010), which contains 196 524 polymorphisms, with a special focus on immunogenetics.

1.6 Network view of complex diseases

Human biology is notoriously complicated, as sheer numbers demonstrate: to form a 70kg man, 3.0×10^{13} cells (Sender, Fuchs, and Milo 2016) assemble and interact to produce and maintain 79 organs. To achieve that level of complexity, human cells depend on their genetic material, carefully tuned by epigenetics and enabled by a favorable environment. In terms of genetics, the object of my work, a human diploid genome is 6.4 billion base pairs long, and encodes 44 393 genes, of which 20 444 encode for a protein and 23 949 are RNA genes. DNA, proteins and RNA are in constant interplay with each other, with the metabolites, and with the environment. Proteins physically interact with each other in highly specific ways (protein-protein interactions or PPIs). If such interactions are stable enough, proteins can assemble into large complexes to carry out particular functions. But proteins also interact with DNA to regulate gene expression (transcription factor - DNA interactions). And so on: enzymes interact with their metabolites, hormones with their receptors, the individual with their environment, etc. At a fundamental level, a person and their traits are just the emerging pattern born from the interaction of all these factors. Hence, biology, from ecosystems to molecular biology, cannot be understood if not as an interplay. Mindful of this, and enabled by the omics technologies of the 21st century, researchers have strove to capture and understand these relationships. The goal of systems biology is achieving a global understanding of the complex interplay that drives biology. Among all the possible biological relationships, in this thesis I focused on protein-protein interactions, including protein complexes, as their coverage is larger and their properties better understood. Such interactions are available in databases like HINT (Das and Yu 2012), The BioGRID (Oughtred et al. 2019) or IntAct (Hermjakob 2004).

Relationships between pairs of entities can be mathematically formalized as a network,

which makes them analytically approachable. In such networks, often proteins or genes are the nodes, which are connected by edges in a pairwise fashion when they are related. Although the edges might have directionality, often PPIN are undirected, as the direction of the edges is often unknown (transitory interactions), or inexistent (co-complexes) (Barabási, Gulbahce, and Loscalzo 2011).

The field of systems biology relies on the assumption that the network accurately captures the context it requires to carry out its biological function. We can distinguish three levels of network properties: properties of individual nodes (local), the joint properties of groups of nodes (mesoscale), and the properties over the whole network (global). Indeed, at all three levels biological networks are structured, very differently from random networks (Barabási, Gulbahce, and Loscalzo 2011; Chaiboonchoe et al. 2013).

At the global level, for instance, the degree, which represents the number of edges per node, follows a power-law distribution (Barabási, Gulbahce, and Loscalzo 2011). This implies that, at the local level, a few genes participate in the majority of the edges. Such nodes are called *hubs*. Importantly, this node property is informative of the gene function: *in utero* essential genes, like knots preventing the network from falling apart, tend to be hubs. An important consequence of the structured degree distribution at the mesoscale level is the emergence of modules, subsets of nodes densely interconnected to each other, and sparsely to the rest. Such modules are often associated with biological function towards which all nodes jointly contribute (Mitra et al. 2013).

Another global property of a graph is the distribution of distances between pairs of nodes, that is, how many edges must be passed to travel from one node to another. Such is the notion of path between nodes. Often we are interested in the *shortest* path, for they reveal the fastest way information can flow from one node to another. An examination of the distribution of shortest paths in biological networks, shows that all nodes are close to each other (Barabási, Gulbahce, and Loscalzo 2011). Such networks are called *small world networks*. This structure makes the flow of information resilient to the removal of nodes or edges (Chaiboonchoe et al. 2013), a cause of biological epistasis (Niel et al. 2015).

Modules are connected groups of nodes which are densely interconnected to each other and sparsely to the rest of the network. A mesoscale property of biological networks is that they have a modular structure i.e. a strong division into modules. Such modules often constitute functional units within the network, where the nodes jointly contribute to a specific function (Mitra et al. 2013).

1.6.1 Networks in disease

Examining how these biological networks relate to disease, and the topology around disease genes produces a nuanced approach to disease: cut the knot in the center of the web that keeps it all together, and it all will fall off; cut a bunch of peripheral, less important nodes around it, the outcome might be the same. In other words, there are many ways of producing the same disease (Leiserson et al. 2013).

The properties of biological networks enumerated above lay down the foundations of the use of biological networks to study disease genes. For instance, they justify the *local hypothesis*, which expects genes involved in a disease to interact with each other (Barabási, Gulbahce, and Loscalzo 2011). They also justify the disease module hypothesis, which expects genes involved in the same disease to share a module. Experimentally, network propagation experiments highlight the differential topological properties of disease genes and biological networks. Network propagation refers to methods that use all the possible paths in the network to re-rank the genes on it (Cowen et al. 2017). They include heat diffusion, random walk, graph kernels, and even Google's search algorithm. In essence, for these methods, association of a node with a phenotype can be thought of as a volume of water: the more strongly associated, the more voluminous. Generically, each node starts with an initial volume which, iteratively, gets distributes its liquid among its neighbors. Equivalently, that node will receive a share from its neighbors'. The expectation is that truly associated genes will be densely interconnected to each other, forming cycles and modules that will keep the water from diffusing to other, uninvolved genes. After a number of steps the volume of water in each node is re-evaluated, and used to re-prioritize the genes. Using network propagation (Huang et al. 2018) recently showcased the differential topological properties of disease genes. Across different biological networks, they were able to retrieve disease-related genes with varying success using only a subset of them and random walk with restart.

Indeed, disease genes exhibit different properties than non-disease genes in disease (Piñero et al. 2016; Cai, Borenstein, and Petrov 2010; Furlong 2013; Barabási, Gulbahce, and Loscalzo 2011). For starters, disease genes are not hubs (Cai, Borenstein, and Petrov 2010; Das and Yu 2012), but tend to be non-essential genes located in the periphery. This is coherent with a evolutionary mindset, where mutations in essential genes would be highly deleterious, even resulting in embryonic lethality. However, this enrichment is driven mainly by cancer genes (Piñero et al. 2016). Additionally, disease genes tend to be bottlenecks i.e. the sole link between many peripheral genes and the rest of the network (Cai, Borenstein, and Petrov 2010), suggesting disease arise when these vulnerable regions of the network break down.

These properties, however, do not affect equally all types of disease genes. Many of the aforementioned properties do not extend to genes identified through GWAS (Cai,

Borenstein, and Petrov 2010). However, differences arise even when comparing genes involved in complex diseases with those involved in Mendelian ones.

1.6.2 Network-guided approaches to disease study

As exemplified in the previous section, networks can be leveraged on to gain insight of the biology of the disease. Below I summarize several ways of doing that I have worked on during my thesis.

1.6.2.1 High-score subnetwork search

One of the focuses of my thesis was the study of networks where the nodes have been scored by their association to the disease. The scores might come from omics experiments, or from *a priori* knowledge of disease genes. In essence, such methods look for connected subnetworks made of nodes with high scores. Lacking a term broadly agreed in the community, I refer to such algorithms as high-score subnetwork search. Although heterogeneous, to some extent all existing approaches are based on the *guilt-by-association principle*: nodes nearby other nodes associated to the disease are suspect of being associated themselves, even if their association is non-significant by conventional standards. However, taking only the genes associated with those associated would be prone to false positives, as often networks include edges that are not relevant for the biological problem at hand. Several high-score subnetwork search methods are available in the literature. Essentially, they differ in the considerations they make on what the solution looks like. In Chapter 2 we describe a representative set of these methods, critically discuss their performance at biomarker discovery on the GENESIS dataset, and discuss their shortcomings.

1.6.2.2 Module detection

In the context of disease, modules of consistently altered nodes might represent the mechanisms that lead to it. Hence, their identification can provide insights into its etiology. In Appendix B we apply module detection techniques to a subnetwork of genes with abnormal splicing in cancer.

1.6.2.3 Aggregation of networks

During my research, often I obtained multiple high-scoring subnetworks using different approaches, which provided complementary perspectives of the same disease. In consequence, I was interested in integrating them into a single subnetwork, which I would analyze. In Chapter 2 we discuss a naive way of aggregating subnetworks from different high-score subnetwork search algorithms, were the edges are unweighted. (Wang et al. 2014) and (Glass et al. 2013) used information theory to integrate find the

closest network to a set of input networks. In Chapter 5, we discuss how to apply such methods to networks coming from different epistasis detection methods, whose edges are weighed by the confidence they exist.

1.7 Contributions

The object of my thesis was the methodological study and application of network methods to GWAS data. In essence, networks contain prior information, which can be traded for statistical astringency in our analysis. In other words, if genes strongly associated to a disease, albeit non-significantly so, are interconnected in an underlying biological network, we are more likely to believe they represent a consistently altered biological mechanism.

In collaboration with Nadine Andrieu and Fabienne Lesueur (Institute Curie), and working closely with Christine Lonjou, I applied six high-score subnetwork search to the GENESIS dataset (Section 1.5.1.1). Our goal was to find new biomarkers for breast cancer susceptibility. My work on GENESIS is explained in Chapter 2. In summary, I applied six different network-based, biomarker discovery methods to the GENESIS dataset. The methods provide a representative view of the high-score subnetwork search field. We performed a methodological comparison and a benchmark of the methods, highlighting their strengths and weaknesses. Finally, we conclude that combining the methods provide a more complete answer than any of the individual solutions. Our network analysis recovers both genes and genomic regions previously found in association with breast cancer susceptibility, as well as new genes. Importantly, all of those genes are tied in a subnetwork, providing a rationale on how alterations on those genes might lead to disease.

My first approach to the problem involved working on SConES (Azencott et al. 2013), one of the high-score subnetwork search method examined in Chapter 2. This algorithm, and my work on it are described in detail in Chapter 3. In summary, I worked on applying SConES to case-control datasets, and strategies to parametrize it. Regarding the former, as SConES implemened exclusively the regression version of SKAT (Wu et al. 2011; Ionita-Laza et al. 2013), I implemented two ways of measuring association between a SNP and a binary phenotype. Regarding the latter, SConES has two parameters, η and λ, which control the sparsity and the inter-connectedness of the selected SNPs, respectively. Originally the parameters that produced the most stable solution were selected. I explored the impact of penalized-likelihood measures like BIC, AIC, and AICc, which score a set of features based on both their sparsity and the accuracy of a linear classifier built on them. The product of my work is published in *martini* (Climente-González and Azencott 2019), an R package published in Bioconductor.

In 2019, when I started working on epistasis detection, we established a collaboration with Kristel Van Steen (University of Liège, Belgium). Specifically, I worked closely with Diane Duroux, a PhD student in her research group. Our goal was to build an epistasis gene network of inflammatory bowel disease. We discuss our efforts in this regard in Chapter 5. For that purpose, we surveyed suitable epistasis detection methods, applied them to the IIBDGC dataset (Section 1.5.2.1), and integrated the solution. However, as a previous step we needed to appropriately map epistatic SNP networks to epistatic gene networks. I describe this work in Chapter 4. In summary, we examined and evaluated four different mappings (physical, eQTL, chromatin, and the three together).

In collaboration with Makoto Yamada (RIKEN AIP, Japan) and Samuel Kaski (Aalto University, Finland), I developed block HSIC Lasso, a general-purpose non-linear feature selector. The work involved modifying an existing algorithm, HSIC Lasso, to reduce its memory consumption. On top of that, we worked on improving its performance, and on solving numerical issues in edge cases. The algorithm is implemented as the Python package pyHSICLasso, available on both PyPI and GitHub. Then, I characterized the algorithm and applied it to several biological datasets. Crucially, three of the datasets were GWAS, which was a milestone in terms of the number of features block HSIC Lasso can handle. Thanks to this work, we analyzed the impact of considering non-redundancy and non-linear models when selecting SNPs for patient classification. We describe the algorithm and our conclusions in the article *Block HSIC Lasso: model-free biomarker detection for ultra-high dimensional data* published in the proceedings of the ISMB/ECCB 2019 (Climente-González et al. 2019). The full manuscript is available in Appendix A. It was made possible by a scholarship from RIKEN AIP, which allowed me to spend 3 months in Makoto Yamada's laboratory in Kyoto.

Additionally, I worked on the involvement of alternative splicing in cancer with Eduardo Eyras (Australian National University), a continuation of my previous research. During 2016 and early 2017, we prepared the answer to reviewers for the article *The functional impact of alternative splicing in cancer*, which was published in Cell Reports in August 2017 (Climente-González et al. 2017). The full manuscript is available in Appendix B. I also explored evidence of epistasis in cancer, looking for mutual exclusivity between alterations in alternative splicing and somatic mutations. We searched for evidence of mutual exclusivity at both the gene and the pathway level. Among others, we found a compelling pattern in GATA3. This finding was published in Cell Reports in March 2018, in the broader article *Systematic analysis of splice-site-creating mutations in cancer* (Jayasinghe et al. 2018). The full manuscript is available in Appendix C. In addition, I put compiled the code required for the analyses in (Climente-González et al. 2017), and created a Python package with a clean interface. The package,

spada, is available in both PyPI and GitHub. This package searches alterations of alternative splicing, and maps them to functional consequences at the protein level. Specifically, such consequences are the removal/addition of functional modules to the protein (e.g. domains) and the loss or gain of protein-protein interactions. Lastly, I applied *spada* to two leukemia datasets from TARGET, an NIH program that aims to understand the molecular basis of several childhood cancers. This last analysis is still on-going.

Lastly, I carried out my research committed to open, reproducible science. As such, all my projects have an associated, version-controlled, laboratory notebook, which includes as much data as I am allowed to share. All such laboratory notebooks are made out public when the paper is. All the scripts I developed are open source, under permissive MIT license. Specifically, I made an effort to develop project-independent tools, which are useful to anyone which needs similar to the ones I had. These tools are published on GitHub https://github.com/hclimente/gwas-tools.

Chapter 2

Combining network-guided GWAS to discover susceptibility mechanisms for breast cancer

---

Joint work with Christine Lonjou, Fabienne Lesueur, the GENESIS investigators, Dominique Stoppa-Lyonnet, Nadine Andrieu and Chloé-Agathe Azencott

Systems biology provides a comprehensive approach to biomarker discovery and biological hypothesis building. It does so by jointly considering the statistical association between a gene and a phenotype, and the biological context of each gene, represented as a network. In this work we study the utility of six network methods to discover new biomarkers for breast cancer susceptibility by searching subnetworks highly associated to a phenotype. We interrogate a familial breast cancer genome-wide association study (GWAS) focused on *BRCA1/2* negative French women. By trading statistical astringency for biological meaningfulness, most network methods get more compelling results than standard SNP- and gene-level analyses, recovering causal subnetworks tightly related to cancer susceptibility. We perform an in-depth benchmarking of the methods with regards to size of the solution subnetwork, their utility as biomarkers, and the stability and the runtime of the methods. Interestingly, a combination of solution subnetworks provided a concise subnetwork of 93 genes, enriched in known breast cancer susceptibility genes (*BABAM1*, *BLM*, *CASP8*, *FGFR2*, and *TOX3*, Fisher's exact test P-value = 7.8×10^{-5}) and more central than average. Additionally, it includes subnetworks of mechanisms related to cancer, like protein folding (*HSPA1A*, *HSPA1B*, and *HSPA1L*) or mitochondrial ribosomes (*MRPS30*, *MRPS31*, *MRPS18B*). We also observed a general dysregulation in the neighborhood of *COPS5*, a gene related to multiple hallmarks of cancer.

2.1 Introduction

As described in Section 1.3, genome-wide association studies (GWAS) aim at quantifying how single-nucleotide polymorphisms (SNPs) predispose to complex diseases, like diabetes or some forms of cancer (Bush and Moore 2012). Despite their successes

(Buniello et al. 2019), GWAS also presents intrinsic challenges (Section 1.3.1). Some of them stem from the high-dimensionality of the problem, as every GWAS to date studies more variants than samples are genotyped. This limits the statistical power of the experiment, as only variants with large and moderate effects can be detected. And it is particularly problematic since the prevailing view is that most genetic architectures involve many variants with small effects (Visscher et al. 2017). Additionally, to avoid false positives, a conservative multiple test correction is applied, typically Bonferroni. However, Bonferroni is known to be overly conservative when the statistical tests are correlated, as is the case in GWAS (Wang, Cordell, and Van Steen 2018). Another open issue is the interpretation of the results, as the functional consequences of most common variants are not well understood. On top of that, recent large-sampled studies suggest that most of the genome contributes to a degree to any complex trait, in accordance with the infinitesimal model (Barton, Etheridge, and Véber 2017). The omnigenic model (Boyle, Li, and Pritchard 2017) explains this by the dense functional inter-relatedness between genes, influencing each other's behavior, which allows alterations in most genes to impact the "core" genes involved in a disease. A comprehensive statistical framework which includes the structure of biological data might address the aforementioned issues.

In this regard many authors turn to network biology to handle the complex interplay of biomolecules that lead to disease (Furlong 2013). As its name suggests, network biology models biology as a network, where the biomolecules under study, often genes, are nodes, and selected functional relationships between them are the edges that link them. These functional relationships come from evidence that the genes jointly contribute to a biological function; for instance, their expression is correlated, or their gene products establish a protein-protein interaction. Under this view, complex diseases are not the consequence of a single altered gene, but of the interaction of multiple interdependent biomolecules (Barabási, Gulbahce, and Loscalzo 2011). In fact, an examination of biological networks shows that disease genes have differential properties (Barabási, Gulbahce, and Loscalzo 2011; Piñero et al. 2016). This is particularly true for cancer driver genes, which tend to be key players in connecting different, densely-connected communities of genes. Additionally, as genes that contribute to a disease tend to participate in similar biological functions, guilt-by-association strategies have proved effective at identifying disease genes (Huang et al. 2018).

Network-based, biomarker discovery methods exploit the guilt-by-association strategy to identify disease genes on GWAS data (Azencott 2016). In essence, each SNP has a measure of association with the disease, given by the experiment, and functionally biological relationships, given by a network built on prior knowledge. Then, the problem becomes finding a functionally-related set of genes that is highly associated with the disease. Different solutions have been proposed to this problem, often stemming from

divergent different mathematical frameworks and considerations of what the optimal solution looks like. Some methods strongly constrain the problem to certain kinds of subnetworks. Such is the extreme case of LEAN (Gwinner et al. 2016), which focuses on star subnetworks, i.e. instances were both a gene and its direct interactors are associated with the disease. Other algorithms, like dmGWAS (Jia et al. 2011) and heinz (M. T. Dittrich et al. 2008), focus on interconnected genes with high association with the disease. However, they differ in their tolerance to the inclusion of lowly associated nodes, and the possible number of disconnected subnetworks in the solution. Lastly, other methods also consider the topology of the network, favoring solutions that are densely interconnected; such is the case of HotNet2 (Leiserson et al. 2015), SConES (Azencott et al. 2013), and SigMod (Liu et al. 2017).

In this work, we analyze the effectiveness of these six methods to discover new biomarkers on GWAS data. We focus on the GENESIS dataset (Sinilnikova et al. 2016), a study of familial breast cancer conducted in the French population. After following a classical GWAS approach, we use these network-based methods to recover additional familial breast cancer biomarkers. Some of them are known, while others are specific to this dataset. Lastly, we carry out a comparison of the solutions obtained by the different methods, and aggregate them to obtain a consensus network of predisposition to familial breast cancer.

2.2 Methods

2.2.1 GENESIS

In this study we used the GENESIS dataset, described in Section 1.5.1.1.

2.2.2 Preprocessing and quality control

We discarded SNPs with a minor allele frequency lower than 0.1%, those not in Hardy - Weinberg equilibrium in controls (P-value <0.001), and those missing on more than 10% of the samples. A subset of 20 duplicated SNPs in *FGFR2* were also removed. In addition, we removed the samples with more than 10% missing genotypes, and an additional 28 samples with TODO. The final dataset included 1 271 controls and 1 280 cases, genotyped over 197 083 SNPs.

We looked for population structure that could create confounding associations. A PCA revealed no differential population structure between cases and controls (Figure 2.1). Independently, we did not find evidence of genomic inflation ($\lambda = 1.05$) either, further confirming the absence of confounding population structure.

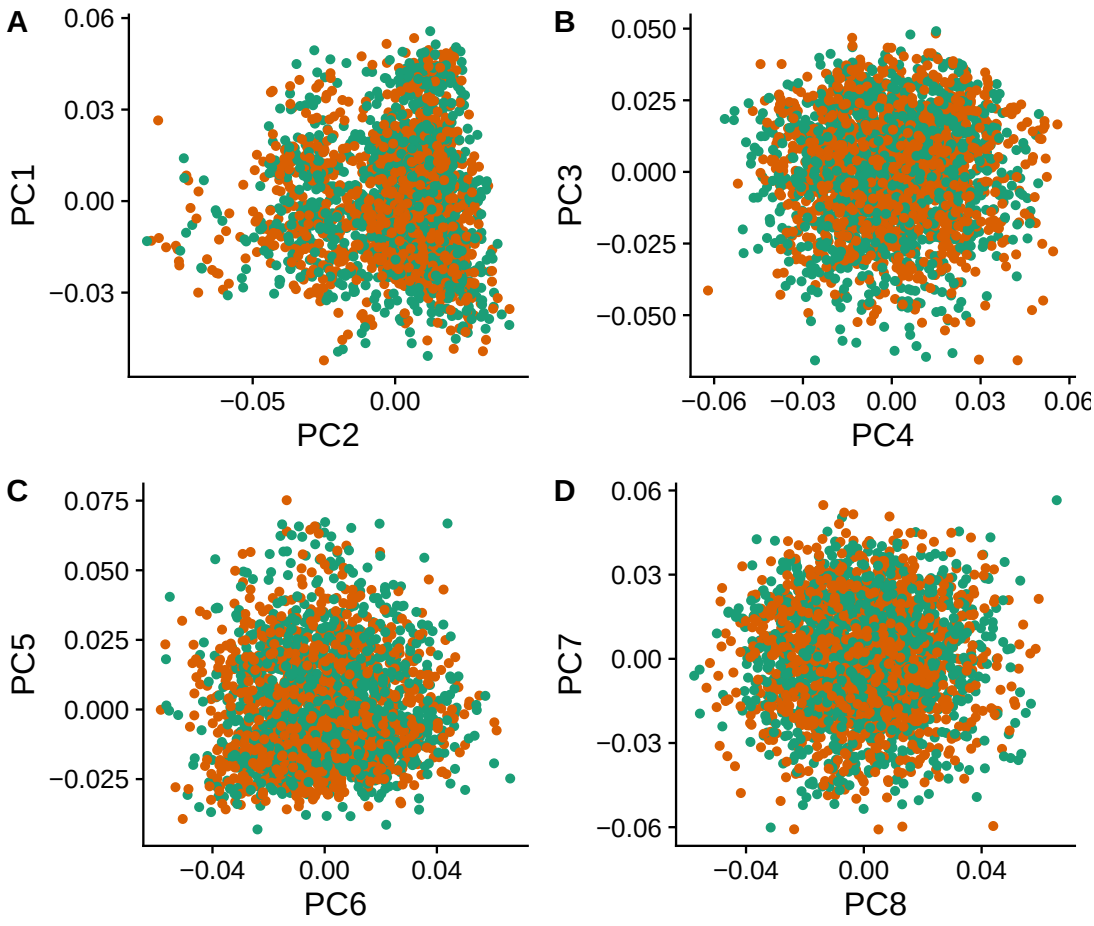


Figure 2.1: GENESIS shows no differential population structure between cases and controls. **(A,B,C,D)** Eight main principal components computed on the genotypes of GENESIS. Cases are colored in green, controls in orange.

2.2.3 High-score subnetwork search algorithms

2.2.3.1 SNP and gene association

To measure association between a genotype and the phenotype, we performed a per-SNP 1df χ^2 allelic test using PLINK v1.90 (Chang et al. 2015). Then, we used VEGAS2v2 to compute the gene-level association score (Mishra and Macgregor 2015) from the SNP P-values. In order to map SNPs to genes we used their overlap on the genome: all SNPs located within the boundaries of a gene, 50 kb, were mapped to that gene. To compute the gene association we used the 10% of SNPs linked to the gene with lowest P-values. We used the 62 193 genes described in GENCODE 31 (Frankish et al. 2019), although only 54 612 could be mapped to at least one SNP. Out of those, we focused exclusively on the 32 767 that could be mapped to an HGNC symbol. Out of the SNPs 197 083 remaining after quality control, 164 037 mapped to at least one of these genes.

We use such mapping to compare the outputs of methods who produce SNP- to those that produce gene-lists, and vice versa. In the former case, we consider any gene that can be mapped to any of the selected SNPs as selected as well. In the latter, we consider all the SNPs that can be mapped to that gene as selected by the method.

2.2.3.2 Mathematical notation

In this chapter, we use undirected, vertex-weighted networks, or graphs, $G = (V, E, w)$. $V = \{v_1, \ldots, v_n\}$ refers to the vertices, with weights $w : V \to \mathbb{R}$. Equivalently, $E \subseteq \{\{x, y\} | x, y \in V \wedge x \neq y\}$ refers to the edges. When referring to a subnetwork S, V_S is the set of nodes in S and E_S is the set of edges in S. A special case of subgraphs are *connected* subgraphs, which occur when every node in the subgraph can be reached from any other node.

On top of a weight, nodes have other properties provided by the topology of the graph. In this chapter we focus on two: degree centrality, and betweenness centrality. The degree centrality, or degree, is the number of edges that a node has. The betweenness centrality, or betweenness, is the number of times a node participates in the shortest paths between two other nodes.

In addition, we use several matrices that describe different properties of a graph. The described matrices are square, and have as many rows and columns as nodes are in the network. The element (i, j) represents a selected relationship between v_i and v_j. The *adjacency matrix* W_G contains a 1 when the corresponding nodes are connected through an edge, and 0 otherwise; the diagonal is zero. The *degree matrix* D_G is a diagonal matrix which contains the degree of the different nodes. Lastly, the *Laplacian matrix* L_G is defined as $L_G = D_G - W_G$.

Table 2.1: Summary statistics on the results of multiple network methods on the gene-gene interaction network. The first row contains the summary statistics on the whole network.

Method	Field	Nodes	Exhaustive	Solution	Components	Input	Scoring	Reference
dmGWAS	GWAS	Genes	No		1	Summary	-log~10~(P)	[@jia_dm
heinz	Omics	Genes	Yes		1	Summary	BUM	[@dittrich
HotNet2	Omics	Genes	Yes	Modular	$\ge$ 1	Summary	Local FDR	[@leiserso
LEAN	Omics	Genes	Yes	Star	$\ge$ 1	Summary	-log~10~(P)	[@gwinne
SConES	GWAS	SNPs	Yes	Modular	$\ge$ 1	Genotypes	χ^2	[@azencot
SigMod	GWAS	Genes	Yes	Modular	1	Summary	-log~10~(P)	[@liu_sig

Note:
/Field/: field in which the algorithm was developed. /Nodes/, the type of network, either gene (protein-pr usually) or a SNP network. /Exhaustive/: whether all the possible solutions given the selected hyper /Solution/: additional properties are enforced on the solution subnetwork, other than being dense in hi /Components/: number of connected subnetworks in the solution. /Input/: genotype data or GWAS summ how SNP/gene P-values are transformed into node scores.

2.2.3.3 Methods used

Beyond the assumption that genes that contribute to the same function will be nearby in the protein-protein interaction network (PPIN), they might be topologically related to each other in diverse ways (densely interconnected modules, nodes around a hub, a path, etc.). That is not the only choice to make: how to score the nodes, whether the affected mechanisms form a single connected component or several, how to frame the problem in a computationally efficient fashion, what is the best network to use, etc. In consequence, multiple solutions have been proposed. In this chapter, we examine six of them: five that explore the protein-protein interaction network, and one which explores SNP networks. We selected methods that were open source, had an implementation available, and an accessible documentation. Their main differences are summarized in Table 2.1.

dmGWAS dmGWAS searches the subgraph with the highest local density in low P-values (Jia et al. 2011). To that end it searches candidate subnetwork solutions using a greedy, "seed and extend", heuristic:

1. Select a seed node.
2. Compute Stouffer's Z-score Z_m for the current subgraph as

$$Z_m = \frac{\sum z_i}{\sqrt{k}}$$

where k is the number of genes in the subgraph, $z_i = \phi^{-1}(1 - \text{P-value}_i)$, and ϕ^{-1} is the inverse normal distribution function.

3. Identify neighboring nodes i.e. nodes at distance $\leq d$. We set d = 2.
4. Add the neighboring nodes whose inclusion increases the Z_{m+1} more than Z_m (1 + r). In our experiments, we set r = 0.1.
5. Repeat 2-4 until no increment Z_m (1 + r) is possible.

Lastly, the module's Z-score is normalized as

$$Z_N = \frac{Z_m - \text{mean}\left(Z_m(\pi)\right)}{\text{SD}\left(Z_m(\pi)\right)}$$

where $Z_m(\pi)$ represent a vector containing 100000 random subsets of the same number of genes.

We used the implementation of dmGWAS in the dmGWAS 3.0 R package (Q. Wang and Jia 2014). We used the function *simpleChoose* to select the solution subnetwork, which aggregates the top 1% modules into the solution subnetwork.

heinz The goal of heinz is to identify the highest-scored connected subgraph on the network (M. T. Dittrich et al. 2008). The authors propose a transformation of the genes' P-value into a score that is negative under no association with the phenotype, and positive value when there is. This transformation is achieved by modelling the distribution of P-values by a beta-uniform model (BUM) parameterized by the desired FDR. Thus formulated, the problem is NP-complete. To solve it efficiently it is re-casted as the Prize-Collecting Steiner Tree Problem (PCST), which seeks to select the connected subnetwork S that maximizes the *profit* p(S):

$$p(S) = \sum_{v \in V_S} p(v) - \sum_{e \in E_S} c(e).$$

were p(v) = w(v) - w' is the *profit* of adding a node, c(e) = w' is the *cost* of adding an edge, and $w' = min_{v \in V_G} w(v)$. All three are positive quantities. heinz implements the algorithm from (Ljubić et al. 2006), which in practice is often fast and optimal, neither is guaranteed. We used BioNet's implementation of heinz, available on Bioconductor (Beisser et al. 2010; M. Dittrich and Beisser 2008).

HotNet2 HotNet2 was developed to find connected subgraphs of genes frequently mutated in cancer (Leiserson et al. 2015). To that end, it considers both the local

topology of the network and the scores of the nodes. The former is captured by an insulated heat diffusion process: at the beginning, the score of the node determines its initial heat; iteratively each node yields heat to its "colder" neighbors, and receives heat from its "hotter" neighbors, while retaining part of its own (hence, *insulated*). This process continues until equilibrium is reached, and results in a similarity matrix F. F is used to compute the similarity matrix E that accounts also for similarities in node scores as

$$E = F\,\mathrm{diag}(w(V)),$$

where $\mathrm{diag}(w(V))$ is a diagonal matrix with the node scores in its diagonal. We scored the nodes as in (Nakka, Raphael, and Ramachandran 2016), assigning a score of 0 for the genes with low probability of being associated to the disease, and -$\log_{10}$(P-value) to those likely to be. In this dataset, the threshold separating both was a P-value of 0.125, which was obtained using a local FDR approach (Scheid and Spang 2005). To obtain densely connected subnetworks, HotNet2 prunes E, only preserving edges such that w(E) $> \delta$. Lastly, HotNet2 evaluates the statistical significance of the subnetworks by comparing their size to the size of networks obtained by permuting the node scores. HotNet2 has two parameters: the restart probability β, and the threshold heat δ. Both parameters are set automatically by the algorithm, and are robust (Leiserson et al. 2015). HotNet2 is implemented in Python (Leiserson et al. 2018).

LEAN LEAN searches disregulated "star" gene subnetworks, that is, subnetworks composed by one central node and all its interactors (Gwinner et al. 2016). By imposing this restriction, LEAN is able to exhaustively test all such subnetworks (one per node). For a particular subnetwork of size m, the P-values corresponding to the involved nodes are ranked as $\mathrm{p_1} \leq \ldots \leq \mathrm{p_m}$. Then, k binomial tests are conducted, to compute the probability of having k out of m P-values lower or equal to $\mathrm{p_k}$ under the null hypothesis. The minimum of these k P-values is the score of the subnetwork. This score is transformed into a P-value through an empirical distribution obtained via a subsampling scheme, where sets of m genes are selected randomly, and their score computed. Lastly, P-values are corrected for multiple testing through a Benjamini-Hochberg correction. We used the implementation of LEAN from the LEANR R package (Gwinner 2016).

SConES SConES searches the minimal, modular, and maximally associated subnetwork in a SNP graph (Azencott et al. 2013). Specifically, it solves the problem

$$\arg\max_{S \subseteq G} \underbrace{\sum_{v \in V_S} w(v)}_{\text{association}} + \underbrace{\lambda \sum_{v \in V_S} \sum_{u \notin V_S} L_{vu}}_{\text{connectivity}} - \underbrace{\eta |V_S|}_{\text{sparsity}} \tag{2.1}$$

where λ and η are hyperparameters that control the sparsity and the connectivity of the model. Given two hyperparameters, the aforementioned problem has a unique solution, that SConES finds using a graph min-cut procedure. We used the version on SConES implemented in the R package martini (Climente-González and Azencott 2019). As in (Azencott et al. 2013), we selected λ and η by cross-validation, choosing the values that produce the most stable solution across folds. Note that the solution to the above problem can consist of several connected subnetworks which are disconnected from each other. In this case, the selected hyperparameters were $\eta = 3.51$, $\lambda = 210.29$ for SConES GS; $\eta = 3.51$, $\lambda = 97.61$ for SConES GM; and $\eta = 3.51$, $\lambda = 45.31$ for SConES GI.

SigMod SigMod aims at identifying the most densely connected gene subnetwork that is most strongly associated to the phenotype (Liu et al. 2017). It addresses an optimization problem similar to that of SConES (Equation (2.1)), but using the Laplacian matrix rather than the adjacency matrix (Section 2.2.3.2), to quantify solutions containing many edges.

$$\arg\max_{S \in G} \underbrace{\sum_{v \in V_S} w(v)}_{\text{association}} + \underbrace{\lambda \sum_{v \in V_S} \sum_{u \in V_S} W_{vu}}_{\text{connectivity}} - \underbrace{\eta |V_S|}_{\text{sparsity}} .$$

As SConES, this optimization problem can also be solved by a graph min-cut approach.

SigMod presents three important differences with SConES. First it is designed for gene-gene networks. Second, by replacing the adjacency by the Laplacian matrix, it favors subnetworks containing many edges. SConES, instead, penalizes connections between the selected selected and unselected nodes. Third, it returns a single connected subnetwork, which it achieves by exploring a grid of hyperparameters and processing their respective solutions. Specifically, for the range of $\lambda = \lambda_{\text{min}}$, ..., λ_{max} for the same η, it prioritizes the solution with the largest change in size from λ_{n} to $\lambda_{\text{n+1}}$. Such a large change implies that the network is strongly interconnected. This results in one candidate solution for each η, which are processed by removing any node not connected to any other. A score is assigned to each candidate solution by summing their node scores and normalizing by size. The candidate solution with the highest standardized score

is the chosen solution. SigMod is implemented in an R package (Y. Liu 2018).

2.2.3.4 Gene-gene network

Out of the six methods tested, five use a gene-gene interaction network (Section 2.2.3.3). Although their respective statistical frameworks are compatible with any type of network (protein interactions, gene coexpression, regulatory, etc.), for practical reasons we focused on a PPIN, as they are interpretable, well characterized, and most of the methods were designed to scale appropriately to it. We built our PPIN from both binary and co-complex interactions stored in the HINT database (release April 2019) (Das and Yu 2012). Unless specified otherwise, we used only interactions coming from high-throughput experiments to avoid biasing the topology of the network by well-studied genes with more known interactions on average. Out of the 146 722 interactions from high-throughput experiments that HINT stores, we were able to map 142 541 to a pair of HGNC symbols. The scoring function for the nodes changed from method to method (Section 2.2.3.3).

Additionally, we compared the results of the aforementioned PPIN with those obtained on another PPIN built using interactions coming from both high-throughput and targeted studies. In that case, out of the 179 332 interactions in HINT, we mapped 173 797 to a pair of HGNC symbols.

2.2.3.5 SNP networks

SConES (Azencott et al. 2013) is the only of the studied methods designed to handle SNP networks. As in gene networks, two SNPs are linked in a SNP network when there is evidence of shared functionality between two SNPs. The authors suggested three ways of building these networks: connecting the SNPs consecutive in the genomic sequence ("GS network"); interconnecting all the SNPs mapped to the same gene, on top of GS ("GM network"); and interconnecting all SNPs mapped to two genes for which a protein-protein interaction exists ("GI network"). We focused on the GI network, as it is the network that fits better the scope of this chapter. However, at different stages of the chapter we also used GS and GM. For the GM network, we used the mapping described in Section @ref(methods:node_score). For the GI network, we used the PPI as described in Section 2.2.3.4. For all three networks the node score used is the association of the individual SNPs with the phenotype; specifically, we used the 1 d.f. χ^2.

2.2.3.6 Consensus network

The different high-weight subnetwork discovery algorithms make different assumptions on the properties of the solutions, and employ different strategies to find them. Hence,

combining the outcome of the different approaches might provide a more complete outlook on the specific alterations on the GENESIS dataset. We built such consensus network by retaining the nodes that were selected by at least two of the methods. We combined the results of 6 methods: dmGWAS, heinz, HotNet2, LEAN, SConES GI, and SigMod.

2.2.4 Evaluation of methods

2.2.4.1 Classification accuracy of selected biomarkers

A desirable solution is one that is sparse, while offering a good predictive power on unseen samples. We evaluated the predicting power of the SNPs selected by the different methods through the performance of an L1-penalized logistic regression trained exclusively on those SNPs to predict the outcome (case/control). The L1 penalty helps to account for LD to reduce the number of SNPs included in the model (size of the active set), while improving the generalization of the classifier. The value of the regularization parameter, which controls both the magnitude and the sparsity of the coefficients, was set by cross-validation. To that end, we used the different network-methods on a random subset of 80% of the samples. On this same subset we trained our classifier exclusively on the SNPs selected by a particular method. When the method retrieved a list of genes (all of them except SConES), we considered as selected all the SNPs mapped to any of those genes were used. Then we evaluated performance of the classifier on the remaining 20% of the dataset. We repeated this procedure 5 times to estimate the average and the deviation of the different performance measures. The different performance measures we used were: size of the solution, size of the active set, specificity, and sensitivity. The size of the active set provides an estimate of a plausible, more sparse solution with a comparable predictive power to the original solution.

Additionally, for each of the methods, we evaluated their stability and their runtime. The stability of an algorithm is its sensitivity to small changes of the input, and is measured using the Pearson's correlation between different runs as suggested in (Nogueira and Brown 2016). To obtain a baseline, we also performed the procedure using all the SNPs. Lastly, another desirable property is that the method retrieves a good candidate causal subnetwork. In consequence, we compared the outcome of each of the methods to the consensus subnetwork of all the solutions (Section 2.2.3.6).

2.2.4.2 Biological relevance of the genes

An alternative way to validate the results is comparing our results to an external dataset. For that purpose, we recovered a list of 153 genes associated to familial breast cancer from DisGeNET (Piñero et al. 2017). Across this chapter we refer to these genes as *breast cancer susceptibility genes*.

Additionally, we used the summary statistics from the Breast Cancer Association Consortium (BCAC) (Michailidou et al. 2015). BCAC has conducted one of the largest efforts in GWAS, involving over 120 000 women of European ancestry. As opposed to GENESIS, samples were not selected based on family history, and hence is enriched in sporadic breast cancers. Another difference is that BCAC is a relatively heterogeneous study on a pan-European sample, while GENESIS is a homogeneous dataset focused on the French population. Despite these differences, there should be shared genetic architecture. On top of that, that overlap should become larger when the results are aggregated at the gene level. For that purpose, we computed the gene association as in Section @ref(methods:node_score). iCOGS array was used for genotyping in BCAC (Sakoda, Jorgenson, and Witte 2013), the same array as for GENESIS (Sinilnikova et al. 2016). Although imputed data is available, we used exclusively the SNPs available on GENESIS after quality control to make the results comparable.

2.2.5 Code availability

This work required developing computational pipelines for several GWAS analyses, such physically mapping SNPs to genes, computing gene scores, and performing six different network analyses. For each of those processes, a streamlined, project-agnostic pipeline with a clear interface was created. They are compiled in the following GitHub repository: https://github.com/hclimente/gwas-tools. The code that applies these pipelines to the GENESIS project, as well as the code that reproduces all the analyses in this chapter are available at https://github.com/hclimente/genewa. Although the GENESIS dataset is not publicly available, the published code should work on any other GWAS dataset. All the produced gene subnetworks were deposited on NDEx (http://www.ndexbio.org), under the UUID e9b0e22a-e9b0-11e9-bb65-0ac135e8bacf.

2.3 Results

2.3.1 A conventional GWAS shows that FGFR2 is strongly associated with familial breast cancer

We conducted association analyses in the GENESIS dataset at both the SNP and the gene levels (Section @ref(methods:node_score)). Two genomic regions have a P-value lower than the Bonferroni threshold in chromosomes 10 and 16 (Figure 2.2 A). The former overlaps with gene *FGFR2*; the latter with *CASC16*, and it is located near the protein-coding gene *TOX3*. Variants in both *FGFR2* and *TOX3* were related to breast cancer susceptibility in other cohorts negative for *BRCA1/2* (Rinella et al. 2013). Only the peak in chromosome 10 replicated in the gene-level analysis, with *FGFR2* just above the threshold of significance (Figure 2.2B).

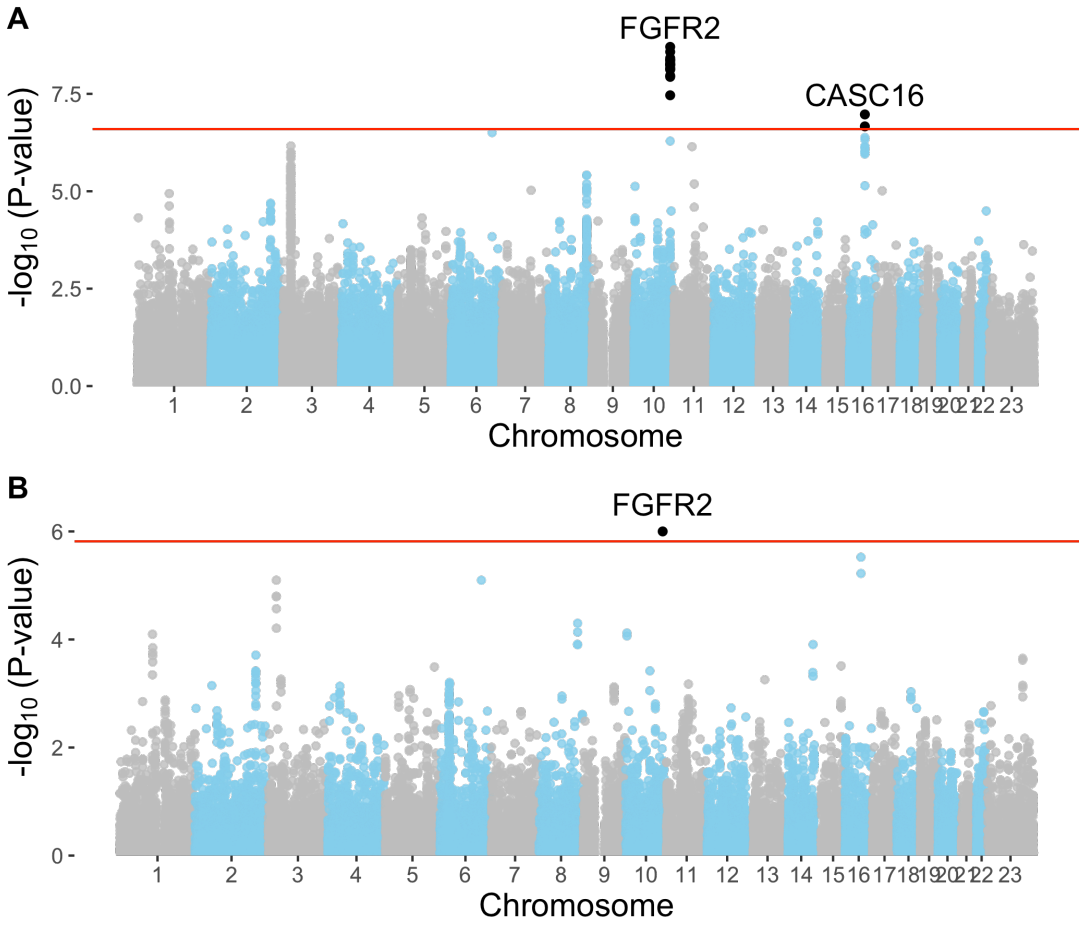


Figure 2.2: Association in GENESIS. The red line represents the Bonferroni threshold. **(A)** SNP association, measured from the outcome of a 1df χ^2 allelic test. Significant SNPs that are within a coding gene, or within 50 kilobases of its boundaries, are annotated. The Bonferroni threshold is 2.54×10^{-7}. **(B)** Gene association, measured by P-value of VEGAS2v2 (Mishra and Macgregor 2015) using the 10% of SNPs with the lowest P-values. The Bonferroni threshold is 1.53×10^{-6}.

Table 2.2: Summary statistics on the results of multiple network methods on the gene-gene interaction network. The first row contains the summary statistics on the whole network.

Network	Num genes	Num edges	$\overline{\mbox{Betweenness}}$	$\hat{\mbox{P}}_{gene}$
HINT HT	13619	142541	16706	0.46
Consensus	55	117	74062	0.0051
dmGWAS	194	450	49115	0.19
heinz	4	3	113633	0.0012
HotNet2	440	374	7739	0.048
LEAN	0	0	NA	NA
SConES GI	0 (1)	0	NA	NA
SigMod	142	249	92603	0.0083

Note:
Num genes: number of genes selected out of those that are part of the PPIN; for SConES GI the including RNA genes, was added in parentheses. *overlinemboxBetweenness*: mean betweenness of th PPIN. *hatmboxP* gene : median P-value of the selected genes. *rho* consensus : Pearson's correlation betw the consensus network.

These results show the overlap in the genetic architecture of the disease between the studied French population sample and other populations, especially at the gene level. In addition, there are other SNPs whose P-values, although higher than the conventional threshold of significance, show a strong association with familial breast cancer. The most prominent of such regions are 3p24 and 8q24, both of which have been associated to breast cancer susceptibility in the past (Brisbin et al. 2011; Ahmed et al. 2009a). This motivates exploring network methods, which trade statistical significance for biological relevance.

2.3.2 Network methods successfully identify genes associated with breast cancer

We applied six network methods to the GENESIS dataset (Section 2.2.3.3), obtaining six solutions (Figure 2.3, Supplementary Files 1 and 2): one for each of the five gene-based methods (Section 2.2.3.4), and one for SConES GI (Section 2.2.3.5). The solutions are very heterogeneous (Table 2.2 and Table 2.3): none of the subnetworks examined by LEAN is significant (adjusted P-value < 0.05), while HotNet2 produced the largest solution subnetwork with 440 genes. SConES GI failed to recover genes in the PPIN, but it recovered one genomic region mapped to RNA gene RNU6-420P. All solution subnetworks except LEAN's are, on average, more strongly associated to breast cancer than the whole PPIN (median P-values $\ll 0.46$), despite containing genes

Table 2.3: Summary statistics on the results of SConES on the three SNP-SNP interaction networks. The first row within each block contains the summary statistics on the whole network.

Network	SNPs	Edges	Subnetworks	X..overline..mbox.Betweenness...	X..hat..mbox.P....
GS	197083	1.97e+05	NA	2.03*(10^7)	
SConES GS	1590	1.58e+03	5	2.52*(10^7)	
GM	197083	6.44e+06	NA	3.99*(10^6)	
SConES GM	1692	1.78e+05	5	4.40*(10^6)	
GI	197083	2.87e+07	NA	1.46*(10^6)	
SConES GI	408	5.39e+02	5	9.33*(10^6)	

Note:
overlinemboxBetweenness: mean betweenness of the selected SNPs in the corresponding full ne
hatmboxP SNP : median P-value of the selected SNPs.

with higher P-values (Figure 2.4). This exemplifies the trade-off between statistical significance and biological relevance. However, there are nuances between solutions: heinz strongly favored highly associated genes, while dmGWAS is less conservative (median gene P-values 0.0012 and 0.19, respectively); SConES tended to select whole LD-blocks; and HotNet2 and SigMod were less likely to select lowly associated genes.

The solution subnetworks present other desirable properties. First, four of the methods succeeded at recovering genes involved in the disease (Figure 2.5), as their subnetworks were enriched in breast cancer susceptibility genes (dmGWAS, heinz, HotNet2, and SigMod, Fisher's exact test one-sided P-value < 0.03). We also compared the outcome of the network methods to the association tests conducted on the population of European ancestry from the Breast Cancer Association Consortium (BCAC) (Michailidou et al. 2015) (Figure 2.6). Encouragingly, every solution subnetwork is enriched in genes or SNPs that are Bonferroni-significant in BCAC. This confirms the capability of network methods to find the same signal as in more powered studies by leveraging on prior knowledge. Second, the genes in four solution subnetworks display on average a higher betweenness centrality than the rest of the genes, a difference that is significant in three solutions (dmGWAS, and SigMod, Wilcoxon rank-sum test P-value $< 1.4\ 10^{-21}$). This agrees with the notion that disease genes are more central than other, non-essential genes (Piñero et al. 2016). We observe that this conclusion holds in this disease, as known breast cancer susceptibility genes have higher betweenness centrality than others (one-tailed Wilcoxon rank-sum test P-value $= 2.64\ 10^{-5}$, Figure 2.7C). Interestingly, SConES' selected SNPs are also more central than the average SNP (Table 2.3), suggesting that causal SNPs are also more central than unrelated SNPs. However, very central nodes are also more likely to be connecting a random pair of nodes, making then

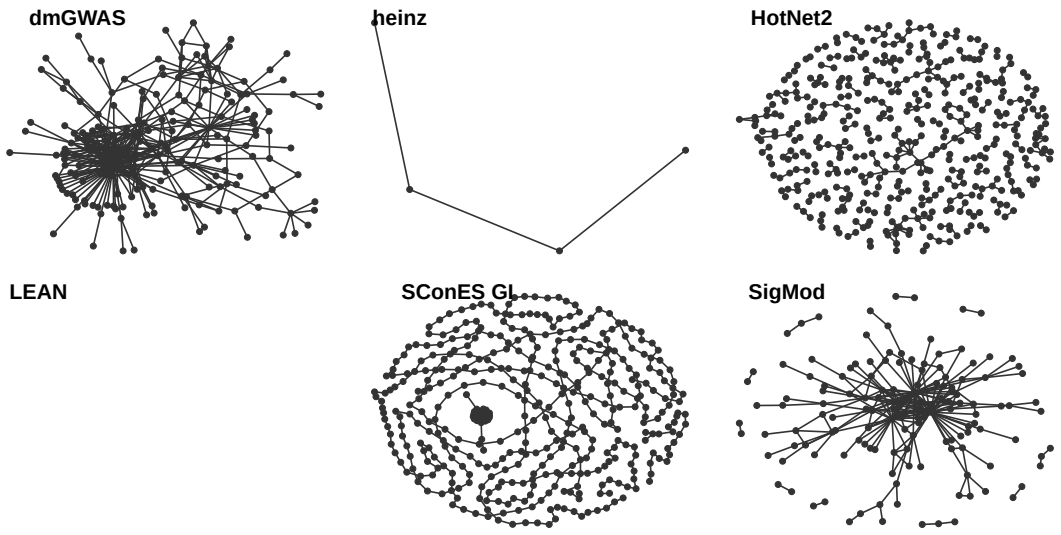


Figure 2.3: Overview of the subnetworks produced by the different network methods. **(dmGWAS, heinz, HotNet2, LEAN, and SigMod)** contain gene subnetworks; **(SConES GI)**, SNP subnetworks.

more likely to be selected by the examined methods. Hence, further work is needed draw conclusions.

As the solutions were quite different from each other it is hard to draw joint conclusions. The 4-gene solution selected by heinz includes the breast cancer susceptibility gene *TOX3*, in region 16q12. By dealing with SNP networks, SConES studies the association of non-coding regions, as well as SNPs in any gene, coding or not. In fact, SConES GI, which adds to GM the interactions between genes, retrieves 4 subnetworks in intergenic regions, and 1 overlapping an RNA gene (*RNU6-420P*). SigMod, despite being related to SConES, produces a vastly different, large solution. On top of recovering three breast cancer susceptibility genes, a keratin-based region of its subnetwork affects the cytoskeleton (*structural constituent of cytoskeleton*, GO enrichment's adjusted P-value = 9.10 10^{-4}), a potentially novel susceptibility mechanism for cancer susceptibility. Interestingly, dmGWAS solution is also related to cytoskeleton (*tubulin binding*, GO enrichment's adjusted P-value = 0.031). But, additionally, it includes a submodule of proteins related to *unfolded protein binding* (GO enrichment's adjusted P-value = 0.045), which has been previously related to cancer susceptibility (Calderwood and Gong 2016). Lastly, HotNet2 produced 135 subnetworks, 115 of which have less than five genes. The second largest subnetwork (13 nodes), contains the two breast cancer susceptibility genes *CASP8* and *BLM*.

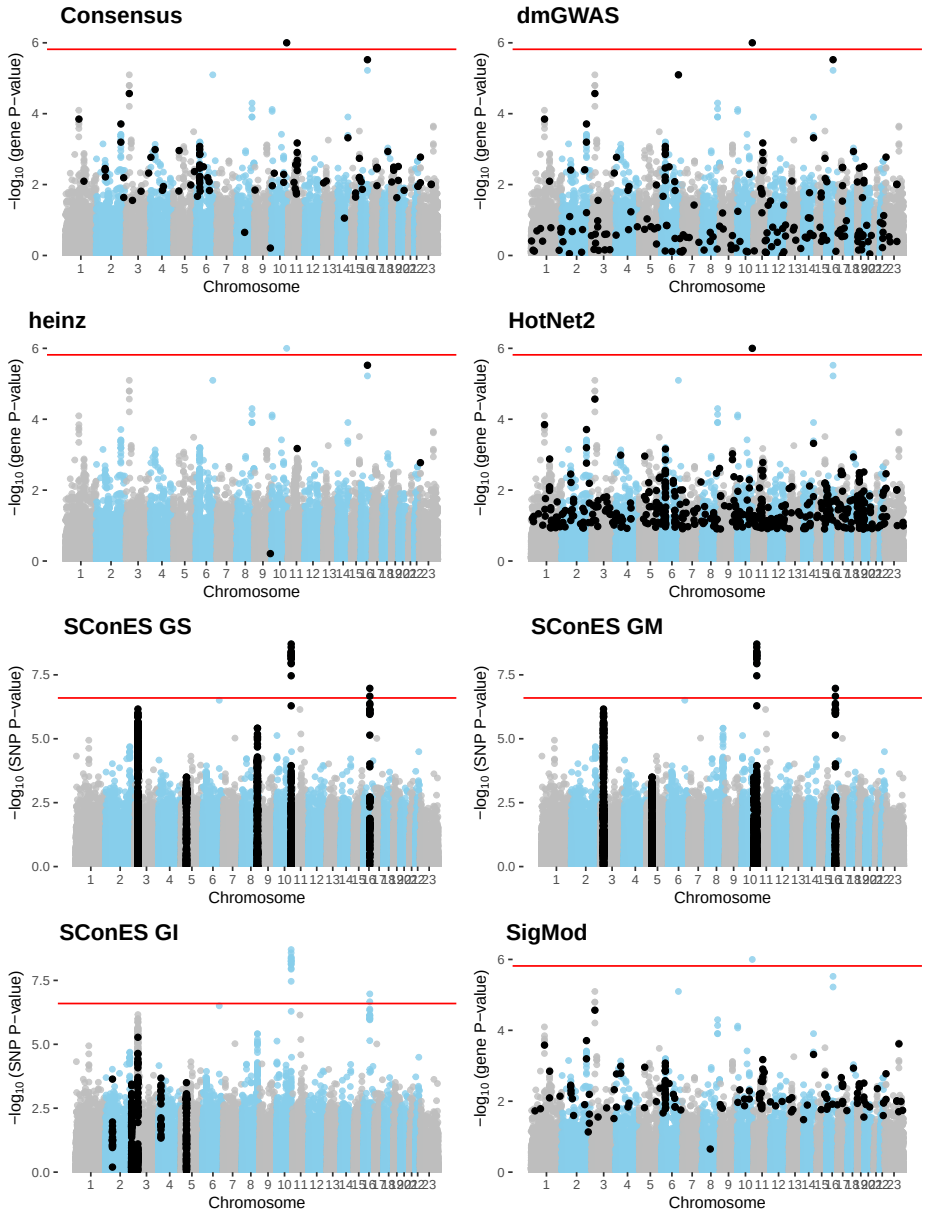


Figure 2.4: Manhattan plots showing the biomolecules selected by each method. In **(Consensus, dmGWAS, heinz, HotNet2, and SigMod)** datapoints are genes; in **(SConES GS, GM, and GI)**, SNPs. LEAN was excluded, as it did not select any gene.

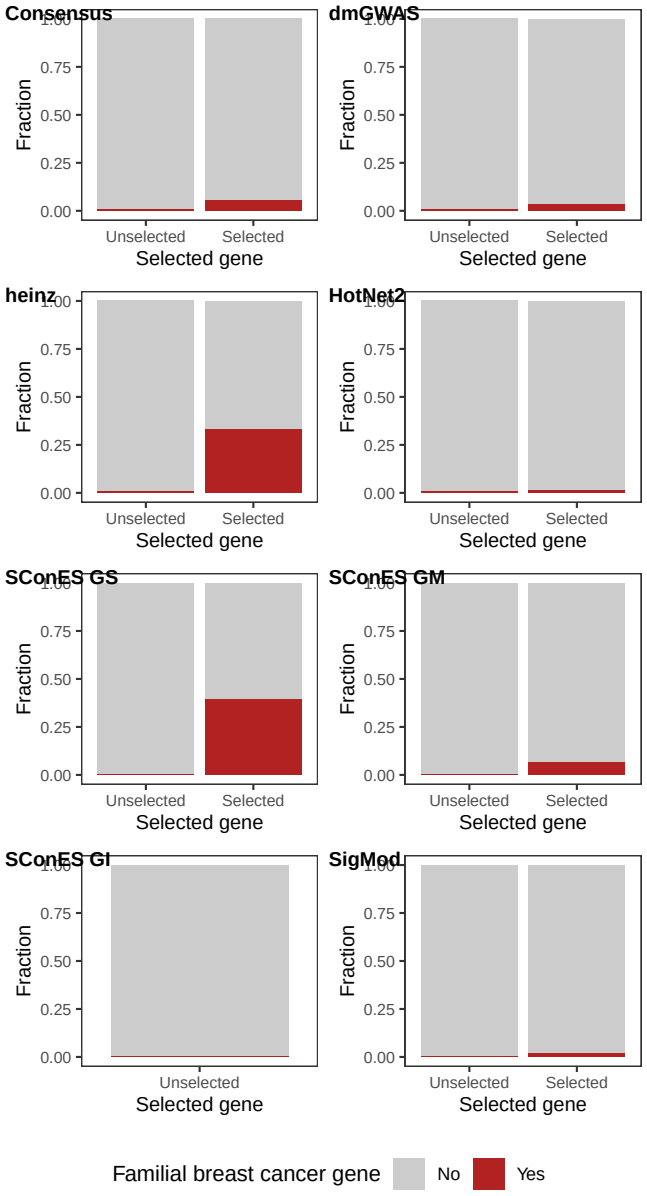


Figure 2.5: Proportion of the selected genes by each of the methods on the GENESIS data that is a known breast cancer susceptibility gene (Section 2.2.4.2). Only genes present in the protein-protein interaction network were considered. LEAN is not displayed as it did not select any gene. The presented network methods recover a higher proportion of breast cancer susceptibility genes than of other genes, despite their lack of significance in GENESIS.

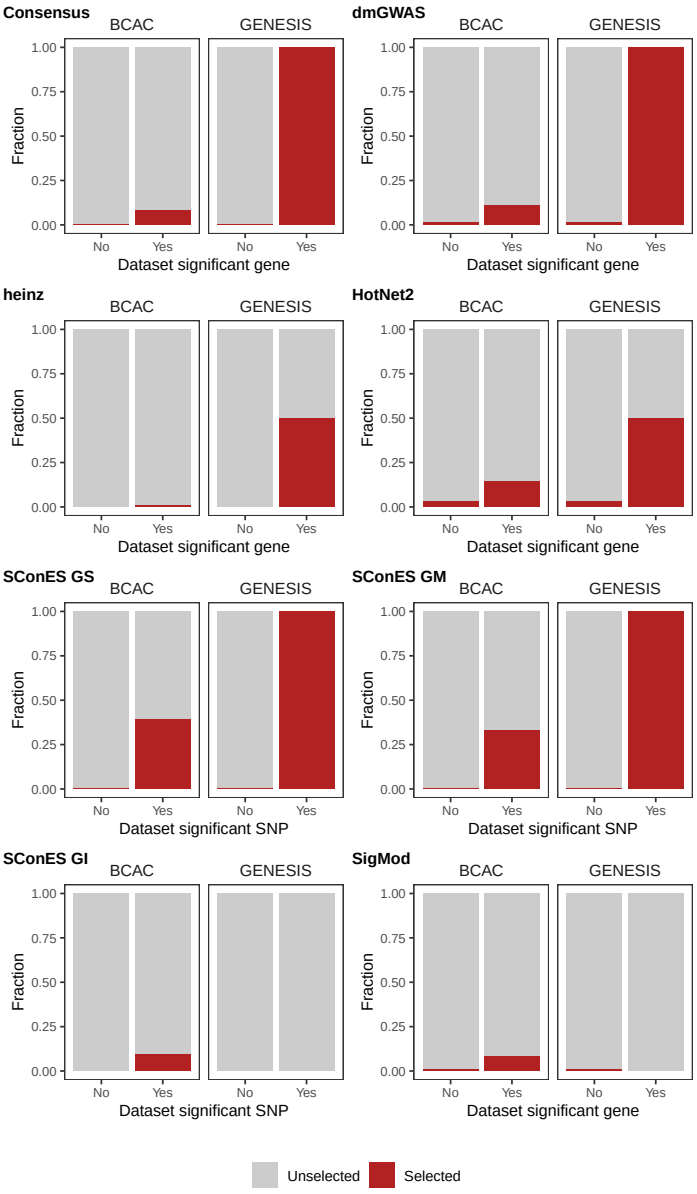


Figure 2.6: Proportion of the Bonferroni significant biomolecules (in either the GENESIS or the BCAC datasets) selected by each of the methods on the GENESIS data. **(Consensus, dmGWAS, heinz, HotNet2, and SigMod)** involve significant genes, only among those present in the protein-protein interaction network. **(SConES GS, GM and GI)** involve significant SNPs. LEAN is not displayed as it did not select any gene. The presented network methods recover a higher proportion of significant genes than of non-significant genes in both datasets, despite their lack of significance in GENESIS.

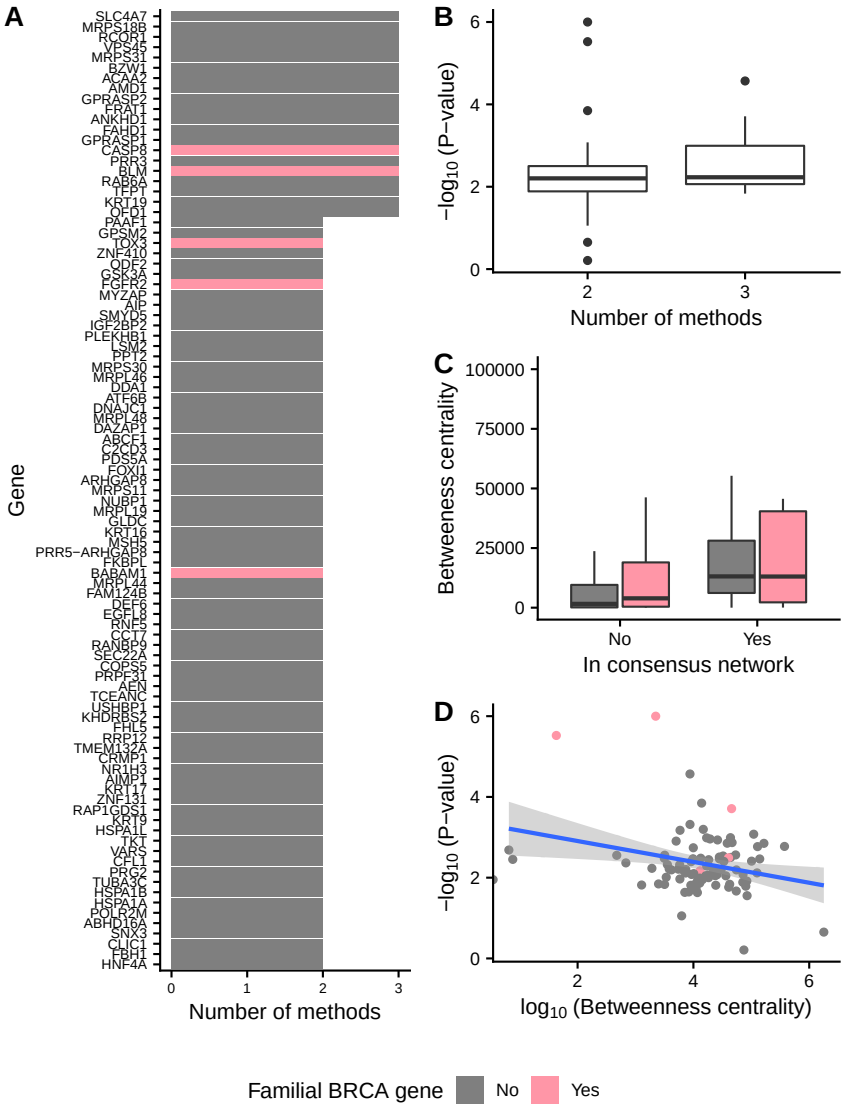


Figure 2.7: Genes on the consensus network. Breast cancer susceptibility genes are colored in pink; the rest are colored in grey. **(A)** Number of methods selecting every gene in the subnetwork. **(B)** VEGAS P-values of association of the genes, with regards to the number of methods that selected them. **(C)** Comparison of betweenness centrality of the genes in the consensus network and the other genes in the PPIN and not in the consensus network. To improve visualization, we removed outliers. **(D)** Relationship between the $\log_{10}$ of the betweenness centrality and the -$\log_{10}$ of the VEGAS P-value of the genes in the consensus network. The blue line represents a fitted generalized linear model.

2.3.3 heinz retrieves a small, highly informative set of biomarkers in a fast and stable fashion

As the employed methods produced such different results, we compared their solutions in a 5-fold subsampling setting (Section 2.2.4.1). Specifically, we measured the following properties (Figure 2.8): (i) size of the solution subnetwork; (ii) sensitivity and specificity of an L1-penalized logistic regression on the selected SNPs; (iii) stability; and (iv) computational runtime.

Both solution size and active set of SNPs selected by Lasso varies greatly between the different methods (Figure 2.8A). heinz has the smallest solutions, with an average of 182 selected SNPs, out of which 5.6% (10.2) are selected by Lasso. The largest solutions come from SConES GI (6256.6 SNPs), and dmGWAS (4255.0 SNPs). Interestingly, heinz has the highest proportion of the selected SNPs that go into the active set (99.9%), although it is high for all the methods ($> 86\%$). This suggests methods are selecting informative SNPs on average.

To determine whether the selected SNPs could be used for patient classification we computed the sensitivity and the specificity of the classifier on the testing data (Figure 2.8B). All classifiers' sensitivities were in the $0.42 - 0.51$ range; the specificities, between 0.54 and 0.62. On average, SigMod had the highest sensitivity (0.51); dmGWAS, the highest specificity (0.52). Both heinz and SigMod had on average better sensitivity than the classifier trained on all the SNPs, but none had superior specificity. However, the differences are negligible, well within the 95% confidence interval.

Another desirable quality of an algorithm is stability (Section 2.2.4.1). Both heinz and LEAN displayed a high stability in our benchmark, consistently selecting the same genes and no genes over the 5 subsamples, respectively (Figure 2.8C). Conversely, the other methods displayed similarly low stabilities.

In terms of computational runtime, the fastest method was heinz (Figure 2.8D), which leverages on its ability to find efficiently the solution in a few seconds. The slowest method was dmGWAS (1 day and 17 hours on average) followed by SConES GI (1 day and 4.32 hours on average). Including the time required to compute the gene scores, however, slows down considerably gene-based methods; on this benchmark, that step took on average 1 day and 9.33 hours. Considering that, it took 3 days and 2.4 hours on average for dmGWAS' to produce results.

2.3.4 No solution is perfect

In practice, and despite their similarities and their involvement in cancer mechanisms, the solutions are remarkably different (Figure 2.10A). That is due to the particularities of the methods which directly or indirectly provide information about the dataset. For

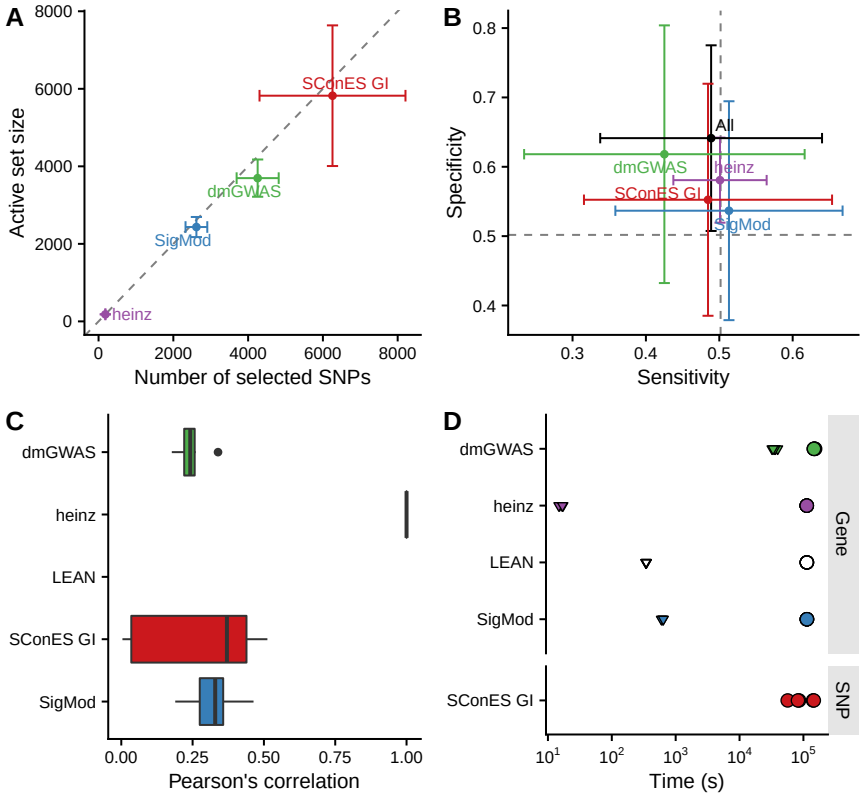


Figure 2.8: Comparison of network-based GWAS methods on GENESIS. Each method was run 5 times of a random subset of the samples, and tested on the remaining samples (Section 2.2.4.1). **(A)** Number of SNPs selected by each method and number of SNPs on the active set used by the Lasso classifier. Points are the average over the 5 runs; lines represent the standard error of the mean. A grey diagonal line with slope 1 is added for comparison. For reference, the active set of Lasso using all the SNPs included, on average, 154 117.4 SNPs. **(B)** Sensitivity and specificity on test set of the L1-penalized logistic regression trained on the features selected by each of the methods. In addition, the performance of the classifier trained on all SNPs is displayed. Points are the average over the 5 runs; lines represent the standard error of the mean. **(C)** Pairwise Pearson's correlations of the solutions used by different methods. A Pearson's correlation of 1 means the two solutions are the same. A Pearson's correlation of 0 means that there is no SNP in common between the two solutions. **(D)** Runtime of the evaluated methods, by type of network used (gene or SNP). For gene network-based methods, inverted triangles represent the runtime of the algorithm itself, and circles the total time, which includes the algorithm themselves and the additional 119 980 seconds (1 day and 9.33 hours) which took VEGAS2v2 on average to compute the gene scores from SNP summary statistics.

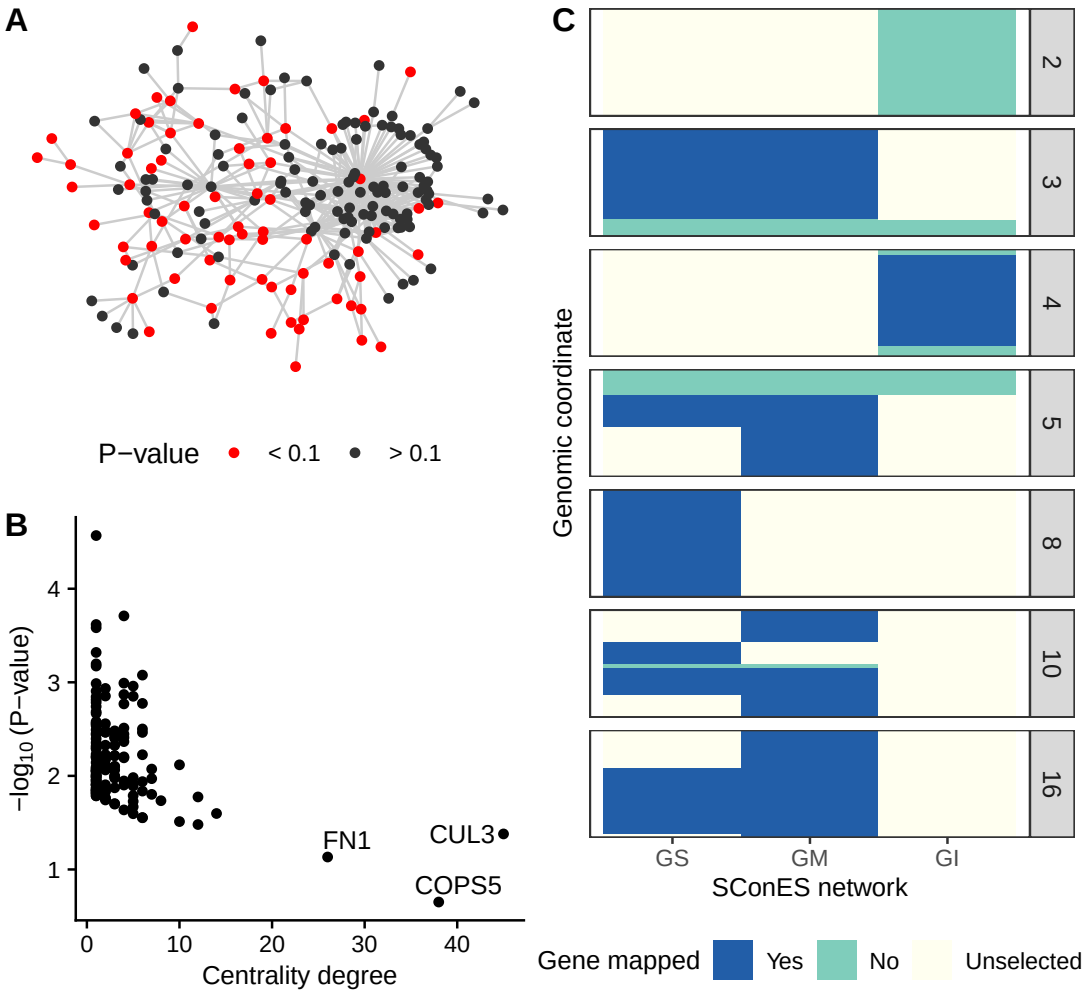


Figure 2.9: Drawbacks confronted when using network guided methods. **(A)** dmGWAS solution subnetwork. Genes with a P-value < 0.1 are highlighted in red. **(B)** Centrality degree and -$\log_{10}$ of the VEGAS P-value for the nodes in SigMod solution subnetwork. **(C)** Genomic regions where either SConES GS, GM or GI select SNPs.

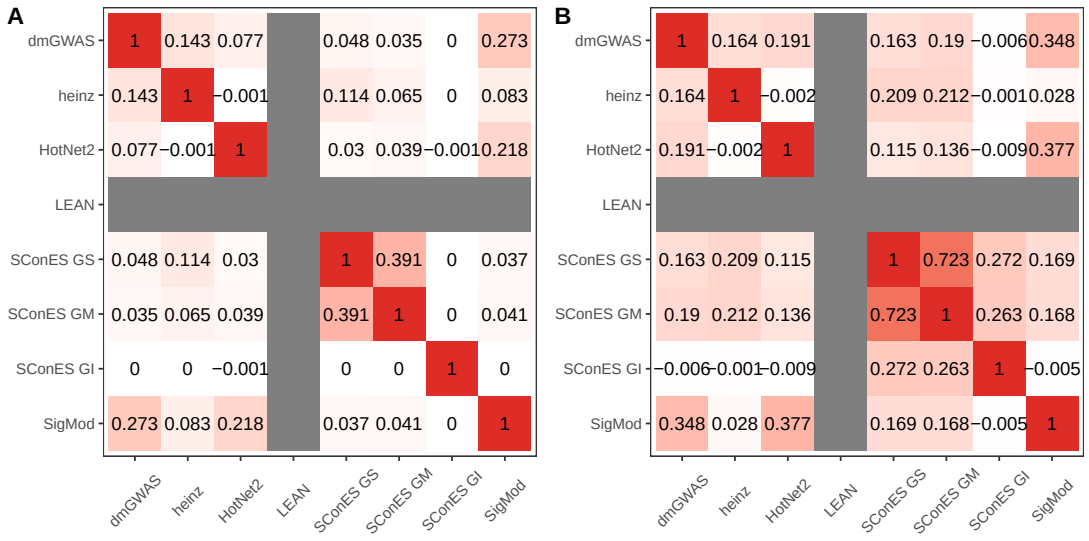


Figure 2.10: Pearson's correlation between the different solution subnetworks. **(A)** Correlation between selected SNPs. **(B)** Correlation between selected genes. In general, the solutions display a very low overlap.

instance, the fact that LEAN did not provide any biomarkers implies that there is no gene such that both itself and its environment are on average strongly associated with the disease.

In this dataset, heinz's solution is very conservative, providing a small solution with the lowest median P-value for the subnetwork (Table 2.2). Due to this parsimonious and highly associated solution, it was the best method to select a set of good biomarkers for classification. (Figure 2.8B). Its conservativeness stems from its preprocessing step, which models the gene p-values as a mixture model of a beta and a uniform distribution, controlled by an FDR parameter. Due to the limited signal at the gene level in this dataset (Figure 2.2B), only 36 of them are retain a positive score after applying the BUM model (Section 2.2.3.3). Hence, heinz's solution subnetwork consists only of 4 genes, which does not provide much insight of the biology of cancer. Importantly, it ignores genes that are strongly associated to cancer in this dataset like *FGFR2*.

On the other end of the spectrum, we have large solutions provided by dmGWAS, HotNet2, and SigMod. dmGWAS' subnetwork is the least associated subnetwork on average. This is due to the greedy framework it uses, which considers all nodes at distance 2 of the examined, and accepts weakly associated genes if they are linked to another, strongly associated one. This is exacerbated when the results of successive greedy searches are aggregated, leading to a large, tightly connected cluster of unassoci-

ated genes (Figure 2.9A). SigMod displays the same tendency, as the most central genes are the least associated to the disease (Figure 2.9B). This relatively low signal-to-noise ratio combined with the large solution requires additional analyses to draw conclusions, such as enrichment analyses. In the same line, HotNet2's subnetwork is even harder to interpret, being composed of 440 genes divided into 135 subnetworks. Lastly, SigMod misses some of the most strongly associated, breast cancer susceptibility genes in the dataset, like *FGFR2* and *TOX3*.

By virtue of using a SNP subnetwork, SConES analyzes each SNP in their context. It therefore selects SNPs in genes none of whose interactors are associated to the disease, as well as SNPs in non-coding regions or in non-interacting genes. In fact, due to linkage disequilibrium, such genes are favored by SConES, as selecting SNPs in an LD block which overlaps with a gene favors selecting the rest of the gene. This might explain why the GS and GM networks, heavily affected by linkage disequilibrium, produce similar results (Figure 2.10B). On the other hand, SConES penalizes selecting SNPs and not their neighbors. This makes it conservative regarding SNPs with many interactions, for instance those mapped to hubs in the PPIN. For this reason, SConES GI did not select any protein coding gene, despite selecting similar regions as SConES GS (Figure 2.9C). In fact SConES GS and SConES GM select regions related to breast cancer, like 16q12 (*TOX3*, Section 2.3.1), 3p24 (*SLC4A7/NEK10* (Ahmed et al. 2009b)), 5p12 (*FGF10*, *MRPS30* (Quigley et al. 2014)), and 10q26 (*FGFR2*, Section 2.3.1). On top of that only SConES GS selects region 8q24 (*POU5F1B* (Breyer et al. 2014)). We hypothesize that the lack of results on the PPIN network of SConES GI and LEAN due to the same cause: the absence of joint association of a module. Although in the case of SConES other hyperparameters could lead to a more informative solution (e.g. lower λ, Section2.2.3.3), it is unclear what is the best strategy to find them. In addition, due to the iCOGS SNP array design, the genome of GENESIS participants has not been unbiasedly surveyed: some regions are fine-mapped — which might distort gene structure in GM and GI networks — while others are under studied — hurting the accuracy with which the GS network captures the genome structure.

2.3.5 Aggregating solutions provides insights into the biology of cancer

To leverage on the strengths of each of the methods and compensate their respective weaknesses, we built a consensus subnetwork that captures the mechanisms most shared among the solution subnetworks (Section 2.2.3.6). The consensus subnetwork (Figure 2.11 contains 93 genes and is enriched in breast cancer susceptibility genes (Fisher's exact test P-value = $7.8\ 10^{-5}$). Due to the limited overlap between methods, only 20 genes were common to more than two of them (Figure 2.7A). Encouragingly, the more methods selected a gene, the higher its association was (Figure 2.7B). Globally,

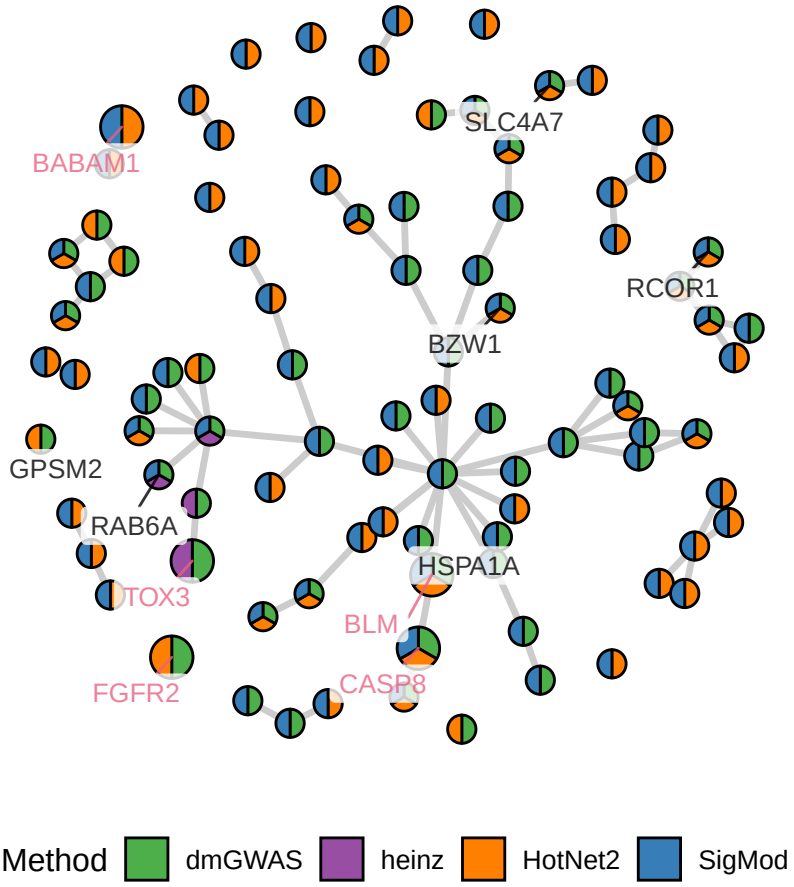


Figure 2.11: Consensus subnetwork on GENESIS (Section 2.2.3.6). Each node is represented by a pie chart, which accounts the methods that selected it. The labeled genes have a VEGAS2v2 P-value < 0.001 and/or are known breast cancer susceptibility genes (colored in pink).

a GO enrichment shows the involvement of two cellular processes: unfolded protein binding, and structural constituent of cytoskeleton (adjusted P-values of 0.001, 0.001, respectively), which were already observed in different solutions (Section 2.3.2). Remarkably, many of the selected genes are related to mitochondrial translation. For instance, MRPS30 (VEGAS P-value = 0.001), encodes a mitochondrial ribosomal protein and was also linked to breast cancer susceptibility (Quigley et al. 2014). Albeit disconnected from MRPS30, the consensus network includes a 2-node subnetwork composed of two mitochondrial ribosomal protein (MRPS31 - VEGAS P-value = 7.67 10^{-3} - and MRPS18B - VEGAS P-value = 7.92 10^{-3}), which suggests an involvement of mitochondrial ribosomes in carcinogenesis (**???**).

We also examined the topological properties of the nodes. The genes in the consensus network have higher betweenness centrality than the rest of the genes (Wilcoxon rank-sum test P-value = 4.29 10^{-18}). Interestingly, within genes in the consensus network, cancer genes are as central as non-cancer genes (Wilcoxon rank-sum test P-value = 0.57). Centrality, however, is weakly anti-correlated with association to the disease (Pearson correlation coefficient = -0.26, Figure 2.7D), which suggests that some highly central genes were selected because they were on the shortest path between two highly associated genes. In view of this, we hypothesize that highly central genes might contribute to the heritability through consistent alterations of their neighborhood, consistent with the omnigenic model of disease (Boyle, Li, and Pritchard 2017). For instance, the most central node in the consensus network is *COPS5* (Figure 2.12), a gene related to multiple hallmarks of cancer and which is overexpressed in multiple tumors, including breast and ovarian cancer (G. Liu et al. 2018). Despite its lack of association in GENESIS (VEGAS P-value = 0.22), its neighbors in the consensus subnetwork have consistently low P-values (median VEGAS P-value = 0.006).

The consensus subnetwork is not completely connected: out of the 93 genes, the largest connected subnetwork includes only 49. A GO enrichment analysis showed that this component is related to three major cellular processes: unfolded protein binding, structural constituent of cytoskeleton, and poly(U) RNA binding (adjusted P-values of 0.01, 0.04, and 0.04, respectively). We found support in the literature of the involvement of each of these functions in the development of cancer, as discussed next. The consensus network also contains a protein directly involved in caspase-mediated apoptosis, *CASP8* (VEGAS P-value = 1.95 10^{-4}). This is related to the enriched activity, *unfolded protein binding*, which inhibits caspase-dependent apoptosis, raising the chances of developing cancer (Calderwood and Gong 2016). It involves three Hsp70 chaperones of the consensus subnetwork: HSPA1A, HSPA1B, and HSPA1L. They genes encoding these proteins are all near each other at 6p21. In fact, out of the 22 SNPs that map to any of these three genes, 9 map to all of them, and 4 to two, making hard to disentangle their association. HSPA1A was the most strongly associated one (VEGAS

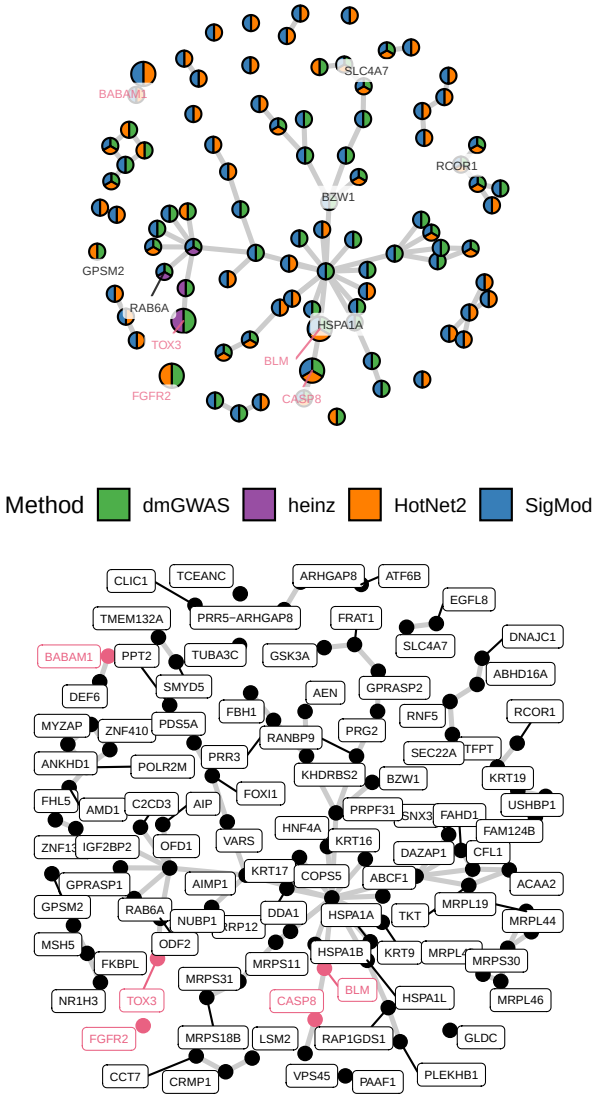


Figure 2.12: Consensus subnetwork on GENESIS (Section 2.2.3.6). **(A)** Each node is represented by a pie chart, which accounts the methods that selected it. The labeled genes have a VEGAS2v2 P-value < 0.001 and/or are known breast cancer susceptibility genes (colored in pink). This panel is equivalent to Figure 2.11. **(B)** The name of every gene is indicated.

P-value = 8.37 10^{-4}). Remarkably, 14 of the 93 genes are in subnetworks of size 1 (isolated) or 2, as they do not have a consistently altered neighborhood. One of them is the well-known breast cancer susceptibility gene *FGFR2* (Section 2.3.1). Another one is the also well-known *SLC4A7* gene (VEGAS P-value = 2.70 10^{5}), which encodes a sodium bicarbonate cotransporter. The genomic region containing both *SLC4A7* and nearby gene *NEK10* (VEGAS P-value = 1.56 10^{-5}) have been consistently associated with breast cancer susceptibility (Ahmed et al. 2009b). *NEK10* is a gene that might be involved in cell-cycle control, but it is absent from the PPIN and hence it could not be studied by gene methods. Despite that, the fact that both dmGWAS, HotNet2 and SigMod link *SLC4A7* in their different subnetwork supports the notion that this gene is the responsible for breast cancer susceptibility.

2.4 Discussion

In this chapter we evaluate the viability of systems biology approaches to GWAS, and examine a GWAS dataset on familial breast cancer focused on BRCA1/2 negative French women. Systems biology addresses two of the largest GWAS issues: interpretability and an overly conservative statistical framework that hinders discovery. This is achieved by considering the biological context of each of the genes and SNPs. Based on divergent considerations of what the desired set of biomarkers is, several methods for network-guided biomarker discovery have been proposed. We reviewed the performance of six of them on GWAS. Despite their differences, most of them produced a relevant subset of biomarkers, recovering known familial breast cancer genes. We also discuss the limitations of such analyses, related to the lack of known interactions around some genes. A crucial step for the gene based methods is the computation of the gene score. In this chapter we used VEGAS2v2 (Mishra and Macgregor 2015) due to the flexibility it offers to use user-specified gene annotations. However, it presents known problems (selection of an appropriate percentage of top SNPs, long runtimes and P-value precision limited to the number of permutations (Nakka, Raphael, and Ramachandran 2016)), other algorithms might have more statistical power.

The network methods we studied differ in what the optimal solution subnetwork looks like. On the one hand, SConES and heinz prefer small highly associated solutions. On the other hand, SigMod and dmGWAS gravitate towards larger, less associated solutions which provide a wide overview of the biological context. While the former provide a reduced set of biomarkers, the latter deepen our understanding of the disease and provide biological hypotheses. They are not exempt of limitations. dmGWAS and SigMod's solution's size require further analyses, which risk oversimplifying their richness. Also, incautious practitioners might be misled by some genes, which are very central in the solution subnetworks, while being weakly associated. Nonetheless, they

are pushed into the solution by their privileged topological properties. On the other end, conservative solutions, like SConES GI and heinz might not shed much light on the etiology of the disease.

To overcome the problems posed by the individual methods while exploiting their strengths, we propose combining them into a consensus subnetwork. We use a straightforward aggregation to generate it, including any node that was recovered by at least two methods. The resulting network is a synthesis of the altered mechanism: it is smaller than the largest solutions (SigMod and dmGWAS), which makes it more manageable, and includes the majority of the strongly associated smaller solutions (SConES and heinz). The consensus subnetwork captures mechanisms and genes known to be related to cancer, recovering known breast cancer susceptibility genes as well as genome regions associated to breast cancer susceptibility. However, thanks to its small size and its network structure, it provides compelling hypotheses of non-canonical mechanisms involved in carcinogenesis, like mitochondrial translation and chaperone activity.

The strength of network-based analyses comes from leveraging prior knowledge to boost discovery. In consequence, they show their shortcomings in front of understudied genes, especially those not in the network. Out of the 32 767 genes that we can map the genotyped SNPs to, 60.7% (19 887) are not in the protein-protein interaction network. The majority of those (14 660) are non-coding genes, mainly lncRNA, miRNA, and snRNA (Figure 2.13). The importance of these genes, like *CASC16*, is highlighted in Section 2.3.1. Among the excluded protein-coding genes we find genes like *NEK10* (P-value 1.6 10^{-5}) or *POU5F1B*, both linked to breast cancer susceptibility (Ahmed et al. 2009b). However, on average protein-coding genes absent from the PPIN are less associated with this phenotype (Wilcoxon rank-sum P-value = 2.79 10^{-8}, median P-values of 0.43 and 0.47). As we are using interactions from high-throughput experiments, such difference cannot be due to well-known genes having more known interactions. As disease genes tend to be more central (Piñero et al. 2016), we hypothesize that it is due to interactions between central genes being more likely. It is worth noting that network approaches that do not use PPIs, like SConES GS and GM, did recover SNPs in *NEK10* and *CASC16*. Lastly, all the methods rely heavily on how SNPs are mapped to genes. In Section 2.3.1 we highlight ambiguities that appear when genes overlap or are in linkage disequilibrium.

As not all databases compile the same interactions, the choice of the PPIN determines the final output. In this work we used exclusively interactions from HINT from high-throughput experiments. This responds to concerns of some authors about biases introduced by adding interactions coming from targeted studies in the literature (Cai, Borenstein, and Petrov 2010; Das and Yu 2012) where a "rich getting richer" phenomenon is observed: popular genes have a higher proportion of their interactions

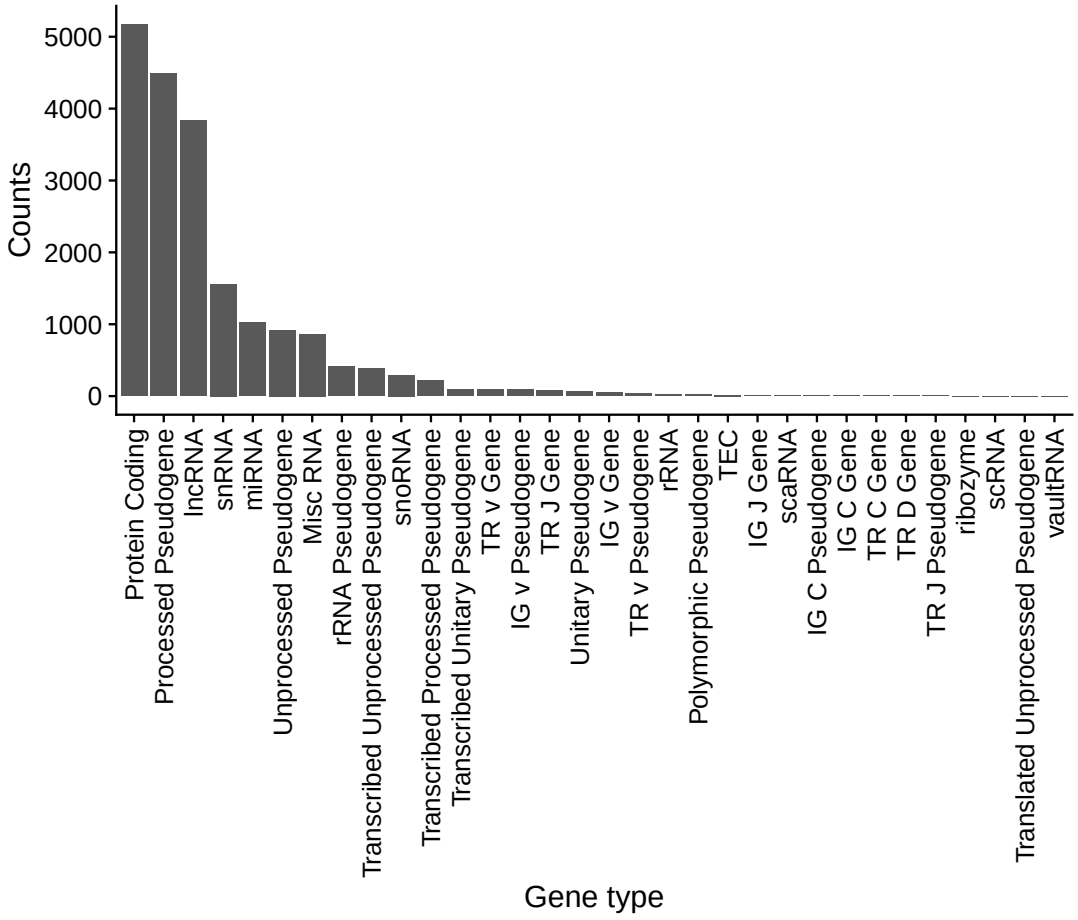


Figure 2.13: Biotypes of genes from the annotation that are not present in the HINT protein-protein interaction network.

described. On the other hand, one study found that the best predictor of the performance of a network for disease gene discovery is the size of the network (Huang et al. 2018). This also supports using the largest amount of interactions. To clarify their impact on this study, we compared the impact of using only physical interactions from high-throughput experiment versus interactions from both high-throughput and the literature (Section 2.2.3.4). We conclude that for most of the methods a larger network did not greatly impact the size or the stability of the solution, the classification accuracy, or the runtime (Figure 2.14).

In order to produce the consensus network, we had to face the different interfaces, preprocessing steps, and unexpected behaviors of the various methods. To facilitate that other authors apply them to new datasets and aggregate their solutions, we built six nextflow pipelines (Di Tommaso et al. 2017) with a consistent interface and, whenever possible, parallelized computation. They are available on GitHub: https://github.com/hclimente/gwas-tools. Importantly, those methods that had a permissive license were compiled into a Docker image for easier use, which is available on Docker Hub hclimente/gwas-tools.

Funding and acknowledgments

This project was supported by funding from Agence Nationale de la Recherche (ANR-18-CE45-0021-01). Financial support for GENESIS resource and genotyping was provided by the Ligue Nationale contre le Cancer (grants PRE05/DSL, PRE07/DSL, PRE11/NA), the French National Institute of Cancer (INCa grant No b2008-029/LL-LC) and the comprehensive cancer center SiRIC, (Site de Recherche Intégrée sur le Cancer: Grant INCa-DGOS-4654).

GENESIS (GENE SISters) is a French national study sponsored by UNICANCER (Sinilnikova et al. BMC Cancer 2016). We wish to thank the genetic epidemiology platform (the PIGE, Plateforme d'Investigation en Génétique et Epidemiologie : Séverine Eon-Marchais, M. Marcou, D. Le Gal, L. Toulemonde, J. Beauvallet, N. Mebirouk, E. Cavaciuti), the biological resource centre (S. Mazoyer, F. Damiola, L. Barjhoux, C. Verny-Pierre, V. Sornin) and all the GENESIS collaborating cancer clinics.

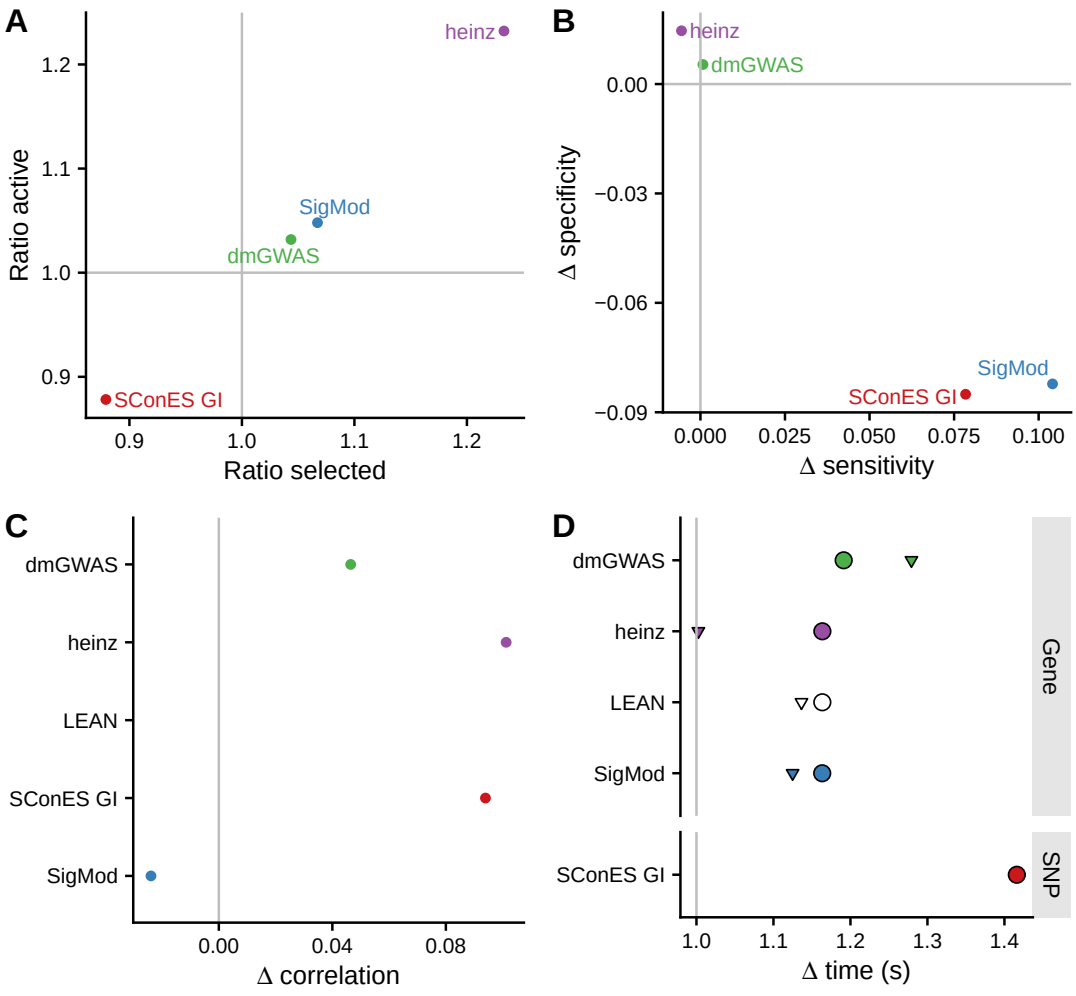


Figure 2.14: Comparison of benchmark on high-throughput interactions to benchmark on both high-throughput and literature curated interactions. Grey lines represent no change between the benchmarks (1 for ratios, 0 for differences). **(A)** Ratios of the selected features between both benchmarks and of the active set. **(B)** Shifts in sensitivity and specificity. **(C)** Shift in Pearson's correlation between benchmarks. **(D)** Ratio between the runtimes of the benchmarks. For gene network-based methods, inverted triangles represent the ratio of runtimes of the algorithms themselves, and circles the total time, which includes the algorithm themselves and the additional 119 980 seconds (1 day and 9.33 hours) which took VEGAS2v2 on average to compute the gene scores from SNP summary statistics. In general, adding additional interactions slightly improves the stability of the solution, but increases the solution size, has mixed effects on the sensitivity and specificity, and impacts negatively the required runtime of the algorithms.

Chapter 3

The *martini* R package

3.1 Introduction

In Chapter 2 I presented six high-score subnetwork search methods, and their application to GWAS. In this chapter I am going to focus on my work on one of them, SConES (Azencott et al. 2013), which was presented with the other methods in Section 2.2.3.3. As a reminder, SConES finds a small set of highly interconnected SNPs associated to the disease by solving the following problem:

$$\arg\max_{S \subseteq G} \underbrace{\sum_{v \in V_S} s_v}_{\text{association}} \underbrace{- \lambda \sum_{v \in V_S} \sum_{u \notin V_S} W_{vu}}_{\text{connectivity}} - \underbrace{\eta |V_S|}_{\text{sparsity}} \tag{3.1}$$

where η and λ are hyperparameters, s_v is the association score of node v, and W is the Laplacian matrix of the network. The mathematical notation is described in Section 2.2.3.2. A particularity of SConES is that it works on a SNP network, where SNPs are linked to each other if there is evidence of shared function For instance, if two SNPs are mapped to the same gene (Section 1.3.1.5), they will share an edge.

With the goal of applying SConES to the GENESIS dataset (Section 1.5.1.1), I worked on a user-friendly version that solved some of SConES' initial shortcomings (explained in Section 3.2). The result was an R package, *martini* (Climente-González and Azencott 2019), which was published in Bioconductor 3.7. Bioconductor is a peer-reviewed R repository. *martini* was the R (user-friendly) version of *gin* (Gwas Incorporating Networks) my C++ re-implementation of SConES based on EasyGWAS' (Grimm et al. 2017). Hence, it combines the accessibility of R and and the Bioconductor environment with the computational efficiency of C++.

3.2 Improvements over SConES

3.2.1 Additional measures of association

SConES scores the relevance of each SNP to the phenotype using the linear SKAT test of association (Wu et al. 2011; Ionita-Laza et al. 2013). We decided to implement additional measures of association, namely χ^2 and logistic regression. The latter allowed *martini* to handle covariates and hence correct for population structure (Section 1.3.1.4).

3.2.2 Hyperparameter optimization

One of the earliest issues we detected in the implementation of SConES was the difficulty to find the most appropriate values for the hyperparameters λ and η. After examining simulated examples were SConES was retrieving suboptimal solutions TODO explain, we implemented the following changes.

3.2.2.1 Selection criterion

SConES chooses the best set of parameters based on consistency across folds. Each λ and η is explored in a 10-fold setting. Each iteration produces a selection vector v, which length is equal to the number of SNPs N. Each of its element is set to 0 if the corresponding SNP was not selected and to 1 if it was. Then, the consistency C between the selection vectors v_i and v_j of two folds for the same λ and η is calculated as

$$C = N\|v_i \cdot v_j\|_0 - \|v_i\|_0\|v_j\|_0.$$

Then a normalized consistency C' is calculated by dividing by the maximum possible consistency C^*, computed as

$$C^* = N \min(\|v_i\|_0, \|v_j\|_0) - \|v_i\|_0\|v_j\|_0.$$

A mean of all pairwise normalized consistencies between selection vectors for a particular η and λ is returned as the consistency score for those hyperparameters. Consistency was the choice for SConES as other model selection approaches displayed proneness to overfitting. For instance, it was examined the performance of a regression model TODO trained over the selected SNPs.

We explored alternatives to consistency as selection criterion, specifically information theory-based criteria. These measures are also known as penalized log-likelihood, as

they take the form

$$L(X, y, \hat{\theta}) - c(\hat{\theta})$$

where $\hat{\theta}$ is a vector of parameters ; $L(X, y, \hat{\theta})$ is the likelihood function of the model; and $c(\hat{\theta})$ is a measurement of model complexity, usually some norm that measures how big $\hat{\theta}$ is (Dziak, Li, and Collins 2005). In general, these measurements take this form:

$$L(X, y, \hat{\theta}) - \lambda p_{in}$$

where λ is a factor that controls the penalty for complexity; and p_{in} is the number of parameters included in the model. We are exploring three measures: the Akaike information criterion(AIC), the Bayesian information criterion (BIC), and the corrected Akaike information criterion (AICc). They are defined as:

$$AIC = 2L(X, y, \hat{\theta}) - 2p_{in},$$

$$BIC = -2L(y|x, \hat{\theta}_{M_i}) - \ln(n)(p_{in} + 2),$$

and

$$AIC_c = AIC + \frac{2p_{in}(p_{in} + 1)}{n - p_{in} - 1} = -2L(X, y, \hat{\theta}) + 2\left(\frac{n}{n - p_{in} - 1}\right) p_{in}.$$

In *martini* it is possible to use either of these three measures to score a particular combination of hyperparameters. As with consistency, every tested η and λ is tested in a 10-fold split of the data. Then, for each fold, a linear model is built, trying to predict the phenotype with the SNPs selected in that fold. These scores how the likelihood of the linear model relates to its size. For a particular combination of hyperparameters, the 10 folds are averaged to get the final score. In this case, the combination of hyperparameters that results in the lowest score is chosen.

3.2.3 Network-based simulations

We propose that SConES will detect biomarker with increased sensibility respect to non-network frameworks. In other words, it should be able to detect causal SNPs for less heritable phenotypes. In order to test this hypothesis, we facilitate the simulation

of phenotypes on real GWAS datasers, setting causal, interconnected SNPs. This simulation tool is implemented in *martini*, and it is broken down into two functions.

The first of them is `simulate_causal_snps()`, which takes a SNP networks in which each SNP has the genes it is mapped to annotated (as can be obtained by `get_GI_network()`, see Section 3.2.4). It takes two additional parameters: the number of genes n involved in the disease, and the proportion of the SNPs p mapped to a causal gene that are causal themselves. Then, it randomly scans the network until it finds maximum of two connected SNP subnetworks that are mapped to n different genes. A fraction p of such subnetwork are selected as causal.

The second function is `simulate_phenotype()`, which re-implements the `--simu-cc` phenotype simulation function of the GCTA suite (Yang et al. 2011). However, when the causal SNPs are the output of `simulate_causal_snps()`, it adds the additional constraint that the causal SNPs are connected in an underlying network. `simulate_phenotype()` requires an existing GWAS experiment (`gwas` parameter), and a set of causal SNPs (`snps` parameter). It also accepts other optional parameters which I describe below. Then, it simulates the quantitative phenotype y_j for patient j using the following additive model:

$$y_j = \sum\nolimits_i w_{ij} u_i + e_j$$

where the weight w_{ij} is the inclination of the genotype i of patient j over the phenotype; the allelic effect of the i-th causal variant u_i in arbitrary units; and the residual effect e_j is the the proportion of the trait not attributable to the genotype. The vector of effect sizes u can be specified by the user via the `effectSize` parameter If it is not, by default it is sampled from a standard Normal distribution.

The weight w_{ij} is calculated as

$$w_{ij} = \frac{x_{ij} - 2p_i}{\sqrt{2p_i(1-p_i)}}$$

where x_{ij} is the number of reference alleles for the i-th causal variant of the j-th individual; and p_i is the frequency of the i-th causal variant. w_{ij} follows a sigmoid like behavior for different p (Figure 3.1): the rarer an allele is, the stronger its impact on the phenotype.

An interesting bit of this simulation is the residual effect e_j. It depends directly on the heritability of the trait, which must be given by the user using the `h2` parameter. Then e_j is generated from a normal distribution with mean of 0 and variance

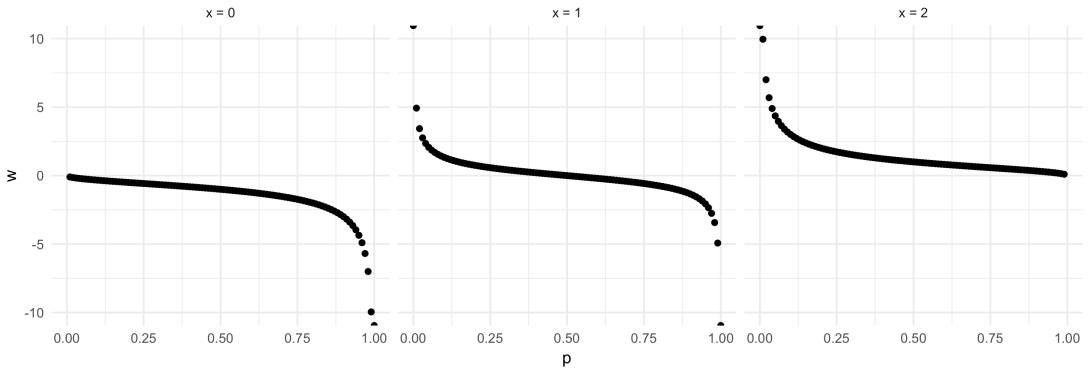


Figure 3.1: Allelic effect i as function of causal allele frequency p for different counts of causal allele in a patient ($x = 0, 1, 2$).

$$\frac{1}{h^2 - 1}\text{var}(\sum\nolimits_i w_{ij} u_i)$$

where w and u are the weight and effect sizes specified above. When all variance is due to genetics ($h^2 = 1$), $e_j = 0$ for all the patients j.

Lastly, a user can request a binary phenotype via the `qualitative = TRUE` option. In this case, the user must also specify three additional parameters: the number of cases `ncases`, the number of controls `ncontrols`, and the prevalence of the trait `prevalence`. With these parameters, `simulate_phenotype()` takes the `ncontrols` samples with the lowest y as controls, and the `ncases` samples with the highest y as cases. However, ncases = prevalence $\times$ $|y|$, where $|y|$ is the total number of samples in the GWAS experiment. This ensures that only the most extreme samples (as defined by the prevalence and the qualitative simulation) are cases.

3.2.4 Interface, documentation and quality assurance

Last, but not least, *martini* includes the two main groups of functions required to run SConES. The first group includes the creation of the SNP networks, which were described in detail in Section 2.2.3.5. They are the `get_GS_network()`, to obtain a network that relates the SNPs based on genomic structure; `get_GM_network()` for a network that, on top of the previous one, relates SNPs mapped to the same gene; and `get_GI_network()` which, on top of the latter, relates SNPs mapped to genes that interact in a user provided list. The second important function is `scones()` which takes a GWAS dataset and a SNP network and runs SConES.

All functions exported by *martini* have a man page, and hence information of the functions arguments, behavior and return value can be obtained via `help(fun)`. Accompanying examples and toy datasets are provided. Additionally, I wrote two vignettes to explain its basic behavior: one to run SConES (Running SConES), and another to simulate network-based phenotypes (Simulating SConES-based phenotypes).

martini was thoroughly subjected to unit tests via the *testthat* package (Wickham 2011). At the moment of writing this text, *martini* had a code coverage of 96%.

3.3 The `scones.nf` pipeline

In addition to the changes implemented in *martini*, I developed a ready-to-use computational pipeline that simplifies its usage. This pipeline just requires genotype data in PLINK binary files and, when needed for the creation of a GM or GI network, a gene annotation file and a protein-protein interaction file. The pipeline `scones.nf` is available on GitHub (https://github.com/hclimente/gwas-tools). In terms of function, the difference with vanilla SConES is that it performs an exhaustive grid-search to optimize both λ and η, as opposed to SConES' single grid search. In the latter, both parameters explore the same range of values, which is calculated from the association scores c (e.g. SKAT score). It creates a linearly spaced n-component vector (n = 10 by default) between $\lfloor \log_{10} \min(c) \rfloor$ and $\lceil \log_{10} \max(c) \rceil$, then explores its powers of 10.

In `scones.nf` we make the grid search finer, because it explores the grid in an iterative way. After the first exploration, the best λ and η according to some selection criterion are picked. Then, a new hyperparameter space ranging from $\log_{10}(\text{best } \lambda) - \Delta$ to $\log_{10}(\text{best } \lambda) + \Delta$, where $\Delta = 0.2(\log_{10} \max \lambda_{explored} - \log_{10} \min \lambda_{explored}$. Prior to these improvements, `scones.nf` was not able, in some instances, to recover the best solution, returning a trivial solution instead.

3.4 Conclusions

In this chapter I introduced *martini* and the `scones.nf` pipeline. Jointly, they make SConES easily applicable to any GWAS dataset. Specially, they provide a wider range of options to the user in terms of how to measure the association between the genotypes and the phenotype, and how to select SConES hyper-parameters. Thanks to these improvements, we were able to obtain the SConES results presented in Chapter 2. Additionally, we provide a network-based phenotype simulation framework. I presented *martini* in a poster entitled "R package for network-guided Genome-Wide Association Studies" in ISMB/ECCB 2017. *martini* (Climente-González and Azencott 2019) is available in Bioconductor (https://www.bioconductor.org/packages/martini); `scones.nf` is available on GitHub (https://github.com/hclimente/gwas-tools).

However, *martini* and SConES still present shortcomings with regards to hyperparameter selection (Sections 2.3.4 and 2.3.3). As we note, SConES solutions were unstable despite using *consistency* for model selection. This requires further examination. Additionally, the selected models do not use the protein-protein interaction network, but other methods do. Although this is, to some extent, expected in the GENESIS dataset, it does not inform much about the biology of the disease. In other words, in this case, SConES should be more tolerant to including unassociated SNPs in order to interconnect subnetworks of strongly associated SNPs. Hence different parameters (lower values of λ) might lead relax the connectivity constraints enough to capture mechanisms that other methods do. In this regard, it would be promising to use topological measures for hyperparameter selection. For instance, favoring the parameters that lead to densely interconnected networks.

Chapter 4

Boosting interpretability and statistical power in epistasis detection by using prior biological knowledge

Joint work with Diane Duroux, Lars Wienbrandt, David Ellinghaus, Chloé-Agathe Azencott and Kristel Van Steen

4.1 Introduction

Genome-wide association studies (GWAS) have identified over 70 000 genetic variants associated with complex traits (Buniello et al. 2019). However, often these variants altogether do not explain the whole genetic architecture of diseases. Possible explanations include a large number of common variants with small effects, rare variants with large effects not covered in GWAS, unaccounted gene-environment interactions, and genetic interactions (Manolio et al. 2009). This project deals with the latter, epistasis (described in Section 1.2). To date, few replicable, functional conclusions have been obtained on Genome-Wide Association Interaction Studies (GWAIS). In this regard, going from SNP to gene epistasis helps converting statistical findings into biological hypotheses (Lehne, Lewis, and Schlitt 2011) and facilitates the functional interpretability of findings (Jorgenson and S Witte 2006). In such gene-level tests, the SNPs mapped to a gene are jointly considered as a set. Aggregation of such SNP statistics into gene statistics is likely to increase the statistical power when dealing with complex diseases (Wu et al. 2010). Yet, specific gene-gene interactions have rarely been identified. This is largely due to the statistical and biological challenges described in Sections 1.3.1 and 1.4, mainly low power, and the difficulty to link tag SNPs to the risk SNP and to genes/functions. Nonetheless, the identification of gene-gene interactions is crucial to properly understand the functional basis of SNP-SNP interactions, and how they relate to disease. For instance, we want to know if two interacting SNPs alter the same gene, and buth alterations are required to modify its function; or if they affect different

genes that participate in the same pathway, or n two pathways that cross-talk.

In this chapter we study the detection of gene-level epistasis from SNP-level epistasis to improve the interpretability of GWAIS. Our analysis compares ways of converting statistical epistasis at the SNP level into a gene-based statistical epistasis network. First, we investigate different functional filters (Ma, Keinan, and Clark 2015) and SNP-to-gene mapping functions. Second, as a bijectivity issue arises because a SNP can be mapped on several genes, we investigate the use of current knowledge on gene-gene interaction to focus on the most promising gene-pairs. We study whether the epistatic interactions driving the phenotype are likely to be currently in existing databases. Third, we used the adaptive truncated product method to estimate gene-pairs significance. Protocols and their associated outputs are compared through networks.

4.2 Materials and methods

4.2.1 Dataset and initial quality control

We investigated the IIBDGC dataset, described in Section 1.5.2.1. Its large sample size helps to overcome the issue of reduced statistical power due, for example, to multiple testing and LD. We performed a quality control as in D. Ellinghaus, Jostins, et al. (2016), reducing the number of SNPs from 196 524 to 130 071.

The IIBDGC dataset aggregates different cohorts, and hence contains confounding population structure (Section 1.3.1.4). PLINK (Purcell et al. 2007) cannot take covariates in the logistic regression used to detect epistasis (Section 4.2.2.3). In consequence, we had to adjust the phenotypes to account for population structure using the top 7 principal components as in D. Ellinghaus, Jostins, et al. (2016). Essentially, we derive adjusted phenotypes from the logistic regression model by subtracting model-fitted values from observed phenotype values (i.e. response residuals).

4.2.2 Gene interaction detection procedure

As we describe below, we applied four different functional filters to the dataset. The functional filter used known interactions between genes, and the three different ways of mapping SNPs to genes described in Section 1.3.1.5, and hence, to these interactions. In essence, in each of the filtered datasets we only tested the interactions between SNPs which can be mapped in a particular way to a pair of interacting genes. The resulting four datasets (including an unfiltered dataset) were analyzed separately. For convenience, we will refer to their associated analyses using the terms *Standard*, *physical*, *eQTL* and *chromatin*. For each dataset, the entire pipeline described below is applied and the four obtained outputs are subsequently compared. An overview of the whole pipeline is available in Figure 4.1.

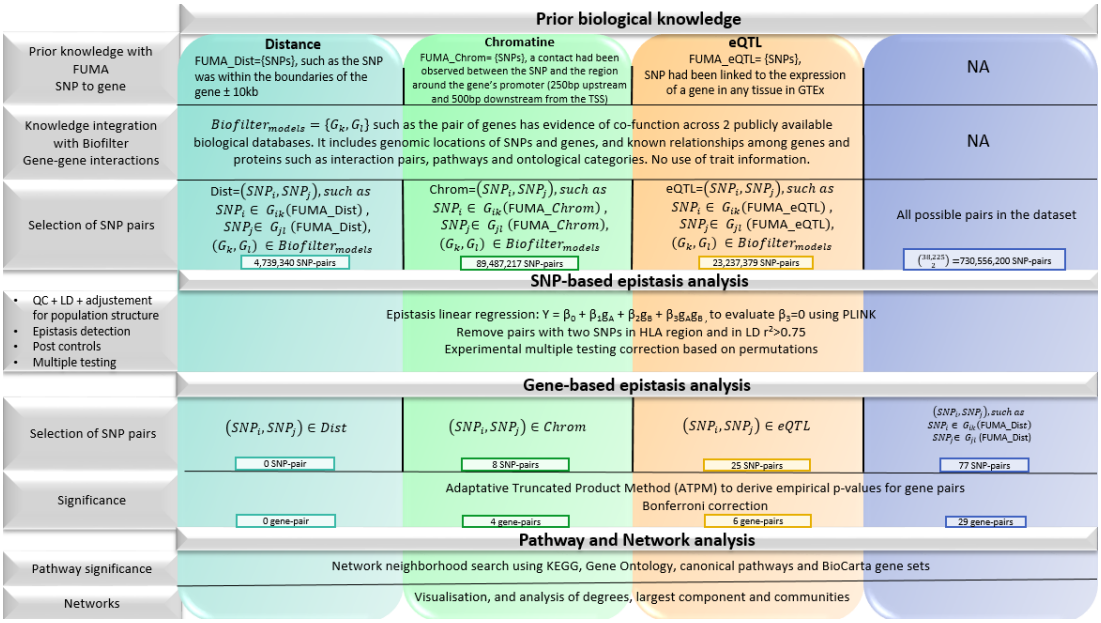


Figure 4.1: Overview of the gene-gene interaction detection procedure. The whole protocol is described in Section 4.2.2.

4.2.2.1 Functional SNP pre-filtering

The initial step of the protocol is a functional SNP pre-filtering, which has three stages. First, we mapped the SNPs in the dataset to genes using FUMA (Watanabe et al. 2017). FUMA is a post-GWAS annotation tool. Its SNP2GENE function takes GWAS summary statistics and maps significant SNPs to genes according to both physical and functional criteria specified by the user. We created an artificial input where every SNP is significant in order to perform such mapping on all the SNPs. We performed three SNP-gene mappings using SNP2GENE: physical, eQTL and 3D chromatin interaction. In the physical mapping, we mapped a SNP to a gene when the former was within the boundaries of the latter ± 10kb. The eQTL mapping uses eQTLs, loci that are significantly associated to the expression of a gene. We obtained eQTL information from GTEx (GTEx Consortium 2017), a project that genotype multiple human subjects, and extracts expression information from their different tissues. We mapped an eQTL SNP to its target gene when the association P-value was significant in any tissue (FDR < 0.05). Lastly, in the 3D chromatin interaction mapping, we mapped a SNP to a gene when a contact had been observed between the former the region around the latter's promoter (250bp upstream and 500bp downstream from the transcription start site) in any of the Hi-C datasets included in FUMA (FDR $< 10^{-6}$). The chro-

matin mapping might contain new, undiscovered, regulatory variants which, as eQTL, regulate the expression of a gene.

Second, after obtaining the SNP-to-gene mappings, we used Biofilter 2.4 (Pendergrass et al. 2013) to obtain the candidate gene-pairs subsequently investigated for epistasis evidence. Biofilter generates pairs of genes with evidence of co-function across multiple publicly available biological databases. It includes genomic locations of SNPs and genes, as well as known relationships among genes and proteins such as interaction pairs, pathways and ontological categories. Biofilter has become popular to reduce burden of tests and hopefully increase interpretability. Notably, it does not use trait information. Specifically, we considered only pairs of genes for which both genes could be mapped, using any of the mappings, to a SNP in our GWAS dataset. We used only gene pairs supported by evidence in at least 2 databases. When the two SNPs of a pair were located in the HLA region, we removed the pair, as this complex genomic region is currently not well understood. Additionally, we removed self-interactions, as detection of within-gene epistasis requires special considerations and is beyond the scope of this paper.

Lastly, we filtered the datasets again to explore exclusively interactions between SNPs mapped to genes known to interact, according to Biofilter. For that purpose we first converted the Biofilter gene-pair models into SNP-pair models separately via each of the FUMA SNP-gene mappings described above. Then, from these SNP pairs sets, we built the four datasets enumerated above: one without any filter (*no filter*); and one for each SNP to gene mapping (*physical*, *eQTL*, *chromatin*). To reduce the number of statistical tests to the minimum, only the SNPs involved in at least one SNP-pair model were included in the respective analyses. It is worth pointing out that SNP-pair models were also built based exclusively on the corresponding mapping e.g. *physical* contains exclusively pairs of SNPs which can be associated to pairs of genes via a physical mapping. This helps interpretability and keeps the number of interactions in the network under control.

4.2.2.2 Post-filtering quality control

Additional quality controls are performed on each of the four generated datasets. As motivated in Gusareva and Van Steen (2014), only common variants (MAF $> 5\%$) and in Hardy–Weinberg equilibrium (P-value > 0.001) are considered for analysis. Also, we pruned SNPs that are in linkage equilibrium ($R^2 > 0.75$). Lastly all risk SNP described in Liu et al. (2015) were included.

4.2.2.3 SNP-level epistasis detection and multiple test correction

We used PLINK 1.9 to detect epistasic association with the phenotype. Specifically, we performed a linear regression on the adjusted phenotypes with the option `-epistasis` to study the association of all the possible SNP pairs:

$$Y = \beta_0 + \beta_1 g_A + \beta_2 g_B + \beta_3 g_A g_B$$

where g_A and g_B are the genotypes under additive encoding for SNPs A and B respectively; β_0, β_1, β_2, β_3, are the regression coefficients. PLINK performs a statistical test to evaluate if $\beta_3 \neq 0$. Crucially, PLINK only returns SNP-pairs with a P-value lower than a specified threshold. We used the default 0.0001.

We only considered pairs of SNPs not in strong LD ($R^2 < 0.75$) and that could be mapped to the corresponding SNP-model obtained from Biofilter, with self-gene interactions removed (Section 4.2.2.1). This impacted the total number of statistical tests. The exception was the *Standard* process, which only underwent LD pruning, but no Biofilter filtering.

To correctly account for multiple testing, the P-value threshold of significance had to be dataset-dependent as the number of tested SNP pairs changed from dataset to dataset (Section 4.2.2.1). We obtained the threshold through a permutation analysis as in Hemani et al. (2014). In essence, for each dataset, we permuted the phenotypes 400 times and measured SNP association as above. This produced a null distribution of the extreme P-values for this number of tests given the LD structure. For each dataset, we took the most extreme P-value from each of the 400 permutations and set the threshold for 5% family-wise error rate (FWER) to be the 95% percentile of these most extreme P-values.

4.2.2.4 From SNP-level to gene-level epistasis

Next we converted SNP-level epistasis into gene-level epistasis. For this purpose, first significant SNP pairs were converted into gene-gene interactions for each dataset. We associated SNP-pairs to gene-pairs using both FUMA and Biofilter, using the same procedure described in 4.2.2.1. The exception was the *Standard* dataset, where SNPs are physically mapped to genes, and no Biofilter restriction on which pairs of genes can interact was applied. Still, self-interactions were removed.

Then, we computed gene-level statistics from the respective SNP-level statistics of the involved SNPs. In this regard, all N pairs of SNPs mapped to a gene pair are taken as a set of tests on the same global null hypothesis H_{0i}, where $i = 1, 2, \ldots, N$. Zaykin et al. (2002) developed the truncated product method (TPM) as a method to

combine P-values on a same global hypothesis. It does so by computing the statistic $W(\tau) = \prod_{i=1}^{N} p_i^{I(p_i \leq \tau)}$ where $I(\cdot)$ is the indicator function and τ is the truncation point. A P-value $\hat{s}(\tau)$ can be the estimated for a given $W(\tau)$. TPM is interesting since it is fast to compute and we do not have P-values for every SNP pair but for the most strongly associated ones only (Section 4.2.2.3). However, TPM requires setting the truncation point τ, a parameter that is arbitrary and might be gene-pair specific. On top of that, the null distribution of $W(\tau)$ is unknown when P-values are correlated, as is the case. To solve these problems the adaptive truncated product method (ATPM) was proposed (Sheng and Yang 2013). ATPM explores different τ, choosing the one that produces the minimum P-value $\hat{s}(\tau)$. We estimated the distribution of the ATPM using permutations as in Ge, Dudoit, and Speed (2003). Specifically, we created $B = 999$ permuted datasets by permuting the phenotype vector. Based on these b^{th} permuted dataset, $1 \leq b \leq B$, we perform the N individual tests. We used three values for τ (0.001, 0.01, 0.05) and set the significance level $\alpha = 0.05$. The specific procedure goes as follows:

1. For each gene-pair in the output of the original data analysis, based on ${p_1}^{(b)}, ..., {p_N}^{(b)}, 0 \leq b \leq B$, calculate the truncated product statistics for each candidate truncation threshold for the original data and B permuted datasets.
2. Based on $W(\tau)_b$, use Ge's algorithm to obtain the estimated P-value $\hat{s}_k^b = \frac{\sum_{l=0}^{B} I(W(\tau)_b \geq W(\tau)_l)}{B+1}$, $1 \leq k \leq K$, $0 \leq b \leq B$
3. Calculate $M_b, 0 \leq b \leq B$, as $M_b = min_{1 \leq k \leq} \hat{s}_k^b$
4. The adjusted P-value for the adaptive truncated product statistic M is estimated as $P_{minP^{(b)}} = \frac{\sum_{b=0}^{B} I(M \geq M_b)}{B+1}$.
5. We reject the joint null hypothesis if the adjusted P-value is smaller than the significance level α.

4.3 Preliminary results

4.3.1 Chromatin contacts map more SNPs per gene than other mappings

We considered three procedures to map SNPs to their gene, as a proxy for their link to functionality (Section 4.2.2.1): *physical*, *eQTL* and *chromatin*. *Chromatin* produced the largest number of mappings (2 394 589), an order of magnitude more than *eQTL* (411 120) and *physical* (174 879) (Table 4.1). Similarly, *chromatin* has the largest number of SNPs mapped to an individual gene, followed by *eQTL* and *physical* (Figure 4.2A). Nonetheless, different genes had very unequal contributions from each of the

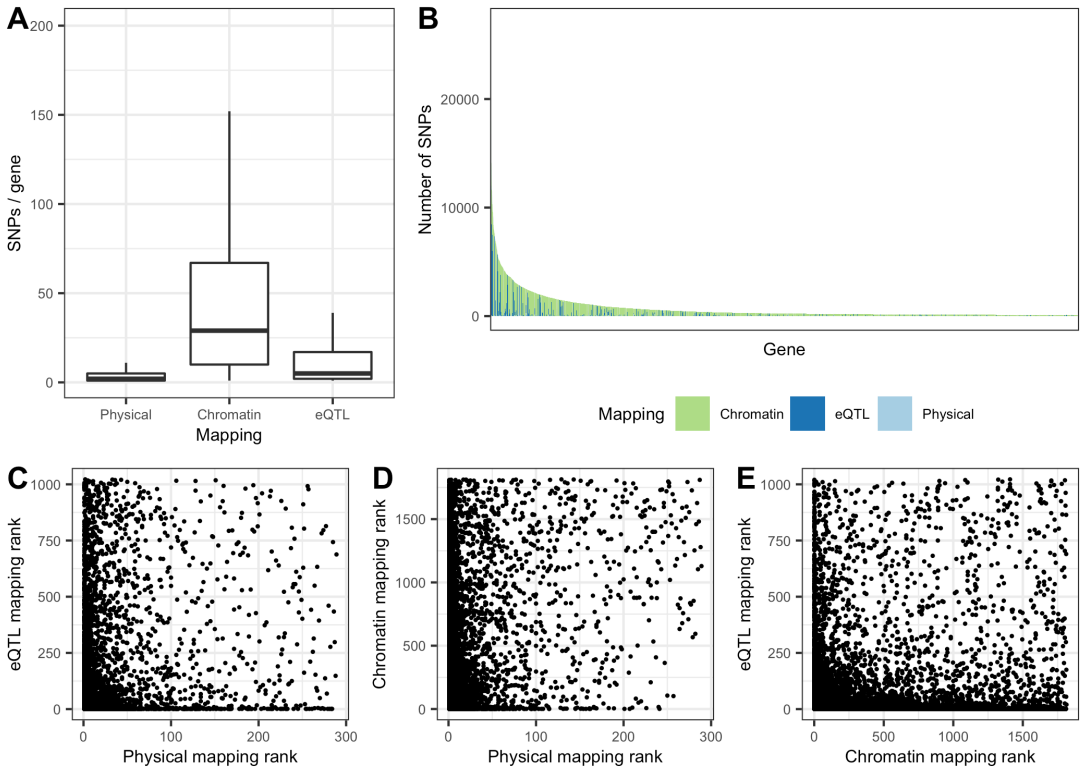


Figure 4.2: **(A)** Number of SNPs per gene for each of the three mappings described in Section 4.2.2.1. Outliers are not displayed to facilitate visualization. **(B)** Ranking of genes with most SNPs mapped using any of the mappings, colored by mapping. Only genes with more than 100 SNPs mapped to it are displayed. **(C,D,E)** Comparison between the rank of each gene according to the number of SNPs mapped to it using each mapping.

mappings (Figure 4.2B). This is consistent with the stark differences between the ranking of genes according to the number of SNPs mapped them (Figures 4.2C, D and E): in general, the genes with the most SNPs mapped using the eQTL mapping had relatively few SNPs mapped in the chromatin mapping, and viceversa.

The number of mappings is directly linked the number of SNPs and interactions tested per dataset (Table 4.1). As it can be observed, restricting our search exclusively to Biofilter-plausible interactions leads to an increase in statistical power with respect to the *Standard* protocol. Specifically, the number of tests are between 1 and 2 orders of magnitude smaller.

Table 4.1: Properties of the different SNP-gene mappings and the filtered datasets.

	Standard	Physical	Chromatin	eQTL
# SNPs	38225	16417	30146	16652
# SNP-gene mappings	NA	1.7e+05	2.4e+06	4.1e+05
# tests	7.2e+08	4.6e+06	8.9e+07	2.2e+07

Table 4.2: Properties of the SNP networks from the different datasets.

	Standard	Physical	Chromatin	eQTL
# significant pairs	57	0	19	64
# nodes	55	0	20	46
# connected components	12	NA	5	6
Size of the largest component	25	NA	11	17
Average degree	2.07	NA	1.9	2.78

4.3.2 The *physical* protocol does not recover any SNP interaction

We searched SNP epistasis in the four datasets (Section 4.2.2.3). The different epistatic SNP-SNP networks are described on Table 4.2 and Figure 4.3. Strikingly, while the Standard protocol generated the largest network (55 nodes), the Chromatin was the largest by number of interactions (64). Physical protocol produced no significant pairs.

4.3.3 Gene-level network

We converted the aforementioned SNP-pair networks into gene-pair epistasis networks, estimating their significance through ARTP (Section 4.2.2.4). Most of the SNP-pairs mapped to exclusively one gene pair in *eQTL* and *Standard*, removing possible sources of ambivalence (Figure 4.4A). That was not the case under the chromatin mapping, where it was more common for the same SNP pair to map to different gene pairs. We then compared the relationship between significant gene-pairs and the number of significant SNP pairs that map to them (Figure 4.4B). Interestingly, we observe that most significant gene pairs are supported by relatively small number of SNPs: either few in number, or few with respect to the total number of SNP pairs for that gene pair.

We built an epistatic gene network from the significant gene pairs (Methods 4.2.2.4), shown in Figure 4.5 and Table 4.3. Overall, the Standard protocol still produces the largest network, and contains more connected components and significant gene-pairs. On the other hand, chromatin and eQTL mappings produce similar networks in terms of sizes, number of gene-pairs and connected components. However, both *chromatin* and *eQTL*'s networks are notably smaller than *Standard*'s (11 and 10 nodes versus 29,

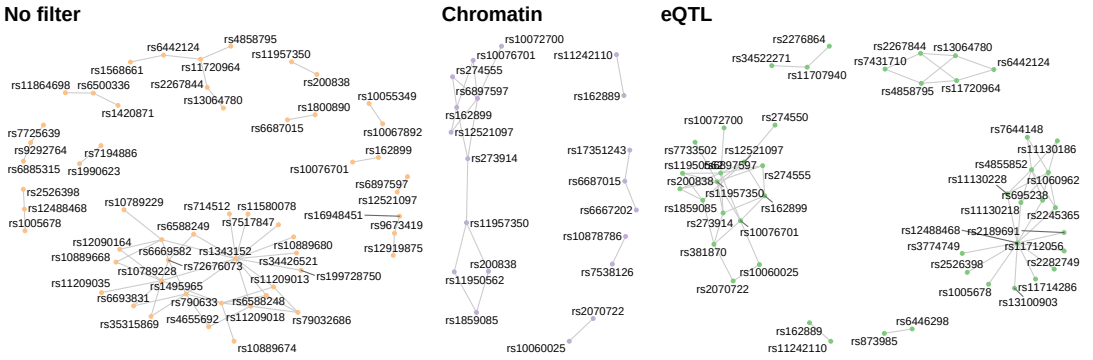


Figure 4.3: SNP-level epistasis networks for *Standard* (orange), *eQTL* (green), and *Chromatin* (violet) (Sections 4.2.2.1 and 4.2.2.3). The *physical* dataset is absent, as no SNP pairs were significant.

Table 4.3: Properties of the gene networks from the different datasets.

	Standard	Physical	Chromatin	eQTL
# significant pairs	26	0	5	7
# nodes	29	0	10	11
# connected components	8	NA	5	5
Size of the largest component	6	NA	2	3
Average degree	1.79	NA	1	1.27

respectively).

Standard's nodes are proportionally more clustered in connected components, while most *eQTL* and *chromatin*'s connected components are composed of only a pair of genes. Although this might reveal the affection of a common mechanism, it is likely a result of the overlap in the genome of multiple genes, which are mapped to highly overlapping sets of SNP.

A hub is node with a number of links that greatly exceeds the average. For this application, we define a hub as a node having a degree strictly superior to three. Only the Standard process contains such hubs: *P4HA2*, *NKD1*, *RNU4ATAC4P*, *C1orf141*, *IL12RB2*, *IL23R* and *USP4*.

Jointly, 38 significant gene pairs are involved in at least one method, involving 46 unique genes. Seven chromosome are involved in epistasis. Notably, 39% of the epistatic genes are located in chromosome 3 and 22% in chromosome 10.

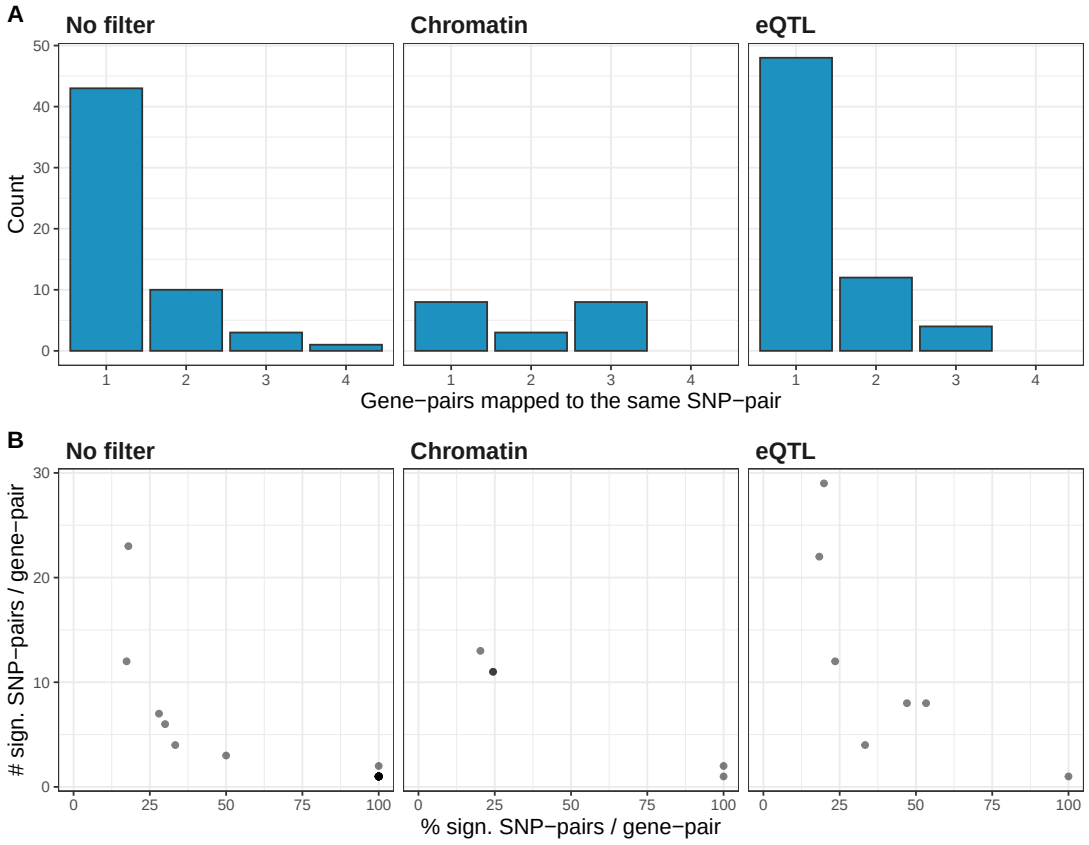


Figure 4.4: Relationship between the number of significant SNP pairs and of significant gene pairs. **(A)** Histogram of the number of significant gene pairs mapped to the same SNP pair. **(B)** Relationship between the total number of SNP pairs mapped to the same gene pair (y-axis), and the percentage of all significant SNP-pairs between all the SNP-pairs mapped to the same gene (x-axis). Data points are semi-transparent, so multiple points stacked result in a darker shade.

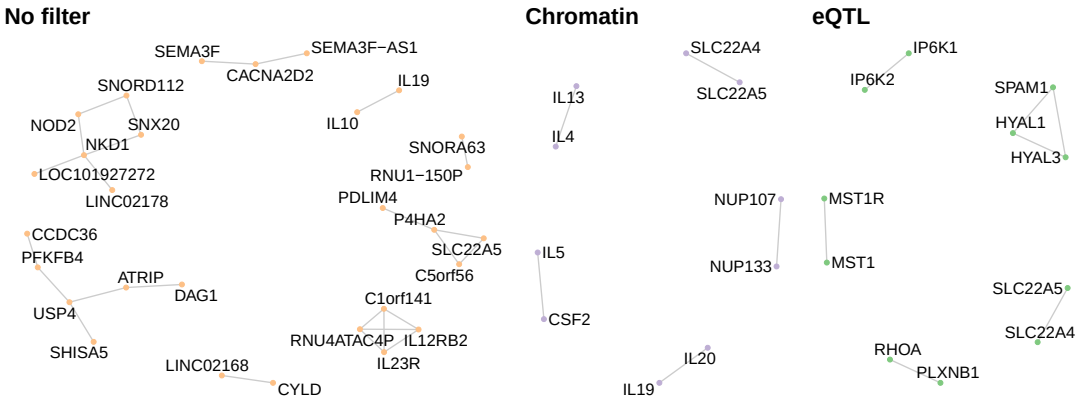


Figure 4.5: (ref:fig-gene-network-caption)

4.3.4 Chromatin and Standard mappings partially replicate previous studies on IBD

Several genetic studies studying epistasis on IBD have been conducted (Lin et al. 2017, 2013; Vermeire et al. 2004; Pedros et al. 2015; McGovern et al. 2009; Glas et al. 2009). We compared them to our results at the gene level, the minimal functional unit at which we expect genetic studies on different populations to converge. For instance several studies showed epistatic alterations involving interleukins, like *IL-10* (Lin et al. 2017), *IL-17* and *IL-23* (McGovern et al. 2009), and *IL-2/IL-21* and *IL-23R* (Glas et al. 2009). Encouragingly, *Standard*'s results include interactions involving both *IL-10* and *IL-23*, although we do not find support for the specific interactions described in the aforementioned studies. In fact, the Standard protocol highlights the relevance of interleukins as hubs (Section 4.3.3). Out of the five gene interactions retrieved in Chromatin pipeline, three of them involve at least one interleukin. Lin et al. (2013) detected interactions involving *NOD2*, with both *IL-23R* and other genes. Our Standard protocol also detects two potentially new epistasis interactions involving *NOD2*.

4.3.5 The type I error of the protocol is controlled

The "multi-stage" nature of the protocol presented in Section 4.2.2 required controlling the type I error. For that purpose, we performed 1 000 permutation analyses for each of the four datasets, permuting the phenotypes and running the entire protocol to detect significant gene pairs. When at least one significant gene-pair is observed in a permutation, that permutation is considered a false positive (FP). This allowed to compute the type I error rate as $\frac{\#\,\mathrm{FP}}{1000}$. We observed that the type I error was under control for

all four datasets (3.6%, 3.7%, 6.1%, and 3.9% for Standard, Physical, Chromatin, and eQTL, respectively).

4.4 Discussion

In this chapter we explore a protocol for functional filtering for epistasis detection on an inflammatory bowel disease dataset (Section 4.2.2). This is expected to bring two advantages. The first one is an increase in statistical power. In GWAIS, the high dimensionality of data requires a conservative approach to multiple testing and limits the detection of epistasis with low effect sizes. The proposed protocol tackles this issue, while controlling for type I error. It does so by limiting the number of tests by filtering the dataset with functional filters. As we observe in Section 4.3.1, the reduction is notable. The second advantage is an improvement of the interpretability of the results, by examining only statistical interactions that map to a known biological interaction. As shown in Section 4.3.4, the proposed eQTL and, specially, chromatin mappings provide results which match the biology of IBD, while corresponding to known interactions. On the other hand, the Standard protocol detects multiple interactions that are hard to interpret. For instance, several interactions involve RNA genes of unknown function (e.g. *LOC101927272* or *LINC02178*). Hence, our results stress that considering the 3D structure of the genome in GWAIS might inform about the susceptibility.

In this chapter, we aim at developing a set of guidelines for the detection of gene epistasis, with an application to IBD. Nonetheless, epistasis detection at the gene level still requires making many choices which were out of the scope of this work. One instance is the choice of encoding for the genotypes. In this work we used the additive encoding, which can lead to an increased false positive rate (Van Steen and Moore 2019). Also we focused on linear regression as epistasis detection algorithm, as it accepts a continuous outcome variable, corrects for main effects, and is computationally efficient. However, other algorithms like MBMDR (Calle et al. 2010), which share these properties but make different assumptions about epistasis, would have been suitable as well. It would be of interest to the GWAIS community for us to provide set of recommendations based on our experience, and the results justifying each of them. Before we reach that point, a few extra experiments are required.

Our current protocol produces compelling hypotheses, and shows the benefits of functional filtering with regards to statistical power and interpretability. However, the multi-stage nature of the process makes it impossible to find out what each of the step brings. For instance, if the detected interactions in any dataset are just a subset of the interactions that could be found without filtering out the interactions not in Biofilter; if they are a subset of the interactions that could be found by a joint *physical+chromatin+eQTL* mapping; or how often are interactions between genes mapped

to SNPs through different mechanisms (e.g. a gene regulated by eQTL and a gene physically mapped). The answers to such questions are relevant to the community, and cannot be answered without isolating their effect from the Biofilter interactions. In other words, it would be useful just map the results of a conventional GWAIS result, to observe how they differ from a conventional mapping. Related to this point, risk SNPs from GWAS are often located in chromatin that is active in the tissues involved in the disease (Boyle, Li, and Pritchard 2017). Hence, the presented protocol might lead to the most biologically plausible epistatic interactions while boosting the power if it focused exclusively on mappings the eQTLs and chromatin mappings obtained in the tissues relevant to IBD (intestines and leukocytes).

It would also be interesting to explore alternative sources of known interactions. In this chapter we worked exclusively on interactions from Biofilter, which compiles multiple databases. The database that Biofilter built contained 37 266 interactions. This is notably smaller than other gene interaction databases, like STRING (Szklarczyk et al. 2019) (11 759 455 interactions). Hence changing databases might result in more, equally interpretable, detected interactions.

Pathway enrichment analyses can inform about the broader framework in which the observed gene epistasis occurs. I would like to adapt the "network neighborhood search" procedure from Yip et al. (2018) to build appropriate gene sets. In summary, given reference biological network (e.g. the Biofilter network), a gene set for a given pair of genes is obtained in three steps:

1. Remove the edge connecting the two genes in the reference network.
2. Find the shortest path between them in the reference network.
3. Create a gene set including the initial two genes and all the genes in the shortest path that are part of the epistasis network as well.

Another important question is which is the null hypothesis we are testing in the pathway enrichment analyses. In this regard, the literature often distinguishes two kinds of test: self-contained and competitive (Wang et al. 2011). As in our study we do not have gene-wise statistics, we are restricted to the former. Those tests compare the overlap between a pathway and the gene set to the expected overlap from taking equally-sized random sets from the universe of genes. This is often tested using a hypergeometric test. However, this approach requires deciding *a priori* what that gene universe is. Selecting all the known genes is not an option, as a GWAS experiment surveys all the genome unbiasedly, but not necessarily so all the genes. This is specially true in an array focused on immunogenomics. Hence, I propose computing the gene background in a dataset specific way. For instance, the *chromatin* results are analyzed in a gene universe where only the genes with a chromatin mapping to the chip are used. However not all genes are surveyed at the same resolution, as we observe in Figure 4.4. If two

genes are equally involved in a disease, we are more likely to find an association in the gene which we are testing more often. Hence, I would like to weight every gene by the number of SNPs that map to it, which should provide a conservative null hypothesis.

Lastly, the protocol presented here is a complex, multi-stage approach which can be useful to any researcher with any GWAIS dataset. In consequence, it would be useful to provide a dataset-agnostic computational pipeline. The user would just need to provide a gene-gene network, a SNP-gene mapping, and a GWAIS dataset, and would be given two epistasic networks, SNP- and gene-based respectively. Generating such a pipeline would also allow us to answer the questions outlined above faster

Acknowledgements

Computational resources have been provided by the Consortium des Équipements de Calcul Intensif (CÉCI), funded by the Fonds de la Recherche Scientifique de Belgique (F.R.S.-FNRS) under Grant No. 2.5020.11 and by the Walloon Region.

Chapter 5

Epistasis networks

Chapter 6

Conclusions

In the last 10 years great progress has been made in the understanding of the genetic architecture of complex diseases. That success would not have been possible without the effort of multiple international consortia which coordinated research groups to tackle specific diseases. Nonetheless, the GWAS and GWAIS experimental settings still present drawbacks, which restrain the applicability of their results. In this work I explore the usefulness of network methods to tackle these challenges.

In **Chapter 2** we apply and critically evaluate different high-score subnetwork search methods to the GENESIS breast cancer GWAS dataset. Such methods are particularly relevant in GWAS, as they address some of the drawbacks of the experimental setting: low statistical power and interpretability. All network methods produced a biologically-plausible answer, which by itself deepens our understanding of the susceptibility mechanisms acting out in this specific dataset. However this methodological comparison context also highlights how radically different the solutions are from each other. This is the product of different mathematical models of the optimal susceptibility mechanism. We explored a combination of the different solutions into a consensus network, which was more manageable than the largest of the individual solutions, and preserved the most important topological and biological properties.

In **Chapter 3** I present *martini*, an R implementation of SConES that addresses some of its initial shortcomings: extending it to case-control phenotypes, adding hyperparameter selection options, and improving user-friendliness. As I show in this chapter, and further elaborate in Chapter 2, SConES is a particularly flexible algorithm among high-score subnetwork search methods. Specifically, it has two hyperparameters that allow to fine-tune the topology and the sparsity of the selected subnetwork. Nonetheless, this flexibility comes at the price of appropriately tuning these parameters, an issue that is exacerbated by the algorithm instability. In this regard, the implemented feature selection scores helped finding more realistic solutions in some simulations, although a final solution to the problem remains to be found. Possibly it will involve scoring the solution using the topology (edge density, centrality betweenness, number of connected components).

In **Chapter 4** TODO

In **Chapter 5** TODO

After completing this work, and having extensively dealt with both GWAS and network methods, it is clear that some challenges remain ahead.

The future of network methods. The network methods I worked with during my PhD are notably heterogeneous. Although that heterogeneity stems from divergences in what different researchers aim to find, being able to obtain different points of view from a disease is a strength. However, clearer language and more exhaustive comparisons to other methods would be well-received in the methods' publications. With the exception of SConES, all high-score subnetwork search methods tested deal with protein-protein interaction networks. SConES on the other hand deals directly with SNP networks. The latter kind of networks are potentially very interesting, as they operate at a lower level than the gene and, hence, potentially can handle more information. For instance, it could contain information about the specific protein residues that participate in a protein-protein interface. Or encode LD blocks, via altering the weights of the edges of the network in a proportional way to the correlation between SNPs. This is highlighted by the positive results of SConES even when no protein-protein information was added to the SNP network. On the other hand, the SNP networks that I handled in this thesis were orders of magnitude more complex than the corresponding protein-protein interaction networks for the same dataset. Probably, new methods (or faster versions of the current ones) are required to explore the potential of such, more informative, SNP networks.

The role of network methods in bioinformatics. My experience with network methods sometimes clashed with other people's impressions of the field during informal exchanges. In particular, some researchers receive network results and methods with a mix puzzlement in face of the overwhelming amount of information they represent, and a subsequent skepticism. In my opinion, this anectodal evidence speak to the state of the field of study of biological networks, which lacks clear, agreed on, protocols and goals. Hence, I believe a multiple-front effort must be made to close this gap. One of them are more accessible tools, with similar interfaces and proper documentation. One of the goals of Chapter 2 was to create such interface at least for the tools I dealt with. Also, better tools to visualizing and manipulating network results. Moves in the right direction are innovative visualizations like hive plots (Krzywinski et al. 2012), and the package *tidygraph* to manipulate networks (Pedersen 2019). Interaction databases are also part of the inaccessibility: (Huang et al. 2018) evaluated 21, often collecting overlapping information, and with unclear definitions of what *interaction* means. Efforts must be unified towards a single database, with a user-friendly interface and clarity about its contents[1]. In this regard, I believe HINT(Das and Yu 2012) is

[1] https://xkcd.com/927/

a step in the right direction. This relates to the issue, also discussed in Chapter 2, of the different types of biological interactions: despite the preponderance of protein-protein interactions, other types of biological interactions need to be better compiled and characterized. Lastly, as long as possible, we should be able to establish meaningful goals in the field, which will allow for descriptors of the networks and protocols on how to achieve them.

The future of GWAS. GWAS has been a sound success in identifying genetic associations with complex traits, and in understandind their genetic basis. However, work remains to be done in the field of functional genomics, that is, finding the cellular function associated to a specific genotype. In this front, network methods can be powerful allies, as shown in Chapter 2, but work in other fields, both experimental and *in silico*, is also required. For instance a proper incorporation of LD patterns to processing analysis, might lead to both an increase in statistical power (by reducind the number of tests to one per independent test), the interpretability (by dealing with the true unit of variation), while accounting for population structure by default (as it is caused by LD). Also, the technology of choice for GWAS is slowly shifting from SNP arrays to the increasingly affordable whole genome-sequencing. This in itself will bring a substantial change to the GWAS scene, by providing a deeper coverage which will include rare variants. Nonetheless, this increase in the number of variants studied makes even more necessary an appropriate treatment of LD.

The future of GWAIS. Epistasis detection is an open and promising field. As in high-score subnetwork search, and as I describe in Chapter 5, epistasis detection enjoys a multiplicity of tools that capture different aspects of the problem. And as with those network methods, we can exploit that multiplicity by collapsing the results into a unified view of disease. Yet, bit challenges remain ahead. The first one is the lack of a standardize protocol for epistasis detection. The second one, is the inability of most methods to account for population structure, either by accepting covariates, or by accepting an adjusted, continuous phenotype. Regarding the latter, an identital problem comes up in the study of continuous phenotypes, which were beyond the scope of this thesis. In this thesis we explored the possible contribution of functional pre-filtering to epistasis detection. In this regard, another compelling field of study are epistasis-detection methods that exploit prior knowledge in the form of a network.

Open science with sensitive data. I would like to finish reflecting about my experience on *open science* as a GWAS researcher. Understandably, GWAS data from human samples requires a careful treatment, as it contains very sensitive information about the participants and their families. For that reason, dealing with GWAS data requires compromising on open data, one of the pillars of open science. TODO IDEAS? (Azencott 2018)

Appendix A

Block HSIC Lasso: model-free biomarker detection for ultra-high dimensional data

Bioinformatics, 35, 2019, i427–i435
doi: 10.1093/bioinformatics/btz333
ISMB/ECCB 2019

Block HSIC Lasso: model-free biomarker detection for ultra-high dimensional data

Héctor Climente-González[1,2,3,4], Chloé-Agathe Azencott[1,2,3], Samuel Kaski[5] and Makoto Yamada[4,6,*]

[1]Institut Curie, PSL Research University, Paris F-75005, France, [2]INSERM, U900, Paris F-75005, France, [3]MINES ParisTech, PSL Research University, CBIO-Centre for Computational Biology, Paris F-75006, France, [4]RIKEN AIP, Tokyo 103-0027, Japan, [5]Department of Computer Science, Aalto University, Espoo, Finland and [6]Department of intelligence science and technology, Kyoto University, Kyoto 606-8501, Japan

*To whom correspondence should be addressed.

Abstract

Motivation: Finding non-linear relationships between biomolecules and a biological outcome is computationally expensive and statistically challenging. Existing methods have important drawbacks, including among others lack of parsimony, non-convexity and computational overhead. Here we propose block HSIC Lasso, a non-linear feature selector that does not present the previous drawbacks.

Results: We compare block HSIC Lasso to other state-of-the-art feature selection techniques in both synthetic and real data, including experiments over three common types of genomic data: gene-expression microarrays, single-cell RNA sequencing and genome-wide association studies. In all cases, we observe that features selected by block HSIC Lasso retain more information about the underlying biology than those selected by other techniques. As a proof of concept, we applied block HSIC Lasso to a single-cell RNA sequencing experiment on mouse hippocampus. We discovered that many genes linked in the past to brain development and function are involved in the biological differences between the types of neurons.

Availability and implementation: Block HSIC Lasso is implemented in the Python 2/3 package pyHSICLasso, available on PyPI. Source code is available on GitHub (https://github.com/riken-aip/pyHSICLasso).

Contact: myamada@i.kyoto-u.ac.jp

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Biomarker discovery, the goal of many bioinformatics experiments, aims at identifying a few key biomolecules that explain most of an observed phenotype. Without a strong prior hypothesis, these molecular markers have to be identified from data generated by high-throughput technologies. Unfortunately, finding relevant molecules is a combinatorial problem: for d features, 2^d binary choices must be considered. As the number of features vastly exceeds the number of samples, biomarker discovery is a high-dimensional problem. The statistical challenges posed by such high-dimensional spaces have been thoroughly reviewed elsewhere (Clarke *et al.*, 2008; Johnstone and Titterington, 2009). In general, due to the *curse of dimensionality*, fitting models in many dimensions and on a small number of samples is extremely hard. Moreover, since biology is complex, a simple statistical model such as a linear regression might not be able to find important biomarkers. Those that are found in such experiments are often hard to reproduce, suggesting overfitting. Exploring the solution space and finding true biomarkers are not only statistically challenging, but also computationally expensive.

In machine learning terms, biomarker discovery can be formulated as a problem of feature selection: identifying the best subset of features to separate between categories, or to predict a continuous response. In the past decades, many feature selection algorithms that deal with high-dimensional datasets have been proposed. Due to the difficulties posed by high-dimensionality, linear methods tend to be the feature selector of choice in bioinformatics. A widely used linear feature selector is the Least Absolute Shrinkage and Selection Operator, or Lasso (Tibshirani, 1996). Lasso fits a linear model between the input features and phenotype by minimizing the sum of the least square loss and an ℓ_1 penalty term. The balance between the least square loss and the penalty ensures that the model explains

the linear combination of features, while keeping the number of features in the model small. However, in many instances biological phenomena do not behave linearly. In such cases, there is no guarantee that Lasso can capture those non-linear relationships or an appropriate effect size to represent them.

In the past decade, several non-linear feature selection algorithms for high-dimensional datasets have been proposed. One of the most widely used, called Sparse Additive Model, or SpAM (Ravikumar *et al.*, 2009), models the outcome as a sparse linear combination of non-linear functions based on kernels. However, since SpAM assumes an additive model over the selected features, it cannot select important features if the phenotype cannot be represented by the additive functions of input features—for example, if there exist a multiplicative relationship between features (Yamada *et al.*, 2014).

Another family of non-linear feature selectors are association-based: they compute the statistical association score between each input feature and the outcome, and rank features accordingly. Since these approaches do not assume any model about the output, they can detect important features as long as an association exists. When using a non-linear association measure, such as the mutual information (Cover and Thomas, 2006) or the Hilbert–Schmidt Independence Criterion (HSIC) (Gretton *et al.*, 2005), they select the features with the strongest dependence with the phenotype. However, association-based methods do not account for the redundancy between the features, which is frequent in biological datasets, since they do not model relationships between features. Hence, many redundant features are typically selected, hindering interpretability. This is important in applications like drug target discovery, where only a small number of targets can be validated, and it is crucial to discriminate the most important target out of many other top-ranked targets.

To deal with the problem of redundant features, Peng *et al.* (2005) proposed the minimum redundancy maximum relevance (mRMR) algorithm. mRMR can select a set of non-redundant features that have high association to the phenotype, while penalizing the selection of mutually dependent features. Ding and Peng (2005) used mRMR to extract biomarkers from microarray data, finding that the selected genes captured better the variability in the phenotypes than those identified by state-of-the-art approaches. However, mRMR has three main drawbacks: the optimization problem is discrete; it must be solved by a greedy approach and the mutual information estimation is difficult (Walters-Williams and Li, 2009). Moreover, it is unknown whether the objective function of mRMR has good theoretical properties such as submodularity (Fujishige, 2005), which would guarantee the optimality of the solution.

Recently, Yamada *et al.* (2014) proposed a kernel-based mRMR algorithm called HSIC Lasso. Instead of mutual information, HSIC Lasso employs the HSIC (Gretton *et al.*, 2005) to measure dependency between variables. In addition, it uses an ℓ_1 penalty term to select a small number of features. This results in a convex optimization problem, for which one can therefore find a globally optimal solution. In practice, HSIC Lasso has been found to outperform mRMR in several experimental settings (Yamada *et al.*, 2014). However, HSIC Lasso is memory intensive: its memory complexity is $O(dn^2)$, where d is the number of features and n is the number of samples. Hence, HSIC Lasso cannot be applied to datasets with thousands of samples, nowadays widespread in biology. A MapReduce version of HSIC Lasso has been proposed to address this drawback, and it is able to select features in ultra-high dimensional settings (10^6 features, 10^4 samples) in a matter of hours (Yamada *et al.*, 2018). However, it requires a large number of computing nodes, inaccessible to common laboratories. Since it relies on the Nyström approximation of Gram matrices (Schölkopf and Smola, 2002), the final optimization problem is no longer convex, and hence finding a globally optimal solution cannot be easily guaranteed.

In this article, we propose block HSIC Lasso: a simple yet effective non-linear feature selection algorithm based on HSIC Lasso. The key idea is to use the recently proposed block HSIC estimator (Zhang *et al.*, 2018) to estimate the HSIC terms. By splitting the data in blocks of size $B \ll n$, the memory complexity of HSIC Lasso goes from $O(dn^2)$ down to $O(dnB)$. Moreover the optimization problem of the block HSIC Lasso remains convex. Through its application to synthetic data and biological datasets, we show that block HSIC Lasso can be applied to a variety of settings and compares favorably with the vanilla HSIC Lasso algorithm and other feature selection approaches, linear and non-linear, as it selects features more informative of the biological outcome. Further considerations on the state of the art and the relevance of block HSIC Lasso can be found in Supplementary File 1.

2 Materials and methods

2.1 Problem formulation

Assume a dataset with n samples described by d real-valued features, each corresponding to a biomolecule (e.g. the expression of one transcript, or the number of major alleles observed at a given SNP), and a label, continuous or binary, describing the outcome of interest (e.g. the abundance of a target protein, or disease status). We denote the ith sample by $\boldsymbol{x}_i = [x_i^{(1)}, x_i^{(2)}, \ldots, x_i^{(d)}]^\top \in \mathbb{R}^d$, where $\top$ denotes transpose; and its label by $y_i \in \mathcal{Y}$, where $\mathcal{Y} = \{0, 1\}$ for a binary outcome, corresponding to a classification problem, and $\mathcal{Y} = \mathbb{R}$ for a continuous outcome, corresponding to a regression problem. In addition, we denote by $\boldsymbol{f}_k = [x_1^{(k)}, x_2^{(k)}, \ldots, x_n^{(k)}]^\top \in \mathbb{R}^n$ the kth feature in the data.

The goal of supervised feature selection is to find m features ($m \ll d$) that are the most relevant for predicting the output y for a sample $\boldsymbol{x}$.

2.2 HSIC Lasso

Measuring the dependence between two random variables X and Y can be achieved by the HSIC (Gretton *et al.*, 2005):

$$\begin{aligned}\mathrm{HSIC}(X, Y) \;&= \mathbb{E}_{x,x',y,y'}[K(x,x')L(y,y')] \\ &+ \mathbb{E}_{x,x'}[K(x,x')]\mathbb{E}_{y,y'}[L(y,y')] \\ &- 2\mathbb{E}_{x,y}[\mathbb{E}_{x'}[K(x,x')]\mathbb{E}_{y'}[L(y,y')]],\end{aligned} \tag{1}$$

where $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ and $L : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ are positive definite kernels, and $\mathbb{E}_{x,x',y,y'}$ denotes the expectation over independent pairs (x, y) and (x', y') drawn from $p(x, y)$. $\mathrm{HSIC}(X, Y)$ is equal to 0 if X and Y are independent, and is non-negative otherwise.

In practice, for a given Gram matrix $\boldsymbol{K}_k \in \mathbb{R}^{n\times n}$, computed from the kth feature, and a given output Gram matrix $\boldsymbol{L} \in \mathbb{R}^{n\times n}$, the normalized variant of HSIC is computed using its V-statistic estimator as (Yamada *et al.*, 2018)

$$\mathrm{HSIC}_v(\boldsymbol{f}_k, \boldsymbol{y}) = \mathrm{tr}(\overline{\boldsymbol{K}_k}\,\overline{\boldsymbol{L}}), \tag{2}$$

where for a Gram matrix $\boldsymbol{K} \in \mathbb{R}^{n\times n}$, $\overline{\boldsymbol{K}}$ is defined as $\overline{\boldsymbol{K}} = \boldsymbol{HKH}/||\boldsymbol{HKH}||_F$ with $\boldsymbol{H} \in \mathbb{R}^{n\times n}$ a centering matrix defined by $H_{ij} = \delta_{ij} - \frac{1}{n}$. Here δ_{ij} is equal to 1 if $i = j$ and 0 otherwise, and tr denotes the trace. Note that we employ the normalized variant of the original empirical HSIC.

The largest the value of $\mathrm{HSIC}_v(\boldsymbol{f}_k, \boldsymbol{y})$, and the more dependent the kth feature and the outcome are. Song *et al.* (2012) therefore

proposed to perform feature selection by ranking the features by descending value of $\mathrm{HSIC}_v(\boldsymbol{f}_k, \boldsymbol{y})$.

With HSIC Lasso, Yamada *et al.* (2014) extend the work of Song *et al.* (2012) so as to avoid selecting multiple redundant features. For this purpose, they introduce a vector $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_d]^\top$ of feature weights and solve the following optimization problem:

$$\max_{\boldsymbol{\alpha} \geq 0} \sum_{k=1}^{d} \alpha_k \mathrm{HSIC}_v(\boldsymbol{f}_k, \boldsymbol{y}) - \frac{1}{2} \sum_{k,k'=1}^{d} \alpha_k \alpha_{k'} \mathrm{HSIC}_v(\boldsymbol{f}_k, \boldsymbol{f}_{k'}) - \lambda ||\boldsymbol{\alpha}||_1. \quad (3)$$

The first term enforces selected features that are highly dependent on the phenotype; the second term penalizes selecting mutually dependent features and the third term enforces selecting a small number of features. The selected features are those that have a non-zero coefficient α_k. Here $\lambda > 0$ is a regularization parameter that controls the sparsity of the solution: the larger λ, the fewer features have a non-zero coefficient.

The HSIC Lasso optimization problem can be rewritten as

$$\min_{\boldsymbol{\alpha} \geq 0} ||\mathrm{vec}(\overline{\boldsymbol{L}}) - [\mathrm{vec}(\overline{\boldsymbol{K}}_1), \ldots, \mathrm{vec}(\overline{\boldsymbol{K}}_d)]\boldsymbol{\alpha}||_2^2 + \lambda ||\boldsymbol{\alpha}||_1,$$

where $\mathrm{vec} : \mathbb{R}^{n \times n} \to \mathbb{R}^{n^2}, \boldsymbol{K} \mapsto [K_{11}, \ldots, K_{1n}, K_{21}, \ldots, K_{nn}]$ is the vectorization operator. Using this formulation, we can solve the problem using an off-the-shelf non-negative Lasso solver.

HSIC Lasso performs well for high-dimensional data. However, it requires a large memory space ($O(dn^2)$), since it stores d Gram matrices. To handle this issue, two approximation methods have been proposed. The first approach uses a memory lookup to dramatically reduce the memory space (Yamada *et al.*, 2014). However, since this method needs to perform a large number of memory lookups, it is computationally expensive. Another approach (Yamada *et al.*, 2018) is to rewrite the problem using the Nyström approximation (Schölkopf and Smola, 2002) and solve the problem using a cluster. However using the Nyström approximation makes the problem non-convex.

2.3 Block HSIC Lasso

In this article, we propose an alternative HSIC Lasso method for large-scale problems, the *block HSIC Lasso*, which is convex and can be efficiently solved on a reasonably sized server.

Block HSIC Lasso employs the block HSIC estimator (Zhang *et al.*, 2018) instead of the V-statistics estimator of Equation (2). More specifically, to compute the block HSIC, we first partition the training dataset into n/B partitions $\{\{(\boldsymbol{x}_i^\ell, \boldsymbol{y}_i^\ell)\}_{i=1}^B\}_{\ell=1}^{n/B}$, where B is the number of samples in each block. Note that the block size B is set to a relatively small number such as 10 or 20 ($B \ll n$). Then, the block HSIC estimator can be written as

$$\mathrm{HSIC}_b(\boldsymbol{f}_k, \boldsymbol{y}) = \frac{B}{n} \sum_{\ell=1}^{n/B} \mathrm{HSIC}_v(\boldsymbol{f}_k^{(\ell)}, \boldsymbol{y}^{(\ell)}),$$

where $\boldsymbol{f}_k^{(\ell)} \in \mathbb{R}^B$ represents the kth feature vector of the ℓth partition. Note that the computation of $\mathrm{HSIC}_v(\boldsymbol{f}_k^{(\ell)}, \boldsymbol{y}^{(\ell)})$ requires $O(B^2)$ memory space. Therefore, the required memory for the block HSIC estimator is $O(nB^2)$, where $nB \ll n^2$.

If we denote by $\overline{\boldsymbol{K}}_k^{(\ell)} \in \mathbb{R}^{B \times B}$ the restriction of $\overline{\boldsymbol{K}}_k$ to the ℓth partition, and by $\overline{\boldsymbol{L}}^{(\ell)} \in \mathbb{R}^{B \times B}$ the restriction of $\boldsymbol{L}$ to the ℓth partition, then

$$\mathrm{HSIC}_v\left(\boldsymbol{f}_k^{(\ell)}, \boldsymbol{y}^{(\ell)}\right) = \mathrm{tr}\left(\overline{\boldsymbol{K}}_k^{(\ell)} \overline{\boldsymbol{L}}^{(\ell)}\right) = \mathrm{vec}\left(\overline{\boldsymbol{K}}_k^{(\ell)}\right)^\top \mathrm{vec}\left(\overline{\boldsymbol{L}}^{(\ell)}\right).$$

Block HSIC Lasso is obtained by replacing the HSIC estimator HSIC_v with the block HSIC estimator HSIC_b in Equation (3):

$$\max_{\boldsymbol{\alpha} \geq 0} \sum_{k=1}^{d} \alpha_k \mathrm{HSIC}_b(\boldsymbol{f}_k, \boldsymbol{y}) - \frac{1}{2} \sum_{k,k'=1}^{d} \alpha_k \alpha_{k'} \mathrm{HSIC}_b(\boldsymbol{f}_k, \boldsymbol{f}_{k'}) - \lambda ||\boldsymbol{\alpha}||_1.$$

Using the vectorization operator, the block estimator is written as

$$\mathrm{HSIC}_b(\boldsymbol{f}_k, \boldsymbol{f}_{k'}) = \boldsymbol{u}_k^\top \boldsymbol{u}_{k'}, \quad \mathrm{HSIC}_b(\boldsymbol{f}_k, \boldsymbol{y}) = \boldsymbol{u}_k^\top \boldsymbol{v},$$

where

$$\boldsymbol{u}_k = \sqrt{\frac{B}{n}} \left[\mathrm{vec}\left(\overline{\boldsymbol{K}}_k^{(1)}\right)^\top, \ldots, \mathrm{vec}\left(\overline{\boldsymbol{K}}_k^{(n/B)}\right)^\top\right]^\top \in \mathbb{R}^{nB},$$
$$\boldsymbol{v} = \sqrt{\frac{B}{n}} \left[\mathrm{vec}\left(\overline{\boldsymbol{L}}^{(1)}\right)^\top, \ldots, \mathrm{vec}\left(\overline{\boldsymbol{L}}^{(n/B)}\right)^\top\right]^\top \in \mathbb{R}^{nB}.$$

Hence, block HSIC Lasso can also be written as

$$\min_{\boldsymbol{\alpha} \geq 0} ||\boldsymbol{v} - \boldsymbol{U}^\top \boldsymbol{\alpha}||_2^2 + \lambda ||\boldsymbol{\alpha}||_1,$$

where $\boldsymbol{U} = [\boldsymbol{u}_1, \ldots, \boldsymbol{u}_d] \in \mathbb{R}^{nB \times d}$.

Since the objective function of block HSIC Lasso is convex, we can obtain a globally optimal solution. As with HSIC Lasso, we can solve block HSIC Lasso using an off-the-shelf Lasso solver. Here, we use the non-negative least angle regression-LASSO, or LARS-LASSO (Efron *et al.*, 2004), to solve the problem in a greedy manner. Rather than setting the hyperparameter λ, for example by cross-validation, which would be computationally intensive, this allows us to use a predefined number of features to select.

The required memory space for block HSIC Lasso is $O(dnB)$, which compares favorably to vanilla HSIC Lasso's $O(dn^2)$; as the block size $B \ll n$, the memory space is dramatically reduced. However, the computational cost of the proposed method is still large when both d and n are large. Thus, we implemented the proposed algorithm using multiprocessing by parallelizing the computation of $\overline{\boldsymbol{K}}_k^{(\ell)}$. Thanks to the combination of block HSIC Lasso and the multiprocessing implementation, we can efficiently find solutions on large datasets with a reasonably sized server.

2.4 Improving selection stability using bagging

Since we need to compute block HSIC of the paired data $\{\{(\boldsymbol{x}_i^\ell, \boldsymbol{y}_i^\ell)\}_{i=1}^B\}_{\ell=1}^{n/B}$ with a fixed partition, the performance can be highly affected by the partition. Thus, we propose to use a bagging version of the block HSIC estimator. Given M random permutations of the n samples, we define *bagging block HSIC* as

$$\mathrm{HSIC}_{bb}(\boldsymbol{f}_k, \boldsymbol{y}) = \frac{1}{M} \sum_{m=1}^{M} \frac{B}{n} \sum_{\ell=1}^{n/B} \mathrm{HSIC}_v\left(\boldsymbol{f}_k^{(\ell,m)}, \boldsymbol{y}^{(\ell,m)}\right) = \overline{\boldsymbol{u}}_k^\top \overline{\boldsymbol{v}},$$

where $\boldsymbol{f}_k^{(\ell,m)}$ is the kth feature vector restricted to the ℓth block as defined by the mth permutation,

$$\overline{\boldsymbol{u}}_k = \sqrt{\frac{1}{M}} \left[\boldsymbol{u}_k^{(1)^\top}, \ldots, \boldsymbol{u}_k^{(M)^\top}\right]^\top \in \mathbb{R}^{nBM},$$

$$\overline{\boldsymbol{v}} = \sqrt{\frac{1}{M}} \left[\boldsymbol{v}^{(1)^\top}, \ldots, \boldsymbol{v}^{(M)^\top}\right]^\top \in \mathbb{R}^{nBM},$$

and $\boldsymbol{u}_k^{(m)} \in \mathbb{R}^{nB}$ and $\boldsymbol{v}_k^{(m)} \in \mathbb{R}^{nB}$ are the vectors of the mth block HSIC Lasso, respectively.

Hence, bagging block HSIC Lasso can be written as

$$\min_{\boldsymbol{\alpha} \geq 0} ||\overline{\boldsymbol{v}} - \overline{\boldsymbol{U}}^\top \boldsymbol{\alpha}||_2^2 + \lambda ||\boldsymbol{\alpha}||_1,$$

where $\overline{\boldsymbol{U}} = [\overline{\boldsymbol{u}}_1, \ldots, \overline{\boldsymbol{u}}_d] \in \mathbb{R}^{nBM \times d}$.

We consider the bagging part to be an integral part of the block HSIC Lasso algorithm. That is why, in this text, every time we mention 'block HSIC Lasso', we refer to bagging block HSIC Lasso.

Note that the memory space $O(dnBM)$ required for $B = 60$ and $M = 1$ is equivalent to $B = 30$ and $M = 2$. Empirically, we found that they were providing equivalent feature selection accuracy (Section 4.4).

2.5 Adjusting for covariates

Data analysis tasks in bioinformatics can often be confounded by technical (e.g. batch) or biological variables (e.g. age), which might mask the relevant variables. To adjust for their effect, we consider the following variant of the block HSIC Lasso:

$$\min_{\boldsymbol{\alpha}\geq 0} ||\boldsymbol{\nu} - \boldsymbol{U}^\top \boldsymbol{\alpha} - \beta \boldsymbol{z}||_2^2 + \lambda ||\boldsymbol{\alpha}||_1,$$

where $\beta \geq 0$ is a tuning parameter and

$$\boldsymbol{z} = \sqrt{\frac{B}{n}}\left[\mathrm{vec}\left(\overline{\boldsymbol{K}}_{\mathrm{cov}}^{(1)}\right)^\top, \ldots, \mathrm{vec}\left(\overline{\boldsymbol{K}}_{\mathrm{cov}}^{(n/B)}\right)^\top\right]^\top \in \mathbb{R}^{nB}$$

contains the covariate information. $\boldsymbol{K}_{\mathrm{cov}}$ is the Gram matrix computed from the covariate input matrix $\boldsymbol{X}_{\mathrm{cov}}$. Since for most purposes in bioinformatics we want to remove all information from the covariates, we set β to

$$\hat{\beta} = \frac{\mathrm{HSIC}_b(\boldsymbol{y}, \boldsymbol{X}_{\mathrm{cov}})}{\mathrm{HSIC}_b(\boldsymbol{X}_{\mathrm{cov}}, \boldsymbol{X}_{\mathrm{cov}})} = \mathrm{HSIC}_b(\boldsymbol{y}, \boldsymbol{X}_{\mathrm{cov}}),$$

which is the solution of $\min_\beta ||\boldsymbol{\nu} - \beta \boldsymbol{z}||_2^2$. Here, we used the property $\mathrm{HSIC}_b(\boldsymbol{X}_{\mathrm{cov}}, \boldsymbol{X}_{\mathrm{cov}}) = 1$.

3 Experimental setup

3.1 Feature selection methods

HSIC Lasso and block HSIC Lasso: We used HSIC Lasso and block HSIC Lasso implemented in the Python 2/3 package *pyHSICLasso*. In block HSIC Lasso, M was set to 3 in all experimental settings; the block size B was set on an experiment-dependent fashion. In all the experiments, when we wanted to select k features, HSIC Lasso versions were required to first retrieve 50 features, and then the top k features were selected as the solution.

In this article, we use the following kernels:

- The RBF Gaussian kernel for pairs of continuous variables, of continuous outcomes, or one of each, and for pairs of a continuous variable and categorical outcome:

$$K : x_i^{(k)}, x_j^{(k)} \mapsto \exp\left(-\frac{||x_i^{(k)} - x_j^{(k)}||_2^2}{2\sigma^2}\right),$$

where $\sigma^2 > 0$ is the bandwidth of the kernel;

- The normalized Delta kernel for categorical variables (or outcomes):

$$L : y_i, y_j \mapsto \begin{cases} \frac{1}{n_c} & \text{if } y_i = y_j = c \\ 0 & \text{otherwise,} \end{cases}$$

where n_c is the number of samples in class c.

mRMR: mRMR selects features that are highly associated with the outcome and are non-redundant (Peng *et al.*, 2005). To that end, it uses mutual information between different variables and between the outcome and the variables.

We used a C++ implementation of mRMR (Peng, 2005). The maximum number of samples and the maximum number of features were set to the actual number of samples and features in the data. In regression problems, discretization was set to binarization.

LARS: LARS is a forward stage-wise feature selector (Efron *et al.*, 2004). It is an efficient way of solving the same problem as Lasso. We used the SPAMS implementation of LARS (Mairal *et al.*, 2010), with the default parameters. Note that this is not the implementation of LARS that we use in (block) HSIC Lasso, which is the non-negative LARS solver implemented in *pyHSICLasso*.

3.2 Evaluation of the selected features

Selection accuracy on simulated data: We simulated high-dimensional data where only a few variables were truly related to the outcome. We used these datasets to evaluate the ability of the tested algorithms to find the true causal variables, instead of others, likely spuriously correlated to the outcome. To that end, we requested each algorithm to retrieve the known number of causal features. Then, we studied how many of them were actually causal.

Classification with a random forest: In classification datasets, we evaluated the amount of information retained in the features selected by a given method by evaluating the performance of a random forest classifier based only on those features. We used random forests because of their ability to handle non-linearities. We split the data between a training and a test set, and selected features on the training set only. We estimated the best parameters by cross-validation on the training set: the number of trees (200, 500), the maximum depth of the threes (4, 6, 8), the number of features to consider ($\sqrt{d}$, $\log_2 d$), and the criterion to measure the quality of the chosen features (Gini impurity, information gain). Then, we trained a model with those parameters on the training set and made predictions on a separate testing set to estimate prediction accuracy.

3.3 Datasets

We evaluated the performance of the different algorithms on synthetic data and four types of real-world high-dimensional datasets (Table 1). In our experiments on real-world datasets, we restricted ourselves to classification problems. All discussed methods can however handle regression problems (continuous-valued outcomes) as well, as we show on synthetic data.

Synthetic data: We simulated random matrices of features $X \sim \mathcal{N}(0, 1)$. A number of variables were selected as related to the phenotype, and functions that are non-linear in the data range were selected (cosine, sine and square) and combined additively to create the outcome vector $\boldsymbol{y}$.

Images: Facial recognition is a classification problem classically used to evaluate non-linear feature selection methods, as only a few of all features are expected to be relevant for the outcome, in a non-linear fashion. We used four face image datasets from the Arizona State University feature selection repository (Li *et al.*, 2018)): pixraw10P, warpAR10P, orlraws10P and warpPIE10P.

Gene expression microarrays: We analyzed four gene expression microarray datasets from Arizona State University feature selection repository (Li *et al.*, 2018). The phenotypes were subtypes of B-cell chronic lymphocytic leukemia (CLL-SUB-111), hepatocyte phenotypes under different diets (TOX-171), glioma (GLIOMA) and smoking-driven carcinogenesis (SMK-CAN-187).

Single-cell RNA-seq: Single-cell RNA-seq (scRNA-seq) measures gene expression at cell resolution, allowing to characterize the diversity in a tissue. We performed feature selection on the three most popular datasets in the Broad Institute's Single Cell Portal, related to

Table 1. Summary description of benchmark datasets

Type	Dataset	Features (d)	Samples (n)	Classes
Image	AR10P	2400	130	10
	PIE10P	2400	210	10
	PIX10P	10 000	100	10
	ORL10P	10 000	100	10
Microarray	CLL-SUB-111	11 340	111	3
	GLIOMA	4434	50	4
	SMK-CAN-187	19 993	187	2
	TOX-171	5748	171	4
	Haber *et al.* (2017)	15 972	7216	19
scRNA-seq	Habib *et al.* (2016)	25 393	13 302	8
	Villani *et al.* (2017)	23 395	1140	10
GWA data	RA versus controls	352 773	3451	2
	T1D versus controls	352 853	3443	2
	T2D versus controls	353 046	3456	2

mouse small intestinal epithelium (Haber *et al.*, 2017), mouse hippocampus (Habib *et al.*, 2016) and human blood cells (Villani *et al.*, 2017). Missing gene expressions were imputed with MAGIC (van Dijk *et al.*, 2018).

GWA datasets: We studied the WTCCC1 datasets (Burton *et al.*, 2007) for rheumatoid arthritis (RA), type 1 diabetes (T1D) and type 2 diabetes (T2D) (2000 samples each), using the 1958BC cohort as control (1504 samples). Affymetrix 500K was used for genotyping. We removed the samples and the SNPs that did not pass WTCCC's quality controls, as well as SNPs in sex chromosomes and those that were not genotyped in both cases and controls. Missing genotypes were imputed with CHIAMO. Lastly, individuals with >10% genotype missing rate, and SNPs with >10% genotype missing rate, MAF < 5% or not in HWE (P-value < 0.001) were removed. The remaining missing genotypes were replaced by the major allele in homozygosis.

Preprocessing: Images, microarrays and scRNA-seq data were normalized feature-wise by subtracting the mean and dividing by the standard deviation. GWAS data did not undergo any normalization.

3.4 Computational resources

We ran the experiments on synthetic data, images, microarrays and scRNA-seq on CentOS 7 machines with Intel Xeon 2.6 GHz and 50 GB RAM memory. For the GWA datasets experiments, we used a CentOS 7 server with 96 core Intel Xeon 2.2 GHz and 1 TB RAM memory.

3.5 Software availability and reproducibility

Block HSIC Lasso was implemented in the Python 2/3 package *pyHSICLasso*. The source code is available on GitHub (https://github.com/riken-aip/pyHSICLasso), and the package can be installed from PyPI (https://pypi.org/project/pyHSICLasso). All analyses in this article and the scripts needed to reproduce them are also available on GitHub (https://github.com/hclimente/nori).

4 Results

4.1 Block HSIC Lasso performance is comparable to state of the art

At first, we worked on synthetic, non-linear data (Section 3.2). We generated synthetic data with combinations of the following experimental parameters: $n = \{100, 1000, 10\,000\}$ samples; $d = \{100, 2500, 5000, 10\,000\}$ features; and 5, 10 and 20 causal features,

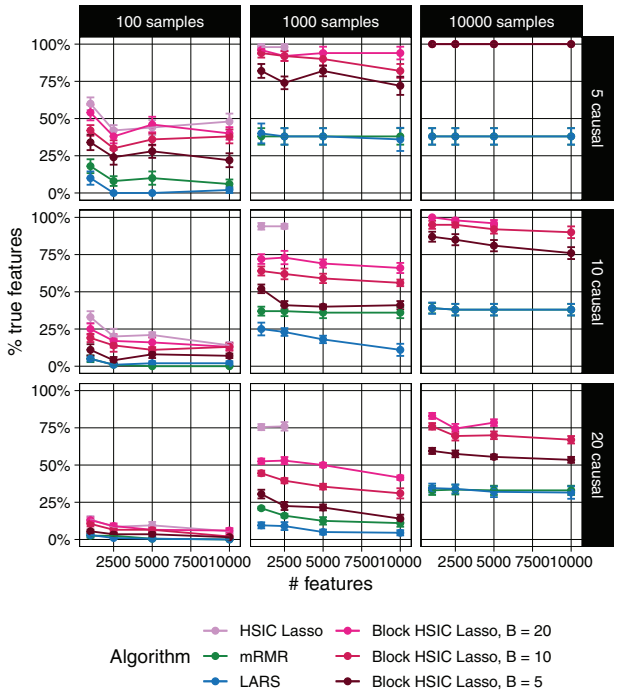


Fig. 1. Percentage of true causal features extracted by different feature selectors. Each data point represents the mean over 10 replicates, and the error bars represent the standard error of the mean. Lines are discontinued when the algorithm required more memory than the provided (50 GB). Note that in some conditions mRMR's line cannot be seen due to the overlap with LARS

that is, features truly related to the outcome. We evaluated the performance of different feature selectors at retrieving the causal features. These conditions range from an ideal setting, where the number of features is smaller than the number of samples, to an ultra-high dimensional scenario, where spurious dependencies among variables, and between those and the outcome are bound to occur.

Each of the methods was required to select as many features as the number of true causal features. In Figure 1, we show the proportion of the causal features retrieved by each method. The different versions of HSIC Lasso outperform the other approaches in virtually all settings. Block HSIC Lasso with decreasing block sizes results in worse performances. As expected, vanilla HSIC Lasso outperforms the block versions in accuracy, but increases memory use. Crucially, block HSIC Lasso on a larger number of samples performs better than vanilla HSIC Lasso on fewer samples. Hence, when the number of samples is in the thousands, it is better to apply block HSIC Lasso on the whole dataset, than to apply vanilla HSIC Lasso on a subsample.

We wanted to test these conclusions using a non-linear, real-world dataset. We selected four image-based face recognition tasks (Section 3.3). In this case, we selected different numbers of features (10, 20, 30, 40 and 50). Then, we trained random forest classifiers on these subsets of the features, and compared the accuracy of the different classifiers on a test set (Supplementary Fig. S1). Block HSIC Lasso displayed a performance comparable to vanilla HSIC Lasso, and comparable or superior to the other methods. This is remarkable, since it shows that, in many practical cases, block HSIC Lasso does not need more samples to achieve vanilla HSIC Lasso performance.

4.2 Adjusting by covariates improves feature selection

To evaluate the impact of covariate adjustment, we worked on a synthetic dataset (Section 3.2) with the following experimental

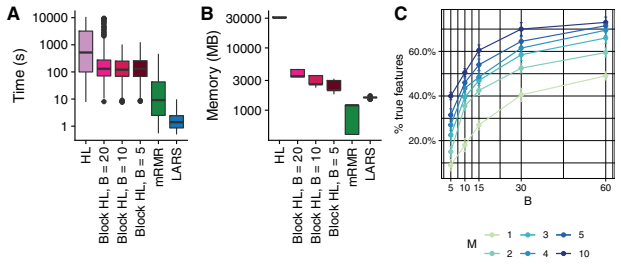


Fig. 2. Computational resources used by the different methods. (**A**) Time elapsed in a multiprocess setting. (**B**) Memory usage in a single-core setting. (**C**) Number of correct features retrieved on synthetic data ($n = 1000$, $d = 2500$, 20 causal features) by block HSIC Lasso at different block sizes B and number of permutations M

parameters: $n = 1\,000$; $d = \{100, 2500, 5000, 10\,000\}$ features; seven causal features. Two covariates were generated by taking two causal features and adding Gaussian noise (mean = 0; standard deviation = 0.5). In the experiment shown in Supplementary Figure S2, we tested the ability of (block) HSIC Lasso to retrieve exclusively the remaining five causal features adjusting for the covariates. We observe that block HSIC Lasso is able to find more relevant features when it adjusts for known covariates.

4.3 Block HSIC Lasso is computationally efficient

In our experiments on synthetic data, vanilla HSIC Lasso runs into memory issues already with 1000 samples (Fig. 1). This experiment shows how block HSIC Lasso keeps the good properties of HSIC Lasso, while extending it to more experimental settings. Block HSIC Lasso with $B = 20$ reaches the memory limit only at 10 000 samples, which is already sufficient for most common bioinformatics applications. If larger datasets need to be handled, it can be done by using smaller block sizes or a larger computer cluster.

We next quantified the computational efficiency improvement the block HSIC estimator brings. We compared the runtime and the peak memory usage in the highest dimensional setting where all methods could run ($n = 1000$, $d = 2500$, 20 causal features) (Fig. 2). We observe how, as expected, block HSIC Lasso requires an order of magnitude less memory than vanilla HSIC Lasso. Block versions also run notoriously faster, thanks to the lower number of operations and the parallelization. mRMR is 10 times faster than block HSIC Lasso, at the expense of a clearly lower accuracy. However, a fraction of this gap is likely due to mRMR having been implemented in C++, while HSIC Lasso is written in Python. In this regard, there is potential for other faster implementations of (block) HSIC Lasso.

4.4 Block HSIC Lasso improves with more permutations

We were interested in the trade-off between the block size and the number of permutations, which affect both the computation time and accuracy of the result. We tested the performance of block HSIC Lasso with $B = \{5, 10, 15, 30, 60\}$ and $M = \{1, 2, 3, 5\}$ in datasets of $n = 1000$, $d = 2500$ and 20 causal features. As expected, causal feature recovery increases with M and B (Fig. 2C), as the HSIC estimator approaches its true value.

The memory usage $O(dnBM)$ of several of the conditions was the same, e.g. $B = 10$, $M = 3$ and $B = 30$, $M = 1$. Such conditions are indistinct from the points of view of both accuracy, and memory requirements. In practice, we found no major differences in runtime between different combinations of B and M. Hence, a reasonable

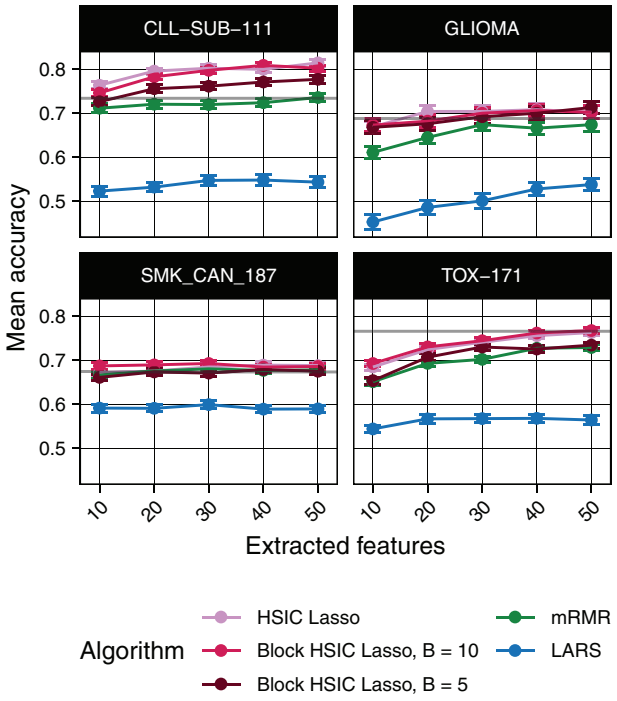


Fig. 3. Random forest classification accuracy of microarray gene expression samples after feature extraction by the different methods. The gray line represents the mean accuracy of 10 classifiers trained on all the dataset

strategy is to fix B to a given size, and tune the M to the available memory/desired amount of information. This strategy, however, should be adapted to fit properties of the data. More specifically, GWAS data are notably sparse, and as result a small block size would result in many blocks consisting entirely of zeros, which would hence be uninformative. In such cases, it might be interesting to prioritize larger block sizes, and fewer permutations.

4.5 Block HSIC Lasso finds more relevant features

We tested the dimensionality reduction potential of different feature selectors. We selected a variable number of features from different multi-class biological datasets, then used a random forest classifier to retrieve the original classes (Section 3.2). The underlying assumption is that only selected features which are biologically relevant will be useful to classify unseen data. To that end, we evaluated the classification ability of the biomarkers selected in four gene expression microarrays (Fig. 3) and three scRNA-seq experiments (Supplementary Fig. S3). Unsurprisingly, we observe that non-linear feature selectors perform notably better than linear selectors. Of the non-linear methods, in virtually all cases block HSIC Lasso showed similar or superior performance to mRMR. Interestingly, as little as 20 selected genes retain enough information to achieve a plateau accuracy in most experiments.

Surveying $10^5 - 10^6$ SNPs in $10^3 - 10^4$ patients, genome-wide association (GWA) datasets are among the most high dimensional in biology, an unbalance which worsens the statistical and computational challenges. We performed the same evaluation on three WTCCC1 phenotypes (Section 3.3). As a baseline, we also computed the accuracy of a classifier trained on all the SNPs (Supplementary Table S1). We observe that a feature selection prior step is not always favorable: LARS worsens the classification accuracy by 5–10%. On top of that, LARS could not select any SNP in 2

out of the 15 experimental settings. On the other hand, non-linear methods improve the classification accuracy by 10%, with mRMR and block HSIC Lasso achieving similar accuracies. In fact, those two selected the same 14 out of 30 SNPs when we selected 10 SNPs in each the three datasets with each method (Supplementary Fig. 5).

4.6 Block HSIC Lasso is robust to ill-conditioned problems

Single-cell RNA-seq datasets differ from microarray datasets in two ways. First, the number of features is larger, equaling the number of genes in the annotation (> 20 000). Second, the expression matrices are very sparse, due to biological variability (genes actually not expressed in a particular cell) and dropouts (genes whose expression levels have not been measured, usually because they are low, i.e. technical zeroes). In summary, the problem is severely ill conditioned, and the feature selectors need to deal with this issue. We observed that block HSIC Lasso runs reliably when faced with variations in the data, even on ill-conditioned problems like scRNA-seq. In the different scRNA-seq datasets, LARS was unable to select the requested number of biomarkers in any of the cases, returning always a lower number (Supplementary Fig. S4). mRMR did in all cases. However, the implementation of mRMR that we used crashed while selecting features on the full Villani *et al.* (2017) dataset.

4.7 Block HSIC Lasso for biomarker discovery

4.7.1 New biomarkers in mouse hippocampus scRNA-seq

To study the potential of block HSIC lasso for biomarker discovery in scRNA-seq data, we focused on the mouse hippocampus dataset from Habib *et al.* (2016), as a list of 1669 known biomarkers for the different cell types is also provided by the authors. We requested block HSIC Lasso, mRMR and LARS to select the best 20 genes for classification of 8 cell types (Supplementary Table S2). The cell types were four different hippocampal anatomical subregions (DG, CA1, CA2 and CA3), glial cells, ependymal, GABAergic and unidentified cells.

The overlap between the genes selected by different algorithms was empty. We compared the selected genes to the known biomarkers. Out of the 20 genes selected by mRMR, 14 are known biomarkers, a number that goes down to 0 in the case of block HSIC Lasso (Supplementary Fig. S4A). Hence, these 20 genes, which are sufficient for accurately separating the cell types, are potential novel biomarkers. However, we have no reason to believe that HSIC Lasso generally has a higher tendency to return novel genes than other approaches; we merely emphasize that it suggests alternative, statistically plausible biological hypotheses that can be worth investigating.

We therefore evaluated whether the novel genes found by block HSIC Lasso participate in biological functions known to be different between the cell classes. To obtain the biological processes responsible for the differences between classes, we mapped the known biomarkers to GO Biological process categories using the GO2MSIG database (Powell, 2014). Then we repeated the process using the genes selected by the different feature selectors, and compared the overlap between them. The overlap between the different techniques increases when we consider the biological process instead of specific genes (Supplementary Fig. S4B). Specifically, one biological process term that is shared between mRMR and block HSIC Lasso, 'Adult behavior' (associated to *Sez6* and *Klhl1*, respectively), is clearly related to hippocampus function. This reinforces the notion that the selected genes are relevant for the studied phenotypes.

Then we focused on potential biomarkers and biologically interesting molecules among those genes selected by block HSIC Lasso. As it is designed specifically to select non-redundant features, often-used GO enrichment analyses are not meaningful: we expect genes belonging to the same GO annotation to be correlated, and HSIC lasso should not accumulate them. Among the top five genes, two mapped to a biological processes known to be involved: the aforementioned *Klhl1* and *Pou3f1* (related to Schwann cell development). *Klhl1* is a gene expressed in seven of the studied cell types and which has been related to neuron development in the past (He *et al.*, 2006). *Pouf1* is a transcription factor which in the past has been linked to myelination, and neurological damage in its absence (Jaegle *et al.*, 1996). The only gene among the top five that was expressed exclusively in one of the clusters is the micro RNA *Mir670*, expressed exclusively in CA1. According to miRDB (Wong and Wang, 2015), *Mir670* top predicted target of its 3' arm is *Pcnt*, which is involved in neocortex development.

4.7.2 GWAS without assumptions on genetic architecture

We applied block HSIC Lasso ($B = 60$, $M = 1$) to three GWA datasets (Section 3.3). It is typical in GWAS to assume a genetic model before performing statistical testing of associations between SNPs and the phenotype. Two common, well-known models are the additive model—the minor allele in homozygosity has twice the effect as the minor allele in heterozygosity—and the dominant model—any number of copies of the minor allele have a phenotypic outcome. Using non-linear models such as block HSIC Lasso to explore the relationship between SNPs and outcome is attractive since no assumptions are needed on how individual SNPs affect the trait. The only assumption is that the phenotype can be explained by a combination of main effects, as block HSIC Lasso does not account for epistasis. On top of that, by penalizing the selection of redundant features, block HSIC Lasso avoids selecting multiple SNPs in high linkage disequilibrium.

In our experiments, we selected 10 SNPs with block HSIC Lasso for each of the three phenotypes. These are the SNPs that best balance high relatedness to the phenotype and not giving redundant information, be it through linkage disequilibrium or through an underlying shared biological mechanism. We compared these SNPs to those selected by the univariate statistical tests implemented in PLINK 1.9 (Chang *et al.*, 2015). Some of them explicitly account for non-linearity by considering dominant and recessive models of inheritance. The number of SNPs that were positive in at least one test were disparate between the studied phenotypes: all 10 in T1D, 5 in RA, and only 2 in T2D.

Specifically, we compared the genome-wide genotypic *P*-values to the SNPs selected by block HSIC Lasso (Fig. 4). In T1D, block HSIC Lasso selected SNPs among those with the most extreme p-values. However, not being constrained by a conservative *P*-value threshold, block HSIC Lasso selects five and eight SNPs in RA and T2D, respectively, with non-Bonferroni significant *P*-values when they improve classification accuracy Interestingly, one of these SNPs can be physically mapped to PFKM (Keildson *et al.*, 2014), a gene previously identified in genome-wide studies of T2D. The selected SNPs are scattered all across the genome, displaying the lack of redundancy between them. This strategy gives a more representative set of SNPs than other approaches common in bioinformatics, like selecting the smallest 10 *P*-values.

5 Discussion

In this work, we presented block HSIC Lasso, a non-linear feature selector. Block HSIC Lasso retains the properties of HSIC Lasso

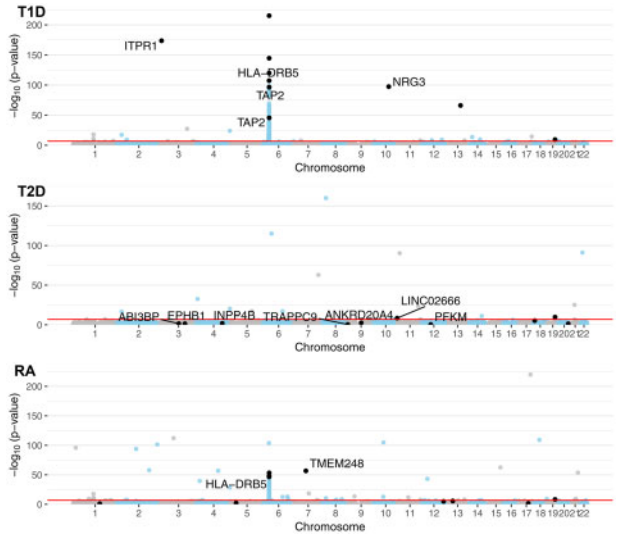


Fig. 4. Manhattan plot of the GWA datasets using *P*-values from the genotypic test. A constant of 10^{-220} was added to all *P*-values to allow plotting *P*-values of 0. SNPs in black are the SNPs selected by block HSIC Lasso ($B = 20$), 10 per phenotype. When SNPs are located within the boundaries of a gene (±50 kb), the gene name is indicated. The red line represents the Bonferroni threshold with $\alpha = 0.05$

while extending its applicability to larger datasets. Among the attractive properties of block HSIC Lasso we find, first, its ability to handle both linear and non-linear relationships between the variables and the outcome. Second, block HSIC Lasso has a convex formulation, ensuring that a global solution exists, and that it is accessible. Third, the HSIC score can be accurately estimated, as opposed to other measures of non-linearity like mutual information. Fourth, block HSIC Lasso's memory consumption scales linearly with respect to both the number of features and the number of samples. In addition, block HSIC Lasso can be easily adapted to different problems via different kernel functions that better capture similarities in new datasets. Lastly, block HSIC Lasso can be adjusted for covariates known to affect the outcome, which helps removing confounding effects from the analysis. Due to all these properties, we show how block HSIC Lasso outperforms all other algorithms in the tested conditions.

Block HSIC Lasso can be applied to different kinds of datasets. As other non-linear methods, block HSIC Lasso is particularly useful when we do not want to make strong assumptions about how the causal variables relate to the outcome. Thanks to the advantages mentioned above, HSIC Lasso and block HSIC Lasso tend to outperform other state-of-the-art approaches in terms of both causal features retrieval in simulated data, and classification accuracy on real-world datasets.

Whereas the Lasso is limited to selecting at most as many features as there are available samples (n), for block HSIC Lasso the limitation is nBM. Hence, even if the number of samples is small, block HSIC Lasso can be used to select a larger number of features. If nBM is still limiting, one could replace the ℓ_1 regularization with an elastic-net regularization. However, in most cases, we expect block HSIC Lasso to be used to select a small number of features.

Regarding its potential in bioinformatics, we applied block HSIC Lasso to images, microarrays, single-cell RNA-seq and GWAS. The two latter involve thousands of samples, making it unfeasible to run vanilla HSIC Lasso on a regular server because of its memory requirements. The selected biomarkers are biologically plausible, agree with the outcome of other methods and provide a good classification accuracy when used to train a classifier. Such a ranking is useful, for instance, when selecting SNPs or genes to assay in *in vitro* experiments.

Block HSIC Lasso's main drawback is the memory complexity, markedly lower than in vanilla HSIC Lasso but still $O(dnB)$. Memory issues might appear in low-memory servers in cases with a large number of samples n, of features d, or both. However, through our work on GWA datasets, the largest type of dataset in bioinformatics, we show that working on these datasets is feasible. Another drawback, which block HSIC Lasso shares with the other non-linear methods, is their black box nature. Block HSIC Lasso looks for biomarkers which, after an unknown, non-linear transformation, would allow a linear separation between the samples. Unfortunately, we cannot access this transformed space and explore it, which makes the results hard to interpret.

Funding

Computational resources and support were provided by RIKEN AIP. H.C-G. was funded by the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie [666003]. S.K. was supported by the Academy of Finland (292334, 319264). M.Y. was supported by the JST PRESTO program JPMJPR165A and partly supported by MEXT KAKENHI 16H06299 and the RIKEN engineering network funding.

Conflict of Interest: none declared.

References

Burton,P.R. *et al.* (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, **447**, 661–678.

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

Clarke,R. *et al.* (2008) The properties of high-dimensional data spaces: implications for exploring gene and protein expression data. *Nat. Rev. Cancer*, **8**, 37–49.

Cover,T.M. and Thomas,J.A. (2006) *Elements of Information Theory*, 2nd edn. John Wiley & Sons, Inc., Hoboken, NJ.

Ding,C. and Peng,H. (2005) Minimum redundancy feature selection from microarray gene expression data. *J. Bioinform. Comput. Biol.*, **3**, 185–205.

Efron,B. *et al.* (2004) Least angle regression. *Ann. Statist.*, **32**, 407–499.

Fujishige,S. (2005) *Submodular Functions and Optimization*, Vol. **58**. Elsevier, Boston.

Gretton,A. *et al.* (2005) Measuring statistical dependence with Hilbert–Schmidt norms. In: *International Conference on Algorithmic Learning Theory (ALT), Singapore*, pp. 63–77.

Haber,A.L. *et al.* (2017) A single-cell survey of the small intestinal epithelium. *Nature*, **551**, 333–339.

Habib,N. *et al.* (2016) Div-Seq: single-nucleus RNA-Seq reveals dynamics of rare adult newborn neurons. *Science*, **353**, 925–928.

He,Y. *et al.* (2006) Targeted deletion of a single Sca8 ataxia locus allele in mice causes abnormal gait, progressive loss of motor coordination, and Purkinje cell dendritic deficits. *J. Neurosci.*, **26**, 9975–9982.

Jaegle,M. *et al.* (1996) The POU factor Oct-6 and Schwann cell differentiation. *Science*, **273**, 507–510.

Johnstone,I.M. and Titterington,D.M. (2009) Statistical challenges of high-dimensional data. *Philos. Trans. Series A Math. Phys. Eng. Sci.*, **367**, 4237–4253.

Keildson,S. *et al.* (2014) Expression of phosphofructokinase in skeletal muscle is influenced by genetic variation and associated with insulin sensitivity. *Diabetes*, **63**, 1154–1165.

Li,J. *et al.* (2018) Feature selection: a data perspective. *ACM Comp. Surveys*, **50**, 94.

Mairal,J. *et al.* (2010) Online learning for matrix factorization and sparse coding. *J. Mach. Learn. Res.*, **11**, 19–60.

Peng,H. (2005) mrmr. http://home.penglab.com/proj/mRMR/ (15 June 2018, date last accessed).

Peng,H. *et al*. (2005) Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans. Pattern Anal. Mach. Intell*., **27**, 1226–1237.

Powell,J.A.C. (2014) GO2MSIG, an automated GO based multi-species gene set generator for gene set enrichment analysis. *BMC Bioinformatics*, **15**, 146.

Ravikumar,P. *et al*. (2009) Sparse additive models. *J. R. Statist. Soc. Series B Statist. Methodol*., **71**, 1009–1030.

Schölkopf,B. and Smola,A.J. (2002) *Learning with Kernels*. MIT Press, Cambridge, MA.

Song,L. *et al*. (2012) Feature selection via dependence maximization. *J. Mach. Learn. Res*., **13**, 1393–1434.

Tibshirani,R. (1996) Regression shrinkage and selection via the Lasso. *J. R. Statist. Soc. Series B Methodol*., **58**, 267–288.

van Dijk,D. *et al*. (2018) Recovering gene interactions from single-cell data using data diffusion. *Cell*, **174**, 716–729.

Villani,A.-C. *et al*. (2017) Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors. *Science*, **356**, 925–928.

Walters-Williams,J. and Li,Y. (2009) Estimation of mutual information: a survey. In: Wen,P. *et al*. (eds), *Rough Sets and Knowledge Technology*. Springer, Berlin, Heidelberg, pp. 389–396.

Wong,N. and Wang,X. (2015) miRDB: an online resource for microRNA target prediction and functional annotations. *Nucleic Acids Res*., **43**, D146–D152.

Yamada,M. *et al*. (2014) High-dimensional feature selection by feature-wise kernelized lasso. *Neural Computation*, **26**, 185–207.

Yamada,M. *et al*. (2018) Ultra high-dimensional nonlinear feature selection for big biological data. *IEEE Trans. Knowl. Data Eng*., **30**, 1352–1365.

Zhang,Q. *et al*. (2018) Large-scale kernel methods for independence testing. *Statist. Comput*., **28**, 113–130.

Appendix B

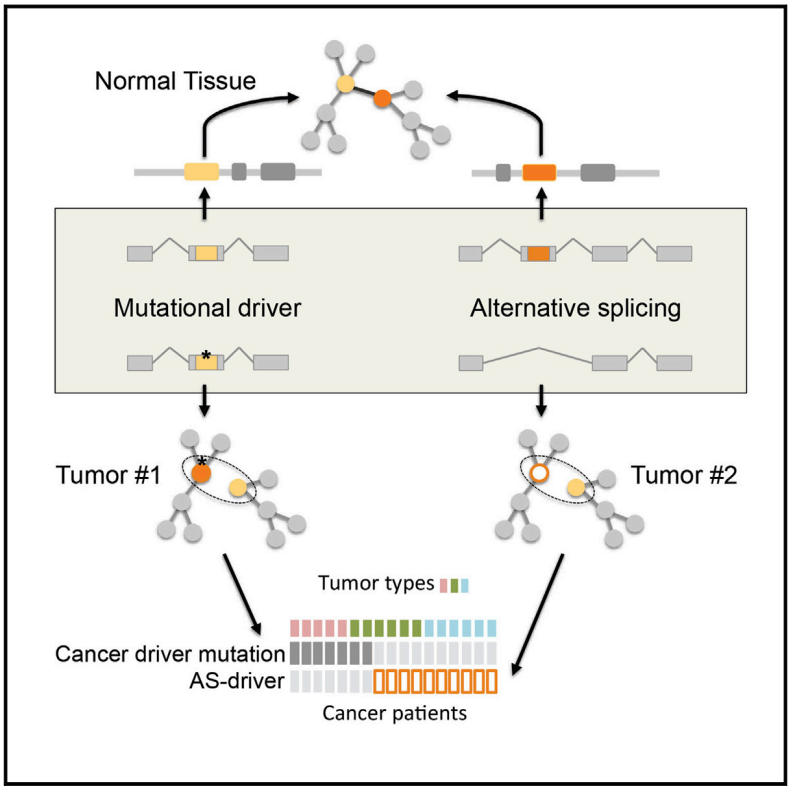
The Functional Impact of Alternative Splicing in Cancer

Cell Reports

Resource

The Functional Impact of Alternative Splicing in Cancer

Graphical Abstract



Authors

Héctor Climente-González, Eduard Porta-Pardo, Adam Godzik, Eduardo Eyras

Correspondence

eduardo.eyras@upf.edu

In Brief

Climente-González et al. show that alternative splicing (AS) changes in tumors are linked to a significant loss of functional domain families that are also frequently mutated in cancer. These domain losses happen independently of somatic mutations and lead to the remodeling of complexes and protein-protein interactions in cancer.

Highlights

- We mapped cancer-associated splicing changes (CASCs) to changes in proteins
- CASCs impact domains classically affected by somatic mutations in different genes
- CASCs remodel protein-protein interactions involving cancer drivers
- A subset of CASCs could represent independent oncogenic processes

Climente-González et al., 2017, Cell Reports *20*, 2215–2226
August 29, 2017
http://dx.doi.org/10.1016/j.celrep.2017.08.012

CellPress

Cell Reports
Resource

The Functional Impact of Alternative Splicing in Cancer

Héctor Climente-González,[1,2,3,4] Eduard Porta-Pardo,[5,6] Adam Godzik,[5] and Eduardo Eyras[1,7,8,*]
[1]Computational RNA Biology Group, Pompeu Fabra University (UPF), 08003 Barcelona, Spain
[2]MINES ParisTech, PSL-Research University, CBIO-Centre for Computational Biology, 77300 Fontainebleau, France
[3]Institut Curie, 75248 Paris Cedex, France
[4]INSERM U900, 75248 Paris Cedex, France
[5]Sanford Burnham Prebys Medical Discovery Institute, La Jolla, CA 92037, USA
[6]Barcelona Supercomputing Centre (BSC), 08034 Barcelona, Spain
[7]Catalan Institution of Research and Advanced Studies (ICREA), 08010 Barcelona, Spain
[8]Lead Contact
*Correspondence: eduardo.eyras@upf.edu
http://dx.doi.org/10.1016/j.celrep.2017.08.012

SUMMARY

Alternative splicing changes are frequently observed in cancer and are starting to be recognized as important signatures for tumor progression and therapy. However, their functional impact and relevance to tumorigenesis remain mostly unknown. We carried out a systematic analysis to characterize the potential functional consequences of alternative splicing changes in thousands of tumor samples. This analysis revealed that a subset of alternative splicing changes affect protein domain families that are frequently mutated in tumors and potentially disrupt protein-protein interactions in cancer-related pathways. Moreover, there was a negative correlation between the number of these alternative splicing changes in a sample and the number of somatic mutations in drivers. We propose that a subset of the alternative splicing changes observed in tumors may represent independent oncogenic processes that could be relevant to explain the functional transformations in cancer, and some of them could potentially be considered alternative splicing drivers (AS drivers).

INTRODUCTION

Alternative splicing provides the potential to generate diversity at RNA and protein levels from an apparently limited number of loci in the genome (Yang et al., 2016). Besides being a critical mechanism during development, cell differentiation, and regulation of cell-type-specific functions (Norris and Calarco, 2012), alternative splicing is also involved in multiple pathologies, including cancer (Chabot and Shkreta, 2016). Many alternative splicing changes recapitulate cancer-associated phenotypes by promoting angiogenesis (Vorlová et al., 2011), inducing cell proliferation (Yanagisawa et al., 2008), or avoiding apoptosis (Karni et al., 2007). Alternative splicing changes may originate from somatic mutations that disrupt splicing regulatory motifs in exons and introns (Jung et al., 2015; Supek et al., 2014), as well as through mutations or expression changes in core and auxiliary splicing factors, which impact the splicing of cancer-related genes (Bechara et al., 2013; Darman et al., 2015; Madan et al., 2015; Zong et al., 2014). Alterations in alternative splicing are also emerging as relevant targets of therapy (Lee and Abdel-Wahab, 2016). For instance, lung tumors with an exon skipping in the proto-oncogene *MET* respond to *MET*-targeted therapies despite not having any other activating alteration in this gene (Frampton et al., 2015; Paik et al., 2015). Alternative splicing is also important in drug resistance. For example, a proportion of non-responders to *BRAF*-targeted therapy express a *BRAF* isoform lacking exons 4–8, which encompass the RAS binding domain (Poulikakos et al., 2011). Similarly, alternative splicing of *CD19* in relation to the aberrant activity of the splicing factor *SRSF3* impairs immunotherapy in leukemia (Sotillo et al., 2015). Thus, specific alterations in splicing induce functional impacts that provide a selective advantage to tumor cells and could represent targets of therapy.

Despite the prevalence of alternative splicing in tumors and its relation to therapy, tumor progression, and metastasis (Lee and Abdel-Wahab, 2016; Lu et al., 2015; Trincado et al., 2016), its functional impacts have not been exhaustively described. Alternative splicing changes can confer radical functional changes (Wang et al., 2005), remodel the network of protein-protein interactions in a tissue-specific manner (Buljan et al., 2012; Ellis et al., 2012), and expand the protein interaction capabilities of genes (Yang et al., 2016). Here, we present a systematic evaluation of the potential functional impacts of alternative splicing changes in cancer samples. We described splicing changes in terms of transcript isoforms switches per tumor sample and determined the protein features and protein-protein interactions they affected. Our analysis revealed a set of isoform switches that affect protein domains from families frequently mutated in tumors, remodel the protein interaction network of cancer drivers, and tend to occur in patients with low number of mutations in cancer drivers. Furthermore, a subset of them has driver-like properties and, hence, could play a role in the neoplastic process independently of or in conjunction with mutations in cancer drivers.

RESULTS

Patient-Specific Definition of Isoform Switches across Multiple Cancer Types

To determine the potential functional impacts of alternative splicing in cancer, we analyzed the expression of human transcript isoforms in 4,542 samples from 11 cancer types from The Cancer Genome Atlas (TCGA) (Supplemental Experimental Procedures). We described splicing changes using transcript isoforms, as they represent the endpoint of transcription and splicing, and ultimately determine the functional capacity of cells. For each gene and each patient sample, we calculated the differential transcript isoform usage between the tumor and normal samples. An isoform switch was defined as a pair of transcripts, the tumor and the normal isoforms, such that the change in relative abundance in a single patient in both isoforms was higher than the observed variability across normal samples. Moreover, the involved gene must not show differential expression between tumor and normal. Additionally, we discarded switches with a significant association with stromal or immune cell content (Supplemental Experimental Procedures). The final set of switches identified and that we kept for further analysis had a mean change in relative abundance of 54% and a SD of 7%.

In all patients, we found a total of 8,122 different isoform switches in 6,442 genes that described consistent changes in the transcriptome of the tumor samples and that would not be observable by simply measuring gene expression changes (Figure 1A; Table S1). These switches occurred in 4,443 patients: each switch in 5 or more patients, with the majority (75%) occurring in 10 or more patients (Table S1). Using SUPPA (Alamancos et al., 2015), we calculated the relation with local alternative splicing events (Supplemental Experimental Procedures). From the 8,122 switches, 5,667 (69.7%) were mapped to one or more local alternative splicing events. Compared with the expected proportion of event types, we observed an enrichment of alternative 5′ss, alternative first exon and retained intron, and a depletion of alternative 3′ss, alternative last exon, mutually exclusive exons, and exon cassette (Figure S1A). Mapping the tumor isoform to either form of the event, we observed that retained intron events are predominantly retained, in agreement with previous observations (Dvinge and Bradley, 2015), whereas exon-cassette events were predominantly skipped (Figure S1B). Interestingly, 30.3% of the switches were not mapped to any event, indicating that transcripts provide a wider spectrum of RNA variation compared to local alternative splicing events.

Isoform Switches in Cancer Are Frequently Associated with Protein Feature Losses

We next studied the proteins encoded by the transcripts involved in switches. Interestingly, annotated proteins in tumor isoforms tended to be shorter than proteins in normal isoforms (Figure S1C). Moreover, whereas for most switches—6,937 (85.41%)—both transcript isoforms coded for protein, the rest had a significantly higher proportion of cases with only the normal isoform as protein-coding, 732 (9.01%) versus 231 (2.8%; binomial test p value < 2.2e−16, using 0.5 as expected frequency; Table S1), suggesting that isoform switches in tumors are associated with the loss of protein coding capacity. To determine the potential functional impact of the isoform switches, we calculated the protein features they affected. Out of the 6,937 switches with both isoforms coding for protein, 5,047 (72.7%) involved a change in at least one of the following features: Pfam domains; Prosite patterns; general disordered regions; and disordered regions with potential to mediate protein-protein interactions (Figure S1D). Interestingly, there was a significant enrichment in protein features losses when compared with a set of 100 sets of simulated switches, controlling for isoform expression (Figure 1B). This enrichment was observed despite the fact that, for simulated switches, the normal protein isoform also tended to be longer than the tumor protein isoform (Figure S1E). This indicates that isoform switches in cancer are strongly associated with the loss of protein function capabilities.

We focused on the 6,004 (73.9%) isoform switches that had a gain or loss in at least one protein feature, which we named "functional switches," as they were likely to impact gene activity (Table S1). These functional switches included 729 (8.9%) and 228 (2.8%) cases, for which only the normal or the tumor isoform, respectively, coded for a protein with one or more protein features. Interestingly, cancer drivers were enriched in functional switches (Fisher's exact test p value = 2.0e−05; odds ratio [OR] = 1.9; Figure S1F). Among the top switches in cancer drivers, we identified one in *RAC1*, which was linked before to tumor initiation and progression (Zhou et al., 2013) and which we predicted to gain an extra Ras family domain, and one in *TP53*, which we predicted to change to a non-coding isoform (Figure 1C).

To characterize how functional switches affected protein function, we calculated the enrichment in gains or losses of specific domain families with respect to their proportions in a reference proteome. To ensure that this was attributed to a switch and not to the co-occurrence of two domains, we requested a minimum of two switches in different genes affecting the domain. We detected 220 and 41 domain families exclusively lost or gained, respectively, and 13 that were both gained and lost, more frequently than expected by chance (Table S2). Domain families that were significantly lost included those involved in regulation of protein activity (Figure 1D), suggesting effects on protein-protein interactions. To further characterize these functional switches, we calculated the proportion of oncogenes or tumor suppressors that contained domain families enriched in gains or losses, compared with the reference proteome. From the 69 cancer drivers with domains enriched in gains, 58 (84%) corresponded to oncogenes (Fisher's exact test p value = 0.0066; OR = 0.4). Although tumor suppressors were not enriched in domain losses, domain families enriched in gains occurred more frequently in oncogenes than in tumor suppressors (Wilcoxon test p value = 9e−04). These results suggest a similarity between our functional isoform switches and oncogenic mechanisms in cancer.

Isoform Switches and Somatic Mutations Affect Similar Domain Families

We conducted various comparisons using our switches and *cis*-occurring mutations from whole-exome sequencing (WES) and whole-genome sequencing (WGS) data (Supplemental

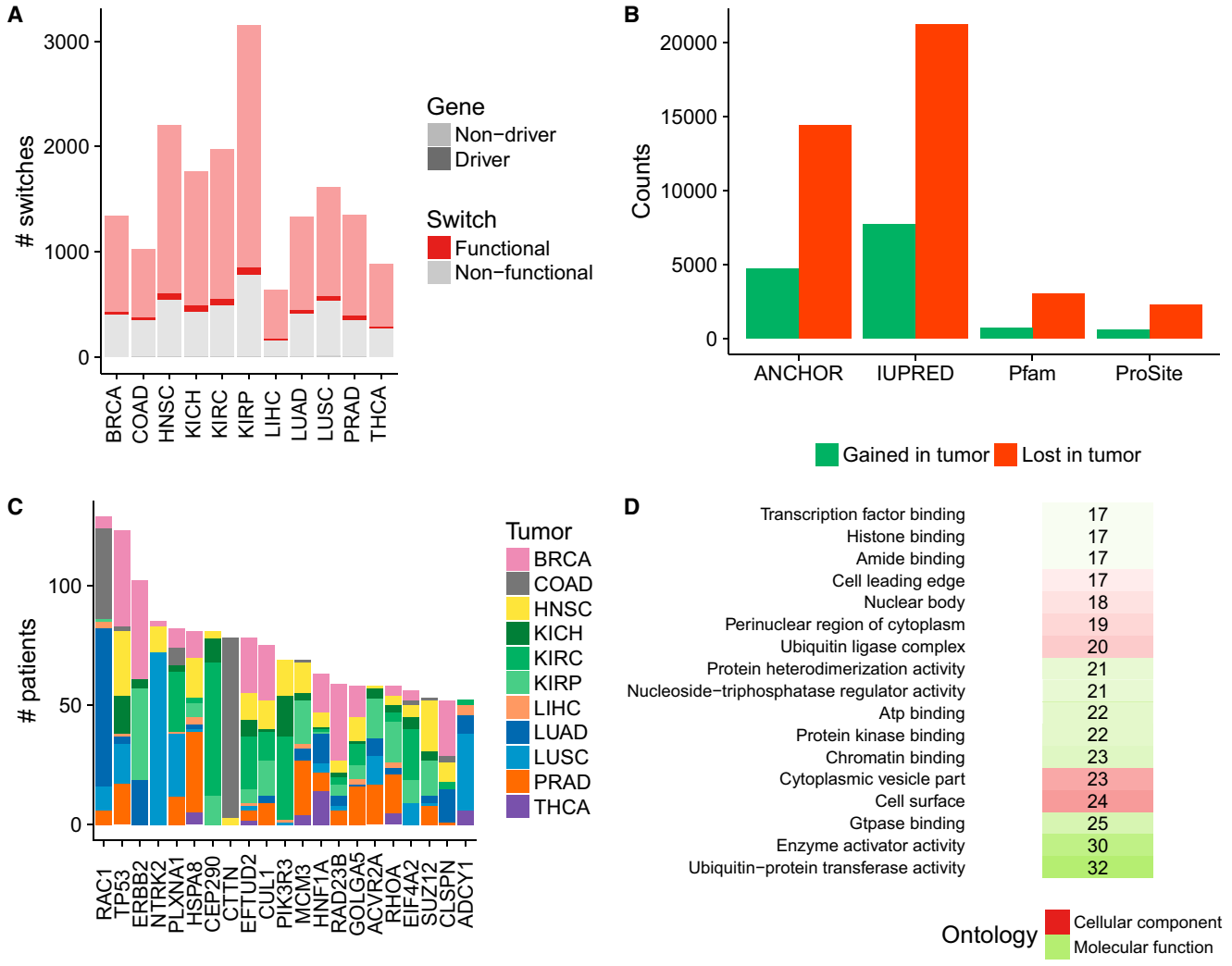


Figure 1. Patient-Specific Definition of Isoform Switches across Multiple Cancer Types

(A) Number of isoform switches (y axis) calculated in each tumor type, separated according to whether the switches affected an annotated protein feature (functional) or not (non-functional) and whether they occurred in cancer gene drivers (driver) or not (non-driver).

(B) Number of different protein feature gains and losses in functional switches for each of the protein features considered, which showed significant enrichment in losses compared to random switches: Pfam (Fisher's exact test p value = 4.4e−23; odds ratio [OR] = 1.5); Prosite (p value = 1.4e−08; OR = 1.3); IUPRED (p value = 1.1e−127; OR = 1.3); and ANCHOR (p value = 7.5e−139; OR = 1.5).

(C) Top 20 functional switches in cancer drivers (x axis) according to patient count (y axis). Tumor types are indicated by color: breast carcinoma (BRCA); colon adenocarcinoma (COAD); head and neck squamous cell carcinoma (HNSC); kidney chromophobe (KICH); kidney renal clear-cell carcinoma (KIRC); kidney papillary cell carcinoma (KIRP); liver hepatocellular carcinoma (LIHC); lung adenocarcinoma (LUAD); lung squamous cell carcinoma (LUSC); prostate adenocarcinoma (PRAD); and thyroid carcinoma (THCA).

(D) Cellular component (red) and molecular function (green) ontologies associated with protein domain families that are significantly lost in functional isoform switches (binomial test; BH-adjusted p value < 0.05). For each functional category, we give the number of switches in which a domain family from this category is lost, which is also indicated by the color shade.

Experimental Procedures). The frequencies of genes or samples with functional switches were similar to those with protein-affecting mutations (PAMs) but smaller than the frequencies for all mutations from WGS data (Figures S2A and S2B), indicating a similar prevalence of switches and PAMs, but not for switches and WGS mutations. Because we calculated switches per patient, we were able to study how these distributed across patients (Supplemental Experimental Procedures). The top cases according to the co-occurrence of WGS somatic mutations with switches across patients included a switch in the cancer driver *CUX1*, although only in 7 patients (Figures S2C and S2D), whereas the top cases according to the number of patients with mutations and switches included *TP53* as well as *FAM19A5*, *DST*, and *FBLN2*, which we already described as isoform switches before (Sebestyén et al., 2015; Figures S2E and S2F). In agreement with the observed low association of mutations and switches (Figure S2G), the number of genes with PAMs and functional switches tended to be inversely correlated

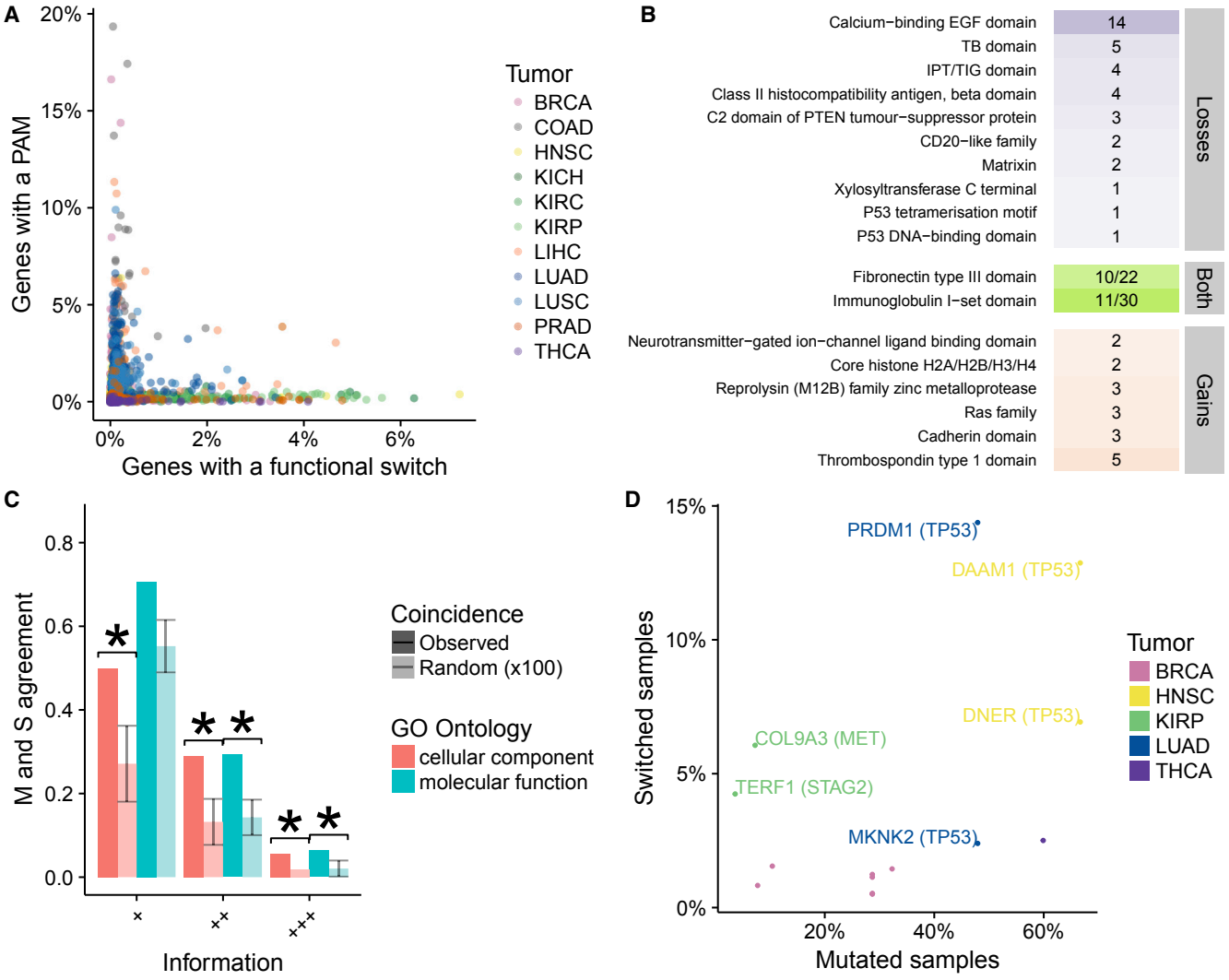


Figure 2. Comparison of Isoform Switches and Somatic Mutations

(A) For each patient sample, color coded according to the tumor type, we indicate the proportion of all genes with protein-affecting mutations (PAMs) (y axis) and the proportion of genes with multiple transcript isoforms that presented a functional isoform switch in the same sample (x axis).

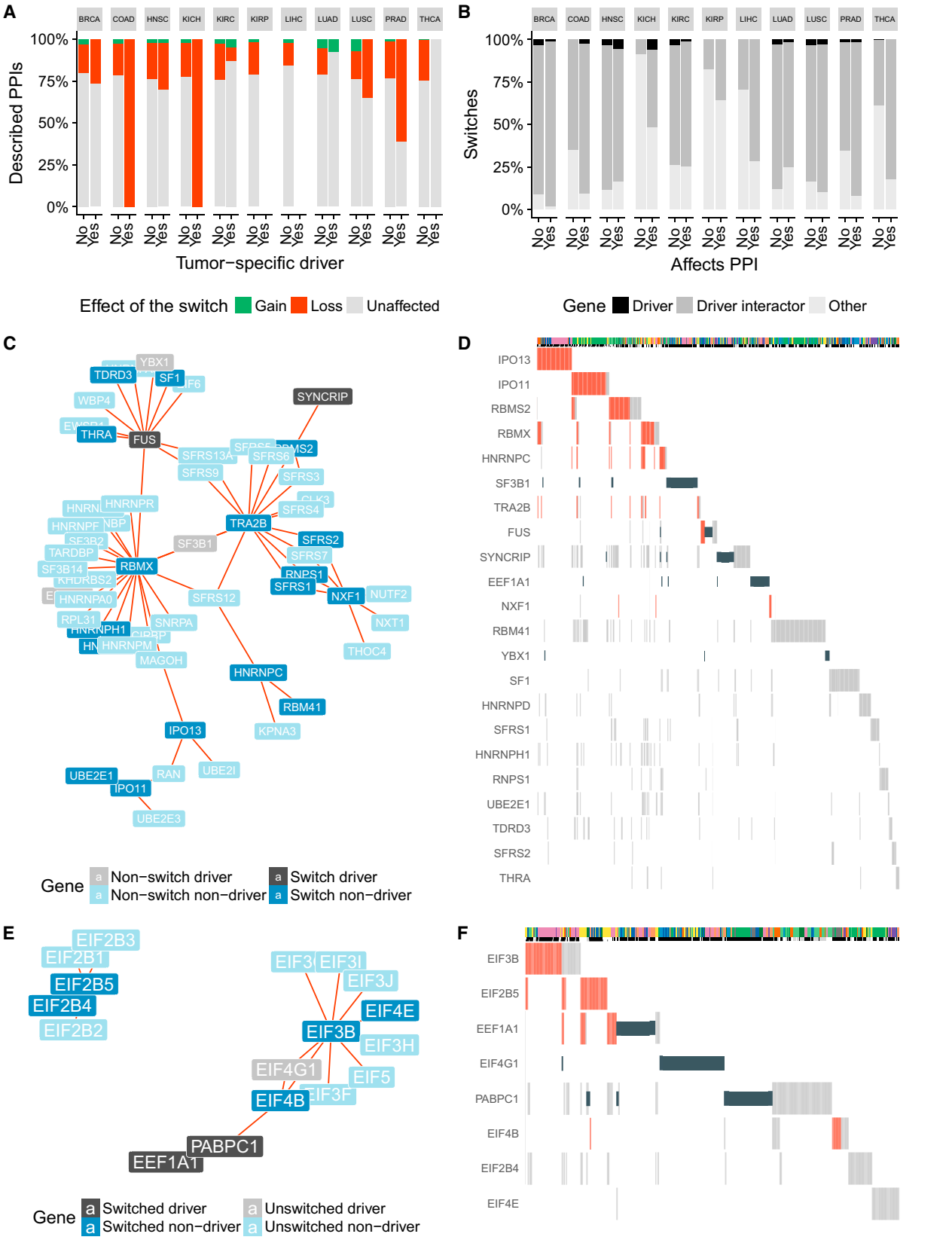
(B) Domain families that were significantly lost or gained in functional isoform switches that are also significantly enriched in protein-affecting mutations in tumors. For each domain class, we indicate the number of different switches in which they occurred. We include here the loss of the P53 DNA-binding and P53 tetramerization domains, which only occurred in *TP53*.

(C) Agreement between protein-affecting mutations and functional switches (y axis) measured in terms of the functional categories of the protein domains they affected (x axis), using two gene ontologies (GOs) at three different GO Slim levels, from most specific (+++) to least specific (+). Random occurrences (plotted in light color) were calculated by sampling 100 times the same number of GO terms from the reference proteome as those enriched in domain families affected by functional switches and in domains families affected by PAMs. Agreement was calculated as the percentage of the union of functional categories from both sets that were common to both. The error bars correspond to the SD calculated from the 100 random samples.

(D) Pairs formed by a cancer driver (in parentheses) and a functional switch from the same pathway and showed significant mutual exclusion (before multiple test correction) between PAMs and switches across patients in at least one tumor type—color-coded by tumor type. The y axis indicates the percentage of samples where the switch occurred, and x axis indicates the percentage of samples where the driver was mutated in the same tumor type.

(Figure 2A), suggesting a complementarity between PAMs and switches affecting protein domains.

We explored this complementarity by checking whether mutations and switches affected the same molecular mechanisms. First, we calculated domain families enriched in PAMs and found 76 domain families across 11 tumor types enriched in mutations (Table S2), which were more frequent in cancer drivers compared to non-drivers (Wilcoxon test p value < 2.2e−16), in agreement with recent reports (Yang et al., 2015). Then, we compared the domain families enriched in mutations with those enriched in gains or losses through switches; we found an overlap of 15 domain families, which was higher than expected by chance given the domains affected by the 6,004 functional switches and the 5,307 domain families observed



(legend on next page)

in the reference proteome (Fisher's test p value = 5.6e−06; OR = 4.7). From the domain families enriched in mutations, 7 showed enrichment in losses, 6 showed enrichment in gains, and 2 showed enrichment in both (Figure 2B; Table S2). The gains included cadherin domains related to switches in *CHD8*, *CDH26*, *FAT1*, *FAT2*, and *FAT3*, whereas the losses included the calcium-binding epidermal growth factor (EGF) domain, which is affected by various switches, including one in *NOTCH4*. A notable case was the loss of the *TP53* DNA-binding domain and the tetramerization motif. Although it occurred in a single switch, its recurrence in 123 patients highlights the relevance of *TP53* alternative splicing (Bourdon, 2007).

We questioned whether the similarity was beyond the coincidence of single-domain families and could affect more generally the function associated to domains. Hence, we calculated the enriched Gene Ontology (GO) terms associated to the domains enriched in mutations and switches separately and then calculated the overlap between both sets. This overlap was compared to the overlap obtained by randomly sampling hundred times from the reference proteome the same number of GO terms found for domains in enriched switches or mutations. Notably, the observed overlap was higher than expected for each GO term and at different GO slim levels (Figure 2C), and the shared functional categories included receptor activity and protein binding. A total of 754 (12.5%) functional switches in 634 genes (47 of them in 37 cancer drivers) affected domain families that were also enriched in mutations, supporting the notion that isoform switches and mutations may impact similar functions in tumors.

If switches and mutations have similar functional impacts, we would expect a tendency toward mutual exclusion of some switches with mutations in cancer drivers. In fact, we identified 292 functional switches that were mutually exclusive with somatic PAMs in three or more cancer drivers (Fisher's test p value < 0.05; Supplemental Experimental Procedures), and 16 of them showed mutual exclusion with at least one cancer gene driver from the same pathway (Table S3). These 16 switches included one in *COL9A3*, which had mutual exclusion with *MET* mutations in kidney renal papillary cell carcinoma (KIRP), and one in *PRDM1*, which showed mutual exclusion with mutations in *TP53* in lung adenocarcinoma (LUAD) (Figure 2D) as well as in *PTEN* In lung squamous cell carcinoma (LUSC) (Figure S2H; Table S3). Despite the observed mutual exclusion, none of the cases was significant after multiple test correction, indicating that the described switches may not provide strong signatures for pan-negative tumors (Saito et al., 2015).

Isoform Switches Affect Protein Interactions with Cancer Drivers

Many of the frequently lost and gained domain families in functional switches were involved in protein-binding activities, indicating a potential impact on protein-protein interactions (PPIs) in cancer. To study this, we used data from five different sources to build a consensus PPI network with 8,142 nodes, each node representing a gene (Figure S3). Then, to determine the effect of switches on the PPI network, we mapped PPIs from this network to domain-domain interactions (DDIs). Domains involved in DDIs were mapped to the specific protein isoforms using their encoded protein sequence. For genes with switches, we then considered those PPIs that could be mapped to DDIs involving domains mapped on either the normal or the tumor isoforms (Figure S4). From the 8,142 genes in the PPI network, 3,243 had at least one isoform switch, and for 1,688 isoform switches (in 1,355 genes), we were able to map at least one PPI to a specific DDI with domains on either the normal or the tumor isoform. A total of 162 of these switches were located in 123 cancer drivers, with the remaining 1,526 in non-driver genes.

For each isoform switch, using the DDI information, we evaluated whether the change between the normal and tumor isoforms would affect a PPI from the network by matching the domains affected by the switch to the domains mediating the interaction, controlling for the expression of the isoforms predicted to be interaction partners. We found that 477 switches (28.3%) in 423 different genes affected domains that mediated protein interactions and thus likely impacted such interactions. Most of these interaction-altering switches (n = 414; 86.8%) caused the loss of the domain that mediated the interaction, whereas a minority (n = 64; 13.2%) led to a gain of the interacting domain. Only a switch in *TAF9* led to gains and losses of interactions with different partners, mediated by the loss of a TIFIID domain and a gain of an AAA domain (Table S4).

Notably, switches in driver genes tended to lose PPIs more frequently than those in non-drivers (Figure 3A). From the 162 switches in drivers, 41 (25.3%) of them altered at least one interaction, either causing loss (33 switches) or gain (8 switches). Moreover, switches that affected domains from families enriched in mutations or that showed frequent mutual exclusion

Figure 3. Potential Impact of Isoform Switches in Protein Interactions with Cancer Drivers

(A) Functional switches were divided according to whether they occurred in tumor-specific drivers (yes) or not (no). For each tumor type, we plot the proportion of PPIs (y axis) that were gained (green), lost (red), or remained unaffected (gray). All comparisons except for KIRC and LUAD were significant (Supplemental Experimental Procedures). Samples from KIRP and LIHC had no PPI-affecting switches in drivers.

(B) Functional switches mapped to PPIs were divided according to whether they affected a PPI (yes) or not (no). For each tumor type, we plotted the proportion of functional switches (y axis) that occurred in cancer drivers (black), in interactors of drivers (dark gray), or in other genes (light gray). All tests for the enrichment of PPIs affected by switches in driver interactors were significant except for KIRC, LUAD, and LUSC (Supplemental Experimental Procedures).

(C) Network for module 11 (Table S6) with PPIs predicted to be lost (red). Cancer drivers are indicated in black or gray if they had a functional switch or not, respectively. Other genes are indicated in dark blue or light blue if they had a functional switch or not, respectively. We do not show unaffected interactions.

(D) OncoPrint for the samples that present protein-affecting mutations (PAMs) in drivers or switches from (C). Mutations are indicated in black, and PPI-affecting switches are indicated in red (loss in this case). Other switches with no predicted effect on the PPI are depicted in gray. The top panel indicates the tumor type of each sample by color (same color code as in previous figures). The second top panel indicates whether the sample harbors a PAM in a tumor-specific driver (black) or not (gray) or whether no mutation data are available for that sample (white).

(E) As in (C) for module 28 (Table S6).

(F) OncoPrint for the switches and drivers from (E). Colors are as in (D).

with mutational drivers also affected PPIs significantly more frequently than other functional switches (Chi-square test p value < 2.2e−16 and p value = 6.8e−08, respectively; Figure S5). Looking at genes annotated as direct interactors of drivers, they tended to affect PPIs more frequently than the rest of functional switches mapped to PPIs (Figure 3B). Additionally, all functional pathways found enriched in PPI-affecting switches were related to cancer (adjusted Fisher's exact test p value < 0.05 and odds ratio > 2; Table S5), reinforcing the functional relevance of these 477 PPI-affecting isoform switches in cancer.

Isoform Switches Remodel Protein Interaction Networks in Cancer

To further characterize the role of switches, we calculated modules in the PPI network (Blondel et al., 2008) using only interaction edges affected by switches (Supplemental Experimental Procedures). This produced 179 modules involving 1,405 genes (Table S6). From these, 52 modules included a cancer driver, and 47 of them included also switches that involved two protein-coding isoforms. We tested for the enrichment of genes belonging to specific protein complexes (Ruepp et al., 2010), complexes related to RNA processing and splicing (Akerman et al., 2015), and cancer-related pathways (Liberzon et al., 2015; Table S6; Supplemental Experimental Procedures). From the 47 modules described above, 8 showed enrichment in pathways and complexes: apoptosis-related pathways (module 109 in Table S6); ubiquitin-mediated proteolysis pathway (module 26); and *ERBB*-signaling pathway (module 169), as well as spliceosomal (module 11); ribosomal (module 170); SMN (module 28); PA700 (module 58); and TFIID (module 66) complexes (Table S6). In particular, module 11 was enriched in splicing factors and RNA-binding proteins and included the cancer drivers *SF3B1*, *FUS*, *SYNCRIP*, *EEF1A1*, and *YBX1* (Figure 3C; Table S6). The module contained a switch in *RBMX* involving the skipping of two exons and the elimination of an RNA recognition motif (RRM) that would impact interactions with *SF3B1*, *EEF1A1*, and multiple RNA binding protein (RBP) genes (Figure 3C) and a switch in *TRA2B* that yielded a non-coding transcript previously described (Stoilov et al., 2004) and would eliminate an interaction with *SF3B1* and other splicing factors. We also found a switch in *HNRNPC*, *TRA2A*, *NXF1*, and *RBMS2* that lost interactions with various serine/arginine-rich (SR)-protein-coding genes. Consistent with a potential functional impact, the PPI-affecting switches showed mutual exclusion with the mutational cancer drivers (Figure 3D). Interestingly, this module also contained switches in the Importin genes *IPO11* and *IPO13*, which affected interactions with ubiquitin-conjugating enzymes *UBE2E1*, *UBE2E3*, and *UBE2I* and which showed mutual exclusion across different tumor types (Figure 3D). These results indicate that the activity of RNA-processing factors may be altered in cancer through the disruption of their PPIs by alternative splicing.

Another interesting case was module 28 (Table S6), with switches in the regulators of translation, *EIF4B*, *EIF3B*, and *EIF4E*, which affected interactions with the drivers *EIF4G1*, *EIF4A2*, and *PABPC1* (Figure 3E). The switch in *EIF4B* caused the skipping of one exon, which we predicted to eliminate an RRM domain and lose interactions with drivers *EIF4G1* and *PABPC1*. The switch in *EIF3B* yielded a non-coding transcript that would lose multiple interactions. Although we did not predict any PPI change for *EIF4E*, this switch lost eight predicted ANCHOR regions (Table S4), suggesting a possible effect on yet to be described interactions. Besides frequent PAMs, *PABPC1* also presented a functional switch that affected 2 disordered regions but did not affect any of the RRMs. In this case, we did not predict any change in PPI, and the possible functional impact remains to be discovered. Moreover, the identified PPI-affecting switches showed mutual exclusion with PAMs in *EIF4G1* and *PABPC1* (Figure 3F). These results suggest that isoform switches may impact translational regulation in tumors through the alteration of PPIs of the corresponding regulators.

Isoform Switches as Potential Drivers of Cancer

Our results provide evidence that a subset of the alternative splicing switches (1) induced a gain or loss of a protein domain from a family frequently mutated in cancer, (2) affected one or more PPIs, (3) displayed some mutual exclusion with drivers, or (4) displayed recurrence across patients. One or more of these properties were fulfilled by 1,662 functional switches, which we hypothesized could define potential alternative splicing drivers (potential AS drivers; Figure 4A; Table S1), with the majority of them (1,080; 65%) affecting mutated domain families and/or PPIs (Figure 4B). To test possible driver-like properties in these switches, we calculated their centrality and distance to mutational drivers in the PPI network, which are considered as defining properties for cancer-relevant genes (Jonsson and Bates, 2006). Potential AS drivers showed greater centrality (Mann-Whitney test p value < 2.2e−16; Figure S6A) and closer distances to tumor-specific drivers (Fisher's exact test p value < 2.2e−16; OR = 1.5; Figure S6B) compared to the rest of switches.

The prevalence of these potential AS drivers varied across samples and tumor types. Considering tumor-specific mutational drivers (Mut drivers) and our set of potential AS drivers, we labeled each patient as AS driver enriched or Mut driver enriched according to whether the proportion of switched potential AS drivers or mutated Mut drivers was higher, respectively. This partition of the samples indicated that, although Mut drivers were predominant in patients for most tumor types, potential AS drivers were predominant for a considerable number of patients across several tumor types and particularly for kidney and prostate tumors (Figure 4C). Additionally, regardless of the tumor type, patients with many mutations in Mut drivers tended to show a low number of switched potential AS drivers and vice versa (Figure 4D). The occurrence of copy number alteration (CNA) drivers also showed a pattern of anti-correlation with our potential AS drivers similar to the one we found between Mut drivers and potential AS drivers (Figure S6C). The patient distribution patterns of candidate AS drivers compared with mutational or CNA drivers bear resemblance with the proposed cancer genome hyperbola between mutations and CNAs (Figure S6D; Ciriello et al., 2013), which supports the notion that a subset of isoform changes represents alternative, yet-unexplored relevant mechanisms that could provide a complementary route to induce similar effects as genetic mutations.

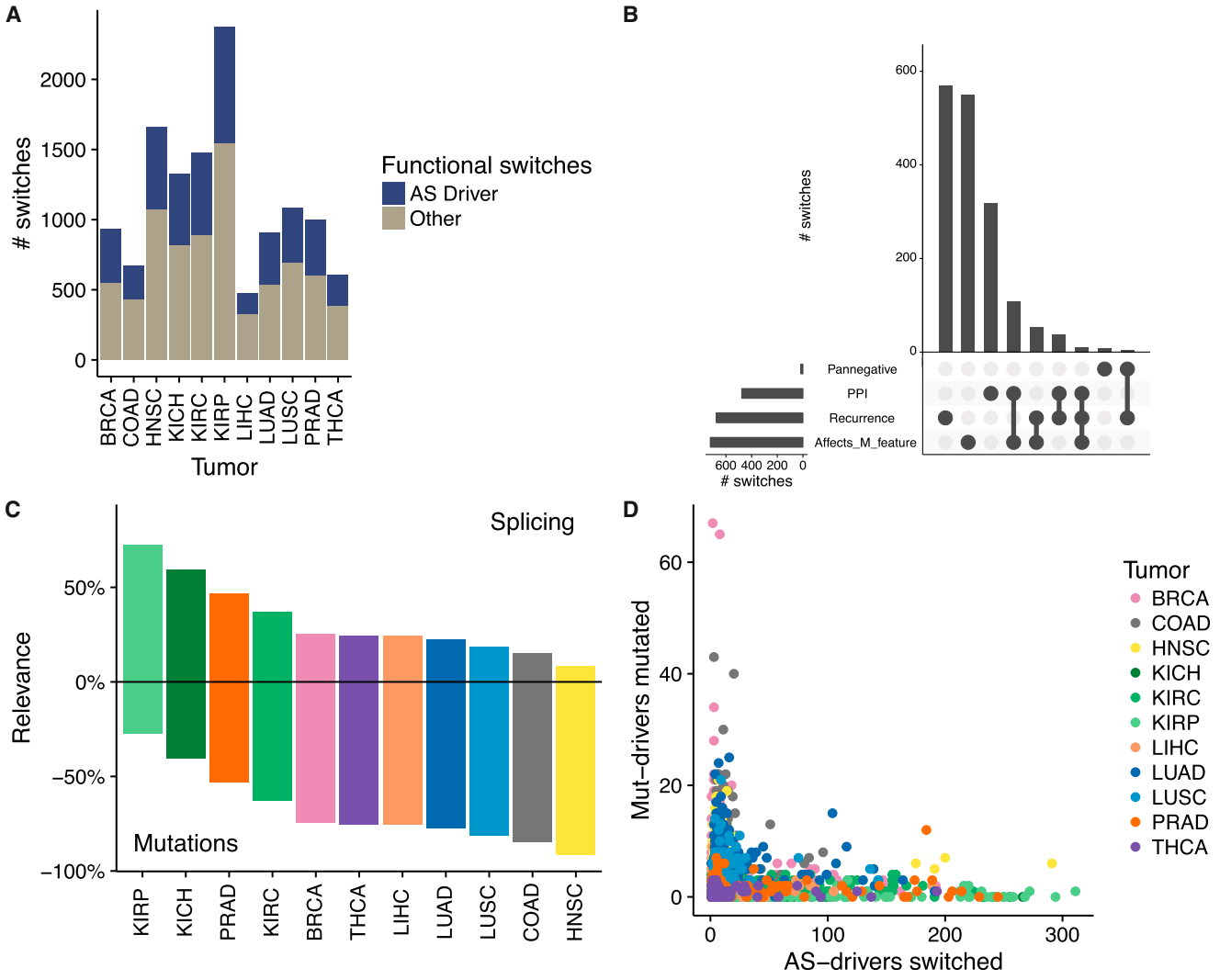


Figure 4. Isoform Switches as Potential Drivers of Cancer

(A) Number of functional isoform switches and potential AS drivers detected in each tumor type.

(B) Candidate potential AS drivers grouped according to their properties: disruption of PPIs; significant recurrence across patients (recurrence); gain or loss of a protein feature that was frequently mutated in tumors (affects M_feature); mutual exclusion; and sharing pathway with cancer drivers (pannegative). Horizontal bars indicate the number of switches for each property. The vertical bars show those in each of the intersections indicated by connected bullet points (Conway et al., 2017).

(C) Classification of samples according to the relevance of potential AS drivers or Mut drivers in each tumor type. For each tumor type (x axis), the positive y axis shows the percentage of samples that had a proportion of switched potential AS drivers higher than the proportion of mutated Mut drivers. The negative y axis shows the percentage of samples in which the proportion of mutated Mut drivers was higher than the proportion of switched potential AS drivers. Only patients with mutation and transcriptome data are shown.

(D) Each of the patients from (C) is represented according to the percentage of mutated Mut drivers (y axis) and the percentage of switched potential AS drivers (x axis).

DISCUSSION

We have identified consistent and recurrent transcript isoform switches that impact the function of affected proteins by adding or removing protein domains that were frequently mutated in cancer or by disrupting or gaining PPIs—possibly also altering the formation of protein complexes—with cancer drivers or in cancer-related pathways. Moreover, we observed that patients with some of these isoform switches tended not to harbor mutations in cancer drivers and the other way around. Recently, an alternative splicing change in *NFE2L2* has been described to lead to the loss of a protein domain and the interaction with its negative regulator *KEAP1*, thereby providing an alternative mechanism for the activation of an oncogenic pathway (Goldstein et al., 2016). Similarly, an isoform change in the gene *ATF2* has been shown to drive melanomagenesis (Claps et al., 2016). These examples, together with the analyses presented here, support a model by which functions and pathways often

altered in cancer through somatic mutations may be affected in a similar way by isoform changes in some patients and therefore contribute to the tumor phenotype. Importantly, these isoform changes could occur without gene expression changes in the host gene and thus provide an independent catalog of functional alterations in cancer.

Functional domains and interactions might not always be entirely lost through a switch, as normal isoforms generally retain some expression in tumors. This could be partly due to the uncertainty in the estimate of transcript abundance from RNA sequencing or to the heterogeneity in the transcriptomes of tumor cells. Still, a relatively small change in transcript abundance has been shown to be sufficient to trigger an oncogenic effect in cells (Anczuków et al., 2015; Bechara et al., 2013; Sebestyén et al., 2016). Additionally, we observed that a number of isoform changes defined a switch from a protein-coding transcript to a non-coding one, possibly undergoing nonsense-mediated decay, which is a widespread mechanism of alternative-splicing-mediated gene expression regulation (Hansen et al., 2009), and could potentially alter function in a way similar to other isoform changes between protein-coding isoforms. The predicted impact on domains and interactions could therefore be indicative of alterations on regulatory networks with variable functional effects.

Our description in terms of transcript isoform switches allowed us to describe more variations in the transcriptome than using local alternative splicing events and to determine the protein features potentially gained or lost through splicing changes. However, this approach has some potential limitations. Accurate determination of differential transcript usage in genes with many isoforms requires high coverage and sufficient samples per condition (Sebestyén et al., 2015), which we expect was mitigated by our use of the variability across normal samples to determine significance. Additionally, because we used annotated transcript isoforms, we may have missed tumor-specific transcripts not present in the annotation. We also only recovered a small fraction of the entire set of PPIs taking place in the cell. For instance, we did not characterize those interactions mediated through low-complexity regions (Buljan et al., 2012; Ellis et al., 2012); hence, many more interactions and protein complexes may be affected in tumors.

The origin of the observed splicing changes remains to be elucidated. We did not find a general association with somatic mutations in *cis*. It is possible that small copy number alterations or indels are responsible for these switches but are still hard to detect with WES and WGS data, and more targeted searches or deeper sequencing are necessary. An alternative explanation is that the majority of the switches described occur through *trans*-acting alterations, such as the expression change in splicing factors (Sebestyén et al., 2016). For instance, mutations in *RBM10* or downregulation of *QKI* lead to the same splicing change in *NUMB* that promotes cell proliferation (Bechara et al., 2013; Zong et al., 2014), and the oncogenic switch in *RAC1* (Zhou et al., 2013) is regulated by expression changes in various splicing factors (Gonçalves et al., 2009; Pelisch et al., 2012), which are controlled by pathways often altered in tumors (Fu and Ares, 2014). Another possibility is that these switches describe signatures of non-genetic variability (Brock et al., 2009). The intra-tumor heterogeneity could allow recapitulating similar transcriptome phenotypes, which would determine the fitness of cells and the progression of tumors independently of somatic mutations. Because natural selection acts on the phenotype rather than on the genotype, an interesting hypothesis is that specific transcript isoform expression patterns could define particular tumor phenotypes that would be closely related to those determined by somatic mutations in drivers, thereby defining an advantageous phenotype such that the selective pressure to develop equivalent adaptations is relaxed. Accordingly, our identified isoform switches could play an important role in the neoplastic process independently of or in conjunction with the already characterized genetic alterations.

EXPERIMENTAL PROCEDURES

Further details and an outline of resources used in this work can be found in Supplemental Experimental Procedures.

Calculation of Significant Isoform Switches per Patient

We modeled splicing alterations in a gene as a switch between two transcript isoforms: one normal and one tumoral. For each transcript, the relative abundance per sample, which we called proportion spliced-in (PSI), was calculated by normalizing its abundance in transcripts per million (TPM) units by the sum of abundances of all transcripts in the same gene. Then, for each transcript and sample, we calculated the change in relative abundance as $\Delta PSI = PSI_{tumor} - PSI_{ref}$, where PSI_{tumor} is the relative abundance in the tumor sample and PSI_{ref} is the normal reference value, which is the value of the paired normal sample, when available, or the median of PSIs in the normal samples for the same tissue type otherwise. We considered significant those changes with $|\Delta PSI| > 0.05$ and with empirical $p < 0.01$ in the comparison of the observed $|\Delta PSI|$ value with the distribution of $|\Delta PSI|$ values obtained by comparing the normal samples pairwise without repetition. We only kept those cases for which the tumor isoform PSI was higher than the normal isoform in the tumor sample and the normal isoform PSI in the normal sample was higher than the value for the tumor isoform. Moreover, we discarded genes that either had an outlier expression in the tumor sample compared to normal tissues—had expression below the bottom 2.5% or above the 97.5% of the values of normal expression—or showed differential expression between the tumor and the normal samples (Wilcoxon test p value < 0.01 using the gene TPM values).

Candidate switches were defined per patient and per gene, and in some samples, the same gene could have different switches. We discarded those switches that contradicted a more frequent switch in the same gene and the same tumor type. Moreover, we discarded any switch that affected a number of patients below the top 99% of the distribution of patient frequency of these contradictory switches in each tumor type. Lastly, we filtered out switches that were significantly lowly recurrent, i.e., they occurred in fewer patients than expected by chance (binomial test; adjusted p value < 0.05, using all tumor types). As a consequence, none of the reported switches occurred in less than 5 samples. Thus, a switch in a patient sample was defined as a pair of transcripts in a gene with no expression change and with significant changes in opposite directions that showed consistency across a minimum number of patients. We aggregated the switches from the different tumor types to get the final list (Table S1).

Simulated Switches

To simulate switches between normal and tumor tissues, we used genes with more than one expressed isoform. For each gene, we selected the isoform with the highest median expression across the normal samples as the normal isoform and an arbitrary different transcript expressed in the tumor samples as the tumor isoform. For each gene, we generated a maximum of five such simulated switches.

Functional Switches

A switch was defined as functional if both isoforms overlapped in genomic extent and there was a change in the encoded protein, including cases where

only one of the isoforms was coding and, moreover, there was a gain or loss of a protein feature: Pfam domains (Finn et al., 2016) mapped with InterProScan (Jones et al., 2014); ProSite patterns (Gattiker et al., 2002); disordered regions from IUPred (Dosztányi et al., 2005); and disordered regions potentially involved in PPIs from ANCHOR (Dosztányi et al., 2009). For IUPred and ANCHOR, we only considered changes involving at least 5 amino acids. Switches without any mapped protein features were not considered. Significance on the enrichment of protein features losses versus gains was calculated by comparing the number of gains and losses in switches with the same numbers in simulated switches (Supplemental Experimental Procedures).

Enrichment of Domain Families in Switches and Mutations

To find protein domain families significantly affected by switches, we first calculated a reference proteome for each tumor type. Using genes with multiple transcripts, we selected those that had at least one isoform with TPM > 0.1 and only kept the isoform with the highest median expression across the normal samples in the same tissue type. Proteins encoded by these isoforms were considered the reference proteome in each tumor type. We aggregated the reference proteomes from all tumor types to form a pan-cancer reference proteome. The expected frequency of a protein feature was then measured as the proportion of this feature in the reference proteome. This expected frequency was then used to calculate the probability of a feature to be affected by a switch using a binomial test with the number of times the feature was gained or lost in switches and the total number of feature gains or losses due to switches (Supplemental Experimental Procedures). We selected cases with Benjamini-Hochberg (BH)-adjusted p value < 0.05. Additionally, to ensure the specificity of the enrichment for each domain class, we considered only domain families affected in at least two switches. To calculate domain families enriched in mutations, we considered again the reference proteome in each tumor type. The expected mutation rate of a domain family was considered to be the proportion of the length of domains in the proteome covered by this domain family. We aggregated all observed mutations falling within each family and calculated the probability of the observed mutations using a binomial test using the mutation count for a domain family and the total mutations in all domain families (Supplemental Experimental Procedures). After correcting for multiple testing, we kept those cases with a BH-adjusted p value < 0.05. GO analysis was performed using DcGO (Fang and Gough, 2013). For the enrichment test, we considered significant those cases with FDR < 0.01 (hypergeometric test).

Protein Interaction Analysis

We created a consensus PPI network using data from PSICQUIC (del-Toro et al., 2013), BIOGRID (Chatr-Aryamontri et al., 2015), HumNet (Lee et al., 2011), STRING (Szklarczyk et al., 2011), and from Rolland et al. (2014). The consensus network was built with interactions appearing in at least four of these five sources, yielding a total of 8,142 nodes with 29,991 interactions. To find PPIs likely altered by isoform switches, we first mapped each PPI in a gene to a specific DDI, using information on DDIs from iPfam (Finn et al., 2014), DOMINE (Raghavachari et al., 2008), and 3did (Mosca et al., 2014). Domains involved in DDIs were then mapped to specific protein isoforms. For the genes with switches, we then considered those PPIs that could be mapped to DDIs involving domains mapped to either the normal or the tumor isoforms. In total, 3,242 genes with 4,219 switches mapped to one or more interactions in the consensus network and 1,688 isoform switches (in 1,355 genes) were mapped to at least one specific DDI. We defined a PPI as lost if it was mapped to one or more DDIs in the isoform expressed in the normal tissue, but not in the isoform expressed in the tumor sample. If multiple domains mediated the same interaction, it was considered lost if at least one of these domains was lost in the switch. We defined a PPI as gained if it was mapped to a DDI only in the tumor isoform, but not in the normal isoform.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures, six figures, and six tables and can be found with this article online at http://dx.doi.org/10.1016/j.celrep.2017.08.012.

AUTHOR CONTRIBUTIONS

E.E. proposed the study. H.C.-G. developed the software and performed the analyses. E.P.-P. built the consensus PPI network and mapped the DDIs. E.E. and A.G. supervised the analyses. E.E. and H.C.-G. wrote the manuscript with essential inputs from E.P.-P. and A.G.

ACKNOWLEDGMENTS

H.C.-G. and E.E. were supported by the MINECO and FEDER (BIO2014-52566-R), Consolider RNAREG (CSD2009-00080), AGAUR (SGR2014-1121), the European ITN Network RNP-Net (ID: 289007), and the Sandra Ibarra Foundation for Cancer (FSI2013). E.P.-P. and A.G. were supported by the SBP CC grant (P30 CA030199). All authors thank The Cancer Genome Atlas project for making their data publicly available. The Computational RNA Biology Group is part of the Research Programme on Biomedical Informatics (GRIB), which is a member of ELIXIR-Excelerate of the European Union Horizon 2020 Programme 2014-2020 (No. 676559) and of the Spanish National Bioinformatics Institute (INB), PRB2-ISCIII and is supported by grant PT13/0001/0023 of the PE I+D+I 2013-2016, funded by ISCIII and FEDER.

Received: June 8, 2017
Revised: July 15, 2017
Accepted: July 26, 2017
Published: August 29, 2017

REFERENCES

Akerman, M., Fregoso, O.I., Das, S., Ruse, C., Jensen, M.A., Pappin, D.J., Zhang, M.Q., and Krainer, A.R. (2015). Differential connectivity of splicing activators and repressors to the human spliceosome. Genome Biol. *16*, 119.

Alamancos, G.P., Pagès, A., Trincado, J.L., Bellora, N., and Eyras, E. (2015). Leveraging transcript quantification for fast computation of alternative splicing profiles. RNA *21*, 1521–1531.

Anczuków, O., Akerman, M., Cléry, A., Wu, J., Shen, C., Shirole, N.H., Raimer, A., Sun, S., Jensen, M.A., Hua, Y., et al. (2015). SRSF1-regulated alternative splicing in breast cancer. Mol. Cell *60*, 105–117.

Bechara, E.G., Sebestyén, E., Bernardis, I., Eyras, E., and Valcárcel, J. (2013). RBM5, 6, and 10 differentially regulate NUMB alternative splicing to control cancer cell proliferation. Mol. Cell *52*, 720–733.

Blondel, V.D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. J. Stat. Mech. Theory Exp. *2008*, P10008.

Bourdon, J.-C. (2007). p53 and its isoforms in cancer. Br. J. Cancer *97*, 277–282.

Brock, A., Chang, H., and Huang, S. (2009). Non-genetic heterogeneity–a mutation-independent driving force for the somatic evolution of tumours. Nat. Rev. Genet. *10*, 336–342.

Buljan, M., Chalancon, G., Eustermann, S., Wagner, G.P., Fuxreiter, M., Bateman, A., and Babu, M.M. (2012). Tissue-specific splicing of disordered segments that embed binding motifs rewires protein interaction networks. Mol. Cell *46*, 871–883.

Chabot, B., and Shkreta, L. (2016). Defective control of pre-messenger RNA splicing in human disease. J. Cell Biol. *212*, 13–27.

Chatr-Aryamontri, A., Breitkreutz, B.J., Oughtred, R., Boucher, L., Heinicke, S., Chen, D., Stark, C., Breitkreutz, A., Kolas, N., O'Donnell, L., et al. (2015). The BioGRID interaction database: 2015 update. Nucleic Acids Res. *43*, D470–D478.

Ciriello, G., Miller, M.L., Aksoy, B.A., Senbabaoglu, Y., Schultz, N., and Sander, C. (2013). Emerging landscape of oncogenic signatures across human cancers. Nat. Genet. *45*, 1127–1133.

Claps, G., Cheli, Y., Zhang, T., Scortegagna, M., Lau, E., Kim, H., Qi, J., Li, J.-L., James, B., Dzung, A., et al. (2016). A transcriptionally inactive ATF2 variant drives melanomagenesis. Cell Rep. *15*, 1884–1892.

Conway, J.R., Lex, A., and Gehlenborg, N. (2017). UpSetR: an R package for the visualization of intersecting sets and their properties. Bioinformatics. Published online June 22, 2017. http://dx.doi.org/10.1093/bioinformatics/btx364.

Darman, R.B., Seiler, M., Agrawal, A.A., Lim, K.H., Peng, S., Aird, D., Bailey, S.L., Bhavsar, E.B., Chan, B., Colla, S., et al. (2015). Cancer-associated SF3B1 hotspot mutations induce cryptic 3′ splice site selection through use of a different branch point. Cell Rep. *13*, 1033–1045.

del-Toro, N., Dumousseau, M., Orchard, S., Jimenez, R.C., Galeota, E., Launay, G., Goll, J., Breuer, K., Ono, K., Salwinski, L., and Hermjakob, H. (2013). A new reference implementation of the PSICQUIC web service. Nucleic Acids Res. *41*, W601–W606.

Dosztányi, Z., Csizmok, V., Tompa, P., and Simon, I. (2005). IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. Bioinformatics *21*, 3433–3434.

Dosztányi, Z., Mészáros, B., and Simon, I. (2009). ANCHOR: web server for predicting protein binding regions in disordered proteins. Bioinformatics *25*, 2745–2746.

Dvinge, H., and Bradley, R.K. (2015). Widespread intron retention diversifies most cancer transcriptomes. Genome Med. *7*, 45.

Ellis, J.D., Barrios-Rodiles, M., Colak, R., Irimia, M., Kim, T., Calarco, J.A., Wang, X., Pan, Q., O'Hanlon, D., Kim, P.M., et al. (2012). Tissue-specific alternative splicing remodels protein-protein interaction networks. Mol. Cell *46*, 884–892.

Fang, H., and Gough, J. (2013). DcGO: database of domain-centric ontologies on functions, phenotypes, diseases and more. Nucleic Acids Res. *41*, D536–D544.

Finn, R.D., Miller, B.L., Clements, J., and Bateman, A. (2014). iPfam: a database of protein family and domain interactions found in the Protein Data Bank. Nucleic Acids Res. *42*, D364–D373.

Finn, R.D., Coggill, P., Eberhardt, R.Y., Eddy, S.R., Mistry, J., Mitchell, A.L., Potter, S.C., Punta, M., Qureshi, M., Sangrador-Vegas, A., et al. (2016). The Pfam protein families database: towards a more sustainable future. Nucleic Acids Res. *44* (*D1*), D279–D285.

Frampton, G.M., Ali, S.M., Rosenzweig, M., Chmielecki, J., Lu, X., Bauer, T.M., Akimov, M., Bufill, J.A., Lee, C., Jentz, D., et al. (2015). Activation of MET via diverse exon 14 splicing alterations occurs in multiple tumor types and confers clinical sensitivity to MET inhibitors. Cancer Discov. *5*, 850–859.

Fu, X.-D., and Ares, M., Jr. (2014). Context-dependent control of alternative splicing by RNA-binding proteins. Nat. Rev. Genet. *15*, 689–701.

Gattiker, A., Gasteiger, E., and Bairoch, A. (2002). ScanProsite: a reference implementation of a PROSITE scanning tool. Appl. Bioinformatics *1*, 107–108.

Goldstein, L.D., Lee, J., Gnad, F., Klijn, C., Schaub, A., Reeder, J., Daemen, A., Bakalarski, C.E., Holcomb, T., Shames, D.S., et al. (2016). Recurrent loss of NFE2L2 exon 2 is a mechanism for Nrf2 pathway activation in human cancers. Cell Rep. *16*, 2605–2617.

Gonçalves, V., Matos, P., and Jordan, P. (2009). Antagonistic SR proteins regulate alternative splicing of tumor-related Rac1b downstream of the PI3-kinase and Wnt pathways. Hum. Mol. Genet. *18*, 3696–3707.

Hansen, K.D., Lareau, L.F., Blanchette, M., Green, R.E., Meng, Q., Rehwinkel, J., Gallusser, F.L., Izaurralde, E., Rio, D.C., Dudoit, S., and Brenner, S.E. (2009). Genome-wide identification of alternative splice forms down-regulated by nonsense-mediated mRNA decay in Drosophila. PLoS Genet. *5*, e1000525.

Jones, P., Binns, D., Chang, H.Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G., et al. (2014). InterProScan 5: genome-scale protein function classification. Bioinformatics *30*, 1236–1240.

Jonsson, P.F., and Bates, P.A. (2006). Global topological features of cancer proteins in the human interactome. Bioinformatics *22*, 2291–2297.

Jung, H., Lee, D., Lee, J., Park, D., Kim, Y.J., Park, W.-Y., Hong, D., Park, P.J., and Lee, E. (2015). Intron retention is a widespread mechanism of tumor-suppressor inactivation. Nat. Genet. *47*, 1242–1248.

Karni, R., de Stanchina, E., Lowe, S.W., Sinha, R., Mu, D., and Krainer, A.R. (2007). The gene encoding the splicing factor SF2/ASF is a proto-oncogene. Nat. Struct. Mol. Biol. *14*, 185–193.

Lee, S.C.-W., and Abdel-Wahab, O. (2016). Therapeutic targeting of splicing in cancer. Nat. Med. *22*, 976–986.

Lee, I., Blom, U.M., Wang, P.I., Shim, J.E., and Marcotte, E.M. (2011). Prioritizing candidate disease genes by network-based boosting of genome-wide association data. Genome Res. *21*, 1109–1121.

Liberzon, A., Birger, C., Thorvaldsdóttir, H., Ghandi, M., Mesirov, J.P., and Tamayo, P. (2015). The Molecular Signatures Database (MSigDB) hallmark gene set collection. Cell Syst. *1*, 417–425.

Lu, Z.X., Huang, Q., Park, J.W., Shen, S., Lin, L., Tokheim, C.J., Henry, M.D., and Xing, Y. (2015). Transcriptome-wide landscape of pre-mRNA alternative splicing associated with metastatic colonization. Mol. Cancer Res. *13*, 305–318.

Madan, V., Kanojia, D., Li, J., Okamoto, R., Sato-Otsubo, A., Kohlmann, A., Sanada, M., Grossmann, V., Sundaresan, J., Shiraishi, Y., et al. (2015). Aberrant splicing of U12-type introns is the hallmark of ZRSR2 mutant myelodysplastic syndrome. Nat. Commun. *6*, 6042.

Mosca, R., Céol, A., Stein, A., Olivella, R., and Aloy, P. (2014). 3did: a catalog of domain-based interactions of known three-dimensional structure. Nucleic Acids Res. *42*, D374–D379.

Norris, A.D., and Calarco, J.A. (2012). Emerging roles of alternative pre-mRNA splicing regulation in neuronal development and function. Front. Neurosci. *6*, 122.

Paik, P.K., Drilon, A., Fan, P.-D., Yu, H., Rekhtman, N., Ginsberg, M.S., Borsu, L., Schultz, N., Berger, M.F., Rudin, C.M., and Ladanyi, M. (2015). Response to MET inhibitors in patients with stage IV lung adenocarcinomas harboring MET mutations causing exon 14 skipping. Cancer Discov. *5*, 842–849.

Pelisch, F., Khauv, D., Risso, G., Stallings-Mann, M., Blaustein, M., Quadrana, L., Radisky, D.C., and Srebrow, A. (2012). Involvement of hnRNP A1 in the matrix metalloprotease-3-dependent regulation of Rac1 pre-mRNA splicing. J. Cell. Biochem. *113*, 2319–2329.

Poulikakos, P.I., Persaud, Y., Janakiraman, M., Kong, X., Ng, C., Moriceau, G., Shi, H., Atefi, M., Titz, B., Gabay, M.T., et al. (2011). RAF inhibitor resistance is mediated by dimerization of aberrantly spliced BRAF(V600E). Nature *480*, 387–390.

Raghavachari, B., Tasneem, A., Przytycka, T.M., and Jothi, R. (2008). DOMINE: a database of protein domain interactions. Nucleic Acids Res. *36*, D656–D661.

Rolland, T., Taşan, M., Charloteaux, B., Pevzner, S.J., Zhong, Q., Sahni, N., Yi, S., Lemmens, I., Fontanillo, C., Mosca, R., et al. (2014). A proteome-scale map of the human interactome network. Cell *159*, 1212–1226.

Ruepp, A., Waegele, B., Lechner, M., Brauner, B., Dunger-Kaltenbach, I., Fobo, G., Frishman, G., Montrone, C., and Mewes, H.W. (2010). CORUM: the comprehensive resource of mammalian protein complexes–2009. Nucleic Acids Res. *38*, D497–D501.

Saito, M., Shimada, Y., Shiraishi, K., Sakamoto, H., Tsuta, K., Totsuka, H., Chiku, S., Ichikawa, H., Kato, M., Watanabe, S., et al. (2015). Development of lung adenocarcinomas with exclusive dependence on oncogene fusions. Cancer Res. *75*, 2264–2271.

Sebestyén, E., Zawisza, M., and Eyras, E. (2015). Detection of recurrent alternative splicing switches in tumor samples reveals novel signatures of cancer. Nucleic Acids Res. *43*, 1345–1356.

Sebestyén, E., Singh, B., Miñana, B., Pagès, A., Mateo, F., Pujana, M.A., Valcárcel, J., and Eyras, E. (2016). Large-scale analysis of genome and transcriptome alterations in multiple tumors unveils novel cancer-relevant splicing networks. Genome Res. *26*, 732–744.

Sotillo, E., Barrett, D.M., Black, K.L., Bagashev, A., Oldridge, D., Wu, G., Sussman, R., Lanauze, C., Ruella, M., Gazzara, M.R., et al. (2015). Convergence of acquired mutations and alternative splicing of CD19 enables resistance to CART-19 immunotherapy. Cancer Discov. *5*, 1282–1295.

Stoilov, P., Daoud, R., Nayler, O., and Stamm, S. (2004). Human tra2-beta1 autoregulates its protein concentration by influencing alternative splicing of its pre-mRNA. Hum. Mol. Genet. *13*, 509–524.

Supek, F., Miñana, B., Valcárcel, J., Gabaldón, T., and Lehner, B. (2014). Synonymous mutations frequently act as driver mutations in human cancers. Cell *156*, 1324–1335.

Szklarczyk, D., Franceschini, A., Kuhn, M., Simonovic, M., Roth, A., Minguez, P., Doerks, T., Stark, M., Muller, J., Bork, P., et al. (2011). The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. Nucleic Acids Res. *39*, D561–D568.

Trincado, J.L., Sebestyén, E., Pagés, A., and Eyras, E. (2016). The prognostic potential of alternative transcript isoforms across human tumors. Genome Med. *8*, 85.

Vorlová, S., Rocco, G., Lefave, C.V., Jodelka, F.M., Hess, K., Hastings, M.L., Henke, E., and Cartegni, L. (2011). Induction of antagonistic soluble decoy receptor tyrosine kinases by intronic polyA activation. Mol. Cell *43*, 927–939.

Wang, P., Yan, B., Guo, J.-T., Hicks, C., and Xu, Y. (2005). Structural genomics analysis of alternative splicing and application to isoform structure modeling. Proc. Natl. Acad. Sci. USA *102*, 18920–18925.

Yanagisawa, M., Huveldt, D., Kreinest, P., Lohse, C.M., Cheville, J.C., Parker, A.S., Copland, J.A., and Anastasiadis, P.Z. (2008). A p120 catenin isoform switch affects Rho activity, induces tumor cell invasion, and predicts metastatic disease. J. Biol. Chem. *283*, 18344–18354.

Yang, F., Petsalaki, E., Rolland, T., Hill, D.E., Vidal, M., and Roth, F.P. (2015). Protein domain-level landscape of cancer-type-specific somatic mutations. PLoS Comput. Biol. *11*, e1004147.

Yang, X., Coulombe-Huntington, J., Kang, S., Sheynkman, G.M., Hao, T., Richardson, A., Sun, S., Yang, F., Shen, Y.A., Murray, R.R., et al. (2016). Widespread expansion of protein interaction capabilities by alternative splicing. Cell *164*, 805–817.

Zhou, C., Licciulli, S., Avila, J.L., Cho, M., Troutman, S., Jiang, P., Kossenkov, A.V., Showe, L.C., Liu, Q., Vachani, A., et al. (2013). The Rac1 splice form Rac1b promotes K-ras-induced lung tumorigenesis. Oncogene *32*, 903–909.

Zong, F.Y., Fu, X., Wei, W.J., Luo, Y.G., Heiner, M., Cao, L.J., Fang, Z., Fang, R., Lu, D., Ji, H., and Hui, J. (2014). The RNA-binding protein QKI suppresses cancer-associated aberrant splicing. PLoS Genet. *10*, e1004289.

Appendix C

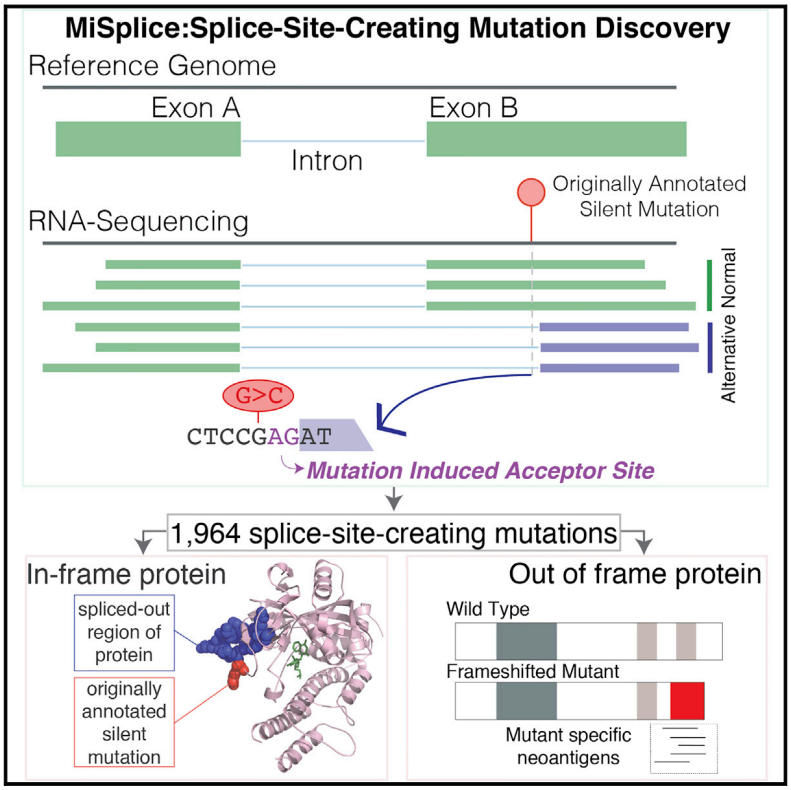
Systematic Analysis of Splice-Site-Creating Mutations in Cancer

Resource

Cell Reports

Systematic Analysis of Splice-Site-Creating Mutations in Cancer

Graphical Abstract



Authors

Reyka G. Jayasinghe, Song Cao, Qingsong Gao, ..., Ilya Shmulevich, Feng Chen, Li Ding

Correspondence

ilya.shmulevich@systemsbiology.org (I.S.), fchen@wustl.edu (F.C.), lding@wustl.edu (L.D.)

In Brief

Jayasinghe et al. identify nearly 2,000 splice-site-creating mutations (SCMs) from over 8,000 tumor samples across 33 cancer types. They provide a more accurate interpretation of previously mis-annotated mutations, highlighting the importance of integrating data types to understand the functional and the clinical implications of splicing mutations in human disease.

Highlights

- MiSplice applied to PanCancer data identifies 1,964 splice-site-creating mutations
- 26% and 11% of SCMs had been previously mis-annotated as missense and silent mutations
- SCMs may be more immunogenic than are missense mutations
- A mini-gene functional assay validates 10 of 11 predicted SCMs

Jayasinghe et al., 2018, Cell Reports *23*, 270–281
April 3, 2018
https://doi.org/10.1016/j.celrep.2018.03.052

CellPress

OPEN ACCESS
CellPress

Cell Reports
Resource

Systematic Analysis of Splice-Site-Creating Mutations in Cancer

Reyka G. Jayasinghe,[1,2,3,20] Song Cao,[1,2,3,20] Qingsong Gao,[1,2,3] Michael C. Wendl,[2,3,4,5] Nam Sy Vo,[6] Sheila M. Reynolds,[7] Yanyan Zhao,[1,2,3] Héctor Climente-González,[8,9,10] Shengjie Chai,[11,12] Fang Wang,[6] Rajees Varghese,[1,13] Mo Huang,[1,2] Wen-Wei Liang,[1,2,3] Matthew A. Wyczalkowski,[1,2,3] Sohini Sengupta,[1,2,3] Zhi Li,[14,15] Samuel H. Payne,[16] David Fenyö,[14,15] Jeffrey H. Miner,[1,13] Matthew J. Walter,[1,17] The Cancer Genome Atlas Research Network, Benjamin Vincent,[11,12] Eduardo Eyras,[18,19] Ken Chen,[6] Ilya Shmulevich,[7,21,*] Feng Chen,[1,13,21,*] and Li Ding[1,2,3,4,17,21,22,*]

[1]Department of Medicine, Washington University in St. Louis, St. Louis, MO 63110, USA
[2]McDonnell Genome Institute, Washington University in St. Louis, St. Louis, MO 63108, USA
[3]Division of Oncology, Washington University in St. Louis, St. Louis, MO 63110, USA
[4]Department of Genetics, Washington University in St. Louis, St. Louis, MO 63110, USA
[5]Department of Mathematics, Washington University in St. Louis, St. Louis, MO 63130, USA
[6]Department of Bioinformatics and Computational Biology, The University of Texas MD Anderson Cancer Center, Houston, TX, USA
[7]Institute for Systems Biology, Seattle, WA 98109, USA
[8]Institut Curie, 75248 Paris Cedex, France
[9]MINES ParisTech, PSL-Research University, CBIO-Centre for Computational Biology, 77300 Fontainebleau, France
[10]INSERM U900, 75248 Paris Cedex, France
[11]Lineberger Comprehensive Cancer Center, The University of North Carolina at Chapel Hill, Chapel Hill, NC, USA
[12]Curriculum in Bioinformatics and Computational Biology, The University of North Carolina at Chapel Hill, Chapel Hill, NC, USA
[13]Division of Nephrology, Washington University in St. Louis, St. Louis, MO 63110, USA
[14]Department of Biochemistry and Molecular Pharmacology, New York University School of Medicine, New York, NY 10016, USA
[15]Institute for Systems Genetics, New York University School of Medicine, New York, NY 10016, USA
[16]Biological Sciences Division, Pacific Northwest National Laboratory, Richland, WA, USA
[17]Siteman Cancer Center, Washington University in St. Louis, St. Louis, MO 63110, USA
[18]Catalan Institution of Research and Advanced Studies (ICREA), 08010 Barcelona, Spain
[19]Computational RNA Biology Group, Pompeu Fabra University (UPF), 08003 Barcelona, Spain
[20]These authors contributed equally
[21]Senior author
[22]Lead Contact
*Correspondence: ilya.shmulevich@systemsbiology.org (I.S.), fchen@wustl.edu (F.C.), lding@wustl.edu (L.D.)
https://doi.org/10.1016/j.celrep.2018.03.052

SUMMARY

For the past decade, cancer genomic studies have focused on mutations leading to splice-site disruption, overlooking those having splice-creating potential. Here, we applied a bioinformatic tool, MiSplice, for the large-scale discovery of splice-site-creating mutations (SCMs) across 8,656 TCGA tumors. We report 1,964 originally mis-annotated mutations having clear evidence of creating alternative splice junctions. *TP53* and *GATA3* have 26 and 18 SCMs, respectively, and *ATRX* has 5 from lower-grade gliomas. Mutations in 11 genes, including *PARP1*, *BRCA1*, and *BAP1*, were experimentally validated for splice-site-creating function. Notably, we found that neoantigens induced by SCMs are likely several folds more immunogenic compared to missense mutations, exemplified by the recurrent GATA3 SCM. Further, high expression of PD-1 and PD-L1 was observed in tumors with SCMs, suggesting candidates for immune blockade therapy. Our work highlights the importance of integrating DNA and RNA data for understanding the functional and the clinical implications of mutations in human diseases.

INTRODUCTION

Large-scale sequencing studies, such as The Cancer Genome Atlas (TCGA) project, have worked to address the functional consequences of genomic mutations in tumors (Dees et al., 2012; Kandoth et al., 2013; Lawrence et al., 2013; Niu et al., 2016), with the larger goal of determining the underlying mechanisms of cancer initiation and progression. Many studies have focused on characterizing (1) non-synonymous somatic mutations that alter amino acid sequence and (2) splice-disrupting mutations at splice donors and acceptors (Jung et al., 2015). Current annotation methods typically classify mutations as disruptors of splicing if they fall on either the consensus intronic dinucleotide splice donor, GT, or the splice acceptor, AG. As a group, splice site mutations have been presumed to be invariably deleterious because of their disruption of the conserved sequences that are used to identify exon-intron boundaries.

While this classification method has been useful, increasing evidence suggests that splice site mutations can lead to transcriptional changes beyond disruption of the canonical junction

(Lim and Fairbrother, 2012; Mort et al., 2014; Rivas et al., 2015; Sauna and Kimchi-Sarfaty, 2011; Steffensen et al., 2014). One such example is the c.190 mutation in *BRCA1*. Conventional annotation had predicted a missense mutation, p.C64G, but our analysis of RNA sequencing (RNA-seq) data in ovarian tumors harboring p.C64G and a published mouse model (Yang et al., 2003) suggested the germline c.190 mutation leads to the creation of an alternative splice junction, resulting in a truncated null protein. Several case studies have reported observations of missense and silent mutations activating cryptic splice sites in *MLH1* (Nyström-Lahti et al., 1999), *LMNA* (Woolfe et al., 2010), *RB1* (Zhang et al., 2008), *RNASEH2A* (Rice et al., 2013), *MECP2* (Sheikh et al., 2013), *BAP1* (Wadt et al., 2012), and *KIT* (Chen et al., 2005), and other studies relate missense and silent mutations to splicing changes (Jung et al., 2015; Kahles et al., 2016; Soemedi et al., 2017; Supek et al., 2014). Despite the broad clinical ramifications of mutation-induced altered splicing, a systematic evaluation of their occurrence and the resultant effects in cancer has yet to be undertaken, and there have not been significant bioinformatics platforms for doing so.

We developed a bioinformatic tool called MiSplice (mutation-induced splicing) that integrates DNA and RNA-seq data across thousands of samples to discover mutations that induce splice site creation. In our large-scale analysis across 8,656 tumor samples, we report 1,964 such somatic mutations that had originally been mis-annotated. Splice-site-creating mutations (SCMs) are enriched in the new introns, with the highest rate at the −3 nt position of acceptors with two-thirds of such events at aGag and agGag repeats by creating an alternative junction 2 nt away. Partial and full splice creation capabilities across these 1,964 sites were evaluated by measuring the fraction of reads supporting the alternative junction, which we termed the "junction allele fraction" (JAF) and which is found to be negatively correlated with distance to the new splice site. In total, 1,607 genes harbor SCMs, with 248 of them having more than one mutation, including *TP53, GATA3, ATRX*, and *NF1*. Recurrent SCMs were found in *TP53*, *GATA3*, *DDX5*, *KDM6A*, *PTEN*, *SETD2*, *SMAD4*, *BCOR*, *SPOP*, and *BAP1*, suggesting an association with cancer development. Broadly speaking, integrated DNA and RNA data can furnish a sound basis for discovering SCMs and for accurately understanding functional consequences of mutations in cancer and in other human diseases.

RESULTS

Splice-Site-Creating Mutation Discovery

We collected high-quality mutation calls from 8,656 tumors across 33 cancer types derived from The Cancer Genome Atlas having available TCGA RNA-seq data (STAR Methods). For every mutation, we defined a set of control samples in the same cancer cohort that lacked the same mutation in the gene of interest. We sought to assess the landscape of SCMs across cancer genomes by evaluating all mutations already having conventional annotations and their potential splice-site-creating effects (Figure 1A). To achieve this goal, we conducted analysis using a bioinformatic tool, MiSplice (mutation-induced splicing), that systematically evaluates mutations in a splicing context using RNA-seq data (Figure 1B).

MiSplice manages large analyses using parallel computation to search for alternative splice junctions within windows of ±20 bp from the mutation of interest. For example, of the 1,416,566 candidate mutations examined here, 4,448 had five or more unique RNA-seq reads supporting the predicted alternative junction in proximity to the mutation. MiSplice then conducts a series of further evaluations, including Ensembl-based filtering of canonical junctions, establishing observational significance by case comparison to a matched set of controls, and assessing score and depth of each cryptic site using MaxEntScan (Yeo and Burge, 2004) and SamTools (Li et al., 2009). From the resultant subset, MiSplice filters out human leukocyte antigen (HLA) genes and sites whose junctions have insufficient difference of expression, as judged from the case-control assessment. Here, we evaluated promising alternative junctions with at least 5% of paired-end RNA-seq reads at the genomic location supporting the alternative junction of interest.

MiSplice processing revealed 2,056 mutations (Table S1) that potentially create an alternative splice site. Manual review indicated a 2.09% false-positive rate, suggesting high specificity of the MiSplice algorithm for discovering these types of mutation-induced splicing events. Of these putative splice events, 1.90% and 0.47% are considered complex and are in highly homologous gene regions, respectively, so they were excluded from further analyses (STAR Methods).

Of the final 1,964 SCMs passing manual review (Table S1), 52% (1,016) are in annotated splice sites, suggesting disruption of the canonical splice site and selection of a the alternative splice site nearby (Figure 1C). Importantly, 26% (513) and 11% (208) of the SCMs had previously been mis-annotated as missense and silent mutations, respectively. In addition, we found 58 insertions or deletions, 46 nonsense, and 123 non-coding region mutations that likewise create cryptic splicing sites.

Molecular and Biological Patterns of SCMs

Next, we characterized the sequence context for the 1,790 SCMs corresponding to single nucleotide mutations. The sequences of each 9-mer (donor) and 23-mer (acceptor) covering the mutation position were extracted for both the mutant and the reference sequences. Their splice scores as potential donor or acceptor sites were then estimated using MaxEntScan (Table S1).

Mutations near the alternative splice junctions show higher mutation rates in the introns for both 5′ ($p < 1 \times 10^{-5}$, binomial test) and 3′ splice site ($p < 1 \times 10^{-6}$) (Figure 2A). More interestingly, we found an enrichment of mutations at the third nucleotide position in the intron, but depletion at the first and second positions (especially for 3′ splice site) (Figure 2A). Comparison of splicing scores between splice-site-creating mutants and reference forms shows that most mutants have stronger splice signals than the reference (Figure 2B). Mutations that create a G or T to produce an alternative 5′ splice site dramatically increase splice site strength. For 3′ splice sites, mutations enriched on the third nucleotide of the newly created intron showed the largest increase of splicing score (Figure 2B). Further examination of the sequence context around mutations at the third

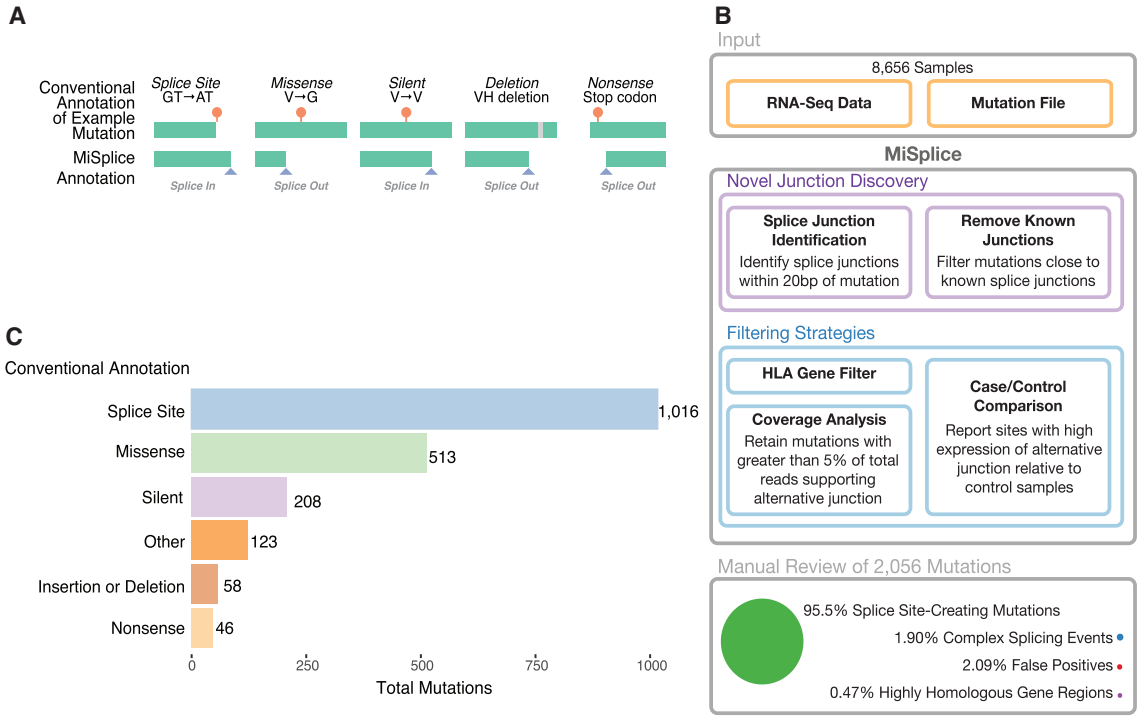


Figure 1. Splice-Site-Creating Mutation Discovery

(A) Examples of splice-site-creating mutations for different conventionally annotated mutation types. Splice-in is defined as mutations contained within the newly created exons, and splice-out is when the mutation is present in the newly created intron.

(B) The MiSplice workflow consists of three steps: alternative junction discovery, filtering, and manual review. First, the user inputs the locations of RNA-seq BAM files along with a mutation file. MiSplice searches the BAM file to identify any alternative splice junctions near the mutation of interest, while filtering out known splice junctions and calculating the number of alternative junction-supporting reads for case and control samples. For the filtering step, the following sites are removed: mutations in HLA genes, a low fraction of reads supporting the alternative splice junction, and sites expressed in controls. Finally, we manually reviewed all sites to validate the *in silico* predictions.

(C) Breakdown of 2,056 manually validated splice-site-creating mutations by conventional annotation.

nucleotide of 3′ splice sites shows that 53% have a mutation on aGag repeats and another 16% are mutated on agGag repeats, all creating alternative junctions 2 nt away from the annotated ones (Figure 2C). Mutations at the −3 position of the alternative acceptor site would potentially enhance *U2AF1* recognition of the acceptor splice site. Previous studies have reported S34F *U2AF1* mutants preferentially skip exons that contain a T nucleotide at the −3 position (Okeyo-Owuor et al., 2015). Of the 192 mutations located at the −3 position from the alternative junction and that contain an AG in the −2 and −1 positions, 56% undergo a G > C transversion (21% G > A, 18% G > T, 3% C > T, 2% A > C, 1% A > T), with C being the preferred base at the −3 position for *U2AF1* binding (Figure 2D).

We also explored the relationship between the alternative and canonical splice junctions. As expected, mutations at splice sites dramatically reduced splice scores of the canonical splice junctions, while strengthening those at the alternative splice junctions in most cases. In contrast, a subset of missense and silent mutations did not drastically alter the canonical junction, but instead preferentially strengthened a nearby alternative splice site (Figure 2E). When analyzing the raw splicing scores (canonical and alternative site before and after mutation), we found that 1,089 out of 1,790 (61%) events showed higher splice score for the alternative splice site than the canonical site, indicating inclination for the alternative sites. Further, while 485 (27%) events saw lower post-mutation alternative splice score, differences between alternative and canonical scores had decreased, suggesting that these mutations are still likely enhancing the preference for the alternative site. Only 214 (12%) events did not show evidence, suggesting increased post-mutational preference for using the alternative site. These cases are a good illustration of the fact that many other genomic splicing features are also relevant, including exonic splicing enhancers (ESE), polypyrimidine tract, branch point, and RNA-binding proteins. They are also consistent with the general view that splice score is not definitive (Jian et al., 2014). We emphasize that all 1,790 alternative splice sites demonstrated usage based on patient RNA-seq data and that 10 out of 11 (>90%) identified SCMs were validated experimentally (see below).

Expressivity and Penetrance of SCMs

In the presence of the mutation, alternative splice junctions exhibited a wide range of expression. To quantify this effect, we measured alternative junction expression as the fraction of alternatively spliced junction spanning reads over the total number of reads at the genomic location, what we refer to as the JAF.

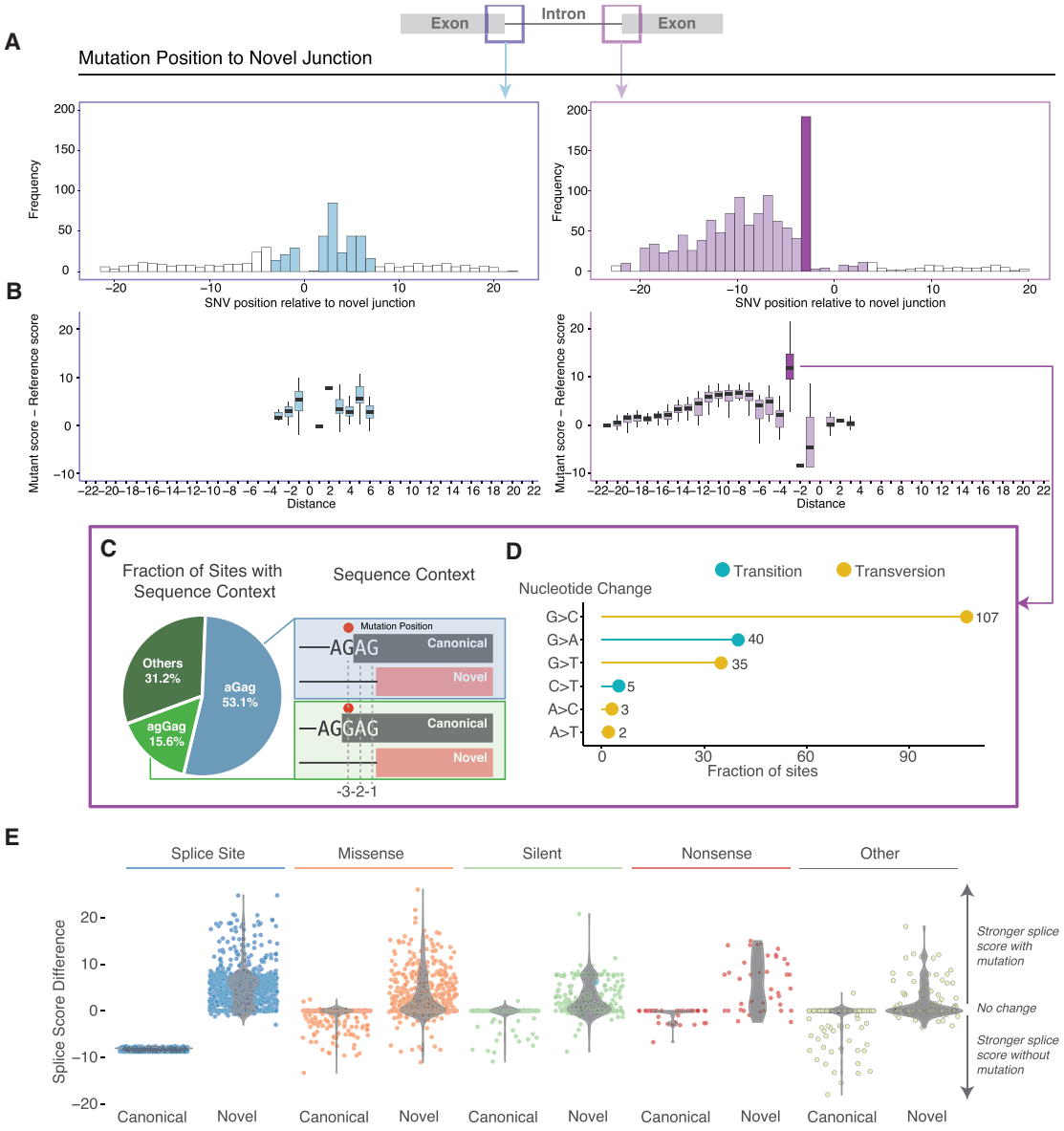


Figure 2. Sequence Contexts and Characteristics of Splice-Site-Creating Mutations

(A) Frequency distribution of splice-site-creating mutations relative to the newly created splice junction, with high frequency shown at the third nucleotide position in the newly created intron.

(B) Comparison of splicing scores for the newly created splice site, before (reference) and after the mutation (mutant). A larger effect of mutations at the third nucleotide position in the intron (especially for the 3′ splice sites) is shown.

(C) Dominant nucleotide sequence context for splice-site-creating mutations at −3 position of the 3′ splice site. Mutation position (red dot) is present 3 base pairs away from the newly created exon.

(D) Transition and transversion rate at the −3 position of the 3′ splice site. Most mutations are G > C transversions, strengthening the consensus sequence of the splicing factor U2AF1.

(E) Comparison of splicing scores between the nearest canonical splice junction with and without a mutation compared to the newly created splice junction with and without a mutation. Most mutations strengthen the alternative splice junction relative to the canonical splice junction.

Figure 3A shows the distribution of JAF's for all high confidence MiSplice predicted alternative junctions, separated by conventional mutation annotations (Figure 3A). Currently, we use a JAF cutoff of 5% for reporting the final high-confidence sites. However, there are some potential alternative sites excluded by this cutoff. Our analysis revealed alternative junction expression varies widely. As expected, DNA variant allele fraction (VAF) and JAF have a generally positive correlation (Figure 3B), with SCMs in *KDM6A* and *FGFR2* having >75% DNA VAF and JAF. However, a SCM in *ARID1A* has a DNA VAF of 23% and JAF of 67%. Such large ranges have been noted for mutations outside of the splice site (Broeks et al., 2003; Clarke et al.,

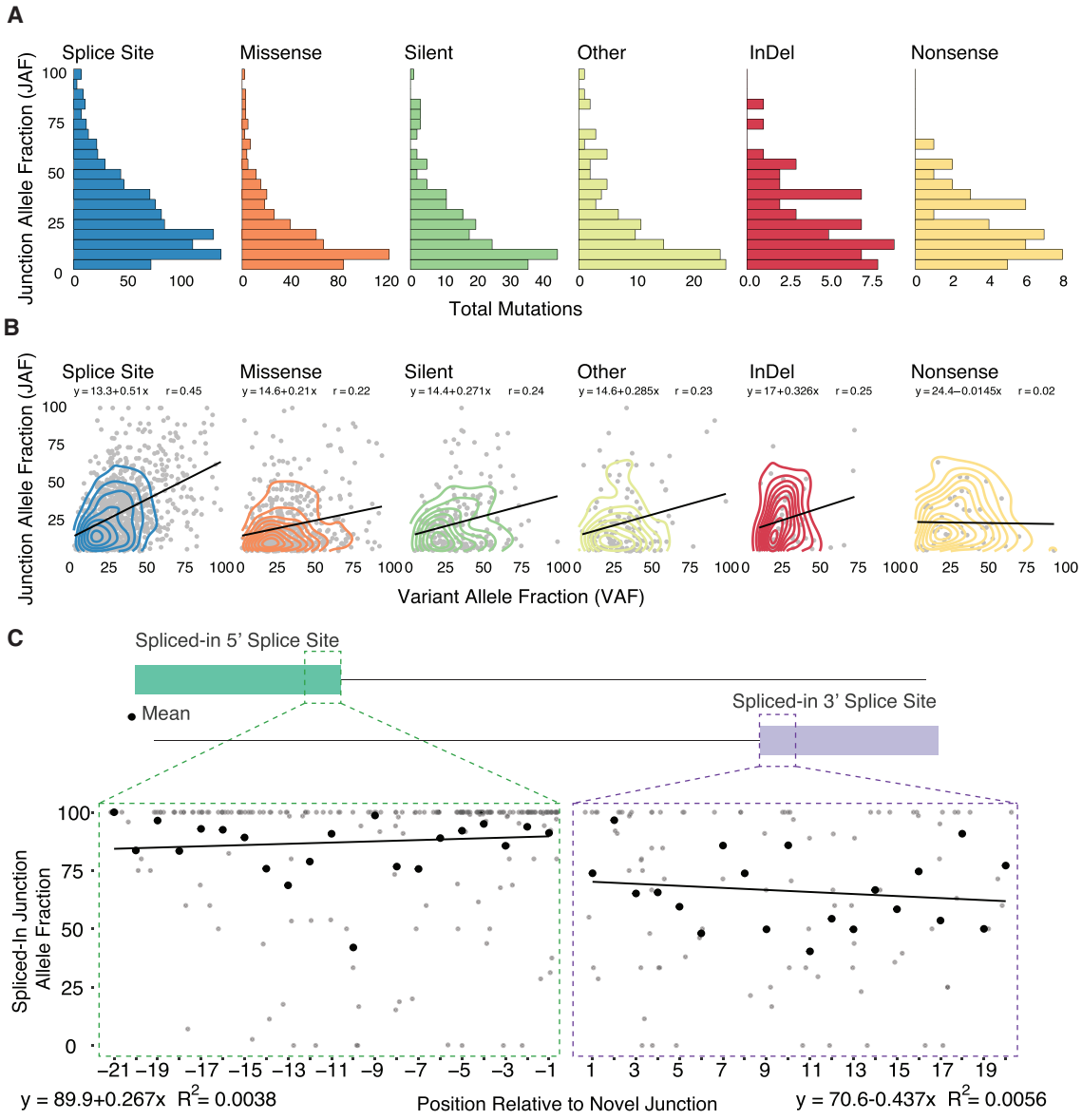


Figure 3. Junction Allele Fraction of Splice-Site-Creating Mutations

(A) The junction allele fraction (JAF) is defined as the number of reads supporting the alternative spliced junction relative to total junction spanning reads. Distribution of JAF values separated by conventional annotation type.

(B) JAF versus DNA variant allele fraction (VAF) comparison by conventional annotation type. Most mutation types show a generally positive correlation between JAF and VAF values.

(C) Splice-site-creating mutations expressed in the newly created exon of the alternative splice junction. Comparison of mutation position relative to the percent of reads supporting the alternative junction and mutation (spliced-in JAF). The mean of each position is highlighted by the black point. For all positions, there is a strong correlation between the presence of the splice-site-creating mutation and the alternative splice junction.

2000; Venables, 2004). Both the truncated and normal spliced products can be observed for many variants, due to either the wild-type allele or leaky splicing, for example, as observed in *RNASEH2A, NFU1, SMN1, CFTR,* and *NF2* (Boerkoel et al., 1995; Caminsky et al., 2014; Ferrer-Cortès et al., 2016; Lohmann and Gallie, 2004; Mautner et al., 1996; Pagani et al., 2003; Rice et al., 2013; Svenson et al., 2001; Vezain et al., 2011).

Next, we considered the expression of mutations that are spliced-in, i.e., mutations located within the exon of the alternatively spliced product. To this end, we determined the ratio of the number of alternative junction reads containing the mutation versus total number of reads supporting the alternative junction (Figure 3C; Table S1). Overall, most of the reads supporting the alternative junction also support the mutation, a finding that suggests a strong association between the mutation and alternative splice junction. Regarding the 5′ splice site, mutations within the first 6 bp of the new exon junction have a much higher fraction of alternative junction reads supporting them; and we see an

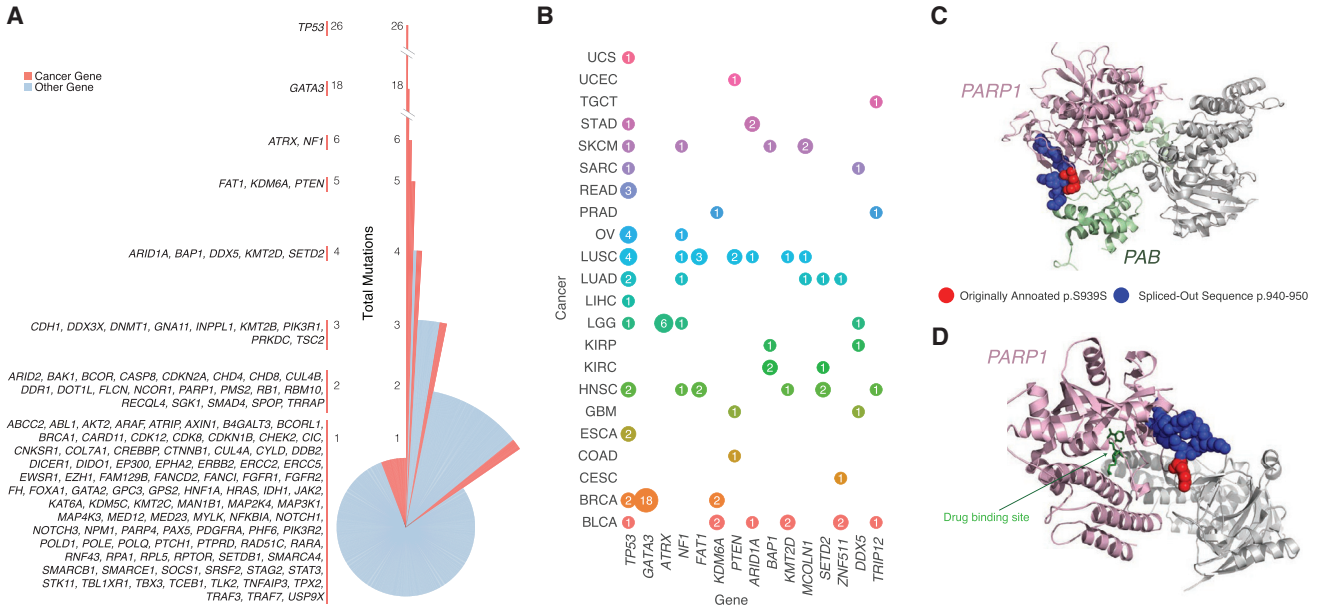


Figure 4. Splice-Site-Creating Mutations across Genes and Cancer Types

(A) Distribution of splice-site-creating mutations in each gene separated by the total number of mutations in each gene. *TP53* has the largest number of splice-site-creating mutations, followed by *GATA3* and *ATRX*.

(B) Genes with the highest number of pancancer splice-site-creating mutations. Circle size correlates with the total number of mutations for each gene (labeled inside circle) and colored by cancer type. Splice-site-creating mutations in *TP53* are present in many cancer types, while mutations in *ATRX* and *GATA3* are specific to LGG and BRCA, respectively.

(C) Proteins Timeless (PAB domain) and PARP1 (chain A) are colored green and pink, respectively. Originally annotated p.S939S mutation (red) and spliced-out sequence (blue) are highlighted on PARP1 (chain A).

(D) 3D protein structure of PARP1 in complex with an inhibitor (PDB ID: 5WRQ). Drug inhibitor and PARP1 (chain A) are indicated in green and pink, respectively.

inverse correlation between the mutation and the junction as the distance between them increases. For the 3′ splice site, we observe a similar trend, although with a higher variability as a function of the distance from the alternative junction.

SCMs across Genes and Cancer Types

A total of 1,607 unique genes harbored SCMs, with 85% (1,359) having one mutation and 15% (248) having two or more. *TP53* contained the greatest number (26), followed by *GATA3* (18). While most SCMs were found outside the current cancer gene compendium (Table S1), Figure 4A shows that a remarkable number of cancer genes harbor splice altering variants, a phenomenon supported in the literature (Sebestyén et al., 2016). A pan-cancer view reveals that *TP53* was the most mutated across cancer types, while 18 *GATA3* mutations and 6 *ATRX* mutations were specific to breast cancer (BRCA) and lower-grade glioma (LGG), respectively.

We observed 137 mutations nearby to one another (±5 bp) which lead to the creation of the same recurrent splice-site-creating events, not only in *TP53* but also in *GATA3*, *DDX5*, *KDM6A*, *SETD2*, *PTEN, SPOP,* and *BAP1*. While some mutations did not occur at the same position, 14 mutations creating the same alternative splice junction were found in the same exon, including 2 mutations in the third exon of *BAK1*. While most mutations in close proximity created the same alternative splice junction, two adjacent SCMs in *CTNND1* and 2 nearby exonic mutations in *ACP2* and *GMPPB* created different alternative junctions.

SCMs can impact protein structure and have potential therapeutic implications. Poly ADP-ribose polymerase 1 (*PARP1*) is an enzyme involved in recruiting protein members of DNA repair pathways including Timeless PAB (PARP1 binding domain) (Figure 4C) (Xie et al., 2015). Since *PARP1* is essential to many cellular processes, including DNA repair, it is commonly targeted by antitumor agents (Malyuchenko et al., 2015). *PARP1* inhibitors targeting the catalytic domain disrupt DNA repair mechanisms thereby increasing the effectiveness of chemotherapeutic agents (Figure 4D). Identifying mutations that disrupt inhibitor binding are essential to properly evaluate treatment options. MiSplice identified a conventionally annotated silent *PARP1* mutation (p.S939S) in a lung squamous cell carcinoma (LUSC) patient that acts as a splice-site-creating variant by creating a *de novo* donor site (Figure 5A). 82 reads supported the *de novo* donor site, which results in a 10 amino acid deletion (p.940-p.950) that falls within the catalytic domain (Figure 4D). Out of 173 LUSC control samples, none contained reads supporting the alternative junction, providing strong evidence that the annotated "silent" mutation is actually a SCM. Previous reports of missense mutations at p.940 are predicted to reduce *PARP1* enzymatic activity by disrupting the binding affinity of *PARP1* to its substrate NAD+ (Alshammari et al., 2014). The in-frame SCM likely disturbs the local structure of *PARP1* and thereby disrupts the interactions between PARP1, its protein binding partners, and drugs binding within the pocket (Figures 4C and 4D).

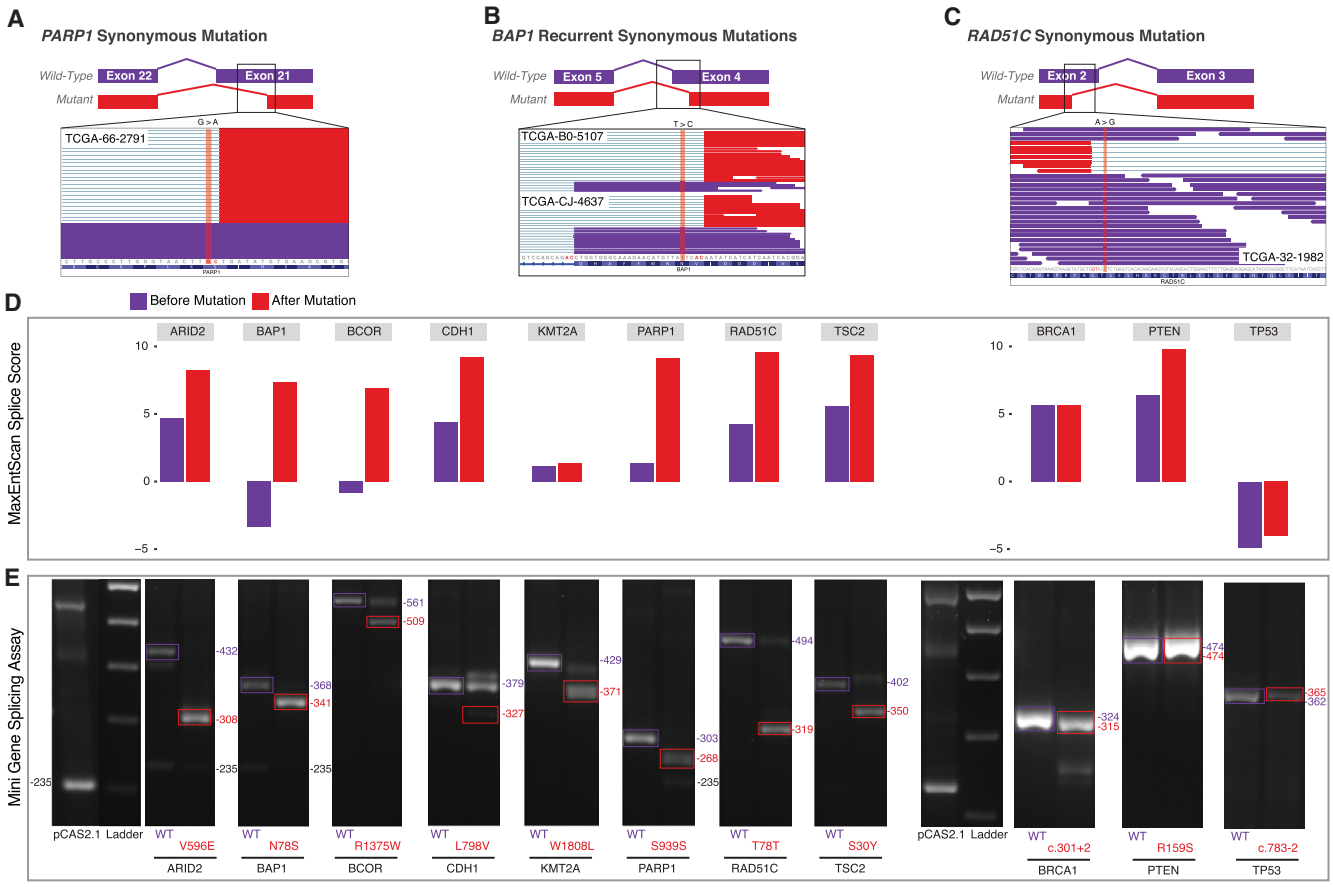


Figure 5. Minigene Functional Assay of Splice-Site-Creating Mutations

(A) Integrative genomics viewer (IGV) screenshot of the conventionally annotated synonymous mutation in *PARP1* in exon 21. RNA-seq reads of the candidate splice-site-creating mutation reveal the creation of an alternative splice site (red reads) created by the conventionally annotated synonymous mutation.

(B) Candidate recurrent splice-site-creating mutations in *BAP1*. Conventionally annotated as synonymous variants, the *BAP1*-mutated region shows alternatively spliced reads (red reads) in the IGV screenshot for each sample with the splice-site-creating mutation.

(C) IGV screenshot of a conventionally annotated synonymous mutation in *RAD51C* in exon 2.

(D) Maximum entropy score of the splice-site-creating variant before (purple) and after (red) the introduced mutation for each variant functionally validated in the mini-gene splicing assay. *In silico* predictions suggest all mutations strengthen the alternative splice site.

(E) Candidate splice-site-creating mutations validated by the mini-gene splicing assay. Exons of interest were cloned into the pCAS2.1 vector and mutant (red); wild-type (purple) plasmids were transfected into 293T cells; and total RNA was extracted to identify mutation-induced alternatively spliced products.

We identified two kidney renal clear cell carcinoma (KIRC) samples having the same conventionally annotated missense mutation (c.233A > G, p.N78S) in *BAP1*, a nuclear deubiquitinase, that created the same spliced-out alternative splicing product (Figure 5B). Inactivation of *BAP1* is prevalent among renal cell carcinomas (Peña-Llopis et al., 2012) and an annotated missense mutation (p.L570V) has been reported to create a cryptic splice site in melanoma (Wadt et al., 2012). At the transcriptional level, the expressions of the case and control samples are relatively comparable, but at the translational level, one case with available protein data (RPPA) showed significantly lower expression ($p = 0.044$, permutation test) relative to the controls (Figure S1; Table S2). This result suggests the conventionally annotated missense mutations in *BAP1* likely create an alternatively spliced transcript that is not readily expressed at the protein level.

We used a pCAS2.1 splicing reporter mini-gene functional assay that was adapted from previous publications (Bonnet et al., 2008; Gaildrat et al., 2010; Malone et al., 2016; Tournier et al., 2008; Vreeswijk and van der Klift, 2012), to validate SCMs in 11 cancer genes, including two originally annotated silent mutations in *PARP1*, *RAD51C*, two splice site mutations in *TP53* and *BRCA1*, and several missense mutations in *ARID2*, *BAP1*, *BCOR*, *CDH1*, *KMT2A, PTEN*, and *TSC2*. Wild-type and mutant exons were cloned into a pCAS2.1 vector (Gaildrat et al., 2010) and transiently transfected into HEK293T cells. Total RNA was extracted to evaluate alternatively spliced products by RT-PCR. Examining the change in the MaxEntScan score for the 11 genes revealed mutations in *ARID2*, *BAP1*, *BCOR*, *CDH1*, *PARP1*, *RAD51C*, *PTEN,* and *TSC2* having dramatically stronger splice scores in the presence of the mutation, while mutations in *BRCA1*, *KMT2A*, and *TP53* did not (Figure 5D). Except for *PTEN*, variants with stronger splice scores showed higher levels of the alternatively spliced product in the mini-gene assay when compared to the wild-type. Variants

with moderate changes in splice score still showed evidence of alternatively spliced transcripts, revealing the importance of utilizing functional assays to evaluate the effect of mutations in a splicing context in addition to *in silico* predictions. The minigene assay confirmed 91% (10/11 genes) splicing alterations in all tested genes and sequencing confirmed the alternatively spliced products (Figure 5E; STAR Methods), suggesting a strong concordance between MiSplice predicted SCMs and the functional assay.

Neoantigens Introduced by SCMs

We have further investigated neoantigens produced by SCMs. By using the RefSeq transcript database, a total of 2,993 protein sequences were translated for transcripts containing mutation-induced alternative splice forms (Table S3). In the translation, we allowed for different transcripts from each SCM. The HLA types for each sample were adopted from the TCGA pancan immune working group (Synapse ID: syn5974636). NetMHC4 and NetMHCpan-3.0 (Andreatta and Nielsen, 2016) were used to predict the binding affinity between epitopes and the major histocompatibility complex (MHC) and showed a high concordance in total predicted neoantigens (Pearson = 0.94; Figure S2). We found that alternative splice forms for some important genes related to tumorigenesis, including *SMARC1*, *KDM6A*, and *NOTCH1,* are highly immunogenic and can contain 40 or more unique neoantigens (Figure 6A) (Dalgliesh et al., 2010; Papadakis et al., 2015). In addition, the mean number of neoantigens across SCMs from NetMHCpan-4.0 and NetMHCpan-3.0 are 2.0 and 2.5, respectively, which are both higher than the average number of around 1 for non-synonymous mutations. Furthermore, 28 genes contain recurrent neoantigen events ($\geq$3) across samples (Figure 6B). In particular, *GATA3* has the highest recurrence and *GATA3* SCMs were mutually exclusive with other mutation types (Figure 6C). The CA deletion at chr8:8111433 disrupts the canonical splice site and an alternative splice site is used for creating the alternative splice form, which results in a frame shifted protein product spanning the Zinc-finger domain (Figures 6D and 6E). 19 unique neoantigen peptide sequences were mapped to the frameshifted protein product for the 16 samples (Figure 6F). We were further able to validate one alternative peptide sequence using mass spectrometry data from a recent proteogenomics study on 77 TCGA breast cancer patients (Mertins et al., 2016). For one sample with the highly recurrent and expressed *GATA3* SCM, we used MSGF+ to search publicly available mass spectrometry data for evidence of alternative GATA3 peptides. Figure 6G shows one identified mass spectrum supporting one alternative *GATA3* peptide, which covers two immunogenic peptides (KPKRRLPG and LIKPKRRLPG) predicted in TCGA-AR-A1AP.

High neoantigen burden is associated with an elevated immune response (Turajlic et al., 2017). To test whether SCMs affect immune response, we compared the expression of T cell markers PD-1, CD8A and CD8B and PD1 immune checkpoint blockades PD-L1 and PD-L2 (Figure 7). We selected six cancer types (BRCA, BLCA, HNSC, LUAD, LUSC, and SKCM) with sufficient samples containing SCMs for adequate statistical power. Both T cell markers (PD-1, CD8A, and CD8B) and immune checkpoint blockade PD-L1 show increased expression in samples with SCMs compared to samples without SCMs (Figure 7), indicating alternative splice forms induced by SCMs increase the overall immunogenicity of these cancers. The highly expressed PD-L1 suggests PD-L1 immunotherapy as potential treatments for samples containing SCMs.

DISCUSSION

In this study, we applied our newly developed bioinformatics tool called MiSplice (mutation-induced splicing) to systematically analyze splice-site-creating events that arise from somatic mutations. Our analysis shows MiSplice reliably identifies SCMs across multiple cancer types. Existing studies have largely focused on splice-disrupting events in known splice sites, but the current study substantially extends our knowledge into the realm of SCMs in human cancer. For instance, we found 1,016 splice site mutations not only disrupt the canonical splice site but also create an alternative splice site. We also found that hundreds of mutations that would traditionally be classified as missense, silent, indel, and nonsense are really acting as SCMs. Many important cancer-related genes harbor these mutations, such as *TP53*, *ATRX*, *BAP1*, *CTNNB1*, *RB1*, etc. It is noteworthy that we found five SCMs in *ATRX* among 288 LGG cases, likely leading to the disruption of ATRX function. A previous study has shown that loss of wild-type *ATRX* is associated with tumor growth in glioma (Koschmann et al., 2016).

Characterization of these alternative splice events show that most SCMs have a higher splice score, as measured by MaxEntScan, in the post-mutation alternative splice site as compared to the reference. These results are consistent with the preferential selection of these alternative sites as new splicing forms. For the splice-site mutation, the splice score associated with the canonical junction is coincidently decreased after mutation. However, while there is no difference in splice scores of canonical junctions before and after missense and silent mutations, the alternative splice site was often strengthened after mutation. This suggests silent and missense mutations instead act as modifiers of splicing by creating or strengthening cryptic sites within the exon as opposed to disrupting the canonical splice site. In addition, we found a significant enrichment of mutations at the -3 position in the 3′ splice site, the two dominant sequence contexts being aGag and agGag, where G is at the -3 position.

In cases in which the mutation is retained in the alternative splice junction, we distinguish mutations with two further categories, splice-in and splice-out. For splice-in mutations, we can characterize the association between mutations and cryptic splicing forms. For example, we found high concordance for RNA-seq reads supporting alternatively spliced junctions and mutations, suggesting the association between mutations and cryptic splicing forms.

The current study has greatly extended insights into the transcriptional ramifications of genomic alterations by identifying nearly 1,964 alternative splice sites introduced by somatic mutations and functionally validating 10 of 11 variants in a mini-gene splicing assay. These events were conventionally annotated as missense, silent, splice site, nonsense, or other mutations when, in fact, we have shown that they often create cryptic splice sites. This relative abundance of the alternative and wild-type product suggests varying levels of junction usage, depending

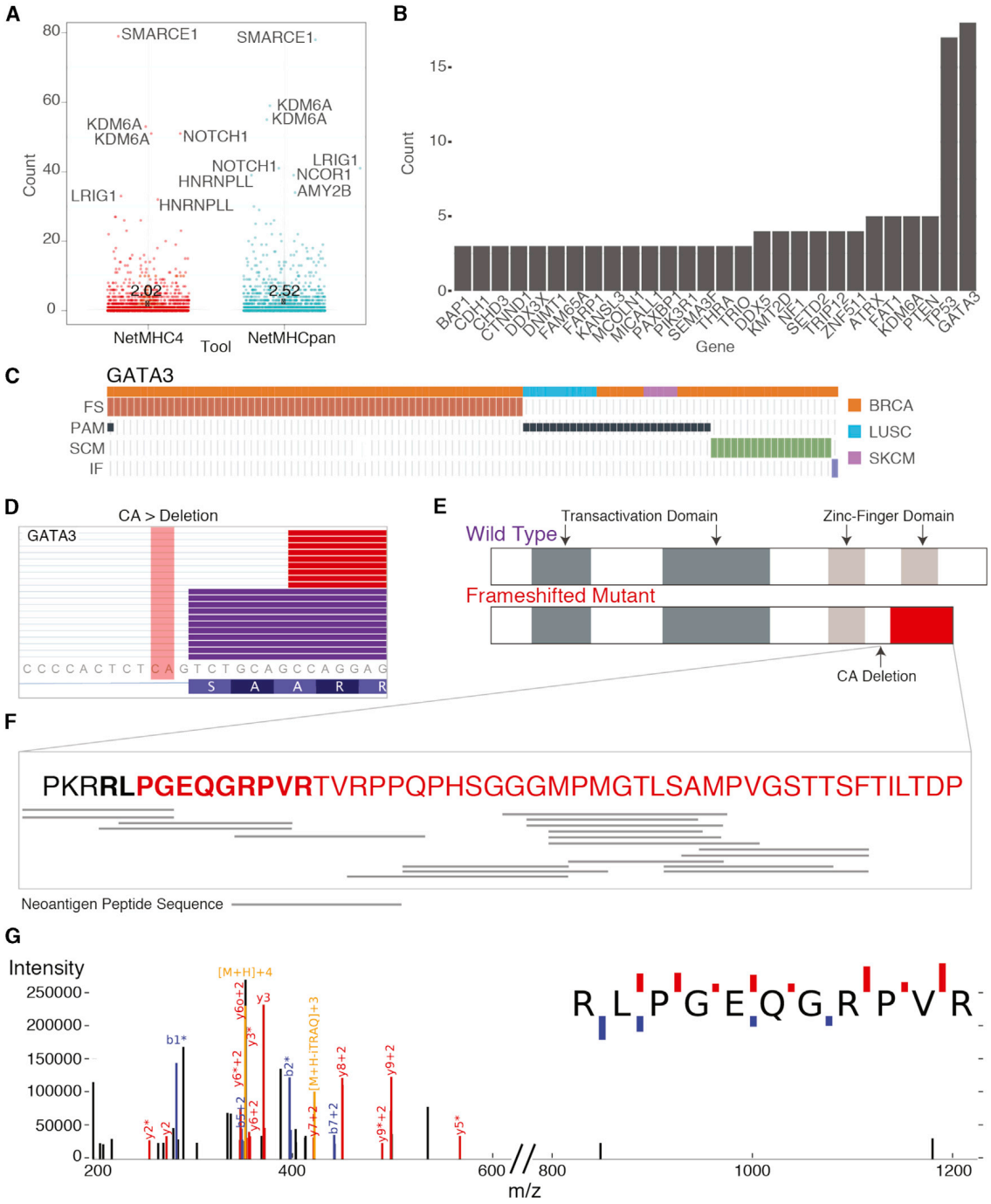


Figure 6. Schematic of GATA3 Splice-Site-Creating Mutations and Neoantigen Predictions

(A) Distribution of neoantigens predicted by NetMHCpan and NetMHC4. Genes with the highest number of neoantigens labeled. Mean value for each tool indicated by X and labeled.

(B) Genes with the largest recurrence of predicted neoantigens across the dataset. *GATA3* shows the highest recurrence.

(C) Mutual exclusivity of protein-affecting mutation (PAM), frameshifting indel (FS), in-frame indel (IF), and splice-site-creating mutations (SCM) in *GATA3*. (D) IGV screenshot of *GATA3* splice-site-creating mutation, which disrupts the canonical splice site and utilizes a cryptic splice site 7 bp downstream. Mutant reads highlighted in red, and normal reads are in purple. CA deletion indicated in the figure.

(E) Predicted functional domains disrupted because of the recurrent splice-site-creating mutation in *GATA3*.

(F) Predicted neoantigen peptide sequences mapped to the frameshifted protein product for samples with *GATA3* SCMs.

(G) Mass spectrum of GATA3 peptide in TCGA-AR-A1AP.

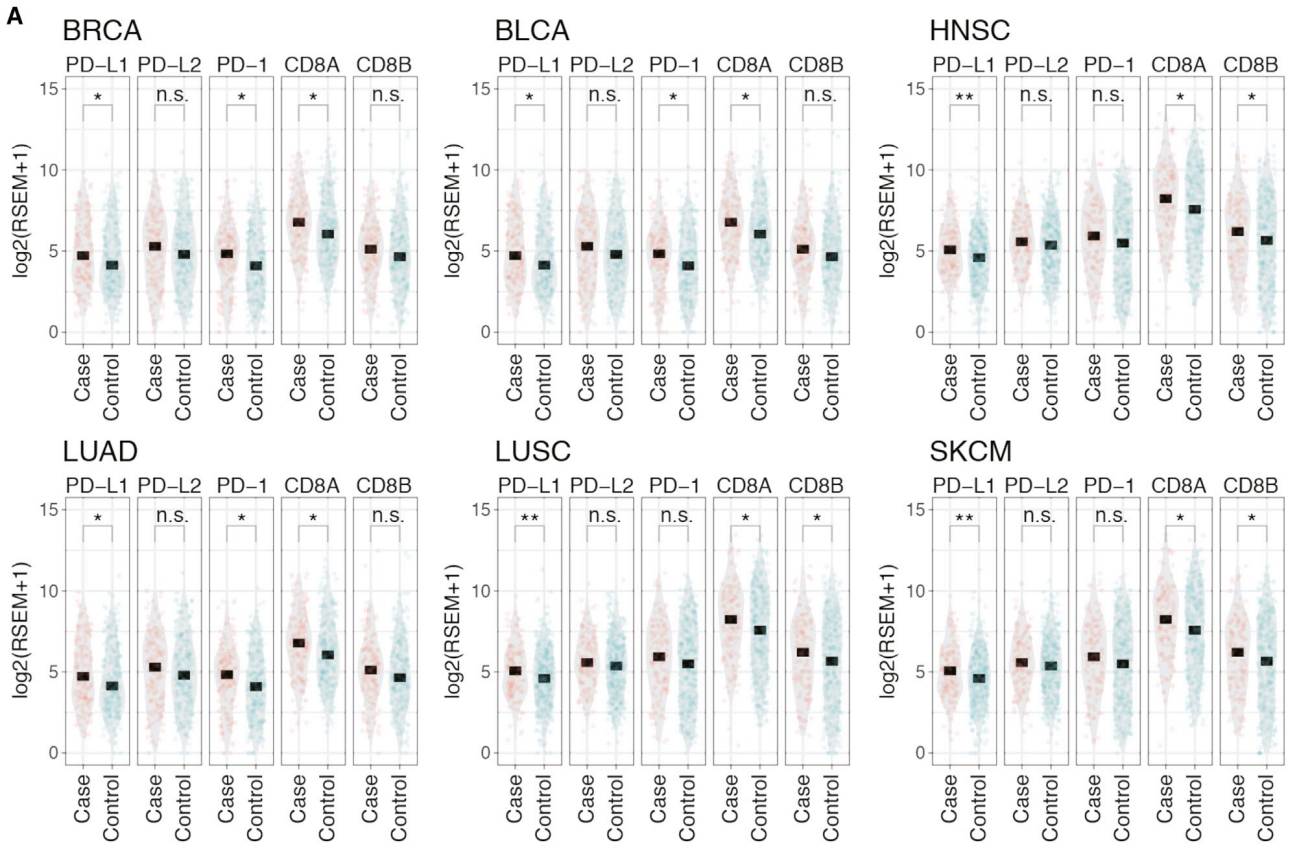


Figure 7. PD-L1, PD-L2, PD-1, CD8A, and CD8B Expression

(A) Expression comparison of PD-L1, PD-L2, and T cell markers PD-1, CD8A, and CD8B between samples with (case) and without (control) SCMs across six cancer types. p values: * less than 0.05; ** < 0.01; and *** < 0.001; ns, not significant.

on the context of the mutation, and emphasizes the importance of validating predictions using a functional assay to understand the full biological consequence. The alternative products may be therapeutically targetable in some cancer patients. For example, targeting neoantigens shows promising results in treating melanoma patients (Carreno et al., 2015). By further evaluating human leukocyte antigen (HLA) genotypes and binding affinities to the MHC, it is likely that new neoantigens from cryptic splice sites may be discovered. The current study reveals that alternative splice forms induced by SCMs are highly immunogenic and correlated with a high T cell immune response and an elevated PD-L1 expression, suggesting the potential for immunotherapy in these samples. Further investigation of the cryptic splice sites by mass spectra or target assay are needed to prioritize therapeutic targets in clinical trials.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- CONTACT FOR REAGENT AND RESOURCE SHARING
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
- METHOD DETAILS
  - Dataset Description
  - MiSplice Pipeline
  - Splice Site Score Estimation
  - Neoantigen Prediction
  - Manual Review
  - Code Availability
  - Mini-gene Splicing Assay
  - Cell Culture
- QUANTIFICATION AND STATISTICAL ANALYSES

SUPPLEMENTAL INFORMATION

Supplemental Information includes six figures and four tables and can be found with this article online at https://doi.org/10.1016/j.celrep.2018.03.052.

ACKNOWLEDGMENTS

Funding supported by U54 HG003273, U54 HG003067, U54 HG003079, U24 CA143799, U24 CA143835, U24 CA143840, U24 CA143843, U24 CA143845, U24 CA143848, U24 CA143858, U24 CA143866, U24 CA143867, U24 CA143882, U24 CA143883, U24 CA144025, and P30 CA016672.

AUTHOR CONTRIBUTIONS

L.D. designed and supervised the research. F.C. supervised the experimental design and the biological evaluations. S. Cao developed the detection scripts for MiSplice. R.G.J. and M.C.W. developed the filtering strategy and the

scripts for MiSplice. Q.G. developed the scoring scripts for MiSplice. R.G.J. and S. Cao performed the discovery of mutation-induced alternative splice sites by using MiSplice. R.G.J., S. Cao, Q.G., W.-W.L., M.H., S.S., H.C.-G., E.E., N.S.V., F.W., Z.L., S.H.P., S.M.R., R.V., M.A.W., J.H.M., S. Chai, and M.C.W. analyzed the data. R.G.J. and Y.Z. conducted the splicing experiments. R.G.J., Q.G., H.C.-G., and S. Cao prepared the figures and the tables. R.G.J., S.C., Q.G., M.C.W., and L.D. wrote the manuscript. F.C., I.S., K.C., E.E., B.V., M.C.W., D.F., M.J.W., and L.D. revised the manuscript.

DECLARATION OF INTERESTS

Michael Seiler, Peter G. Smith, Ping Zhu, Silvia Buonamici, and Lihua Yu are employees of H3 Biomedicine, Inc. Parts of this work are the subject of a patent application: WO2017040526 titled "Splice variants associated with neomorphic sf3b1 mutants." Shouyoung Peng, Anant A. Agrawal, James Palacino, and Teng Teng are employees of H3 Biomedicine, Inc. Andrew D. Cherniack, Ashton C. Berger, and Galen F. Gao receive research support from Bayer Pharmaceuticals. Gordon B. Mills serves on the External Scientific Review Board of Astrazeneca. Anil Sood is on the Scientific Advisory Board for Kiyatec and is a shareholder in BioPath. Jonathan S. Serody receives funding from Merck, Inc. Kyle R. Covington is an employee of Castle Biosciences, Inc. Preethi H. Gunaratne is founder, CSO, and shareholder of NextmiRNA Therapeutics. Christina Yau is a part-time employee/consultant at NantOmics. Franz X. Schaub is an employee and shareholder of SEngine Precision Medicine, Inc. Carla Grandori is an employee, founder, and shareholder of SEngine Precision Medicine, Inc. Robert N. Eisenman is a member of the Scientific Advisory Boards and shareholder of Shenogen Pharma and Kronos Bio. Daniel J. Weisenberger is a consultant for Zymo Research Corporation. Joshua M. Stuart is the founder of Five3 Genomics and shareholder of NantOmics. Marc T. Goodman receives research support from Merck, Inc. Andrew J. Gentles is a consultant for Cibermed. Charles M. Perou is an equity stock holder, consultant, and Board of Directors member of BioClassifier and GeneCentric Diagnostics and is also listed as an inventor on patent applications on the Breast PAM50 and Lung Cancer Subtyping assays. Matthew Meyerson receives research support from Bayer Pharmaceuticals; is an equity holder in, consultant for, and Scientific Advisory Board chair for OrigiMed; and is an inventor of a patent for EGFR mutation diagnosis in lung cancer, licensed to LabCorp. Eduard Porta-Pardo is an inventor of a patent for domainXplorer. Han Liang is a shareholder and scientific advisor of Precision Scientific and Eagle Nebula. Da Yang is an inventor on a pending patent application describing the use of antisense oligonucleotides against specific lncRNA sequence as diagnostic and therapeutic tools. Yonghong Xiao was an employee and shareholder of TESARO, Inc. Bin Feng is an employee and shareholder of TESARO, Inc. Carter Van Waes received research funding for the study of IAP inhibitor ASTX660 through a Cooperative Agreement between NIDCD, NIH, and Astex Pharmaceuticals. Raunaq Malhotra is an employee and shareholder of Seven Bridges, Inc. Peter W. Laird serves on the Scientific Advisory Board for AnchorDx. Joel Tepper is a consultant at EMD Serono. Kenneth Wang serves on the Advisory Board for Boston Scientific, Microtech, and Olympus. Andrea Califano is a founder, shareholder, and advisory board member of DarwinHealth, Inc. and a shareholder and advisory board member of Tempus, Inc. Toni K. Choueiri serves as needed on advisory boards for Bristol-Myers Squibb, Merck, and Roche. Lawrence Kwong receives research support from Array BioPharma. Sharon E. Plon is a member of the Scientific Advisory Board for Baylor Genetics Laboratory. Beth Y. Karlan serves on the Advisory Board of Invitae.

Received: November 6, 2017
Revised: February 21, 2018
Accepted: March 13, 2018
Published: April 3, 2018

REFERENCES

Alshammari, A.H., Shalaby, M.A., Alanazi, M.S., and Saeed, H.M. (2014). Novel mutations of the PARP-1 gene associated with colorectal cancer in the Saudi population. Asian Pac. J. Cancer Prev. *15*, 3667–3673.

Andreatta, M., and Nielsen, M. (2016). Gapped sequence alignment using artificial neural networks: application to the MHC class I system. Bioinformatics *32*, 511–517.

Boerkoel, C.F., Exelbert, R., Nicastri, C., Nichols, R.C., Miller, F.W., Plotz, P.H., and Raben, N. (1995). Leaky splicing mutation in the acid maltase gene is associated with delayed onset of glycogenosis type II. Am. J. Hum. Genet. *56*, 887–897.

Bonnet, C., Krieger, S., Vezain, M., Rousselin, A., Tournier, I., Martins, A., Berthet, P., Chevrier, A., Dugast, C., Layet, V., et al. (2008). Screening BRCA1 and BRCA2 unclassified variants for splicing mutations using reverse transcription PCR on patient RNA and an ex vivo assay based on a splicing reporter minigene. J. Med. Genet. *45*, 438–446.

Broeks, A., Urbanus, J.H.M., de Knijff, P., Devilee, P., Nicke, M., Klöpper, K., Dörk, T., Floore, A.N., and van't Veer, L.J. (2003). IVS10-6T>G, an ancient ATM germline mutation linked with breast cancer. Hum. Mutat. *21*, 521–528.

Caminsky, N., Mucaki, E.J., and Rogan, P.K. (2014). Interpretation of mRNA splicing mutations in genetic disease: review of the literature and guidelines for information-theoretical analysis. F1000Res. *3*, 282.

Carreno, B.M., Magrini, V., Becker-Hapak, M., Kaabinejadian, S., Hundal, J., Petti, A.A., Ly, A., Lie, W.R., Hildebrand, W.H., Mardis, E.R., and Linette, G.P. (2015). Cancer immunotherapy. A dendritic cell vaccine increases the breadth and diversity of melanoma neoantigen-specific T cells. Science *348*, 803–808.

Chen, L.L., Sabripour, M., Wu, E.F., Prieto, V.G., Fuller, G.N., and Frazier, M.L. (2005). A mutation-created novel intra-exonic pre-mRNA splice site causes constitutive activation of KIT in human gastrointestinal stromal tumors. Oncogene *24*, 4271–4280.

Clarke, L.A., Veiga, I., Isidro, G., Jordan, P., Ramos, J.S., Castedo, S., and Boavida, M.G. (2000). Pathological exon skipping in an HNPCC proband with MLH1 splice acceptor site mutation. Genes Chromosomes Cancer *29*, 367–370.

Dalgliesh, G.L., Furge, K., Greenman, C., Chen, L., Bignell, G., Butler, A., Davies, H., Edkins, S., Hardy, C., Latimer, C., et al. (2010). Systematic sequencing of renal carcinoma reveals inactivation of histone modifying genes. Nature *463*, 360–363.

Dees, N.D., Zhang, Q., Kandoth, C., Wendl, M.C., Schierding, W., Koboldt, D.C., Mooney, T.B., Callaway, M.B., Dooling, D., Mardis, E.R., et al. (2012). MuSiC: identifying mutational significance in cancer genomes. Genome Res. *22*, 1589–1598.

Ferrer-Cortès, X., Narbona, J., Bujan, N., Matalonga, L., Del Toro, M., Arranz, J.A., Riudor, E., Garcia-Cazorla, A., Jou, C., O'Callaghan, M., et al. (2016). A leaky splicing mutation in NFU1 is associated with a particular biochemical phenotype. Consequences for the diagnosis. Mitochondrion *26*, 72–80.

Gaildrat, P., Killian, A., Martins, A., Tournier, I., Frébourg, T., and Tosi, M. (2010). Use of splicing reporter minigene assay to evaluate the effect on splicing of unclassified genetic variants. Methods Mol. Biol. *653*, 249–257.

Jian, X., Boerwinkle, E., and Liu, X. (2014). In silico prediction of splice-altering single nucleotide variants in the human genome. Nucleic Acids Res. *42*, 13534–13544.

Jung, H., Lee, D., Lee, J., Park, D., Kim, Y.J., Park, W.Y., Hong, D., Park, P.J., and Lee, E. (2015). Intron retention is a widespread mechanism of tumor-suppressor inactivation. Nat. Genet. *47*, 1242–1248.

Kahles, A., Ong, C.S., Zhong, Y., and Rätsch, G. (2016). SplAdder: identification, quantification and testing of alternative splicing events from RNA-Seq data. Bioinformatics *32*, 1840–1847.

Kandoth, C., McLellan, M.D., Vandin, F., Ye, K., Niu, B., Lu, C., Xie, M., Zhang, Q., McMichael, J.F., Wyczalkowski, M.A., et al. (2013). Mutational landscape and significance across 12 major cancer types. Nature *502*, 333–339.

Koschmann, C., Calinescu, A.A., Nunez, F.J., Mackay, A., Fazal-Salom, J., Thomas, D., Mendez, F., Kamran, N., Dzaman, M., Mulpuri, L., et al. (2016). ATRX loss promotes tumor growth and impairs nonhomologous end joining DNA repair in glioma. Sci. Transl. Med. *8*, 328ra28.

Lawrence, M.S., Stojanov, P., Polak, P., Kryukov, G.V., Cibulskis, K., Sivachenko, A., Carter, S.L., Stewart, C., Mermel, C.H., Roberts, S.A., et al. (2013). Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature *499*, 214–218.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R.; 1000 Genome Project Data Processing Subgroup (2009). The sequence alignment/map format and SAMtools. Bioinformatics *25*, 2078–2079.

Lim, K.H., and Fairbrother, W.G. (2012). Spliceman–a computational web server that predicts sequence variations in pre-mRNA splicing. Bioinformatics *28*, 1031–1032.

Lohmann, D.R., and Gallie, B.L. (2004). Retinoblastoma: revisiting the model prototype of inherited cancer. Am. J. Med. Genet. C. Semin. Med. Genet. *129C*, 23–28.

Malone, A.F., Funk, S.D., Alhamad, T., and Miner, J.H. (2016). Functional assessment of a novel COL4A5 splice region variant and immunostaining of plucked hair follicles as an alternative method of diagnosis in X-linked Alport syndrome. Pediatr. Nephrol. *32*, 997–1003.

Malyuchenko, N.V., Kotova, E.Y., Kulaeva, O.I., Kirpichnikov, M.P., and Studitskiy, V.M. (2015). PARP1 Inhibitors: antitumor drug design. Acta Naturae *7*, 27–37.

Mautner, V.F., Baser, M.E., and Kluwe, L. (1996). Phenotypic variability in two families with novel splice-site and frameshift NF2 mutations. Hum. Genet. *98*, 203–206.

Mertins, P., Mani, D.R., Ruggles, K.V., Gillette, M.A., Clauser, K.R., Wang, P., Wang, X., Qiao, J.W., Cao, S., Petralia, F., et al.; NCI CPTAC (2016). Proteogenomics connects somatic mutations to signalling in breast cancer. Nature *534*, 55–62.

Mort, M., Sterne-Weiler, T., Li, B., Ball, E.V., Cooper, D.N., Radivojac, P., Sanford, J.R., and Mooney, S.D. (2014). MutPred Splice: machine learning-based prediction of exonic variants that disrupt splicing. Genome Biol. *15*, R19–R19.

Nielsen, M., and Andreatta, M. (2016). NetMHCpan-3.0; improved prediction of binding to MHC class I molecules integrating information from multiple receptor and peptide length datasets. Genome Med. *8*, 33.

Niu, B., Scott, A.D., Sengupta, S., Bailey, M.H., Batra, P., Ning, J., Wyczalkowski, M.A., Liang, W.-W., Zhang, Q., McLellan, M.D., et al. (2016). Protein-structure-guided discovery of functional mutations across 19 cancer types. Nat. Genet. *48*, 827–837.

Nyström-Lahti, M., Holmberg, M., Fidalgo, P., Salovaara, R., de la Chapelle, A., Jiricny, J., and Peltomäki, P. (1999). Missense and nonsense mutations in codon 659 of MLH1 cause aberrant splicing of messenger RNA in HNPCC kindreds. Genes Chromosomes Cancer *26*, 372–375.

Okeyo-Owuor, T., White, B.S., Chatrikhi, R., Mohan, D.R., Kim, S., Griffith, M., Ding, L., Ketkar-Kulkarni, S., Hundal, J., Laird, K.M., et al. (2015). U2AF1 mutations alter sequence specificity of pre-mRNA binding and splicing. Leukemia *29*, 909–917.

Pagani, F., Stuani, C., Tzetis, M., Kanavakis, E., Efthymiadou, A., Doudounakis, S., Casals, T., and Baralle, F.E. (2003). New type of disease causing mutations: the example of the composite exonic regulatory elements of splicing in CFTR exon 12. Hum. Mol. Genet. *12*, 1111–1120.

Papadakis, A.I., Sun, C., Knijnenburg, T.A., Xue, Y., Grernrum, W., Hölzel, M., Nijkamp, W., Wessels, L.F., Beijersbergen, R.L., Bernards, R., and Huang, S. (2015). SMARCE1 suppresses EGFR expression and controls responses to MET and ALK inhibitors in lung cancer. Cell Res. *25*, 445–458.

Peña-Llopis, S., Vega-Rubín-de-Celis, S., Liao, A., Leng, N., Pavía-Jiménez, A., Wang, S., Yamasaki, T., Zhrebker, L., Sivanand, S., Spence, P., et al. (2012). BAP1 loss defines a new class of renal cell carcinoma. Nat. Genet. *44*, 751–759.

Rice, G.I., Reijns, M.A., Coffin, S.R., Forte, G.M., Anderson, B.H., Szynkiewicz, M., Gornall, H., Gent, D., Leitch, A., Botella, M.P., et al. (2013). Synonymous mutations in RNASEH2A create cryptic splice sites impairing RNase H2 enzyme function in Aicardi-Goutières syndrome. Hum. Mutat. *34*, 1066–1070.

Rivas, M.A., Pirinen, M., Conrad, D.F., Lek, M., Tsang, E.K., Karczewski, K.J., Maller, J.B., Kukurba, K.R., DeLuca, D.S., Fromer, M., et al.; GTEx Consortium; Geuvadis Consortium (2015). Human genomics. Effect of predicted protein-truncating genetic variants on the human transcriptome. Science *348*, 666–669.

Robinson, J.T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G., and Mesirov, J.P. (2011). Integrative genomics viewer. Nat. Biotechnol. *29*, 24–26.

Sauna, Z.E., and Kimchi-Sarfaty, C. (2011). Understanding the contribution of synonymous mutations to human disease. Nat. Rev. Genet. *12*, 683–691.

Sebestyén, E., Singh, B., Miñana, B., Pagès, A., Mateo, F., Pujana, M.A., Valcárcel, J., and Eyras, E. (2016). Large-scale analysis of genome and transcriptome alterations in multiple tumors unveils novel cancer-relevant splicing networks. Genome Res. *26*, 732–744.

Sheikh, T.I., Mittal, K., Willis, M.J., and Vincent, J.B. (2013). A synonymous change, p.Gly16Gly in MECP2 Exon 1, causes a cryptic splice event in a Rett syndrome patient. Orphanet J. Rare Dis. *8*, 108.

Soemedi, R., Cygan, K.J., Rhine, C.L., Wang, J., Bulacan, C., Yang, J., Bayrak-Toydemir, P., McDonald, J., and Fairbrother, W.G. (2017). Pathogenic variants that alter protein code often disrupt splicing. Nat. Genet. *49*, 848–855.

Steffensen, A.Y., Dandanell, M., Jønson, L., Ejlertsen, B., Gerdes, A.-M., Nielsen, F.C., and Hansen, T.V. (2014). Functional characterization of BRCA1 gene variants by mini-gene splicing assay. European journal of human genetics. Eur. J. Hum. Genet. *3*, 1–7.

Supek, F., Miñana, B., Valcárcel, J., Gabaldón, T., and Lehner, B. (2014). Synonymous mutations frequently act as driver mutations in human cancers. Cell *156*, 1324–1335.

Svenson, I.K., Ashley-Koch, A.E., Pericak-Vance, M.A., and Marchuk, D.A. (2001). A second leaky splice-site mutation in the spastin gene. Am. J. Hum. Genet. *69*, 1407–1409.

Tournier, I., Vezain, M., Martins, A., Charbonnier, F., Baert-Desurmont, S., Olschwang, S., Wang, Q., Buisine, M.P., Soret, J., Tazi, J., et al. (2008). A large fraction of unclassified variants of the mismatch repair genes MLH1 and MSH2 is associated with splicing defects. Hum. Mutat. *29*, 1412–1424.

Turajlic, S., Litchfield, K., Xu, H., Rosenthal, R., McGranahan, N., Reading, J.L., Wong, Y.N.S., Rowan, A., Kanu, N., Al Bakir, M., et al. (2017). Insertion-and-deletion-derived tumour-specific neoantigens and the immunogenic phenotype: a pan-cancer analysis. Lancet Oncol. *18*, 1009–1021.

Venables, J.P. (2004). Aberrant and alternative splicing in cancer. Cancer Res. *64*, 7647–7654.

Vezain, M., Gérard, B., Drunat, S., Funalot, B., Fehrenbach, S., N'Guyen-Viet, V., Vallat, J.M., Frébourg, T., Tosi, M., Martins, A., and Saugier-Veber, P. (2011). A leaky splicing mutation affecting SMN1 exon 7 inclusion explains an unexpected mild case of spinal muscular atrophy. Hum. Mutat. *32*, 989–994.

Vreeswijk, M.P., and van der Klift, H.M. (2012). Analysis and interpretation of RNA splicing alterations in genes involved in genetic disorders. Methods Mol. Biol. *867*, 49–63.

Wadt, K., Choi, J., Chung, J.Y., Kiilgaard, J., Heegaard, S., Drzewiecki, K.T., Trent, J.M., Hewitt, S.M., Hayward, N.K., Gerdes, A.M., and Brown, K.M. (2012). A cryptic BAP1 splice mutation in a family with uveal and cutaneous melanoma, and paraganglioma. Pigment Cell Melanoma Res. *25*, 815–818.

Woolfe, A., Mullikin, J.C., and Elnitski, L. (2010). Genomic features defining exonic variants that modulate splicing. Genome Biol. *11*, R20–R20.

Xie, S., Mortusewicz, O., Ma, H.T., Herr, P., Poon, R.Y., Helleday, T., and Qian, C. (2015). Timeless interacts with PARP-1 to promote homologous recombination repair. Mol. Cell *60*, 163–176.

Yang, Y., Swaminathan, S., Martin, B.K., and Sharan, S.K. (2003). Aberrant splicing induced by missense mutations in BRCA1: clues from a humanized mouse model. Hum. Mol. Genet *12*, 2121–2131.

Yeo, G., and Burge, C.B. (2004). Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. J. Comput. Biol. *11*, 377xcn394.

Zhang, K., Nowak, I., Rushlow, D., Gallie, B.L., and Lohmann, D.R. (2008). Patterns of missplicing caused by RB1 gene mutations in patients with retinoblastoma and association with phenotypic expression. Hum. Mutat. *29*, 475–484.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Experimental Models: Cell Lines		
Human: HEK293T cells	ATCC	https://www.atcc.org/products/all/CRL-3216.aspx
Oligonucleotides		
Primers for cDNA amplification pCAS-KO1-(5′-TGACGTCGCCGCCCATCAC-3′) pCAS-R (5′-ATTGGTTGTTGAGTTGGTTGTC-3′)	This paper	N/A
Primers for Q5 mutagenesis and restriction enzyme primers for amplifying exons of interest see Table S6	This paper	N/A
Recombinant DNA		
Plasmid: pCAS2	Inserm Laboratory	N/A
Software and Algorithms		
MaxEntScan	Yeo and Burge, 2004	http://genes.mit.edu/burgelab/maxent/Xmaxentscan_scoreseq.html
Samtools	Li et al., 2009	http://samtools.sourceforge.net/
MiSplice	In preparation	https://github.com/ding-lab/misplice
Integrative Genomics Viewer	Robinson et al., 2011	http://software.broadinstitute.org/software/igv/
Chemicals, Peptides, and Recombinant Proteins		
Nucleospin PCR Cleanup	Macherey-Nagel	740609.10
DNA Clean and Concentrator-5 Kit	Zymo Research	D4003
BamHI	New England Biomedicine	R0136S
MluI	New England Biomedicine	R0198S
T4 DNA Ligase	New England Biomedicine	M0202S
Q5 Site Directed Mutagenesis	New England Biomedicine	E0554S
Lipofectamine 2000	Thermofisher Scientific	12566014
Superscript III First-Strand Synthesis System	Thermofisher Scientific	18080051
Qiaquick Gel Extraction Kit	QIAGEN	28704
Other		
Public MC3 MAF	In preparation	https://gdc.cancer.gov
MSGF+	N/A	https://www.ncbi.nlm.nih.gov/pubmed/?term=25358478
Mass Spectra Data from 77 TCGA Breast Cancer Patients	N/A	https://cptac-data-portal.georgetown.edu/cptac/s/S029

CONTACT FOR REAGENT AND RESOURCE SHARING

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Li Ding (lding@wustl.edu).

EXPERIMENTAL MODEL AND SUBJECT DETAILS

The Cancer Genome Atlas (TCGA) collected both tumor and non-tumor biospecimens from 10,224 human samples (https://cancergenome.nih.gov/abouttcga/policies/informedconsent). Here, we use variants from a publicly available mutation annotation file (MAF) complied by the MC3 working group (syn7824274).

METHOD DETAILS

Dataset Description

Aligned RNA-seq bam files were analyzed using the ISB google. These cancer types are Acute Myeloid Leukemia [LAML], Adrenocortical carcinoma [ACC], Bladder Urothelial Carcinoma [BLCA], Brain Lower Grade Glioma [LGG], Breast invasive carcinoma [BRCA], Cervical squamous cell carcinoma and endocervical adenocarcinoma [CESC], Cholangiocarcinoma [CHOL], Colon adenocarcinoma [COAD], Esophageal carcinoma [ESCA], Glioblastoma multiforme [GBM], Head and Neck squamous cell carcinoma [HNSC], Kidney Chromophobe [KICH], Kidney renal clear cell carcinoma [KIRC], Kidney renal papillary cell carcinoma [KIRP], Liver hepatocellular carcinoma [LIHC], Lung adenocarcinoma [LUAD], Lung squamous cell carcinoma [LUSC], Lymphoid Neoplasm Diffuse Large B cell Lymphoma [DLBC], Mesothelioma [MESO], Ovarian serous cystadenocarcinoma [OV], Pancreatic adenocarcinoma [PAAD], Pheochromocytoma and Paraganglioma [PCPG], Prostate adenocarcinoma [PRAD], Rectum adenocarcinoma [READ], Sarcoma [SARC], Skin Cutaneous Melanoma [SKCM], Stomach adenocarcinoma [STAD], Testicular Germ Cell Tumors [TGCT], Thymoma [THYM], Thyroid carcinoma [THCA], Uterine Carcinosarcoma [UCS], Uterine Corpus Endometrial Carcinoma [UCEC], Uveal Melanoma [UVM]

MiSplice Pipeline

The MiSplice pipeline was developed to detect mutation-induced splicing events from RNA-seq data. It is written in Perl and incorporates two standard tools, samtools and MaxEntScan. The pipeline is fully automated and can run multiple jobs in parallel on LSF cluster. It executes the following steps:

1) Splitting large maf file into multiple smaller files with less mutations (currently, the default setting is 200).
2) Discovering splicing junctions within 20bps of the mutation with at least 5 supporting reads with mapping quality Q20 and then filtering canonical junctions by using the Ensembl 37.75 database. We selected 20bp as a cut-off since it is the farthest distance from the splice junction in a splice region.
3) Computing the number of supporting reads of above cryptic splice sites for control samples without mutations (Table S1).
4) Calculating the splicing scores for the cryptic splice sites via MaxEntScan.
5) Reporting the depth of each cryptic splice site via Samtools.
6) Filtering cryptic sites which fall in HLA loci or less than 5% of reads at the genomic location supporting the alternative junction of interest.
7) Further filtering cryptic sites by comparing the supporting reads in control samples. The final reported cryptic sites must stand as top 5% for the number of supporting reads in the case (with mutation).

Splice Site Score Estimation

For each cryptic splice site and nearby canonical splice site, the corresponding nucleotide sequences were first extracted for both the mutant and reference sequences (9-mer and 23-mer for donor and acceptor, respectively). Their splice scores as potential donor or acceptor sites were then estimated using MaxEntScan.

Neoantigen Prediction

For each predicted SCM, we use a curated RefSeq transcript database (version 20130722) to obtain the translated protein sequences for transcript containing alternative splice forms induced by SCMs. Different length of epitopes (8-mer, 9-mer, 10-mer and 11-mer) are constructed from the translated protein sequence. We use NetMHC3pan (Nielsen and Andreatta, 2016) and NetMHC4 (Andreatta and Nielsen, 2016) to predict the binding affinity between epitopes and MHC. Epitopes with binding affinity $\leq$ 500nM which are also not present in the wild-type transcript are extracted from the following neoantigen analysis.

Manual Review

All splice-site-creating mutations were manually reviewed using the integrative genomics viewer (http://software.broadinstitute.org/software/igv/). Mutations were placed into one of three categories: Pass, Complex, and No Support. Mutations were classified as complex if more than one alternatively spliced product was observed for the mutated sample.

Code Availability

MiSplice is written in Perl and is freely available from GitHub at https://github.com/ding-lab/misplice under the GNU general public license. MiSplice uses several independent tools and packages, including SamTools and MaxEntScan, all of which are likewise freely available, but which must be obtained from their respective developers. The MiSplice documentation contains complete instructions for obtaining and linking these applications into MiSplice.

Mini-gene Splicing Assay

Exons of interest and approximately 150 bp of their flanking intron sequences were PCR amplified from HEK293T genomic DNA using primers carrying restriction enzyme sites for BamH1 and MluI. PCR products were cleaned up using NucleoSpin PCR Cleanup

(Macherey-Nagel) or DNA Clean and Concentrator-5 Kit (Zymo Research) and digested with BamHI and MluI. The digested pCAS2.1 vector and PCR products were ligated using T4 DNA Ligase (NEB). Mutations were introduced via Q5 Site-Directed Mutagenesis (NEB). WT and MUT constructs were confirmed by sequencing of the insert region. The plasmids were transiently transfected into HEK293T cells using Lipofectamine 2000 (ThermoFisher Scientific). 24 hr post transfection, cDNA was synthesized using 2 to 3 ug of total RNA with the Superscript III First-Strand Synthesis System (ThermoFisher Scientific) and priming with Oligo(dT)20. Finally, cDNA was amplified using pCAS-KO1-(5′-TGACGTCGCCGCCCATCAC-3′) and pCAS-R (5′-ATTGGTTGTTGAGTTGGTTGTC-3′) and the alternative splicing patterns were evaluated on a 2.5% agarose gel with ethidium bromide. Qiaquick Gel Extraction Kit (QIAGEN) was used to purify bands for sequencing (Figures S3, S4, S5, and S6; Tables S5, S6, and S7).

Cell Culture

HEK293T cells were cultured in Dulbecco's modified Eagle's medium (DMEM) supplemented with fetal bovine serum (FBS) and penicillin streptomycin.

QUANTIFICATION AND STATISTICAL ANALYSES

MiSplice assesses the significance of the number of reads supporting the predicted alternative splice junction by comparing to read counts from a control cohort. Specifically, a frequency distribution is constructed from the control cohort, from which threshold values for 5% and 95% tails on the left and right, respectively, are determined. A series of logic tests is then conducted to discern the best explanation of the data. Possible verdicts are low or high expression if the datum is outside the 5% or 95% thresholds, respectively, average expression if no thresholds are exceeded, or no expression in this tissue if the thresholds are zero.

Appendix D

Susceptibility genes to breast cancer

Nearly all known HBOC susceptibility genes encode tumor suppressors that participate in genome stability pathways (homologous recombination repair, replication fork stability, transcription-replication collisions, mismatch repair, and DNA damage signaling, checkpoints and cell death).

D.1 Homologous recombination repair

The homologous recombination repair pathway (HRR) deals with double strand DNA breaks by using the undamaged chromosome as template for error-free repair. After a DSB occurs, the MRN complex (MRE11, RAD50 and NDN) detects and binds the free DNA ends. Then, it promotes DNA damage checkpoint signaling.

HRR involves BRCA1, BRCA2 and, actually, most of the HBOC genes. Because of its ability to interact with a wide range of proteins, BRCA1 is hypothetized to act as a recruitment scaffold. A deficiency of BRCA1 is linked to the inability to trigger HRR. Mutations in the MRN complex have also been clinically associated to breast cancer, although dubiously so in the case of RAD50 variants. Reassuringly, some other HBOC genes are interactors of the MRN complex and BRCA1/2.

D.2 Replication fork stability

BRCA1 and BRCA2 protect newly synthesized DNA and promote the restart of stalled forks in an HRR-independent manner. In the absence of these proteins, newly synthesized DNA in a stalled fork would get degraded, leading to genome instability and increasing the risk of cancer.

D.3 Transcription-replication collisions

Collisions between transcription and replication are emerging as a source of genome instability. In particular, RNA-DNA hybrids called R-loops can form between the nascent transcript and the DNA template. They can lead to double-strand breaks and mutations. Both BRCA1 and BRCA2 participate in the resolution of R-loops, preventing their accumulation. In consequence, BRCA-deficient cells tend to suffer transcriptional stress that leads to genome instability. Nonetheless the relationship between this mechanism and proneness to HBOC is yet to be proven, and the genes involved further investigated.

D.4 Mismatch repair

DNA mismatch repair (MMR) corrects base-base mispairs. When MMR is faulty, accumulations point mutations and genetic changes in repeated nucleotide sequences (microsatellite instability) occur. MMR also plays a role in error-free HRR.

D.5 DNA damage signaling, checkpoints and cell death

Pathways involved in genome maintenance, cell cycle checkpoints and cell death usually eliminate cells with damaged DNA. When proteins involved in them are not active, some processes such as cell cycle arrest, apoptosis and senescence will not occur. In consequence, cells that undergo genomic alterations are allowed to proliferate. The most famous case of HBOC in this pathway is TP53, which coordinates the transcriptional induction of many genome stability factors.

\begin{sidewaystable} \caption{Overview of HBOC genes: estimated lifetime risk of breast cancer (age in years) and tumorogenic molecular mechanisms that involves them: homologous recombination repair (HRR), replication fork stability, transcription-replication collisions, mismatch repair (MMR), DNA damage signaling, checkpoints and cell death, and/or others. Adapted from (Nielsen, Overeem Hansen, and Sørensen 2016).

Gene	Lifetime risk	HRR	Rep. fork stab.	Tr.-rep. clash	MMR	DNA Damage, ap
ATM	60% by age 80	✓				✓
BARD1	Unknown	✓				
BLM	Unknown		✓			
BRCA1	57-65% by age 70	✓	✓	✓		✓
BRCA2	45-55% by age 70	✓	✓	✓		✓
BRIP1	OR: <2.0					
CDH1	42% by age 80					
CHEK2	37% by age 70					✓
FAM175A	Unknown	✓				
FANCC	Unknown		✓			
FANCM	Unknown		✓			
MLH1	∼19% by age 70		✓		✓	
MRE11	Unknown				✓	
MSH2	∼11% by age 70				✓	
NBN	OR: 3.0	✓				
NF1	6.5-fold up ages 30-39					
PALB2	35% by age 70	✓	✓			
PMS2	SIR: 3.8				✓	
PTEN	85% by age 70					
RAD51B	Unknown	✓				
RAD51C	Unknown	✓				
RAD51D	Unknown	✓				
RECQL	Unknown		✓			
RINT1	Unknown	✓				
STK11	32% by age 60					
TP53	25% by age 70					✓

\end{sidewaystable}

Ahmed, Shahana, Gilles Thomas, Maya Ghoussaini, Catherine S Healey, Manjeet K Humphreys, Radka Platte, Jonathan Morrison, et al. 2009a. "Newly Discovered Breast Cancer Susceptibility Loci on 3p24 and 17q23.2." *Nature Genetics* 41 (5): 585–90. https://doi.org/10.1038/ng.354.

———. 2009b. "Newly Discovered Breast Cancer Susceptibility Loci on 3p24 and 17q23.2." *Nature Genetics* 41 (5): 585–90. https://doi.org/10.1038/ng.354.

Azencott, C.-A. 2018. "Machine Learning and Genomics: Precision Medicine Versus Patient Privacy." *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 376 (2128): 20170350. https://doi.org/10.1098/rsta.2017.0350.

Azencott, Chloé-Agathe. 2016. "Network-Guided Biomarker Discovery." In *Machine Learning for Health Informatics*, 9605:319–36. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-50478-0_16.

Azencott, Chloé-Agathe, Dominik Grimm, Mahito Sugiyama, Yoshinobu Kawahara, and Karsten M. Borgwardt. 2013. "Efficient Network-Guided Multi-Locus Association Mapping with Graph Cuts." *Bioinformatics* 29 (13): i171–i179. https://doi.org/10.1093/bioinformatics/btt238.

Barabási, Albert-László, Natali Gulbahce, and Joseph Loscalzo. 2011. "Network Medicine: A Network-Based Approach to Human Disease." *Nature Reviews Genetics* 12 (1): 56–68. https://doi.org/10.1038/nrg2918.

Barton, N.H., A.M. Etheridge, and A. Véber. 2017. "The Infinitesimal Model: Definition, Derivation, and Implications." *Theoretical Population Biology* 118 (December): 50–73. https://doi.org/10.1016/j.tpb.2017.06.001.

Beisser, D., G. W. Klau, T. Dandekar, T. Muller, and M. T. Dittrich. 2010. "BioNet: An R-Package for the Functional Analysis of Biological Networks." *Bioinformatics* 26 (8): 1129–30. https://doi.org/10.1093/bioinformatics/btq089.

Boyle, Evan A., Yang I. Li, and Jonathan K. Pritchard. 2017. "An Expanded View of Complex Traits: From Polygenic to Omnigenic." *Cell* 169 (7): 1177–86. https://doi.org/10.1016/j.cell.2017.05.038.

Breyer, Joan P., Daniel C. Dorset, Travis A. Clark, Kevin M. Bradley, Tiina A. Wahlfors, Kate M. McReynolds, William H. Maynard, et al. 2014. "An Expressed Retrogene of the Master Embryonic Stem Cell Gene POU5F1 Is Associated with Prostate Cancer Susceptibility." *The American Journal of Human Genetics* 94 (3): 395–404. https://doi.org/10.1016/j.ajhg.2014.01.019.

Brisbin, Abra G, Yan W Asmann, Honglin Song, Ya-Yu Tsai, Jeremiah A Aakre, Ping Yang, Robert B Jenkins, et al. 2011. "Meta-Analysis of 8q24 for Seven Cancers Reveals a Locus Between NOV and ENPP2 Associated with Cancer Development." *BMC Medical Genetics* 12 (1): 156. https://doi.org/10.1186/1471-2350-12-156.

Buniello, Annalisa, Jacqueline A L MacArthur, Maria Cerezo, Laura W Harris, James Hayhurst, Cinzia Malangone, Aoife McMahon, et al. 2019. "The NHGRI-EBI GWAS Catalog of Published Genome-Wide Association Studies, Targeted Arrays and Summary Statistics 2019." *Nucleic Acids Research* 47 (D1): D1005–D1012. https://doi.org/10.1093/nar/gky1120.

Bush, William S., and Jason H. Moore. 2012. "Chapter 11: Genome-Wide Association Studies." Edited by Fran Lewitter and Maricel Kann. *PLoS*

Computational Biology 8 (12): e1002822. https://doi.org/10.1371/journal.pcbi.1002822.

Button, Katherine S., John P. A. Ioannidis, Claire Mokrysz, Brian A. Nosek, Jonathan Flint, Emma S. J. Robinson, and Marcus R. Munafò. 2013. "Power Failure: Why Small Sample Size Undermines the Reliability of Neuroscience." *Nature Reviews Neuroscience* 14 (5): 365–76. https://doi.org/10.1038/nrn3475.

Cai, James J., Elhanan Borenstein, and Dmitri A. Petrov. 2010. "Broker Genes in Human Disease." *Genome Biology and Evolution* 2 (January): 815–25. https://doi.org/10.1093/gbe/evq064.

Calderwood, Stuart K., and Jianlin Gong. 2016. "Heat Shock Proteins Promote Cancer: It's a Protection Racket." *Trends in Biochemical Sciences* 41 (4): 311–23. https://doi.org/10.1016/j.tibs.2016.01.003.

Calle, M Luz, Víctor Urrea, Núria Malats, and Kristel Van Steen. 2010. "Mbmdr: An R Package for Exploring Gene–Gene Interactions Associated with Binary or Quantitative Traits." *Bioinformatics* 26 (17). Oxford University Press: 2198–9.

Chaiboonchoe, Amphun, Wiktor Jurkowski, Johann Pellet, Enrico Glaab, Alexey Kolodkin, Antonio Raussel, Antony Le Béchec, et al. 2013. "On Different Aspects of Network Analysis in Systems Biology." In *Systems Biology: Integrative Biology and Simulation Tools*, edited by Aleš Prokop and Béla Csukás, 181–207. Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-94-007-6803-1_6.

Chang, Christopher C, Carson C Chow, Laurent CAM Tellier, Shashaank Vattikuti, Shaun M Purcell, and James J Lee. 2015. "Second-Generation PLINK: Rising to the Challenge of Larger and Richer Datasets." *GigaScience* 4 (1): 7. https://doi.org/10.1186/s13742-015-0047-8.

Cho, J. H., D. L. Nicolae, L. H. Gold, C. T. Fields, M. C. LaBuda, P. M. Rohal, M. R. Pickles, et al. 1998. "Identification of Novel Susceptibility Loci for Inflammatory Bowel Disease on Chromosomes 1p, 3q, and 4q: Evidence for Epistasis Between 1p and IBD1." *Proceedings of the National Academy of Sciences* 95 (13): 7502–7. https://doi.org/10.1073/pnas.95.13.7502.

Climente-González, Héctor, and Chloé-Agathe Azencott. 2019. "Martini." https://www.bioconductor.org/packages/martini/.

Climente-González, Héctor, Chloé-Agathe Azencott, Samuel Kaski, and Makoto Yamada. 2019. "Block HSIC Lasso: Model-Free Biomarker Detection for Ultra-High Dimensional Data." *Bioinformatics* 35 (14): i427–i435. https://doi.org/10.1093/bioinformatics/btz333.

Climente-González, Héctor, Eduard Porta-Pardo, Adam Godzik, and Eduardo Eyras. 2017. "The Functional Impact of Alternative Splicing in Cancer." *Cell Reports* 20 (9): 2215–26. https://doi.org/10.1016/j.celrep.2017.08.012.

Combarros, Onofre, Mario Cortina-Borja, A. David Smith, and Donald J. Lehmann. 2009. "Epistasis in Sporadic Alzheimer's Disease." *Neurobiology of Aging* 30 (9): 1333–49. https://doi.org/10.1016/j.neurobiolaging.2007.11.027.

Cortes, Adrian, and Matthew A Brown. 2010. "Promise and Pitfalls of the Immunochip." *Arthritis Research & Therapy* 13 (1): 101. https://doi.org/10.1186/ar3204.

Cowen, Lenore, Trey Ideker, Benjamin J. Raphael, and Roded Sharan. 2017. "Network Propagation: A Universal Amplifier of Genetic Associations." *Nature Reviews Genetics* 18 (9): 551–62. https://doi.org/10.1038/nrg.2017.38.

Das, Jishnu, and Haiyuan Yu. 2012. "HINT: High-Quality Protein Interactomes and Their Applications in Understanding Human Disease." *BMC Systems Biology* 6 (1): 92. https://doi.org/10.1186/1752-0509-6-92.

Di Tommaso, Paolo, Maria Chatzou, Evan W Floden, Pablo Prieto Barja, Emilio Palumbo, and Cedric Notredame. 2017. "Nextflow Enables Reproducible Computational Workflows." *Nature Biotechnology* 35 (4): 316–19. https://doi.org/10.1038/nbt.3820.

Dittrich, Marcus, and Daniela Beisser. 2008. "BioNet." https://bioconductor.org/packages/BioNet/.

Dittrich, M. T., G. W. Klau, A. Rosenwald, T. Dandekar, and T. Muller. 2008. "Identifying Functional Modules in Protein-Protein Interaction Networks: An Integrated Exact Approach." *Bioinformatics* 24 (13): i223–i231. https://doi.org/10.1093/bioinformatics/btn161.

Dziak, John, Runze Li, and Linda Collins. 2005. "Critical Review and Comparison of Variable Selection Procedures for Linear Regression," 1–69.

Ek, Weronica E, Mauro D'Amato, and Jonas Halfvarson. 2014. "The History of Genetics in Inflammatory Bowel Disease." *Annals of Gastroenterology* 27 (4): 294–303.

Ellinghaus, David, Luke Jostins, Sarah L Spain, Adrian Cortes, Jörn Bethune, Buhm Han, Yu Rang Park, et al. 2016. "Analysis of Five Chronic Inflammatory Diseases Identifies 27 New Associations and Highlights Disease-Specific Patterns at Shared Loci." *Nature Genetics* 48 (5). Nature Publishing Group: 510.

Ellinghaus, David, Sarah L Spain, Adrian Cortes, Jörn Bethune, Buhm Han, Yu Rang

Park, Soumya Raychaudhuri, et al. 2016. "Analysis of Five Chronic Inflammatory Diseases Identifies 27 New Associations and Highlights Disease-Specific Patterns at Shared Loci." *Nature Genetics* 48 (5): 510–18. https://doi.org/10.1038/ng.3528.

Frankish, Adam, Mark Diekhans, Anne-Maud Ferreira, Rory Johnson, Irwin Jungreis, Jane Loveland, Jonathan M Mudge, et al. 2019. "GENCODE Reference Annotation for the Human and Mouse Genomes." *Nucleic Acids Research* 47 (D1): D766–D773. https://doi.org/10.1093/nar/gky955.

Furlong, Laura I. 2013. "Human Diseases Through the Lens of Network Biology." *Trends in Genetics* 29 (3): 150–59. https://doi.org/10.1016/j.tig.2012.11.004.

Ge, Youngchao, Sandrine Dudoit, and Terence P Speed. 2003. "Resampling-Based Multiple Testing for Microarray Data Analysis." *Test* 12 (1). Springer: 1–77.

Glas, Jürgen, Johannes Stallhofer, Stephan Ripke, Martin Wetzke, Simone Pfennig, Wolfram Klein, Jörg T Epplen, et al. 2009. "Novel Genetic Risk Markers for Ulcerative Colitis in the Il2/Il21 Region Are in Epistasis with Il23r and Suggest a Common Genetic Background for Ulcerative Colitis and Celiac Disease." *The American Journal of Gastroenterology* 104 (7). Nature Publishing Group: 1737.

Glass, Kimberly, Curtis Huttenhower, John Quackenbush, and Guo-Cheng Yuan. 2013. "Passing Messages Between Biological Networks to Refine Predicted Interactions." Edited by Szabolcs Semsey. *PLoS ONE* 8 (5): e64832. https://doi.org/10.1371/journal.pone.0064832.

Grimm, Dominik G., Damian Roqueiro, Patrice A. Salomé, Stefan Kleeberger, Bastian Greshake, Wangsheng Zhu, Chang Liu, et al. 2017. "easyGWAS: A Cloud-Based Platform for Comparing the Results of Genome-Wide Association Studies." *The Plant Cell* 29 (1): 5–19. https://doi.org/10.1105/tpc.16.00551.

GTEx Consortium. 2017. "Genetic Effects on Gene Expression Across Human Tissues." *Nature* 550 (7675): 204–13. https://doi.org/10.1038/nature24277.

Gusareva, Elena S., and Kristel Van Steen. 2014. "Practical Aspects of Genome-Wide Association Interaction Analysis." *Human Genetics* 133 (11): 1343–58. https://doi.org/10.1007/s00439-014-1480-y.

Gwinner, Frederik. 2016. "LEANR." https://cran.r-project.org/web/packages/LEANR/.

Gwinner, Frederik, Gwénola Boulday, Claire Vandiedonck, Minh Arnould, Cécile Cardoso, Iryna Nikolayeva, Oriol Guitart-Pla, et al. 2016. "Network-Based Analysis of Omics Data: The LEAN Method." *Bioinformatics*, October, btw676. https://doi.org/10.1093/bioinformatics/btw676.

Hemani, Gibran, Konstantin Shakhbazov, Harm-Jan Westra, Tonu Esko, Anjali K. Henders, Allan F. McRae, Jian Yang, et al. 2014. "Detection and Replication of Epistasis Influencing Transcription in Humans." *Nature* 508 (7495): 249–53. https://doi.org/10.1038/nature13005.

Hermjakob, H. 2004. "IntAct: An Open Source Molecular Interaction Database." *Nucleic Acids Research* 32 (90001): 452D–455. https://doi.org/10.1093/nar/gkh052.

Huang, Justin K., Daniel E. Carlin, Michael Ku Yu, Wei Zhang, Jason F. Kreisberg, Pablo Tamayo, and Trey Ideker. 2018. "Systematic Evaluation of Molecular Networks for Discovery of Disease Genes." *Cell Systems* 6 (4): 484–495.e5. https://doi.org/10.1016/j.cels.2018.03.001.

Ionita-Laza, Iuliana, Seunggeun Lee, Vlad Makarov, Joseph D. Buxbaum, and Xihong Lin. 2013. "Sequence Kernel Association Tests for the Combined Effect of Rare and Common Variants." *The American Journal of Human Genetics* 92 (6): 841–53. https://doi.org/10.1016/j.ajhg.2013.04.015.

Jayasinghe, Reyka G., Song Cao, Qingsong Gao, Michael C. Wendl, Nam Sy Vo, Sheila M. Reynolds, Yanyan Zhao, et al. 2018. "Systematic Analysis of Splice-Site-Creating Mutations in Cancer." *Cell Reports* 23 (1): 270–281.e3. https://doi.org/10.1016/j.celrep.2018.03.052.

Jia, Peilin, Siyuan Zheng, Jirong Long, Wei Zheng, and Zhongming Zhao. 2011. "dmGWAS: Dense Module Searching for Genome-Wide Association Studies in Protein–Protein Interaction Networks." *Bioinformatics* 27 (1): 95–102. https://doi.org/10.1093/bioinformatics/btq615.

Jorgenson, Eric, and John S Witte. 2006. "A Gene-Centric Approach to Genome-Wide Association Studies." *Nature Reviews. Genetics* 7 (December): 885–91. https://doi.org/10.1038/nrg1962.

Jostins, Luke, Stephan Ripke, Rinse K. Weersma, Richard H. Duerr, Dermot P. McGovern, Ken Y. Hui, James C. Lee, et al. 2012. "Host–Microbe Interactions Have Shaped the Genetic Architecture of Inflammatory Bowel Disease." *Nature* 491 (7422): 119–24. https://doi.org/10.1038/nature11582.

Kimura, L., C. B. Angeli, M. T. B. M. Auricchio, G. R. Fernandes, A. C. Pereira, J. P. Vicente, T. V. Pereira, and R. C. Mingroni-Netto. 2012. "Multilocus Family-Based Association Analysis of Seven Candidate Polymorphisms with Essential Hypertension in an African-Derived Semi-Isolated Brazilian Population." *International Journal of Hypertension* 2012: 1–8. https://doi.org/10.1155/2012/859219.

Krzywinski, M., I. Birol, S. J. Jones, and M. A. Marra. 2012. "Hive Plots–Rational

Approach to Visualizing Networks." *Briefings in Bioinformatics* 13 (5): 627–44. https://doi.org/10.1093/bib/bbr069.

Lehne, Benjamin, Cathryn M Lewis, and Thomas Schlitt. 2011. "From Snps to Genes: Disease Association at the Gene Level." *PloS One* 6 (6). Public Library of Science: e20133.

Leiserson, Mark DM, Eldridge, Jonathan V, Ramachandran, Sohini, and Raphael, Benjamin J. 2013. "Network Analysis of GWAS Data," 9.

Leiserson, Mark D M, Fabio Vandin, Hsin-Ta Wu, Jason R Dobson, Jonathan V Eldridge, Jacob L Thomas, Alexandra Papoutsaki, et al. 2015. "Pan-Cancer Network Analysis Identifies Combinations of Rare Somatic Mutations Across Pathways and Protein Complexes." *Nature Genetics* 47 (2): 106–14. https://doi.org/10.1038/ng.3168.

———. 2018. "HotNet2." https://github.com/raphael-group/hotnet2.

Li, Wentian, and Jens Reich. 2000. "A Complete Enumeration and Classification of Two-Locus Disease Models." *Human Heredity* 50 (6): 334–49. https://doi.org/10.1159/000022939.

Lin, Zhenwu, John P Hegarty, Gerrit John, Arthur Berg, Zhong Wang, Rishabh Sehgal, Danielle M Pastor, et al. 2013. "NOD2 Mutations Affect Muramyl Dipeptide Stimulation of Human B Lymphocytes and Interact with Other Ibd-Associated Genes." *Digestive Diseases and Sciences* 58 (9). Springer: 2599–2607.

Lin, Zhenwu, Zhong Wang, John P Hegarty, Tony R Lin, Yunhua Wang, Sue Deiling, Rongling Wu, Neal J Thomas, and Joanna Floros. 2017. "Genetic Association and Epistatic Interaction of the Interleukin-10 Signaling Pathway in Pediatric Inflammatory Bowel Disease." *World Journal of Gastroenterology* 23 (27). Baishideng Publishing Group Inc: 4897.

Liu, Boxiang, Michael J. Gloudemans, Abhiram S. Rao, Erik Ingelsson, and Stephen B. Montgomery. 2019. "Abundant Associations with Gene Expression Complicate GWAS Follow-up." *Nature Genetics*, May. https://doi.org/10.1038/s41588-019-0404-0.

Liu, Guohong, Francois X. Claret, Fuling Zhou, and Yunbao Pan. 2018. "Jab1/COPS5 as a Novel Biomarker for Diagnosis, Prognosis, Therapy Prediction and Therapeutic Tools for Human Cancer." *Frontiers in Pharmacology* 9 (February): 135. https://doi.org/10.3389/fphar.2018.00135.

Liu, Jimmy Z, Suzanne Van Sommeren, Hailiang Huang, Siew C Ng, Rudi Alberts, Atsushi Takahashi, Stephan Ripke, et al. 2015. "Association Analyses Identify 38

Susceptibility Loci for Inflammatory Bowel Disease and Highlight Shared Genetic Risk Across Populations." *Nature Genetics* 47 (9). Nature Publishing Group: 979.

Liu, Ta-Chiang, and Thaddeus S. Stappenbeck. 2016. "Genetics and Pathogenesis of Inflammatory Bowel Disease." *Annual Review of Pathology: Mechanisms of Disease* 11 (1): 127–48. https://doi.org/10.1146/annurev-pathol-012615-044152.

Liu, Yuanlong. 2018. "SigMod V2." https://github.com/YuanlongLiu/SigMod.

Liu, Yuanlong, Myriam Brossard, Damian Roqueiro, Patricia Margaritte-Jeannin, Chloé Sarnowski, Emmanuelle Bouzigon, and Florence Demenais. 2017. "SigMod: An Exact and Efficient Method to Identify a Strongly Interconnected Disease-Associated Module in a Gene Network." *Bioinformatics*, January, btx004. https://doi.org/10.1093/bioinformatics/btx004.

Ljubić, Ivana, René Weiskircher, Ulrich Pferschy, Gunnar W. Klau, Petra Mutzel, and Matteo Fischetti. 2006. "An Algorithmic Framework for the Exact Solution of the Prize-Collecting Steiner Tree Problem." *Mathematical Programming* 105 (2-3): 427–49. https://doi.org/10.1007/s10107-005-0660-x.

Loddo, Italia, and Claudio Romano. 2015. "Inflammatory Bowel Disease: Genetics, Epigenetics, and Pathogenesis." *Frontiers in Immunology* 6 (November). https://doi.org/10.3389/fimmu.2015.00551.

Ma, Li, Alon Keinan, and Andrew G Clark. 2015. "Biological Knowledge-Driven Analysis of Epistasis in Human Gwas with Application to Lipid Traits." In *Epistasis*, 35–45. Springer.

Macdonald, M. 1993. "A Novel Gene Containing a Trinucleotide Repeat That Is Expanded and Unstable on Huntington's Disease Chromosomes." *Cell* 72 (6): 971–83. https://doi.org/10.1016/0092-8674(93)90585-E.

Manolio, Teri A., Francis S. Collins, Nancy J. Cox, David B. Goldstein, Lucia A. Hindorff, David J. Hunter, Mark I. McCarthy, et al. 2009. "Finding the Missing Heritability of Complex Diseases." *Nature* 461 (7265): 747–53. https://doi.org/10.1038/nature08494.

McGovern, Dermot PB, Jerome I Rotter, Ling Mei, Talin Haritunians, Carol Landers, Carrie Derkowski, Deb Dutridge, et al. 2009. "Genetic Epistasis of Il23/Il17 Pathway Genes in Crohn's Disease Dermot." *Inflammatory Bowel Diseases* 15 (6). Oxford University Press Oxford, UK: 883–89.

Michailidou, Kyriaki, Jonathan Beesley, Sara Lindstrom, Sander Canisius, Joe Dennis, Michael J Lush, Mel J Maranian, et al. 2015. "Genome-Wide Association

Analysis of More Than 120,000 Individuals Identifies 15 New Susceptibility Loci for Breast Cancer." *Nature Genetics* 47 (4): 373–80. https://doi.org/10.1038/ng.3242.

Michailidou, Kyriaki, Sara Lindström, Joe Dennis, Jonathan Beesley, Shirley Hui, Siddhartha Kar, Audrey Lemaçon, et al. 2017. "Association Analysis Identifies 65 New Breast Cancer Risk Loci." *Nature* 551 (7678): 92–94. https://doi.org/10.1038/nature24284.

Mishra, Aniket, and Stuart Macgregor. 2015. "VEGAS2: Software for More Flexible Gene-Based Testing." *Twin Research and Human Genetics* 18 (1): 86–91. https://doi.org/10.1017/thg.2014.79.

Mitra, Koyel, Anne-Ruxandra Carvunis, Sanath Kumar Ramesh, and Trey Ideker. 2013. "Integrative Approaches for Finding Modular Structure in Biological Networks." *Nature Reviews Genetics* 14 (10): 719–32. https://doi.org/10.1038/nrg3552.

Moore, Jason H., Joshua C. Gilbert, Chia-Ti Tsai, Fu-Tien Chiang, Todd Holden, Nate Barney, and Bill C. White. 2006. "A Flexible Computational Framework for Detecting, Characterizing, and Interpreting Statistical Patterns of Epistasis in Genetic Studies of Human Disease Susceptibility." *Journal of Theoretical Biology* 241 (2): 252–61. https://doi.org/10.1016/j.jtbi.2005.11.036.

Moore, Jason H., and Scott M. Williams. 2005. "Traversing the Conceptual Divide Between Biological and Statistical Epistasis: Systems Biology and a More Modern Synthesis." *BioEssays* 27 (6): 637–46. https://doi.org/10.1002/bies.20236.

Nakka, P., B. J. Raphael, and S. Ramachandran. 2016. "Gene and Network Analysis of Common Variants Reveals Novel Associations in Multiple Complex Diseases." *Genetics* 204 (2): 783–98. https://doi.org/10.1534/genetics.116.188391.

Ng, Siew C, Hai Yun Shi, Nima Hamidi, Fox E Underwood, Whitney Tang, Eric I Benchimol, Remo Panaccione, et al. 2017. "Worldwide Incidence and Prevalence of Inflammatory Bowel Disease in the 21st Century: A Systematic Review of Population-Based Studies." *The Lancet* 390 (10114): 2769–78. https://doi.org/10.1016/S0140-6736(17)32448-0.

Niel, Clément, Christine Sinoquet, Christian Dina, and Ghislain Rocheleau. 2015. "A Survey About Methods Dedicated to Epistasis Detection." *Frontiers in Genetics* 6 (September). https://doi.org/10.3389/fgene.2015.00285.

Nielsen, Finn Cilius, Thomas van Overeem Hansen, and Claus Storgaard Sørensen. 2016. "Hereditary Breast and Ovarian Cancer: New Genes in Confined Pathways." *Nature Reviews Cancer* 16 (9): 599–612. https://doi.org/10.1038/nrc.2016.72.

Nogueira, Sarah, and Gavin Brown. 2016. "Measuring the Stability of Feature

Selection." In *Machine Learning and Knowledge Discovery in Databases*, 9852:442–57. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-46227-1_28.

Oughtred, Rose, Chris Stark, Bobby-Joe Breitkreutz, Jennifer Rust, Lorrie Boucher, Christie Chang, Nadine Kolas, et al. 2019. "The BioGRID Interaction Database: 2019 Update." *Nucleic Acids Research* 47 (D1): D529–D541. https://doi.org/10.1093/nar/gky1079.

Pedersen, Thomas Lin. 2019. *Tidygraph: A Tidy Api for Graph Manipulation*. https://CRAN.R-project.org/package=tidygraph.

Pedros, Christophe, Guillaume Gaud, Isabelle Bernard, Sahar Kassem, Marianne Chabod, Dominique Lagrange, Olivier Andréoletti, et al. 2015. "An Epistatic Interaction Between Themis1 and Vav1 Modulates Regulatory T Cell Function and Inflammatory Bowel Disease Development." *The Journal of Immunology* 195 (4). Am Assoc Immnol: 1608–16.

Pendergrass, Sarah A, Alex Frase, John Wallace, Daniel Wolfe, Neerja Katiyar, Carrie Moore, and Marylyn D Ritchie. 2013. "Genomic Analyses with Biofilter 2.0: Knowledge Driven Filtering, Annotation, and Model Development." *BioData Mining* 6 (1). https://doi.org/10.1186/1756-0381-6-25.

Piñero, Janet, Ariel Berenstein, Abel Gonzalez-Perez, Ariel Chernomoretz, and Laura I. Furlong. 2016. "Uncovering Disease Mechanisms Through Network Biology in the Era of Next Generation Sequencing." *Scientific Reports* 6 (1): 24570. https://doi.org/10.1038/srep24570.

Piñero, Janet, Àlex Bravo, Núria Queralt-Rosinach, Alba Gutiérrez-Sacristán, Jordi Deu-Pons, Emilio Centeno, Javier García-García, Ferran Sanz, and Laura I. Furlong. 2017. "DisGeNET: A Comprehensive Platform Integrating Information on Human Disease-Associated Genes and Variants." *Nucleic Acids Research* 45 (D1): D833–D839. https://doi.org/10.1093/nar/gkw943.

Price, Alkes L, Nick J Patterson, Robert M Plenge, Michael E Weinblatt, Nancy A Shadick, and David Reich. 2006. "Principal Components Analysis Corrects for Stratification in Genome-Wide Association Studies." *Nature Genetics* 38 (8): 904–9. https://doi.org/10.1038/ng1847.

Purcell, Shaun, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel AR Ferreira, David Bender, Julian Maller, et al. 2007. "PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses." *The American Journal of Human Genetics* 81 (3). Elsevier: 559–75.

Quigley, David A., Elisa Fiorito, Silje Nord, Peter Van Loo, Grethe Grenaker Alnaes, Thomas Fleischer, Jorg Tost, et al. 2014. "The 5p12 Breast Cancer Susceptibility Locus Affects MRPS30 Expression in Estrogen-Receptor Positive Tumors." *Molecular Oncology* 8 (2): 273–84. https://doi.org/10.1016/j.molonc.2013.11.008.

Rinella, Erica S., Yongzhao Shao, Lauren Yackowski, Sreemanta Pramanik, Ruth Oratz, Freya Schnabel, Saurav Guha, et al. 2013. "Genetic Variants Associated with Breast Cancer Risk for Ashkenazi Jewish Women with Strong Family Histories but No Identifiable BRCA1/2 Mutation." *Human Genetics* 132 (5): 523–36. https://doi.org/10.1007/s00439-013-1269-4.

Ritchie, Marylyn D., and Kristel Van Steen. 2018. "The Search for Gene-Gene Interactions in Genome-Wide Association Studies: Challenges in Abundance of Methods, Practical Considerations, and Biological Interpretation." *Annals of Translational Medicine* 6 (8): 157–57. https://doi.org/10.21037/atm.2018.04.05.

Romagnoni, Alberto, Simon Jégou, Kristel Van Steen, Gilles Wainrib, and Jean-Pierre Hugot. 2019. "Comparative Performances of Machine Learning Methods for Classifying Crohn Disease Patients Using Genome-Wide Genotyping Data." *Scientific Reports* 9 (1): 10351. https://doi.org/10.1038/s41598-019-46649-z.

Sakoda, Lori C, Eric Jorgenson, and John S Witte. 2013. "Turning of COGS Moves Forward Findings for Hormonally Mediated Cancers." *Nature Genetics* 45 (4): 345–48. https://doi.org/10.1038/ng.2587.

Scheid, S., and R. Spang. 2005. "Twilight; a Bioconductor Package for Estimating the Local False Discovery Rate." *Bioinformatics* 21 (12): 2921–2. https://doi.org/10.1093/bioinformatics/bti436.

Segrè, Ayellet V., DIAGRAM Consortium, MAGIC investigators, Leif Groop, Vamsi K. Mootha, Mark J. Daly, and David Altshuler. 2010. "Common Inherited Variation in Mitochondrial Genes Is Not Enriched for Associations with Type 2 Diabetes or Related Glycemic Traits." Edited by Peter M. Visscher. *PLoS Genetics* 6 (8): e1001058. https://doi.org/10.1371/journal.pgen.1001058.

Sender, Ron, Shai Fuchs, and Ron Milo. 2016. "Revised Estimates for the Number of Human and Bacteria Cells in the Body." *PLOS Biology* 14 (8): e1002533. https://doi.org/10.1371/journal.pbio.1002533.

Sheng, Xuguang, and Jingyun Yang. 2013. "An Adaptive Truncated Product Method for Combining Dependent P-Values." *Economics Letters* 119 (2). Elsevier: 180–82.

Sinilnikova, Olga M., Marie-Gabrielle Dondon, Séverine Eon-Marchais, Francesca Damiola, Laure Barjhoux, Morgane Marcou, Carole Verny-Pierre, et al. 2016.

"GENESIS: A French National Resource to Study the Missing Heritability of Breast Cancer." *BMC Cancer* 16 (1): 13. https://doi.org/10.1186/s12885-015-2028-9.

Spielmann, Malte, Darío G. Lupiáñez, and Stefan Mundlos. 2018. "Structural Variation in the 3D Genome." *Nature Reviews Genetics* 19 (7): 453–67. https://doi.org/10.1038/s41576-018-0007-0.

Szklarczyk, Damian, Annika L Gable, David Lyon, Alexander Junge, Stefan Wyder, Jaime Huerta-Cepas, Milan Simonovic, et al. 2019. "STRING V11: Protein–Protein Association Networks with Increased Coverage, Supporting Functional Discovery in Genome-Wide Experimental Datasets." *Nucleic Acids Research* 47 (D1): D607–D613. https://doi.org/10.1093/nar/gky1131.

The 1000 Genomes Project Consortium, Richard A. Gibbs, Eric Boerwinkle, Harsha Doddapaneni, Yi Han, Viktoriya Korchina, Christie Kovar, et al. 2015. "A Global Reference for Human Genetic Variation." *Nature* 526 (7571): 68–74. https://doi.org/10.1038/nature15393.

Van Steen, Kristel, and JH Moore. 2019. "How to Increase Our Belief in Discovered Statistical Interactions via Large-Scale Association Studies?" *Human Genetics* 138 (4). Springer: 293–305.

Vermeire, Severine, Paul Rutgeerts, Kristel Van Steen, Sofie Joossens, G Claessens, Marie Pierik, Marc Peeters, and Robert Vlietinck. 2004. "Genome Wide Scan in a Flemish Inflammatory Bowel Disease Population: Support for the Ibd4 Locus, Population Heterogeneity, and Epistasis." *Gut* 53 (7). BMJ Publishing Group: 980–86.

Visscher, Peter M., William G. Hill, and Naomi R. Wray. 2008. "Heritability in the Genomics Era — Concepts and Misconceptions." *Nature Reviews Genetics* 9 (4): 255–66. https://doi.org/10.1038/nrg2322.

Visscher, Peter M., Naomi R. Wray, Qian Zhang, Pamela Sklar, Mark I. McCarthy, Matthew A. Brown, and Jian Yang. 2017. "10 Years of GWAS Discovery: Biology, Function, and Translation." *The American Journal of Human Genetics* 101 (1): 5–22. https://doi.org/10.1016/j.ajhg.2017.06.005.

Wang, Bo, Aziz M Mezlini, Feyyaz Demir, Marc Fiume, Zhuowen Tu, Michael Brudno, Benjamin Haibe-Kains, and Anna Goldenberg. 2014. "Similarity Network Fusion for Aggregating Data Types on a Genomic Scale." *Nature Methods* 11 (3): 333–37. https://doi.org/10.1038/nmeth.2810.

Wang, Lily, Peilin Jia, Russell D. Wolfinger, Xi Chen, and Zhongming Zhao. 2011. "Gene Set Analysis of Genome-Wide Association Studies: Methodological Issues and Perspectives." *Genomics* 98 (1): 1–8. https://doi.org/10.1016/j.ygeno.2011.04.006.

Wang, Maggie Haitian, Heather J. Cordell, and Kristel Van Steen. 2018. "Statistical Methods for Genome-Wide Association Studies." *Seminars in Cancer Biology*, May. https://doi.org/10.1016/j.semcancer.2018.04.008.

Wang, Quan, and Peilin Jia. 2014. "DmGWAS 3.0." https://bioinfo.uth.edu/dmGWAS/.

Watanabe, Kyoko, Erdogan Taskesen, Arjen van Bochoven, and Danielle Posthuma. 2017. "Functional Mapping and Annotation of Genetic Associations with FUMA." *Nature Communications* 8 (1). https://doi.org/10.1038/s41467-017-01261-5.

Wickham, Hadley. 2011. "Testthat: Get Started with Testing." *The R Journal* 3 (1): 5. https://doi.org/10.32614/RJ-2011-002.

Wray, Naomi R., Jian Yang, Ben J. Hayes, Alkes L. Price, Michael E. Goddard, and Peter M. Visscher. 2013. "Pitfalls of Predicting Complex Traits from SNPs." *Nature Reviews Genetics* 14 (7): 507–15. https://doi.org/10.1038/nrg3457.

Wu, Michael C., Seunggeun Lee, Tianxi Cai, Yun Li, Michael Boehnke, and Xihong Lin. 2011. "Rare-Variant Association Testing for Sequencing Data with the Sequence Kernel Association Test." *The American Journal of Human Genetics* 89 (1): 82–93. https://doi.org/10.1016/j.ajhg.2011.05.029.

Wu, Xuesen, Hua Dong, Li Luo, Yun Zhu, Gang Peng, John D Reveille, and Momiao Xiong. 2010. "A Novel Statistic for Genome-Wide Interaction Analysis." *PLoS Genetics* 6 (9). Public Library of Science: e1001131.

Yang, Jian, S. Hong Lee, Michael E. Goddard, and Peter M. Visscher. 2011. "GCTA: A tool for genome-wide complex trait analysis." *American Journal of Human Genetics* 88 (1). The American Society of Human Genetics: 76–82. https://doi.org/10.1016/j.ajhg.2010.11.011.

Yip, Danny Kit-Sang, Landon L Chan, Iris K Pang, Wei Jiang, Nelson LS Tang, Weichuan Yu, and Kevin Y Yip. 2018. "A Network Approach to Exploring the Functional Basis of Gene–Gene Epistatic Interactions in Disease Susceptibility." *Bioinformatics* 34 (10). Oxford University Press: 1741–9.

Zaykin, Dmitri V, Lev A Zhivotovsky, Peter H Westfall, and Bruce S Weir. 2002. "Truncated Product Method for Combining P-Values." *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society* 22 (2). Wiley Online Library: 170–85.

Zuk, O., E. Hechter, S. R. Sunyaev, and E. S. Lander. 2012. "The Mystery of Missing Heritability: Genetic Interactions Create Phantom Heritability." *Proceedings of the*

National Academy of Sciences 109 (4): 1193–8.
https://doi.org/10.1073/pnas.1119675109.

RÉSUMÉ

Cuius acerbitati uxor grave accesserat incentivum, germanitate Augusti turgida supra modum, quam Hannibaliano regi fratris filio antehac Constantinus iunxerat pater, Megaera quaedam mortalis, inflammatrix saevientis adsidua, humani cruoris avida nihil mitius quam maritus; qui paulatim eruditiores facti processu temporis ad nocendum per clandestinos versutosque rumigerulos conpertis leviter addere quaedam male suetos falsa et placentia sibi discentes, adfectati regni vel artium nefandarum calumnias insontibus adfligebant.

Saraceni tamen nec amici nobis umquam nec hostes optandi, ultro citroque discursantes quicquid inveniri poterat momento temporis parvi vastabant milvorum rapacium similes, qui si praedam dispexerint celsius, volatu rapiunt celeri, aut nisi impetraverint, non inmorantur.

Vita est illis semper in fuga uxoresque mercenariae conductae ad tempus ex pacto atque, ut sit species matrimonii, dotis nomine futura coniunx hastam et tabernaculum offert marito, post statum diem si id elegerit discessura, et incredibile est quo ardore apud eos in venerem uterque solvitur sexus.

Sed tamen haec cum ita tutius observentur, quidam vigore artuum inminuto rogati ad nuptias ubi aurum dextris manibus cavatis offertur, inpigre vel usque Spoletium pergunt. haec nobilium sunt instituta.

MOTS CLÉS

Caesar licentia post honoratis haec adhibens urbium honoratis nullum Caesar.

ABSTRACT

Verum ad istam omnem orationem brevis est defensio. Nam quoad aetas M. Caeli dare potuit isti suspicioni locum, fuit primum ipsius pudore, deinde etiam patris diligentia disciplinaque munita. Qui ut huic virilem togam deditÅ¡nihil dicam hoc loco de me; tantum sit, quantum vos existimatis; hoc dicam, hunc a patre continuo ad me esse deductum; nemo hunc M. Caelium in illo aetatis flore vidit nisi aut cum patre aut mecum aut in M. Crassi castissima domo, cum artibus honestissimis erudiretur.

Et eodem impetu Domitianum praecipitem per scalas itidem funibus constrinxerunt, eosque coniunctos per ampla spatia civitatis acri raptavere discursu. iamque artuum et membrorum divulsa conpage superscandentes corpora mortuorum ad ultimam truncata deformitatem velut exsaturati mox abiecerunt in flumen.

Erat autem diritatis eius hoc quoque indicium nec obscurum nec latens, quod ludicris cruentis delectabatur et in circo sex vel septem aliquotiens vetitis certaminibus pugilum vicissim se concidentium perfusorumque sanguine specie ut lucratus ingentia laetabatur.

Ego vero sic intellego, Patres conscripti, nos hoc tempore in provinciis decernendis perpetuae pacis habere oportere rationem. Nam quis hoc non sentit omnia alia esse nobis vacua ab omni periculo atque etiam suspicione belli?

KEYWORDS

gwas network epistasis machine-learning