A Framework for Protein-Level Interpretation of Genetic Associations and Integration With Large-Scale DNA Sequencing Analysis

A framework for protein-level interpretation of genetic associations and integration with large-scale DNA sequencing analysis

A dissertation presented

by

Mykyta Artomov

to

The Department of Chemistry and Chemical Biology

In partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

In the subject of

Chemistry

Harvard University

Cambridge, Massachusetts

May 2017

Dissertation Advisor: Professor Mark J. Daly Mykyta Artomov

A framework for protein-level interpretation of genetic associations and integration with large-scale DNA sequencing analysis

Abstract

With recent rapid decrease in exome and genome sequencing price amount of the available sequencing data has dramatically increased. While analysis of common genetic variation has succeeded with GWAS and fine-mapping methodology, systematic large-scale approach to rare protein-coding DNA variation analysis and interpretation is still in its early days. Rare variation, unlike GWAS, enables deep insight into the personalized disease predisposing factors and better understanding of underlying biology and, thus, facilitates potential new drug discoveries. In this thesis, we have focused on developing methods for interpretation of the genetic association results using protein-protein interaction models to aid the prioritization of disease risk genes and provide insights into involved biological pathways.

We created a composite approach for rare DNA variation analysis in case-control cohorts. Our approach was initially tested in the medium-sized cohort of focal segmental glomerulosclerosis patients, identifying several new risk genes that were validated using proof-of-concept mouse model. This methodology was then extended to the large-scale analysis of the germline cancer cohort (over

2,000 samples matched to more than 7,000 controls). We identified common features shared by known cancer predisposing genes and created a strategy for identification of the new cancer driving genes. List of novel candidate genes was created for several cancer phenotypes and some of the candidates were subjected to validation in mouse model successfully proving tumor suppressor activity of the encoded proteins.

Analysis of the genetic risk factors provides only unstructured pieces of information about the biology of a disorder. Generally, after identification of the associated loci massive follow-up studies are required to, first, prove the causal relationship, and, most importantly, understand the molecular mechanism of causality. Which locus should be prioritized for protein-level studies is currently determined based on empirical knowledge of protein function. Integration of the experimentally proven individual proteins functionality is then aimed to identify pathways affected by disease. Alternatively to this extensive approach, we developed a statistical framework that integrates genetic association data from multiple sources (GWAS, RVAS, etc.) and finds the protein-protein network returning the best cumulative association score. Using Bayesian model association results are then refined with evidence of the specific gene appearance in the best network. Our method provides a ranked list of genes prioritized based on both association strength and integration in the functional pathway. Such approach is essential for understanding biology of the disorders where it is impossible to build adequate animal model – autism, schizophrenia and other neuropsychiatric diseases.

Table of Contents

Acknowledgments

I am very greatfull to my advisor Mark J. Daly for being a great mentor. My best expectations from my graduate school experience were exceeded dramatically due to Mark accepting me to be a part of his lab. Projects described in this thesis sometimes were not easy to move forward, but his guidance and involvement helped to reach the finish line.

I am also lucky to have tight collaboration with Hensin Tsao and his lab. Hensin, throughout the years of our joint work on melanoma projects was supporting myself and taking all necessary time to make sure that I understand the clinical peculiarities of the cancer genetics.

It was wonderful to be a part of ATGU team. I was fortunate to have amazing colleagues with whom I was sharing the office throughout the time in the graduate school – Elaine Lim, Kaitlin Samocha, Jack Kosmicki and Jackie Goldstein.

I am also extremely thankful to my family for their deep care and support as things sometimes have gotten really tough during my time in the graduate school. In particular my wife Marina and my brother Max have made their best effort to help me achieve the time of completion of my PhD.

Chapter 1
Introduction

Linking genetic variation to phenotype

Human medical genetics studies are focused on finding associations between DNA variation and phenotype. According to central dogma of molecular biology alterations in the protein functions or expression levels that are encoded in DNA are main causes of a phenotype. If a phenotype could be observed due to an alteration of a single gene (and as a result – corresponding protein), such trait is called mendelian. First experimental methods provided very limited information about DNA sequence. Restriction sites in the DNA sequence were used as genetic markers linked to neighboring mutations. Inheritance of such markers could be traced within family. Respectively, markers linked with disease predisposing mutations should segregate with observed phenotype. While finding a disease linked marker is a challenge, it is even more challenging to map marker sequence to a specific location in the genome.

Inheritance of severe mendelian traits is relatively easy to trace within kindred and it is reasonable that first research efforts involved large families with multiple affected and unaffected members. The first gene was mapped to phenotype in 1983 by James Gusella and colleagues[1]. Family of more than 5,000 members with history of Huntington's disease was discovered in Central America. After decade of intensive phenotyping and analyzing DNA marker segregation *HD* gene was discovered with a variable length of trinucleotide repeats. However, mendelian disorders like Huntington's disease or cystic fibrosis, which could be exactly mapped to a specific fully penetrant mutation are extremely rare[2]. Majority of diseases are complex – associated with contributions from multiple genes in combination with environmental factors. Mapping of complex disorder markers within a family is largely impeded because of incomplete penetrance and significantly smaller effect size compared with mendelian traits. This fact promoted methods development for study of genetic markers frequencies in population moving the study design towards case-control approach.

With improvements in sequencing and genotyping technology the list of discovered genetic associations is rapidly expanding. Interpretation of association signals in the scope of disease biology yet remains to be a challenge. Thus, a unified integrative framework reconstructing the trajectory of how a mutation in the DNA sequence affects protein functionality and discovering involved pathways could provide the most valuable information about potential therapeutic target.

Architecture of genetic studies

Almost every disease has contribution of genetic, epigenetic and environmental factors. Mendelian disorders, like cystic fibrosis and sickle cell anemia are results of a single DNA mutation. Another extreme is pathogenic disorders, like HIV, which are caused by environmental factors. However, for vast majority of diseases conclusions could be made only about relative contributions to overall risk from each of factors. Case-control study design for complex genetic disorders relies on a careful assortment of individuals for case and control groups. General approach is to maximize number of individuals with early onset of severe disease symptoms and family history of a disorder in a case group and minimize in controls[3].

Next step is identification of genotypes in both affected and non-affected individuals. Methods, like linkage studies and GWAS were successful in identification of the risk loci, however, they do not have sufficient resolution to highlight individual genes or specific mutations and largely limited to common variation. Success of 1000 genomes[4] and HapMap[5] projects has contributed important tool for mapping of association signal to a limited set of genes, although not sufficient to identify causal mutations. It is important to mention, that complex disorders usually require great statistical power (i.e. large cohort size) to achieve significant results due to small effect size carried by common variation. Introduction of next generation sequencing methods increased resolution of genetic studies up to individual base pairs. In 2010 this led to the first precise determination of genetic cause of disease of a particular patient. Two mutations in Charcot-Marie-Tooth (CMT) patient were identified in *SH3TC2* gene, known cause of CMT. Further sequencing studies showed, that majority of disease associated variants are located in the exome – 1% of DNA sequence that encodes proteins[6]. Exome sequencing provided cheaper and efficient alternative to full genome studies. Traditionally, study design for exome sequencing studies could be based on familial or population data. Familial approach is similar to linkage studies with main difference that now individual protein-coding mutations are used as genetic markers. Once the marker linked to a phenotype is identified problem of association signal mapping to a gene highly challenging in linkage studies is naturally solved due to usage of coding variation for analysis. Though, main advancement brought by next-generation sequencing is ability to study rare variation in population, similarly to common variant analysis in GWAS.

Analysis of rare variation requires more complex approach than GWAS. While individual variants are analyzed in GWAS, contrarily in rare variant studies because of low frequency it is unlikely for any of the individual variants to achieve statistically significant association. Rare variation in a given gene is pooled in case and controls group and burden of alternative alleles is then assessed[7]. There are several pitfalls in this approach. First, there is no well-defined threshold for allele frequency to be called “rare”. Commonly used cutoff is 1%, however, frequency of actual causal variation depends on disease genetic profile and prevalence. Second, the stronger effect on the protein functionality is carried by allele the stronger should be selection pressure on the carriers of such alleles. As a consequence frequency of large-effect mutations is expected to be smaller than those with modest effect size. Uniform pooling of rare variants in a gene based association test looses this valuable information.

Several statistical approaches were proposed to overcome these problems. Instead of using hard allele frequency threshold for pooling of the variants, it was proposed to assign statistical weight to each variant base on observed minor allele frequency and vary the frequency threshold defining which variants should be pooled[7]. Variable threshold test was confirmed to have greater statistical power both in simulated and empirical data. Interestingly, variable threshold approach does not make implicit assumptions about the relationship between allele frequency and odds ratio, though naturally recovers this property.

However, even in phenotypically relevant genes, many variants will be neutral. Statistical power of rare variant analysis has been even further enhanced with idea of comparing variant distributions between cases and controls[8]. Analysis of distribution allows binning of variants into several groups. Variants increasing risk should be more common in cases, protective variants should be more common in controls and neutral ones should be observed in both cohorts at a similar frequency. C-alpha test was developed to detect neutral variation and focus on the most likely functional. While providing sufficient power improvement for rare variation analysis it is not suitable for singletons – variants that are observe only once in the dataset. They do not have a distribution of alleles between cases and controls and in this case C-alpha statistic becomes just a burden test.

Composite approach of common variants GWAS and target sequencing for rare-variation analysis provide more complete information about disease genetics. Genetic studies of Chron's disease are a great illustration of multi-model association testing[9]. It is a complex disorder caused by combination of pathogen and immune factors in genetically predisposed individuals. More than 71 susceptibility loci were discovered through GWAS, suggesting a complex genetic structure of the disease. In 2011, through pooled sequencing of 56 genes identified in GWAS 70 low-frequency protein-altering variants associated with

Chron’s disease were discovered. Completing of allele spectrum in complex disease with targeted rare variation study increases fraction of explained trait heritability.

Identifying type of disease causing variation is critical for most powerful analysis. Numerous variation annotation tools became essential for understanding of the effect carried by a variant. Protein-truncating mutations could carry both loss-of-function effect if found in the beginning of a gene and gain-of-function effect once found in the last exon. However, interpretation of missense variants – resulting in change of a single amino acid, is challenging. As advanced association tests are focusing in identifying and eliminating from analysis likely benign variants, measuring of individual variant functional significance becomes critical. Tools like PolyPhen-2[10] and SIFT[11] use information about protein structure and amino acid conservation to predict severity of disruption in protein functionality. Integration of such predictions with expression data enables transcript-specific annotation.

Development of computational tools for joint variant calling (GATK3[12–14]) made possible assembly of the large datasets of human DNA coding variation. Exome Aggregation Consortium pioneered in this field by producing dataset of more than 60,000 samples of multiple ancestries that were processed through the unified alignment, variant calling and quality check pipeline[15]. Estimates of the observed singleton variants in each gene resulted in creation of the evolutionary conservation metrics. Probability of loss-of-function intolerance and missense constrain Z-score demonstrated great agreement with ClinVar and achieved great performance in filtering out likely benign variation[16].

Final step is data interpretation. Despite critical importance of understanding biology underlying observed genetic association this step up until recently was left for experimental molecular biology research.

Protein network analysis

With the exponential growth of expression data numerous disease pathways were discovered – DNA reparation in cancer, insulin secretion and TGF-beta signaling pathways in type II diabetes, etc. For newly discovered risk genes of particular interest becomes identification of functional modules, sharing common cellular function beyond the classical disease pathways.

Protein-protein interaction databases are important reference set for building novel disease pathways. Recently, it has become feasible to experimentally map large-scale protein-protein interaction networks. Reliability of the reported interactions is essential component enabling usage of this data as reference for pathway discovery. High-throughput methods of screening aided assembly of large experimentally derived interactomes like BioPlex[17]. However, experimental data is usually biased due to usage of a specific cell type. Also, probability of detecting an interaction between two proteins depends on their localization within the cell and level of expression.

Group of Kasper Lage at performed extensive analysis of the protein-protein interaction data reported in the literature to create as scored interactome. Scores, in this case, represent reliability of the observed interaction and are derived from number of times that interaction was observed in independent experiments. Such approach does not fully eliminate experimental biases and noise, however, it does provide an estimate of the data robustness. Most recent version of the database – InWeb_InBioMap has several fold more interactions more interactions and better biological relevance than comparable resources. Integration of the protein-protein interactions with GTEX database of gene expression data resulted in construction of tissue-specific references that are better powered for interpretation of the biological data[18].

It has previously been observed that different genes harboring causal mutations for the same Mendelian disease often physically interact. To evaluate the degree to which this is true of genes within strongly associated loci in complex disease, computational tool DAPPLE was developed[19]. It was observed that number of direct protein-protein interactions between genes found in genome-wide significant loci is significantly greater than in non-associated loci. However, choice of genes that DAPPLE should be seeded on was chosen subjectively with no formal metric evaluating what threshold for association signal should be used.

In fact, problem of finding functional modules within biological networks is common in systems biology. General approach was developed in statistical package BioNet[20,21]. Algorithm that uses Steiner decision tree and exactly solves problem of finding the best scoring module of nodes from the reference network was proposed. Originally, this method was used for analysis of protein-protein interactions in microarray data. However, further developments were made in the field of metabolomics. Reference pathways and interactions in metabolomics are known at a much better level of confidence than protein-protein interactions. For example, using original Dittrich algorithm, analysis of the metabolic networks identified key modules responsible for polarization of macrophages[22]. Search algorithm uses p-values of individual instances (genes, metabolites, etc.) to construct node and edge weighted graph. Until emergence of the gene-based association tests mapping of the genetic association signal to an individual genes was limited. Because of this reason network analysis in genetic studies was not widely used up to date.

Summary

Emergence of the large exome and genome sequencing datasets made analysis of the rare variation in population possible. However, lack of methodology for analysis led to quite few successful rare variant association studies up to date. Moreover, in case of identification of a novel risk gene further interpretation and functional credentialing was left for molecular biologist. For this thesis, we wanted to create a systematic pipeline for analysis of the rare coding variation integrating multiple statistical models and create methods for biological interpretation of the observed associations.

We first developed approach for rare variant association tests incorporating multiple models of causal variation. We used separate analysis of protein truncating variants, missense and ultra-rare (filtered with ExAC database allele frequencies) variation for analysis of group of cases with focal segmental glomerulosclerosis. Proof-of-concept analysis was validated using mouse model and confirmed that even with modest size of case and control cohorts composite approach identifies novel susceptibility genes (Chapter 2).

Next, we tested our rare variant association methodology on the large case-control dataset (about 37,000 samples) of germline variation in cancer. Analysis of the multiple phenotypes of both sporadic and genetically selected cases identified common features of variation observed in known cancer risk genes. Specifically, in cutaneous melanoma cohort we identified a novel causal gene and functional testing confirmed its tumor suppressive activity. To fully explore rare causal variation we looked at blood somatic variation in adult-onset cancers and found intriguing association between solid tumor cancer and burden of mosaic protein truncating variants in blood (Chapter 3).

Rare variation association studies have significant advantage over GWAS – they are focused on the coding variation and mapping of the association signal to a gene becomes straightforward. At the same time, complex disorders require large cohorts to gain enough statistical power for association detection, which cannot be achieved due to sequencing cost limitations. Integration of large GWAS statistical power and mapping simplicity of RVAS aligned on the map of protein interactions is a missing resource that will link genetic data to disease biology that can explain observed phenotype. We then focused on the integration of the available solutions for microarray data with multiple interactome permutation schemes that would test hypotheses of network non-randomness (Chapter 4).

In summary, we developed a composite multi-model rare variant association test methodology and integrated it with novel approach for interpreting and refining results of genetic association studies. Our methodology uses protein interactions data to find the most associated subset of connected genes.

Bibliography

1. Gusella, J. F. *et al.* A polymorphic DNA marker genetically linked to Huntington's disease. *Nature* **306,** 234–8

2. Tsui, L. *et al.* Cystic fibrosis locus defined by a genetically linked polymorphic DNA marker. *Science (80-.).* **230,** (1985).

3. Zondervan, K. T. & Cardon, L. R. Designing candidate gene and genome-wide case-control association studies. *Nat. Protoc.* **2,** 2492–501 (2007).

4. Consortium, T. 1000 G. P. A global reference for human genetic variation. *Nature* **526,** 68–74 (2015).

5. Gibbs, R. A. *et al.* The International HapMap Project. *Nature* **426,** 789–796 (2003).

6. Lupski, J. R. *et al.* Whole-Genome Sequencing in a Patient with Charcot–Marie–Tooth Neuropathy. *N. Engl. J. Med.* **362,** 1181–1191 (2010).

7. Price, A. L. *et al.* Pooled association tests for rare variants in exon-resequencing studies. *Am. J. Hum. Genet.* **86,** 832–8 (2010).

8. Neale, B. M. *et al.* Testing for an Unusual Distribution of Rare Variants. *PLoS Genet.* **7,** e1001322 (2011).

9. Rivas, M. A. *et al.* Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. *Nat. Genet.* **43,** 1066–1073 (2011).

10. Adzhubei, I. A. *et al.* A method and server for predicting damaging missense mutations. *Nat. Methods* **7,** 248–9 (2010).

11. Ng, P. C. & Henikoff, S. SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res.* **31,** 3812–4 (2003).

12. Van der Auwera, G. A. *et al.* From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr. Protoc. Bioinforma.* **43,** 11.10.1-33 (2013).

13. DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43,** 491–8 (2011).

14. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20,**

1297–303 (2010).

15. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536,** 285–291 (2016).

16. Samocha, K. E. *et al.* A framework for the interpretation of de novo mutation in human disease. *Nat Genet* **46,** 944–950 (2014).

17. Huttlin, E. L. *et al.* The BioPlex Network: A Systematic Exploration of the Human Interactome. *Cell* **162,** 425–440 (2015).

18. Li, T. *et al.* A scored human protein–protein interaction network to catalyze genomic interpretation. *Nat. Methods* **14,** 61–64 (2016).

19. Rossin, E. J. *et al.* Proteins Encoded in Genomic Regions Associated with Immune-Mediated Disease Physically Interact and Suggest Underlying Biology. *PLoS Genet.* **7,** e1001273 (2011).

20. Dittrich, M. T., Klau, G. W., Rosenwald, A., Dandekar, T. & Muller, T. Identifying functional modules in protein-protein interaction networks: an integrated exact approach. *Bioinformatics* **24,** i223–i231 (2008).

21. Beisser, D., Klau, G. W., Dandekar, T., Muller, T. & Dittrich, M. T. BioNet: an R-Package for the functional analysis of biological networks. *Bioinformatics* **26,** 1129–1130 (2010).

22. Jha, A. K. *et al.* Network Integration of Parallel Metabolic and Transcriptional Data Reveals Metabolic Modules that Regulate Macrophage Polarization. *Immunity* **42,** 419–430 (2015).

Chapter 2
Composite model for gene based association tests. A role for genetic susceptibility in sporadic focal segmental glomerulosclerosis

Work presented in this chapter was published as:

> Artomov M., *et al.* A role for genetic susceptibility in sporadic focal segmental glomerulosclerosis. *Journal of Clinical Investigations;* 3,126. 2016.

Abstract

Focal segmental glomerulosclerosis (FSGS) is a syndrome that involves kidney podocyte dysfunction and causes chronic kidney disease. Multiple factors including chemical toxicity, inflammation, and infection underlie FSGS; however, highly penetrant disease genes have been identified in a small fraction of patients with a family history of FSGS. Variants of apolipoprotein L1 (*APOL1*) have been linked to FSGS in African Americans with HIV or hypertension, supporting the proposal that genetic factors enhance FSGS susceptibility. Here, we used sequencing to investigate whether genetics plays a role in the majority of FSGS cases that are identified as primary or sporadic FSGS and have no known cause. Given the limited number of biopsy-proven cases with ethnically matched controls, we devised an analytic strategy to identify and rank potential candidate genes and used an animal model for validation. Nine candidate FSGS susceptibility genes were identified in our patient cohort, and three were validated using a high-throughput mouse method that we developed. Specifically, we introduced a podocyte-specific, doxycycline-inducible transactivator into a murine embryonic stem cell line with an FSGS-susceptible genetic background that allows shRNA-mediated targeting of candidate genes in the adult kidney. Our analysis supports a broader role for genetic susceptibility of both sporadic and familial cases of FSGS and provides a tool to rapidly evaluate candidate FSGS-associated genes.

Introduction

The glomerulus of the kidney is a specialized capillary bed that generates an ultrafiltrate that, after modification by the kidney tubule system, becomes urine. Diseases of the glomerulus often lead to chronic kidney disease, a major health care problem affecting between 5% and 10% of the adult population in developed countries[1]. Treatment options are limited, in part owing to the poor understanding of the pathogenesis of glomerular disease. Better insights into the root cause of this disease will offer hope for improvement of this situation.

One of the most common glomerular syndromes is focal segmental glomerulosclerosis (FSGS). The pathologic change of FSGS is a scar that develops focally (in some but not all glomeruli) and segmentally (in only part of a glomerulus). While originally considered a disease, FSGS is now thought to consist of a variety of different syndromes. These include primary (idiopathic) FSGS, which is thought to be caused by a circulating factor, and secondary FSGS, which may be caused by viruses, medications, and genetic mutations. The most common form of secondary FSGS follows glomerular hyperfiltration arising from a mismatch between metabolic load and glomerular capacity and is associated with obesity, low birth weight, reduced renal mass, and other causes. Genetic mutations alone can be sufficient to cause disease (Mendelian) or may increase susceptibility to FSGS by potentiating the effects of environmental factors.

The glomerulus is composed of 3 different cell types: endothelial cells, mesangial cells, and epithelial cells known as podocytes. The podocyte is an unusual cell that covers the outside of the capillary wall and interdigitates with other podocytes to create small slits that allow the passage of fluid and small solutes into the urinary space. It is now clear that podocyte dysfunction is responsible for FSGS as well as other glomerular diseases such as minimal change disease, membranous glomerulopathy, and congenital nephrotic syndrome. Current models suggest that increased podocyte loss is the primary lesion in FSGS[2–5].

Over the past 10 years, various genetic approaches have identified mutations in over 20 podocyte genes as causative or leading to increased susceptibility to FSGS[6,7]. Mutations in these genes, however, explain only a small fraction of familial and sporadic FSGS cases[8–10]. A larger fraction of cases may involve non-Mendelian forms of FSGS that could involve variants in multiple genes that interact together to generate susceptibility to podocyte injury and loss. Further gene discovery in oligogenic disease is challenged, however, by the fact that mutations will be distributed across many genes and be difficult to distinguish from numerous neutral gene variants[11,12]. A greater understanding of genetic causes of FSGS has the potential to elucidate molecular pathways that are involved in the disease.

In terms of the number of people affected, the most significant genetic contributor to FSGS susceptibility identified to date is *APOL1*. FSGS-associated alleles of *APOL1*, called G1 and G2, are common in West African populations, likely as a consequence of providing resistance to trypanosomiasis[13–15]. The presence of 2 variant alleles significantly increases the risk of arterionephrosclerosis (hypertensive nephropathy) (odds ratio [OR] = 7), FSGS (OR = 17), or HIV-associated nephropathy (OR = 29) in African Americans[15,16] and in South Africans (OR = 89)[17]. Approximately 13% of African Americans carry 2 variant alleles and are at increased risk for chronic kidney disease. These variants by themselves largely explain the increased frequency of FSGS among African Americans. Despite this, the mechanisms by which *APOL1* variants cause or predispose individuals to glomerular damage remain unknown. As these variants are absent from individuals lacking any African ancestry, they are not documented to play a role in FSGS susceptibility in individuals of other ancestries.

Here, we used high-throughput sequencing of DNA from FSGS patients of Northern European ancestry to identify genes that are potentially involved in susceptibility to the disease. The challenge of studying the genetics of sporadic FSGS is the possibility that a large number of genes may be involved and the likelihood that each gene contributes only a small amount of risk for the disease. In addition, the relatively low incidence of FSGS in adult and pediatric populations (~5/million/year)[1] and the even fewer number of cases that are confirmed by kidney biopsy preclude the assembly of a cohort of the size required for standard genetic approaches like GWAS, whole-genome sequencing, or exome sequencing[18]. Thus, most genetic studies of FSGS have been family studies.

Here, we sequenced DNA from 214 patients of European ancestry with biopsy-confirmed FSGS and tested a variety of analytic approaches to mitigate our limited sample size. Since FSGS is considered a disease of podocytes, we focused our sequencing analysis on 2,500 genes that are highly and/or specifically expressed in podocytes. This approach significantly reduced the multitest penalty. We also developed a robust analytic pipeline permitting the use of individuals sequenced for other genetic studies as controls. Since there is no in vitro assay for FSGS, we developed a screening method using mice. Our system is based on a murine embryonic stem (ES) cell line with an FSGS-susceptible genetic background that allows for efficient, targeted delivery of shRNAs to generate mice that are nearly 100% derived from the ES cells, eliminating the need for subsequent breeding. This method allowed us to rapidly test 6 candidate genes and validate 3 new FSGS susceptibility genes. We expect that our system will allow for large numbers of candidate genes constituting the network of FSGS genes to be validated and that it will provide critical insight into the pathogenesis of this disease syndrome. In addition, our experimental approach should be broadly applicable to studying other uncommon diseases in which susceptibility genes are suspected.

Results

We conducted high-throughput DNA-sequencing studies focusing on 2,500 genes (~7 Mb) that are highly expressed in podocytes, reasoning that the genetic susceptibility would be intrinsic to the podocyte (**Figure 2.1**).

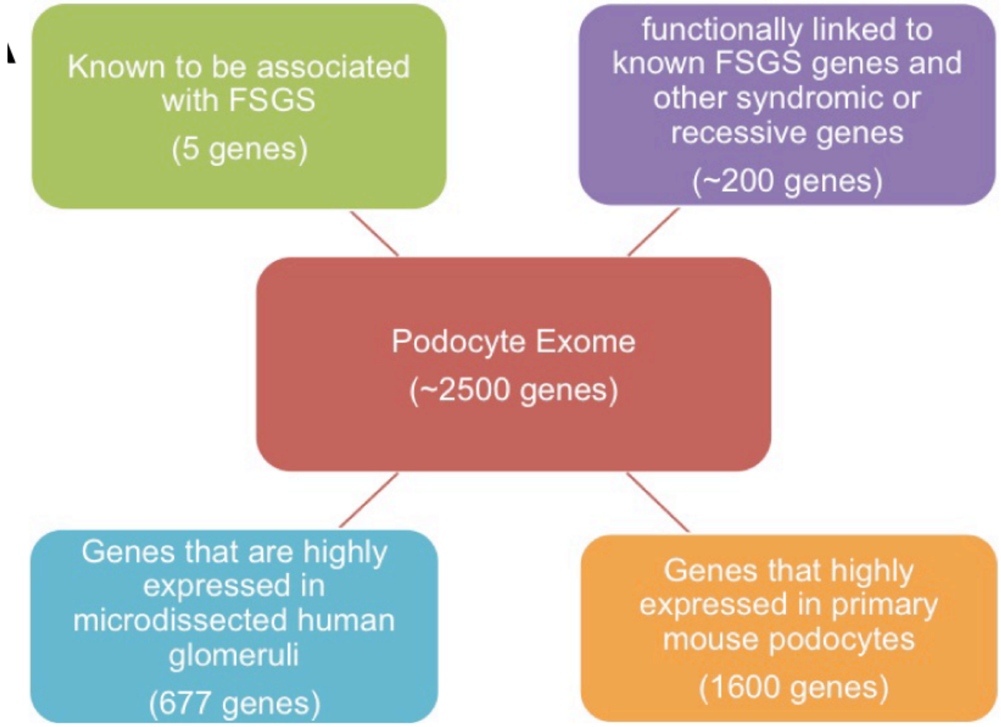

Known to be associated with FSGS
(5 genes)
functionally linked to known FSGS genes and other syndromic or recessive genes
(~200 genes)
Podocyte Exome
(~2500 genes)
Genes that are highly expressed in microdissected human glomeruli
(677 genes)
Genes that highly expressed in primary mouse podocytes
(1600 genes)

Figure 2.1. 2,500 genes clustered and defined as "podocyte exome".

The list of genes that we sequenced included most genes currently implicated in familial FSGS[19–23], approximately 200 genes that are functionally linked to these genes, 677 genes chosen on the basis of their high expression in microdissected human glomeruli[24], and 1,600 human orthologs of highly expressed genes identified by DNA microarrays of mouse podocytes[25–27].

We performed sequencing of DNA from 214 patients of European ancestry with biopsy-confirmed FSGS, including 192 patients with sporadic FSGS and 22 with familial FSGS. DNA samples were obtained from patients participating in a multicenter NIH study of biopsy-confirmed FSGS[16] and from patients diagnosed at Washington University. All subjects provided informed consent for the genetic studies. We focused on patients of European ancestry, because a well-characterized control set used for a genetic study of autism but unascertained for kidney disease was available that had a similar genetic ancestry[28]. A similar control dataset for African or African admixture patients was not available at the time we performed this sequencing study, which prevented us from including these patients in our analysis.

To validate that our patient sequencing data were comparable to those of our control group, we processed data for both patients and controls in a single batch, with raw data aligned to the human genome[29–31]. The depth of coverage was compared between patients and controls, and only those exons covered adequately (>20 times) and similarly in both patients and controls were advanced to the analysis stage. In summary, 16,784 exons and 2,769,942 bp were confidently covered in both patient and control cohorts, resulting in 16,008 SNPs and 1,724 genes analyzed in the final dataset. SNP calls were equally represented in patients and controls.

Thirty-two patients were removed from the study but reserved for follow-up because trace Hispanic ancestry was detected by principal component analysis (PCA) (**Figure 2.2**, **A-C**). Three patients were removed because the call rate of

SNPs was less than 95%. The remaining FSGS patients (157 sporadic and 22 familial) had a similar number of SNPs, heterozygous genotypes, and genotypes containing an alternative allele per sample (**Figure 2.2**, **E-G**), allowing us to proceed to association analysis.

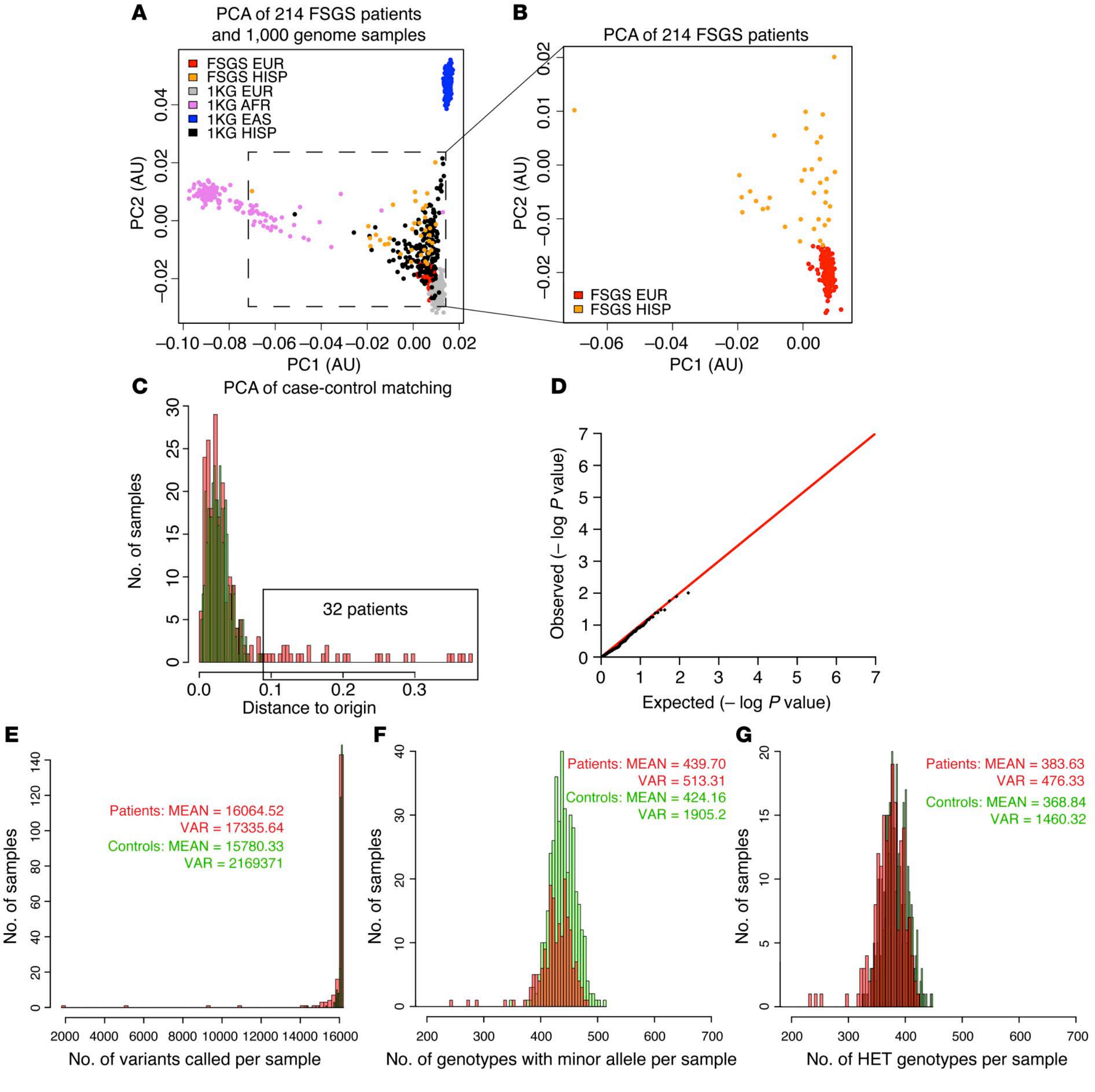


Figure 2.2. (A) PCA plot of FSGS patients and 1,000 genome samples. The inset shows the distribution of putative Northern European FSGS patients in the PCA plot in relationship to 1,000 genome samples. (B) Magnified view of the inset area in A. (**C**) PCA analysis of patients and controls is depicted as the distance from the origin. Thirty-two patients with a highly similar variant profile but with a distance of more than 0.9 were removed and used as a follow-up group. (**D**) Fisher's exact test of the common (MAF >5%) variants showed the

Figure 2.2 (Continued) absence of stratification and confirmed the validity and quality of our method for case-control matching. (**E**) Comparison of the total number of variants per sample showed that patients and controls were similar. (**F**) Comparison of the total number of heterozygous genotypes showed that patients and controls were similar. (**G**) Comparison of the total number of heterozygous and homozygous genotypes containing an alternative allele showed that patients and controls were similar. EUR, European; HISP, Hispanic; AFR, African; EAS, East Asian; VAR, variants; HET, heterozygous; PC1, principal component 1; PC2, principal component 2; KG, from 1000 Genomes Database.

Our final dataset contained 179 patients and 378 controls and included 157 sporadic and 22 familial FSGS patients. The accuracy of our analysis strategy was confirmed by resequencing key SNPs using Sanger sequencing and by showing that sequencing the same sample at both the Broad Institute and Washington University gave similar results.

An association test examining single variants (minor allele frequency [MAF] >1%) was performed using Fisher's exact test. No variants were detected with a *P* value below the multitest threshold ($2x10^{-5}$). The lack of significance is not surprising, given the relatively small size of our sample. This analysis did confirm that the distribution of synonymous and missense variants was similar between patients and controls (**Figure 2.2 D**). Table 1 shows a list of the 10 highest-scoring variants. All the variants are missense variants. As a follow-up, we analyzed the 32 samples with Hispanic admixture, combined with 23 additional European ancestry samples that were sequenced from the Nephrotic Syndrome Study Network (NEPTUNE) cohort[32]. This confirmed enrichment of 3 of the missense sequence variants in *WNK4*, *KANK1*, and *ARHGEF17* (**Table**

2.1). Interestingly, *KANK1* was recently identified as a susceptibility gene for familial nephrotic syndrome[33].

Table 2.1. Single variants enriched in patients versus controls and in the follow-up cohort. POS, chromosome position of the variant; REF, reference allele; ALT, alternative allele; MINA, number of alternative alleles in patients; SMINA, number of alternative alleles in cases with sporadic FSGS; FMINA, number of alternative alleles in cases with familial FSGS; MINU, number of alternative alleles in controls; ESP EA, allele frequencies in European Americans in the NHLBI Exome Sequencing Project; ESP AA, allele frequencies in African Americans in the NHLBI Exome Sequencing Project; P, Fisher's exact test p-value; FLW ALT, number of alternative alleles in follow-up cohort; FLW AF, allele frequency in the follow-up cohort. The frequency of single variants (MAF>1%) was assessed in patients versus controls and high-scoring variants with odds ratio greater than 2.5 are shown in this table ranked by p-value.

POS	Gene Name	REF	ALT	MINA	S MINA	F MINA	MINU	ESP EA	ESP AA	P	FLW ALT	FLW AF
chr17 40947320	WNK4	C	T	6	6	0	0	1.3x10^{-3}	0.5	1.00x10^{-3}	2	1.8x10^{-2}
chr9 710966	KANK1	G	A	15	15	0	8	8.0x10^{-3}	0.5	1.10x10^{-3}	14	0.13
chr2 113737630	IL36G	C	A	5	5	0	0	1.0x10^{-3}	0.12	1.50x10^{-3}	0	0
chr11 73020633	ARHGEF17	G	C	5	5	0	0	2.0x10^{-3}	0.13	2.80x10^{-3}	9	8.20x10^{-2}
chr17 40939855	WNK4	G	T	5	5	0	0	1.2x10^{-3}	0.2	3.20x10^{-3}	2	1.80x10^{-2}
chr22 36661906	APOL1	A	G	5	5	0	0	3.4x10^{-4}	0.2	3.20x10^{-3}	7	0.64

Rare variant analysis identified 6 new potential FSGS susceptibility genes. We analyzed the rare variants (MAF <1%) using tests that compared the total numbers of rare variants between patients and controls. Currently, it is believed that there is an inverse correlation between the frequency of the allele and the potential risk. Thus, at each gene, we tested 2 distinct modes of inheritance for FSGS: the increased presence of extremely rare alleles that are predicted to be highly damaging and therefore highly penetrant (model 1), or the presence of low-frequency and less-damaging alleles with risk and protective variation intermingled with neutral variation (model 2). To discriminate between these 2 models, we analyzed 2 subsets of variants. For the first model, we used the Exome Aggregation Consortium (ExAC)[34] browser to identify 5,662 missense and loss-of-function variants in our dataset that are present at a frequency of less than 0.01% in the European population. We then tested the burden of these rare variants in FSGS patients versus controls[35]. Using this analysis, we found that no genes reached a level of statistical significance for rare, highly penetrant variants under this model.

To examine the second model involving low-frequency risk and protective variants, we selected all missense and loss-of-function variants with a MAF of less than 1% and compared their distribution between patients and controls using 2 different rare variant tests: the variable threshold (VT)[36] and the C-α test[37]. Because the effect sizes of variants differ, the accuracy of each method can vary depending on the specific situation. Using a P value of less than 0.05 (Bonferroni-corrected $P \leq 3\text{x}10^{-5}$) as a cutoff, no genes were identified that exceeded this value, but 8 genes (*WNK4*, *APOL1*, *DLG5*, *GCC1*, *XYLT1*, *KAT2B*, *BPTF*, and *COL4A4*) had P values close to the Bonferroni-corrected value ($P < 6\text{x}10^{-5}$ to $P < 8\text{x}10^{-4}$, **Table 2.2**). Since the Bonferroni test tends to be conservative and *APOL1*[13–15] and *COL4A4*[38] are known FSGS genes, we selected these genes for further analysis as potential FSGS susceptibility genes.

Our analysis of extremely rare variants (MAF <0.01%) in this set of 8 genes showed enrichment in 3 of these genes: *GCC1*, *APOL1*, and *COL4A4*. Examination of our follow-up group (55 samples) confirmed enrichment of a subset of the same rare variants found in our larger cohort in all of the genes except *COL4A4* (**Table 2.2**), supporting the findings of our rare variant analysis.

Table 2.2. Top genes identified by rare variant analyses. Rare, missense and nonsense variants (MAF<1%) were pooled for rare variant analysis using variable threshold (VT) and C-α tests. The top genes identified by each test are shown ranked by P-value. Genes with P-values of less than 8×10^{-4} were selected for further analysis.

Gene	C-α test	Variable threshold test	Allele counts	
	P-value	P-value	Patients (N=178)	Controls (N=378)
XYLT1	1.74×10^{-4}	1.38×10^{-3}	29	18
APOL1	3.36×10^{-4}	1.78×10^{-3}	8	2
KAT2B	4.37×10^{-4}	8.59×10^{-2}	8	5
WNK4	7.63×10^{-4}	3.10×10^{-4}	40	18
BPTF	7.68×10^{-4}	2.58×10^{-3}	24	27
COL4A4	2.34×10^{-2}	6.75×10^{-5}	22	9
DLG5	2.96×10^{-3}	7.71×10^{-5}	50	38
GCC1	2.91×10^{-3}	4.84×10^{-4}	14	5

Since *APOL1* and *COL4A4* were already known[13–15,38], the remainder of the identified genes (*BPTF*, *DLG5*, *GCC1*, *KAT2B*, *WNK4*, and *XYLT1*) could represent 6 new potential FSGS susceptibility genes. Notably, *WNK4* was also identified by single-variant analysis.

Interestingly, 4 patients with sporadic FSGS carried the *APOL1* G1 variant (G1), a known risk variant for FSGS that is present in 29% of African Americans but is a rare variant (0.03%) in European Americans. This finding was also seen in the follow-up set, in which 7 of 55 additional samples were identified with this mutation. Since the allele frequency of the *APOL1* G1 allele in approximately

5,500 Hispanic samples in the ExAC dataset was only 0.5%, this represents significant enrichment, regardless of the Hispanic admixture.

Family studies have identified nearly 30 genes that cause familial FSGS[7]. To determine whether a set of 20 of these genes are also involved in sporadic FSGS, we assessed the frequencies of predicted damaging, rare coding variants (missense and loss-of-function with a MAF <1%) in these genes in patients and controls. Approximately 36.9% of patients (66 of 179) had at least 1 predicted deleterious rare variant in these 20 genes compared with 3.4% of controls (13 of 378). The distribution of variants between familial and sporadic cases was similar and consistent with previous studies showing that approximately 30% of steroid-resistant nephrotic syndrome patients presenting before the age of 25 have a variant in one of the known disease genes[39]. There was also a difference in the total number of unique rare variants identified in patients (32.9%, 59 variants in 179 patients) versus controls (3.9%, 15 variants in 378 control subjects). The significance of this finding was tested using a permutation analysis of differences in variants between patients and controls in all potential random groups of 20 genes chosen from our database of 1,724 genes. This showed, however, that 27% of the random sets of 20 genes had a similar or higher burden of rare variants compared with the set of 20 FSGS genes. This suggests that our patient dataset contains additional novel FSGS susceptibility genes with strong genetic effects.

Since FSGS cannot be modeled in vitro and most confirmatory studies are performed today in zebrafish[33,40–42], we developed a genetic system in mice to examine the function of candidate genes in vivo. Our strategy involved inhibiting the expression of candidate genes in podocytes from mice on a genetic background that is prone to develop FSGS. Mice that are heterozygous for 2 podocyte genes, *Cd2ap* and *Synpo*, develop FSGS with an incomplete penetrance (~25%–50%) and significant albuminuria occurring at about 6 months of age[43]. Assuming that FSGS is an oligogenic disease, we reasoned that knocking down a bona fide disease gene in this background would accelerate disease onset.

We generated a mouse ES cell line that was *Cd2ap+/–* and *Synpo+/–* using standard methods. The ES cell also expresses a podocyte-specific and doxycycline-inducible (DOX-inducible) transactivator (*Nphs1*-rtTA3G) that allows inducible expression of an shRNA[44] (**Figure 2.3 A**). We reasoned that inducible RNAi would allow us to study the role of a gene in the mature kidney without worrying about developmental effects. The new method using laser-assisted microinjection into 8-cell embryos[45] allowed us to generate mice that were nearly 100% derived from the ES cells without further breeding (**Figure 2.3 B**). Consistent with mice generated by conventional breeding, approximately 50% of the mice generated from these ES cells developed mild proteinuria after 12 to 16 weeks of age (**Figure 2.3 C**).

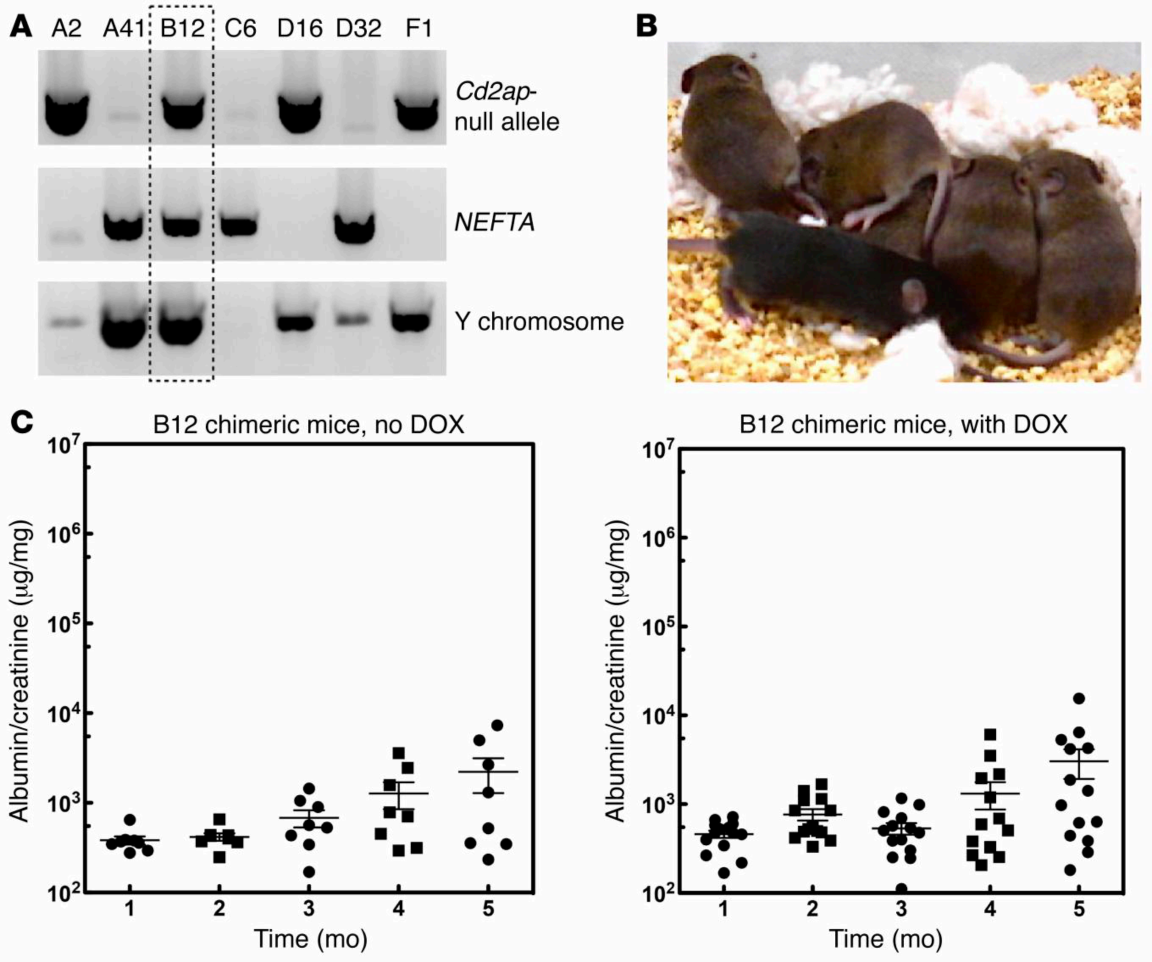


Figure 2.3. Development of ES cells sensitized for FSGS. (**A**) Identification of FSGS-sensitized ES cells. Our breeding strategy predicted that 1 of 8 embryos would have the correct genotype. ES cells were generated using standard approaches and genotyped for *Cd2ap* heterozygosity (upper panel), the *NEFTA* transgene (middle panel), and the Y chromosome (lower panel). (**B**) Laser-assisted injection generated mice with high chimerism. In the example shown, the ES cell line (agouti) was injected into 8-cell C57/BL6 (black) embryos. Compared with noninjected embryos (resulting in the 2 black mice shown on the bottom), all of the injected embryos generated pups that were close to being purely agouti. Injection of ES cells into C57/BL6 albino embryos resulted in completely agouti animals (not shown). (**C**) Mice generated from ES cells developed mild proteinuria after 4 months, with no DOX treatment. Fifteen mice were generated from the sensitized ES cells and treated with or without DOX in the drinking water. Urine was tested every month by measuring the albumin/creatinine ratios. Mice developed low-level proteinuria at 4 months of age, but the level of proteinuria was not affected by DOX treatment.

To eliminate variability introduced by random integration of an RNAi transgene, we targeted a single copy of the RNAi transgene into the mouse *Hprt1* locus[46] that allows the use of 6-thioguanine for efficient selection (**Figure 2.4 A**).

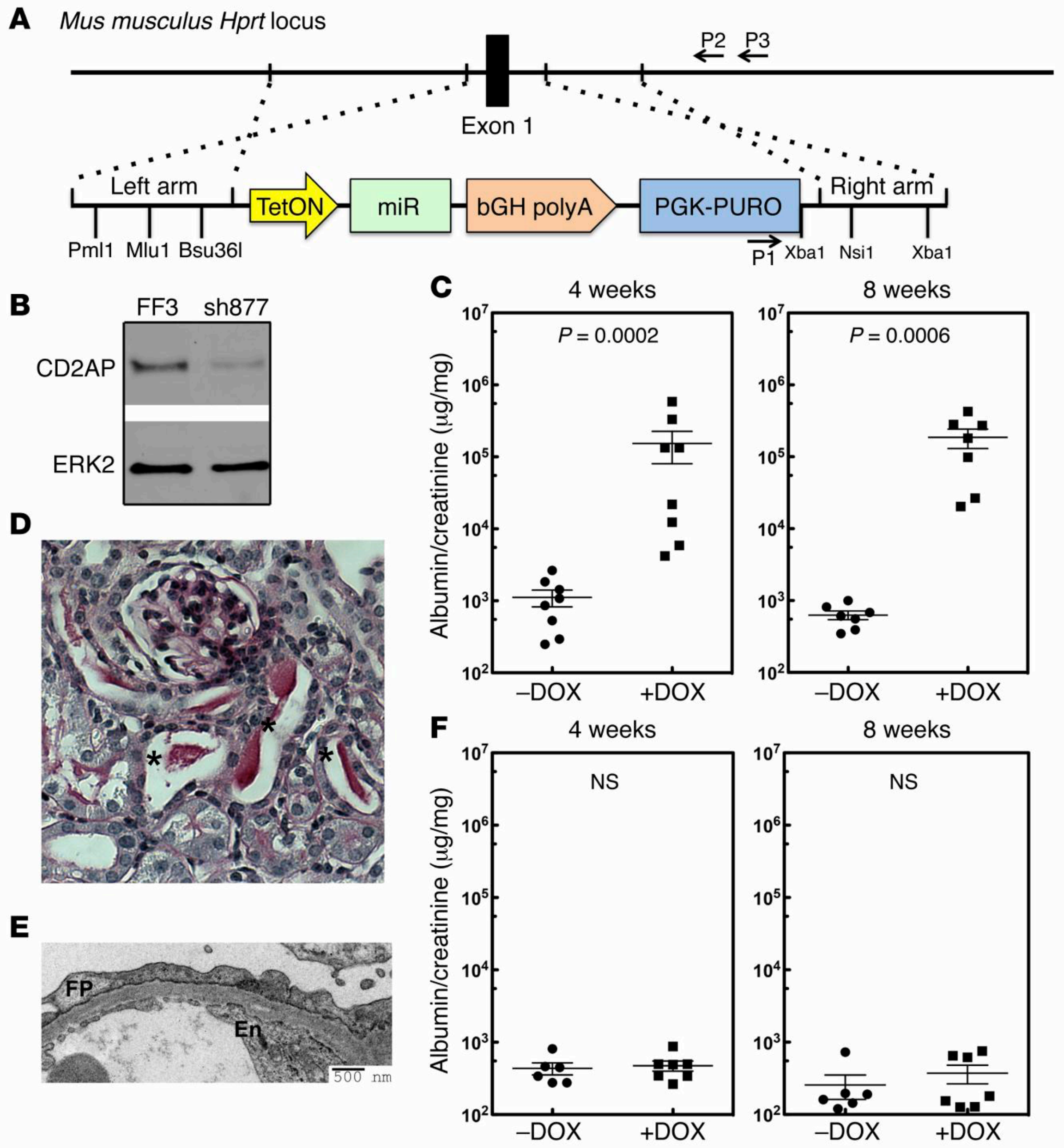


Figure 2.4. Validation of the system using CD2AP knockdown. (**A**) Targeting strategy used to integrate a miR30-shRNA transgene into the *Hprt1* locus. (**B**) Knockdown efficiency of a miR30-shRNA for *Cd2ap* (sh877). Immunoblot shows endogenous CD2AP in NIH3T3 cells stably transduced with FF3 (control shRNA)

Figure 2.4 (Continued) or sh877. Panel **B** represents multiple experiments (*n* = 3) conducted to test the efficiency of the RNAi. (**C**) Sixteen mice generated with ES cells with the *Cd2ap* shRNA that was targeted to the *Hprt1* locus were treated with or without DOX, and urine was analyzed by measuring the urine albumin/creatinine ratio at 4 and 8 weeks. (**D**) Histology from a representative *Cd2ap* RNAi mouse treated with DOX showing protein casts (indicated with asterisks; *n* >5). (**E**) Representative electron microscopic image from a *Cd2ap* RNAi mouse treated with DOX shows podocyte foot process (FP) effacement. En, endothelial cells (*n* = 9). (**F**) Thirteen control mice were generated with a control luciferase RNAi targeted to the *Hgprt* locus. Mice were treated with (*n* = 6) or without (*n* = 7) DOX, and urine was analyzed by measuring the albumin/creatinine ratio at 4 and 8 weeks. A 2-tailed Mann-Whitney *U* test was used to calculate the *P* values in **C** and **F**. A *P* value of less than 0.05 was considered statistically significant.

Since *CD2AP* is an FSGS disease gene[47] and knockout (KO) mice develop severe proteinuria[48], we validated our system by generating *Cd2ap* RNAi mice. Multiple *Cd2ap*-specific RNAis were tested for their ability to inhibit *Cd2ap* expression (**Figure 2.4 B**), and the RNAi showing the greatest inhibition (sh877) was embedded into a miR30 sequence that allows for DOX-inducible expression[49]. An RNAi for the firefly luciferase gene (FF3) was used as a control. Half of the founder (F0) animals were treated with DOX at 2 weeks of age to induce shRNA transgene expression. All of the DOX-treated mice developed sustained proteinuria that was over 150-fold higher than that seen in the control animals (**Figure 2.4 C**). Histological analysis of the kidneys revealed protein casts in the tubules (**Figure 2.4 D**). Electron microscopic examination of the kidney showed widespread foot process effacement, a marker of proteinuria (**Figure 2.4 E**), validating that our RNAi strategy could be used to test candidate FSGS genes. Interestingly, the proteinuric mice recovered after removal of DOX treatment. In contrast, FF3-RNAi mice did not show proteinuria after treatment with DOX for 8 weeks (**Figure 2.4 F**), and no abnormalities were detected by electron microscopy or histology.

Three of the six genes, *WNK4*, *DLG5*, and *KAT2B,* identified by rare variant analysis were chosen for testing. We also chose the 3 single-variant candidates, *KANK1*, *WNK4*, and *ARHGEF17*. Since *WNK4* was present on both lists, a total of 5 genes were selected for analysis. Because the exact mouse ortholog for human *KANK1* is unknown, because *Kank2* is more highly expressed in mouse podocytes[50], and because *Kank1* and *Kank2* were recently identified as susceptibility genes for nephrotic syndrome, we targeted both *Kank1* and *Kank2*. Multiple shRNAs were generated for the 6 candidate genes. Their efficacy was validated in vitro, and the best one was targeted to the *Hprt1* locus (**Figure 2.5 A**). Two independent clones for each candidate gene were selected, and 15–30 mice were generated by laser-assisted microinjection.

Half of each cohort was given DOX, and proteinuria (albumin/creatinine), an indicator of podocyte function, was assessed at 4 and 8 weeks after DOX treatment (**Figure 2.5 B–G**). All 3 RNAi transgenes, *Wnk4*, *Arhgef17*, and *Kank2*, induced substantial proteinuria, with a level of proteinuria that was significantly higher than that seen in the control mice (**Figure 2.5 B–E**). In contrast, the *Dlg5*, *Kat2b*, and *Kank1* RNAi mice did not show statistically significant elevations of proteinuria after 4 or 8 weeks of DOX treatment. Because there was a slight trend toward increased proteinuria in the *Kat2b* and *Kank1* mice, we followed the proteinuria levels for an additional 4 weeks. After 12 weeks, *Kank1* mice had a clear proteinuric phenotype, while the proteinuria present in *Kat2b* mice was still not significant. Thus, *Wnk4*, *Arhgef17*, *Kank1*, and *Kank2* mice were positive for proteinuria, while *Dlg5* and *Kat2b* mice were negative for proteinuria.

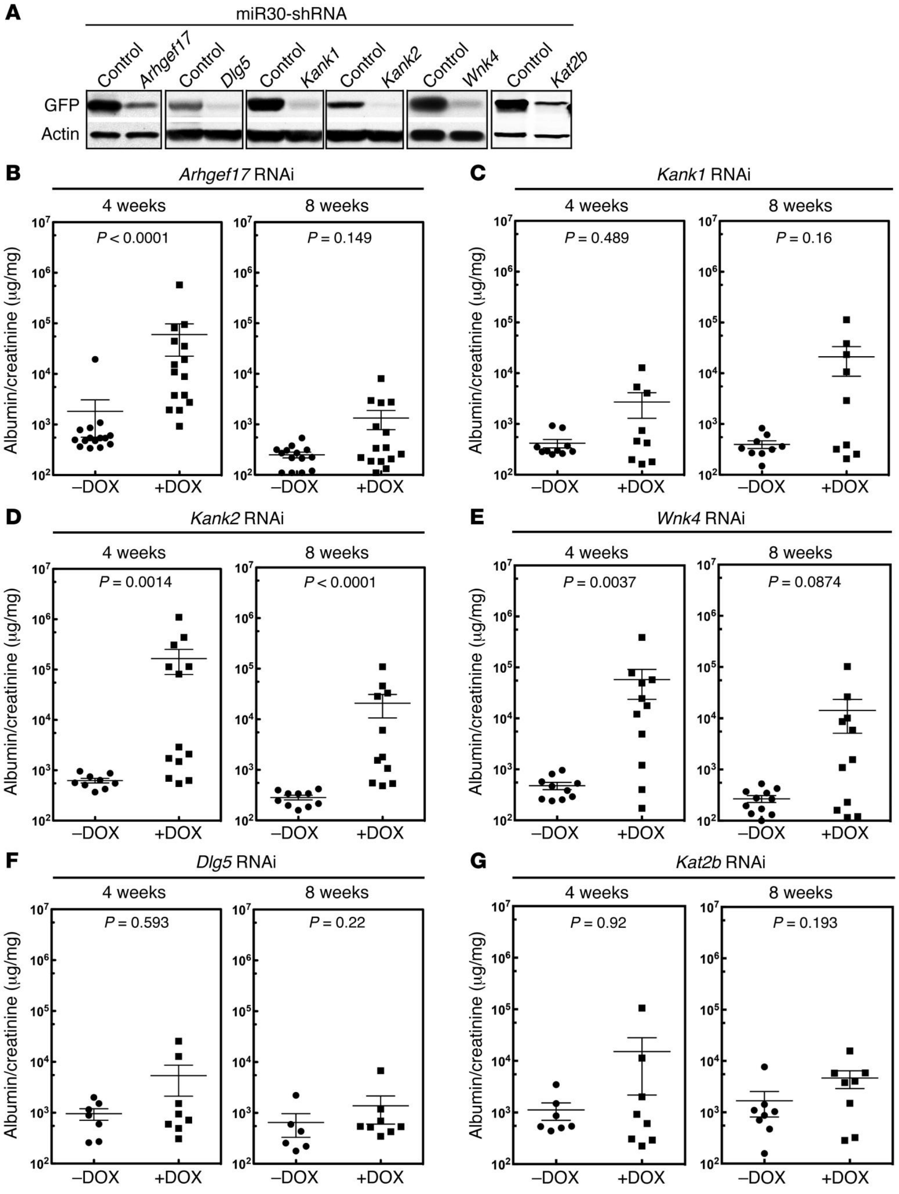


Figure 2.5. Validation of 5 candidate FSGS disease genes. (A) Validation of shRNAs for *Arhgef17*, *Dlg5*, *Kank1*, *Kank2*, *Wnk4*, and *Kat2b*. As described in Methods, shRNAs were tested for the ability to inhibit a target sequence fused to GFP in 293 cells. GFP immunoblotting was used to measure the degree of inhibition. Each immunoblot is representative of at least 3 independent experiments measuring RNAi efficiency. (B–G) Mouse validation screening for candidate FSGS genes. ES cells were generated with the specific RNAis

Figure 2.5 (Continued) targeted to the *Hgprt* locus. Essentially pure chimeric mice were generated by laser-assisted microinjection of ES cells into C57BL6 8-cell embryos. Injections generally resulted in cohorts of 14 to 30 animals; smaller cohorts of animals were not used. Mice were divided into 2 groups and treated with or without DOX to induce expression of the RNAi transgene. Urine albumin/creatinine ratios were measured 4 and 8 weeks after DOX treatment. Albumin/creatinine ratios are shown for each cohort of mice at the indicated time points. A 2-tailed Mann-Whitney *U* test was used to calculate the *P* values for B–G. A *P* value of less than 0.0083 was considered statistically significant (multitest penalty was used).

We confirmed the *Dlg5* result by obtaining *Dlg5*-KO mice[51] and generating *Dlg5*, *Cd2ap*, and *Synpo* triple-heterozygous mice using conventional breeding. No kidney dysfunction was detected, confirming our RNAi result. As expected, electron microscopic examination of the kidneys showed podocyte foot process effacement in *Arhgef17*, *Kank1* (12 week-time point), *Kank2*, and *Wnk4* mice, but not in *Dlg5* RNAi mice. While the overall morphology was normal, some focal areas of mild foot process effacement could be seen in the *Kat2b* mice.

We added *KANK1*, *WNK4*, and *ARHGEF17* to the list of the 20 known FSGS genes and reanalyzed differences in the number of rare variants between patients and controls. Approximately 53.5% of patients (84 of 179) had at least 1 predicted deleterious rare variant in these 23 genes compared with 5.6% of controls (21 of 378) (**Table 2.3**). We separated the patients by sporadic and familial FSGS and found a similar distribution, with 50% of sporadic FSGS patients and 64% of familial FSGS patients having variants in these 23 genes.

Table 2.3. Variant distribution from sequencing analyses.

	EA Sporadic FSGS	**EA Familial FSGS**	**Total EA FSGS**	**Replication Cohort**	**Total # of Controls**
Total samples sequenced	171	11	182	32	378
Samples that passed QC	168	11	179	32	378
Patients with deleterious variants in 20 known FSGS genes	60	6	66	11	13
Patients with known *APOL1* G1 allele	4	0	4	7	0
Patients with rare deleterious variants in 8 genes*	52	3	55	16	43
Patients with variants in 3 validated genes**	32	3	35	10	12
Percentage of patients with variants in 3 validated genes	19.0%	27.3%	19.6%	31.3%	3.2%
Patients with variants in 20 known + 3 validated genes	84	7	91	18	21
Percentage of patients with variants in 20 known + 3 validated genes	50.0%	63.6%	50.8%	56.0%	5.6%

* - Genes identified by rare variants analyses (Table 2.2).
** - Genes validated by mouse model (*ARGHEF17, KANK2, WNK4*)
EA – European American

We tested random sets of 23 genes by permutation analysis of patients and controls, which showed that only 0.67% ($P < 1.6 \times 10^{-27}$) of random sets of 23 genes chosen from the controls equaled or matched the burden of rare variants seen in the patients for these 23 genes. This supports the idea that genetic variants in these 23 genes account for most of the disparity between patients and controls in the numbers of rare variants. This also supports the idea that a specific subset of genes may function more broadly to create a susceptible background for the development of sporadic FSGS.

Discussion

FSGS is a syndrome of diverse etiology that shares a common histologic pattern of focal and segmental glomerular scarring, together with glomerular proteinuria and progressive loss of renal function. The majority of FSGS cases involve primary FSGS, adaptive FSGS, or APOL1 FSGS; less common are viral FSGS, Mendelian FSGS, and medication-associated FSGS. As there are no validated methods to specifically distinguish sporadic (non-familial) FSGS, the present study included subjects with both primary and adaptive FSGS as well as subjects with familial FSGS. Because of the strong predictive power of family history, and because only a small percentage of individuals affected by known etiological factors develop FSGS, the genetic background of the individual is thought to play an important role[6].

The critical locus of injury in FSGS is now thought to be the podocyte[2], a terminal-differentiated cell that has limited replication potential[52]. In the normal kidney, small numbers of podocytes are continuously lost over time[4], and when podocyte numbers drop below a critical level, kidney failure inexorably ensues[2,5,53]. Environmental insults and genetic susceptibility are thought to enhance the rate of podocyte loss, and this increases the probability of developing FSGS. Interpreted this way, the FSGS "lesion" likely represents the common outcome of a wide variety of pathogenetic causes.

In validating genetic susceptibility in sporadic FSGS, a significant challenge is the likelihood that a large number of genes may be involved and that each gene contributes only a small amount of risk for the disease. Additionally, the role of mutations in a specific gene may affect only a small number of patients. This substantially increases the challenge of gene identification in any large genetic study. An additional complication is that sporadic FSGS is relatively uncommon, and most patients do not have a biopsy-confirmed diagnosis. This currently precludes the assembly of a large enough cohort for strong statistical analysis. Because of this, most of the FSGS disease genes identified to date are from family studies, from the sequencing of candidate genes based on the phenotype of mouse models, or from admixture linkage studies of African Americans[15,19–23,25,47,54–57].

Here, we used next-generation sequencing to identify FSGS susceptibility genes. Because of our relatively modest sample size, we adjusted our analytic approach to maximize our ability to identify candidate genes. As both rare and common variants have allele frequencies that are determined by ancestry, well-matched controls for ancestry are required. Since large control datasets for individuals of European ancestry are already available, we focused on FSGS subjects of European ancestry. Because DNA variant calling can be different between institutions and between platforms, we established a pipeline to validate that the datasets were comparable. FSGS is more common in African Americans, but the complex genetic admixture in this population will require a large and complex control dataset that is currently not available. Our focus on genes expressed in podocytes allowed us to focus on higher-likelihood genes and minimized the multitest penalty.

Genetic analysis of FSGS is challenging because of the potentially broad genetic heterogeneity of the disease and the relatively small number of subjects available for analysis when the subjects' ancestry needs to be controlled. Rare variant analysis in ethnically admixed populations such as those found in the United States will require new statistical approaches and the development of large, ancestrally matched control datasets. Nonetheless, our work shows that current statistical approaches, combined with focused sequence analysis, can identify candidate genes from a relatively small sample for a syndrome like FSGS that has widely divergent etiologies. While our sample size was sufficient to extract a list of candidate genes using rare variant analysis, a sample size of at least one order of magnitude larger would be necessary to generate statistically significant data for the single variants[18]. Here, we used next-generation sequencing to identify FSGS susceptibility genes. Because of our relatively modest sample size, we adjusted our analytic approach to maximize our ability to identify candidate genes. As both rare and common variants have allele frequencies that are determined by ancestry, well-matched controls for ancestry are required. Since large control datasets for individuals of European ancestry are already available, we focused on FSGS subjects of European ancestry. Because DNA variant calling can be different between institutions and between platforms, we established a pipeline to validate that the datasets were comparable. FSGS is more common in African Americans, but the complex genetic admixture in this population will require a large and complex control dataset that is currently not available. Our focus on genes expressed in podocytes allowed us to focus on higher-likelihood genes and minimized the multitest penalty.

Genetic analysis of FSGS is challenging because of the potentially broad genetic heterogeneity of the disease and the relatively small number of subjects available for analysis when the subjects' ancestry needs to be controlled. Rare variant analysis in ethnically admixed populations such as those found in the United States will require new statistical approaches and the development of large, ancestrally matched control datasets. Nonetheless, our work shows that current statistical approaches, combined with focused sequence analysis, can identify candidate genes from a relatively small sample for a syndrome like FSGS that has widely divergent etiologies. While our sample size was sufficient to extract a list of candidate genes using rare variant analysis, a sample size of at least one order of magnitude larger would be necessary to generate statistically significant data for the single variants.

An innovation of our approach was the development of a robust pipeline that allowed the use of data on individuals sequenced for other studies as controls. The ability to combine datasets generated at different institutions for different types of studies will become increasingly important and powerful as sequencing becomes more widespread. In our initial studies, we found that batch effects caused by different approaches used for sequencing among different institutions could be a confounding factor precluding the use of analyzed data generated at 2 different institutions. However, by applying the same sequencing read alignment and variant calling pipelines to the primary sequencing data from both patients and controls, we were able to eliminate this variable. We validated our approach by establishing a method for case-control genotype matching and removal of any stratification as well as verifying that primary sequencing data from 2 different institutions using the same control DNA sample gave similar results.

Using a *P* value of less than 0.05, no genes were identified by rare or single-variant analysis that reached genome-wide significance because of Bonferroni's multiple test correction. Because the Bonferroni test tends to be conservative, we assembled a list of the top 8 genes identified by rare variant analysis and the top 3 genes identified by single-variant analysis with *P* values that were close to the Bonferroni corrected *P* value. Supporting the veracity of this analysis, 3 of the genes, *APOL1*, *COL4A4*, and *KANK1*, were already known FSGS susceptibility genes[15,16,33,38], and *WNK4* was identified on both lists.

We initially sequenced over 700 biopsy-confirmed FSGS samples, but most of these samples were genetically admixed, preventing further analysis because of the lack of matched controls. We therefore focused only on the patients of European ancestry as defined by PCA. Because the number of patients of European ancestry with biopsy-confirmed FSGS is extremely small, it is not possible to assemble a true replication study. Also, because of cost, case-control studies with replication using whole-exome or whole-genome sequencing have, in general, been extremely limited and are not yet commonplace in the literature. As a confirmatory approach, we used the 33 samples that we had eliminated because of Hispanic admixture and 23 additional European ancestry samples that we had sequenced subsequent to the original analysis as a confirmatory or follow-up dataset. Our analysis of this second dataset confirmed an increased burden of rare variants in the 6 listed rare variants as well as an increase in 3 common variants (**Table 2.1**). Since *WNK4* was identified by both processes and *APOL1*, *COL4A4*, and *KANK1* are known genes, at least 7 new candidate genes were identified by our sequencing analysis. While the groups were small, the distribution of variants did not seem to differ significantly between the sporadic and familial FSGS cases.

We were surprised to identify the *APOL1* G1 variant in 4 of our subjects and in 7 of the subjects in our follow-up set, as it is rare in non-African populations. The enrichment of this variant in 11 of 208 of our non-African subjects suggests that this particular allele may interact with other variants, leading to susceptibility to FSGS. This is supported by the enrichment of rare, predicted deleterious *APOL1* variants in our subjects. The enrichment in our European American subjects with variants that are common in African Americans, but rare in European Americans, was also found in *WNK4*, *KANK1*, and *ARHGEF17*. The absence of neighboring African SNPs suggests that these are ancestral variants and not due to admixture.

With the availability of large-scale DNA sequencing of human populations, the identification of disease candidate genes and potential disease-associated variants will become more common. How these candidate genes and variants will be validated is unclear. Here, we demonstrate a pipeline that uses common and rare variant association analyses to identify candidate genes from sporadically affected unrelated individuals and control sequence data that were previously generated. We then developed a method to allow these candidates to be tested in vivo. Our method relied on generating an ES cell line that was sensitized for the development of FSGS. It should be possible to generate ES cells on other disease-specific backgrounds to allow for validation studies that will be required to facilitate the discovery of genetic variants associated with both rare and common diseases.

Materials and Methods

Sample preparation and sequencing were carried out using standard protocols for targeted capture and Illumina sequencing. In brief, genomic DNA was fragmented to 150 to 200 bp using a Covaris E220 Focused Ultrasonicator. The ends of the fragmented DNA were repaired using a mixture of T4 DNA polymerase, Klenow polymerase, and T4 polynucleotide kinase. Subsequently, adapters for Illumina sequencing were ligated onto the fragments. These libraries were then hybridized to biotinylated DNA probes from regions of interest (manufactured by MyGenostics). After washing away DNA libraries that bound nonspecifically to the probes, DNA of interest was recovered using Dynabeads MyOne Streptavidin T1 (Life Technologies). Resulting DNA libraries were amplified, if needed, for sequencing on an Illumina HiSeq 2500.

We performed alignment of the raw sequencing data and variant calling according to GATK best practices with the BWA/Picard/GATK software pipeline of the Broad Institute. To insure that gene loci were equally covered in both patients and controls, we performed quality control on patients' and controls' genotypes separately, applying the following filters: (a) retention of only variants that PASS all GATK quality filters; (b) retention of genotypes with DP>10,GQ>30,AB for hets 0.3<AB<0.7, for homozygous alternative AB<0.3; and (c) retention of all variants with less than 5% missing genotypes. After applying these filters, variants were combined from patients and controls, and only those variants with less than 5% missing genotypes in both patients and controls were kept for further analysis. Our final dataset contained 16,108 SNPs in 1,874 genes. The sequencing data were deposited in the NCBI's Sequence Read Archive (http://www.ncbi.nlm.nih.gov/sra/), under accession number SRP067711.

PCA was performed with Eigenstrat software using the common (MAF >5%) variants found in autosomes only. We computed a Euclidean distance from each point on the PCA plot to the origin and plotted distributions of this parameter for both patients and controls. Using the 3-sigma rule, 30 samples of mixed Hispanic ancestry were identified as outliers and removed from the dataset.

Sample statistics and case-control–matching metrics were computed using PLINK/SEQ analysis software. We used the number of variants called per sample, the number of heterozygous genotypes per sample, and the number of genotypes with minor allele per sample as a metric representing the genetic background of the cohort. The similarity between the genetic background of patients and controls was established by matching the mean and variance of patient and control distributions for every metric. We tested the validity of this approach by running Fisher's exact test on the common variation and QQ-plot of the *P* values. This showed no inflation, confirming the absence of any population stratification in the case-control dataset.

Mouse strains. Cd2ap+/– mice were generated previously[48]. *Synpo*–/– mice were obtained from Peter Mundel's laboratory[58]. The *Nphs1-rtTA3G* (*NEFTA*) strain was a gift from Jeffrey Miner's laboratory[44]. The *Dlg5*+/– mouse strain was a gift from Valeri Vasioukhin's laboratory[51]. All mouse strains were genotyped by established methods.

Generation of a male Cd2ap+/–, Synpo+/–, NEFTA+ ES cell line. To generate a male ES cell line that was sensitized to FSGS, we bred *Cd2ap+/– Synpo–/–* males with *NEFTA+* females. The females were superovulated using standard methods. After mating, the embryos were isolated at the 8-cell stage (morulae) and cultured overnight in EmbryoMax KSOM medium (MR-121-D; EMD Millipore) microdrops overlaid with mineral oil at 5% CO2 and 37°C. Blastocysts were transferred, 1 per well, into 48-well plates with γ-irradiated mouse embryonic fibroblast (MEF) feeders and standard ES cell media containing 15% ES-qualified FBS (SH30070.03E; Hyclone). The inner cell mass (ICM) was allowed to grow out and was trypsinized after 5 to 7 days, depending on the size and shape of the outgrowth. Cells were cultured until ES colonies were identified. The colonies were expanded and genotyped using standard methods.

Generation of miR30-shRNA knockin transgenic mice. Integration of a single-copy transgene into the *Hprt1* locus using 6-thioguanine was performed as we previously described[46] and was modified by the addition of a puromycin resistance cassette to increase the efficiency of selection of a positive ES clone. A PGK-Puro cassette was inserted between the left and right arm of the pHPRT targeting vector. The miR30-based shRNA-expressing transgene that was driven by the tetracycline-responsive promoter (*TRE*) was inserted between the left arm and the PGK-Puro cassette. The linearized targeting vector was transfected into ES cells. Twenty-fours hours after transfection, the ES cells were treated with 1 μg/ml puromycin for 48 hours. After passaging once, the ES cells were treated with 6-thioguanine (5 µg/ml) for an additional 48 hours. Surviving ES cell colonies were selected, expanded, and examined by genomic PCR across the right arm (forward primer: 5′-CAAGCCCGGTGCCTGATCTAGATCATAATC-3′; reverse primer: 5′-CTGTAAAGGTCTCTGAACTACCAATTGCAC-3′). Positive ES cells were then stocked for injection.

Laser-assisted microinjection. The ES cells were maintained at the expansion phase before injection. Eight ES cells were injected into a recipient embryo at the 8-cell stage by following a protocol published previously[45]. Since the ES cell line produces mice with agouti coat color, albino B6 (C57BL/6J-*Tyrc–2J*) mice were used as hosts to allow for direct evaluation of chimerism by coat color.

Cell culture and lentiviral infection. Immortalized murine podocytes were maintained and differentiated as described previously[25]. To examine the knockdown efficiency of CD2AP-sh877, podocytes were infected with lentiviral vectors encoding miR30-sh877. A control lentiviral vector encoding miR30-FF3 that targets firefly luciferase cDNA was used as a control. CD2AP expression was examined by immunoblot analysis of the whole-cell lysates.

Design and validation of the miR30-shRNA constructs for genes of interest. The shRNA oligo sequences were chosen using an online algorism (http://katahdin.cshl.org/siRNA/RNAi.cgi?type=shRNA) as described previously (48). The miR30-shRNA backbone was subcloned by PCR from pPRIME-CMV-GFP-FF3 (https://www.addgene.org/11663/) and inserted into a pcDNA3.1-Zeo(+) vector to generate the pcMIR vector. To examine knockdown efficiency, the miR30-shRNA construct and its artificial target were cotransfected into HEK293T cells at a molar ratio of 5:1. The expression of EGFP in whole-cell lysates was examined by immunoblot analysis.

Abs. The Abs used for immunoblotting were mouse anti-XFP (632381; 1:10,000 dilution; Clontech); rabbit anti-ERK2 (sc-154; 1:5,000 dilution; Santa Cruz Biotechnology Inc.); mouse anti–β-actin (A2228; 1:10,000 dilution; Sigma-Aldrich); and rabbit anti-CD2AP (generated in our previous study; 1:10,000 dilution).

Albumin-creatinine assay. Mouse urine samples were collected at the time points indicated in the figures, and urinary albumin (E90-134; Bethyl Laboratories Inc.) and creatinine (DICT-500; BioAssay Systems) were quantified by ELISA according to the manufacturers' protocols.

Transmission electron microscopy. Portions of kidney cortex were fixed with 2% paraformaldehyde and 2% glutaraldehyde. Specimen processing, ultrathin sectioning, and imaging were performed by the Electron Microscopy (EM) Core Facility at Washington University.

Statistics. P values of all albumin/creatinine ratio plots (**Figure 2.4**, **C** and **F**, and **Figure 2.5**, **B–G**) were calculated using a 2-tailed Mann-Whitney *U* test. For Figure 3, C and F, a *P* value of less than 0.05 was considered statistically significant. For **Figures 2.5**, **B–G**, a *P* value of less than 0.0083 was considered statistically significant (the multitest penalty was applied). All error bars represent the mean ± SEM.

Study approval. All animal experiments were conducted with the approval of the Washington University Animal Studies Committee. Because all of the patient samples were deidentified, the Washington University IRB deemed these studies exempt from IRB approval.

Author Contributions

Mykyta Artomov: sequencing data processing, quality-check and analysis, statistical models development, writing.

Haiyang Yu: conceived general experimental scheme, generated ES cells, devised the RNAi validation screen and the ES cell-targeting strategy, validation of the candidate genes by testing RNAis, generating ES cells, and phenotyping mice.

Sebastian Brahler: validation of the candidate genes by testing RNAis, generating ES cells, and phenotyping mice.

Andrey S. Shaw, Robi D. Mitra, M. Christine Stander, Samjay Jain, Ghaidan Shamsan: piloted and completed methods for exome capture and sequencing.

Metthew G. Sampson, Matthias Kretzler, Jeffrey B. Kopp, Andrey S. Shaw: determined list of genes to be sequenced.

J. Michael White: generated ES cells, performed laser-assisted microinjection.

Cheryl A. Winkle, Jeffrey B. Kopp, Samjay Jain: provided patient samples and their selection for this study.

Mark J. Daly: Overall guidance, writing

Andrey S. Shaw: conceived general experimental scheme, overall guidance, writing.

Bibliography

1. D'Agati, V. D., Kaskel, F. J. & Falk, R. J. Focal Segmental Glomerulosclerosis. *N. Engl. J. Med.* **365,** 2398–2411 (2011).

2. Wiggins, R.-C. The spectrum of podocytopathies: A unifying view of glomerular diseases. *Kidney Int.* **71,** 1205–1214 (2007).

3. Wickman, L. *et al.* Urine Podocyte mRNAs, Proteinuria, and Progression in Human Glomerular Diseases. *J. Am. Soc. Nephrol.* **24,** 2081–2095 (2013).

4. Wharram, B. L. *et al.* Podocyte Depletion Causes Glomerulosclerosis: Diphtheria Toxin-Induced Podocyte Depletion in Rats Expressing Human Diphtheria Toxin Receptor Transgene. *J. Am. Soc. Nephrol.* **16,** 2941–2952 (2005).

5. Kim, Y. H. *et al.* Podocyte depletion and glomerulosclerosis have a direct relationship in the PAN-treated rat. *Kidney Int.* **60,** 957–968 (2001).

6. Pollak, M. R. The genetic basis of FSGS and steroid-resistant nephrosis. *Semin. Nephrol.* **23,** 141–146 (2003).

7. Rood, I. M., Deegens, J. K. J. & Wetzels, J. F. M. Genetic causes of focal segmental glomerulosclerosis: implications for clinical practice. *Nephrol. Dial. Transplant.* **27,** 882–890 (2012).

8. Laurin, L.-P. *et al.* Podocyte-associated gene mutation screening in a heterogeneous cohort of patients with sporadic focal segmental glomerulosclerosis. *Nephrol. Dial. Transplant.* **29,** 2062–2069 (2014).

9. Barua, M. *et al.* Mutations in the INF2 gene account for a significant proportion of familial but not sporadic focal and segmental glomerulosclerosis. *Kidney Int.* **83,** 316–322 (2013).

10. Pollak, M. R. Inherited podocytopathies: FSGS and nephrotic syndrome from a genetic viewpoint. *J. Am. Soc. Nephrol.* **13,** 3016–23 (2002).

11. Cirulli, E. T. & Goldstein, D. B. Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nat. Rev. Genet.* **11,** 415–425 (2010).

12. Rabbani, B., Tekin, M. & Mahdieh, N. The promise of whole-exome sequencing in medical genetics. *J. Hum. Genet.* **59,** 5–15 (2014).

13. Thomson, R. *et al.* Evolution of the primate trypanolytic factor APOL1. *Proc. Natl. Acad. Sci.* **111,** E2130–E2139 (2014).

14. Friedman, D. J., Kozlitina, J., Genovese, G., Jog, P. & Pollak, M. R. Population-Based Risk Assessment of APOL1 on Renal Disease. *J. Am. Soc. Nephrol.* **22,** 2098–2105 (2011).

15. Genovese, G. *et al.* Association of Trypanolytic ApoL1 Variants with Kidney Disease in African Americans. *Science (80-.).* **329,** 841–845 (2010).

16. Kopp, J. B. *et al.* APOL1 Genetic Variants in Focal Segmental Glomerulosclerosis and HIV-Associated Nephropathy. *J. Am. Soc. Nephrol.* **22,** 2129–2137 (2011).

17. Kasembeli, A. N. *et al.* APOL1 Risk Variants Are Strongly Associated with HIV-Associated Nephropathy in Black South Africans. *J. Am. Soc. Nephrol.* **26,** 2882–2890 (2015).

18. Altshuler, D., Daly, M. J. & Lander, E. S. Genetic Mapping in Human Disease. *Science (80-.).* **322,** 881–888 (2008).

19. Reiser, J. *et al.* TRPC6 is a glomerular slit diaphragm-associated channel required for normal renal function. *Nat. Genet.* **37,** 739–744 (2005).

20. Winn, M. P. *et al.* A Mutation in the TRPC6 Cation Channel Causes Familial Focal Segmental Glomerulosclerosis. *Science (80-.).* **308,** 1801–1804 (2005).

21. Kaplan, J. M. *et al.* Mutations in ACTN4, encoding α-actinin-4, cause familial focal segmental glomerulosclerosis. *Nat. Genet.* **24,** 251–256 (2000).

22. Brown, E. J. *et al.* Mutations in the formin gene INF2 cause focal segmental glomerulosclerosis. *Nat. Genet.* **42,** 72–76 (2010).

23. Kim, J. M. *et al.* CD2-Associated Protein Haploinsufficiency Is Linked to Glomerular Disease Susceptibility. *Science (80-.).* **300,** 1298–1300 (2003).

24. Lindenmeyer, M. T. *et al.* Systematic Analysis of a Novel Human Renal Glomerulus-Enriched Gene Expression Dataset. *PLoS One* **5,** e11545 (2010).

25. Akilesh, S. *et al.* Arhgap24 inactivates Rac1 in mouse podocytes, and a mutant form is associated with familial focal segmental glomerulosclerosis. *J. Clin. Invest.* **121,** 4127–4137 (2011).

26. Boerries, M. *et al.* Molecular fingerprinting of the podocyte reveals novel gene and protein regulatory networks. *Kidney Int.* **83,** 1052–1064 (2013).

27. Brunskill, E. W., Georgas, K., Rumballe, B., Little, M. H. & Potter, S. S. Defining the Molecular Character of the Developing and Adult Kidney Podocyte. *PLoS One* **6,** e24640 (2011).

28. Neale, B. M. *et al.* Patterns and rates of exonic de novo mutations in autism

spectrum disorders. *Nature* **485,** 242–245 (2012).

29. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25,** 1754–1760 (2009).

30. DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43,** 491–8 (2011).

31. Van der Auwera, G. A. *et al.* From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr. Protoc. Bioinforma.* **43,** 11.10.1-33 (2013).

32. Sampson, M. G., Hodgin, J. B. & Kretzler, M. Defining nephrotic syndrome from an integrative genomics perspective. *Pediatr. Nephrol.* **30,** 51–63 (2015).

33. Gee, H. Y. *et al.* KANK deficiency leads to podocyte dysfunction and nephrotic syndrome. *J. Clin. Invest.* **125,** 2375–2384 (2015).

34. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536,** 285–291 (2016).

35. Li, B. & Leal, S. M. Discovery of Rare Variants via Sequencing: Implications for the Design of Complex Trait Association Studies. *PLoS Genet.* **5,** e1000481 (2009).

36. Price, A. L. *et al.* Pooled association tests for rare variants in exon-resequencing studies. *Am. J. Hum. Genet.* **86,** 832–8 (2010).

37. Neale, B. M. *et al.* Testing for an Unusual Distribution of Rare Variants. *PLoS Genet.* **7,** e1001322 (2011).

38. Voskarides, K. *et al.* COL4A3/COL4A4 Mutations Producing Focal Segmental Glomerulosclerosis and Renal Failure in Thin Basement Membrane Nephropathy. *J. Am. Soc. Nephrol.* **18,** 3004–3016 (2007).

39. Sadowski, C. E. *et al.* A Single-Gene Cause in 29.5% of Cases of Steroid-Resistant Nephrotic Syndrome. *J. Am. Soc. Nephrol.* **26,** 1279–1289 (2015).

40. Gee, H. Y. *et al.* ARHGDIA mutations cause nephrotic syndrome via defective RHO GTPase signaling. *J. Clin. Invest.* **123,** 3243–3253 (2013).

41. Zhou, W. & Hildebrandt, F. Inducible Podocyte Injury and Proteinuria in Transgenic Zebrafish. *J. Am. Soc. Nephrol.* **23,** 1039–1047 (2012).

42. Ashraf, S. *et al.* ADCK4 mutations promote steroid-resistant nephrotic syndrome through CoQ10 biosynthesis disruption. *J. Clin. Invest.* **123,** 5179–5189 (2013).

43. Huber, T. B. *et al.* Bigenic mouse models of focal segmental glomerulosclerosis involving pairwise interaction of CD2AP, Fyn, and synaptopodin. *J. Clin. Invest.* **116,** 1337–1345 (2006).

44. Lin, X., Suh, J. H., Go, G. & Miner, J. H. Feasibility of Repairing Glomerular Basement Membrane Defects in Alport Syndrome. *J. Am. Soc. Nephrol.* **25,** 687–692 (2014).

45. Poueymirou, W. T. *et al.* F0 generation mice fully derived from gene-targeted embryonic stem cells allowing immediate phenotypic analyses. *Nat. Biotechnol.* **25,** 91–99 (2007).

46. Yu, H. *et al.* Rac1 Activation in Podocytes Induces Rapid Foot Process Effacement and Proteinuria. *Mol. Cell. Biol.* **33,** 4755–4764 (2013).

47. Löwik, M. M. *et al.* Focal segmental glomerulosclerosis in a patient homozygous for a CD2AP mutation. *Kidney Int.* **72,** 1198–1203 (2007).

48. Shih, N. Y. *et al.* Congenital nephrotic syndrome in mice lacking CD2-associated protein. *Science* **286,** 312–5 (1999).

49. Stegmeier, F., Hu, G., Rickles, R. J., Hannon, G. J. & Elledge, S. J. A lentiviral microRNA-based system for single-copy polymerase II-regulated RNA interference in mammalian cells. *Proc. Natl. Acad. Sci.* **102,** 13212–13217 (2005).

50. Xu, X. *et al.* Expression of Novel Podocyte-Associated Proteins sult1b1 and ankrd25. *Nephron Exp. Nephrol.* **117,** e39–e46 (2011).

51. Nechiporuk, T., Fernandez, T. E. & Vasioukhin, V. Failure of Epithelial Tube Maintenance Causes Hydrocephalus and Renal Cysts in Dlg5−/− Mice. *Dev. Cell* **13,** 338–350 (2007).

52. Greka, A. & Mundel, P. Cell Biology and Pathology of Podocytes. *Annu. Rev. Physiol.* **74,** 299–323 (2012).

53. D'Agati, V. D. Podocyte injury in focal segmental glomerulosclerosis: Lessons from animal models (a play in five acts). *Kidney Int.* **73,** 399–406 (2008).

54. Antignac, C. *et al.* NPHS2, encoding the glomerular protein podocin, is mutated in autosomal recessive steroid-resistant nephrotic syndrome. *Nat. Genet.* **24,** 349–354 (2000).

55. Gigante, M. *et al.* CD2AP mutations are associated with sporadic nephrotic syndrome and focal segmental glomerulosclerosis (FSGS). *Nephrol. Dial. Transplant.* **24,** 1858–1864 (2009).

56. Kao, W. H. L. *et al.* MYH9 is associated with nondiabetic end-stage renal disease in African Americans. *Nat. Genet.* **40,** 1185–1192 (2008).

57. Kopp, J. B. *et al.* MYH9 is a major-effect risk gene for focal segmental glomerulosclerosis. *Nat. Genet.* **40,** 1175–1184 (2008).

58. Deller, T. *et al.* Synaptopodin-deficient mice lack a spine apparatus and show deficits in synaptic plasticity. *Proc. Natl. Acad. Sci.* **100,** 10494–10499 (2003).

Chapter 3
Large-scale exome sequencing data analysis in cancer

Work presented in this chapter was published as:

Artomov M., *et al.* Rare variant, gene-based association study of hereditary melanoma using whole exome sequencing. *Journal of National Cancer Institute.*
Accepted.

Gupta S., *et al.* Gender disparity and mutation burden in metastatic melanoma. *JNCI;* 107(11); 2015.

Artomov M., *et al.* Mosaic mutations in blood dna sequence are associated with solid tumor cancers. *npj Genomic Medicine.*
Accepted

Artomov M., *et al.* A strategy for large-scale systematic pan-cancer germline rare variation analysis.
Available on BioRxiv

A strategy for large-scale systematic pan-cancer germline rare variation analysis

Abstract

Vast majority of cancer risk genes was identified with tumor-normal tissue comparison. While somatic mutations undoubtedly are the main drivers of the disease onset in sporadic and late-onset cases, patients with early onset and/or familial history of cancer are likely to carry significant inherited risk in their germline DNA. Here we sought to analyze large dataset of genetically enriched cancer cases and unselected cancer cases cohorts with cutaneous and ocular melanoma, colon, breast cancers and identify common features of the risk variation in the germline DNA. We observe that almost entirely statistical signal was driven by singleton protein-truncating variants in the genes tolerant to loss-of-function mutations that followed autosomal dominant inheritance pattern. Interestingly, both unselected and genetically enriched cases show burden of risk variation compared to large pool of matched controls.

Introduction

Analysis of inherited predisposition to cancer usually involves cohorts of early disease onset patients or large kindreds. Here we analyzed a large cohort of genetically enriched (early onset and/or familial) and unselected cases of breast cancer, colon cancer and cutaneous and ocular melanomas (in total about

2,000 cases matched to more than 7,000 non-cancer controls) to systematically search for novel germline cancer risk genes. By first analyzing known cancer predisposition genes, we demonstrate that protein truncating, rather than missense, mutations are the main driver of inherited cancer predisposition and generally these occur in genes tolerant of loss-of-function mutations – distinct from the highly constrained genes more often somatically mutated and found to be drivers in tumors. Interestingly, we find that unselected cancer cases have a significant burden of protein-truncating variants in known cancer risk genes, similar to that observed in genetically enriched (familial and early-onset, herein referred to as 'selected') patients. Using these observations to design our search for new cancer genes, we analyzed individual cancer cohorts with matched controls and constructed a ranked list of new potential candidate risk genes.

Results

Cohort and overview. For this study, germline DNA from selected "genetically-enriched" cases (individuals with familial cancer or onset of the disorder at age of 35 or earlier) of breast cancer, colon cancer, cutaneous and ocular melanomas, and Li-Fraumeni syndrome (with primary breast cancer) was collected (Inclusion criteria is included in Supplementary Methods). For a discovery set a total of 273 cutaneous melanoma (M/F 128/145), 99 ocular melanoma (M/F 46/53), 355 breast cancer (M/F 1/354), 43 Li-Fraumeni syndrome (M/F 7/36), 75 colon cancer (M/F 27/48) and 7924 controls (M/F 5689/2235) passed quality-check and were included in the subsequent analysis. Germline DNA sequences from TCGA were used as "unselected" cancer cases (not controlling for family history or age of onset): 820 breast cancer (M/F 9/811), 250 colon cancer (M/F 135/115), 379 cutaneous melanoma (M/F 233/146) and 47 ocular melanoma (M/F 27/20).

In order to ensure close ancestral matching, we performed principal component analysis (PCA; **Figure 3.1A**) of the case and control cohorts. Since there were few samples outside the large cluster representing predominantly European ancestry, our analysis was then restricted to this cluster of samples only. Within European-ancestry samples we performed relatedness analysis and removed all duplicates and first-degree relatives. Examination of common synonymous variants (MAF>5%) revealed a null-distribution of the test statistic between cases and controls (**Figure 3.1B**).

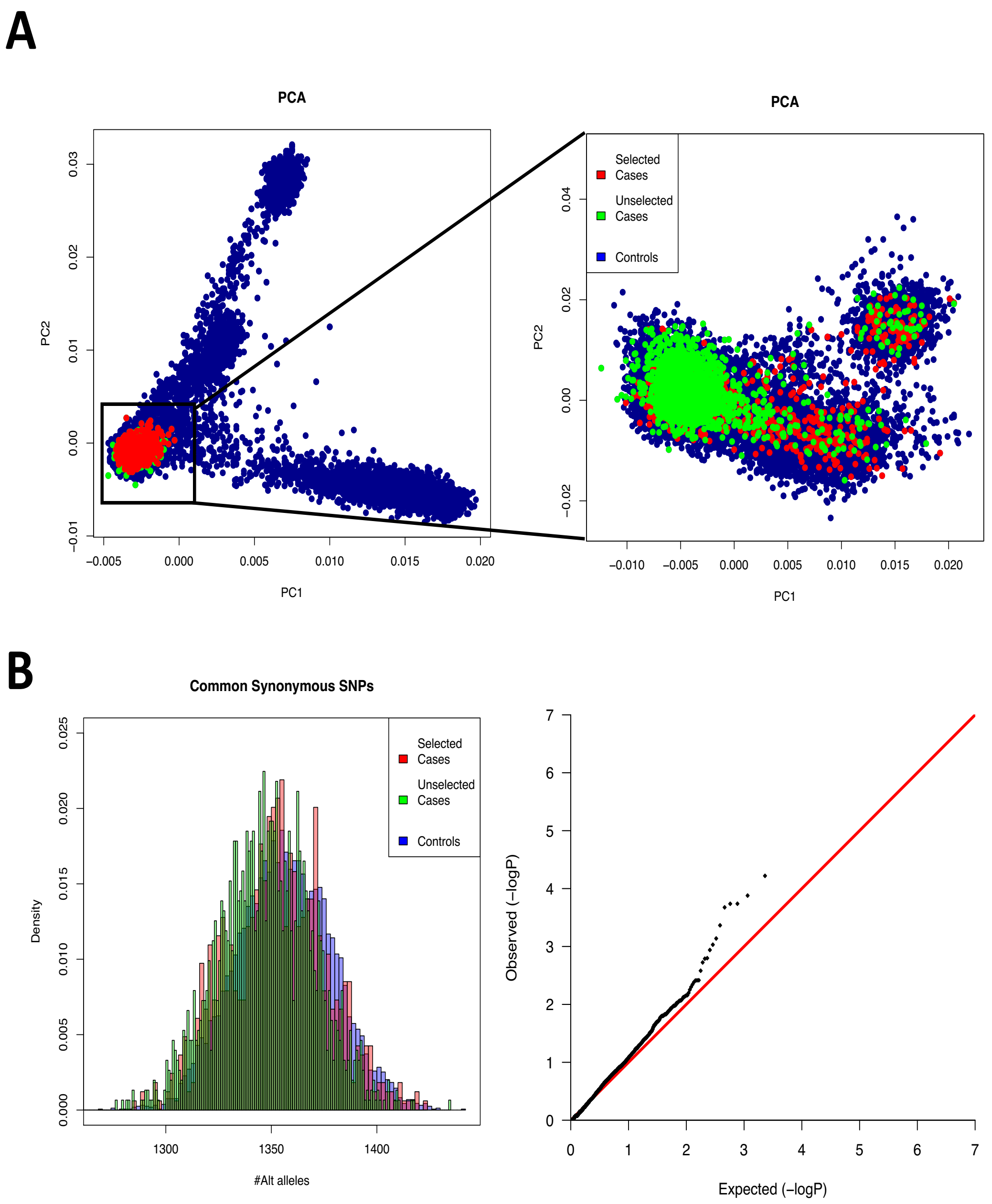


Figure 3.1. Case-control matching results. (**A**) Principal component analysis and closer image of European ancestry cluster. (**B**) Analysis of common synonymous mutations in cases and controls. QQ-plot shows null-distribution.

Search strategy for new cancer risk genes. Recent studies of multiple pediatric cancer phenotypes provide insights into the prevalence of inherited pathogenic mutations in known cancer genes[1]. In order to define an exome-wide strategy to search for new cancer predisposition genes, we began by analyzing rare genetic variation in known risk genes. Specifically, we examined the prevalence of risk alleles in lists of known genes subdivided by reported model of inheritance to identify features common to genes in each list and also to compare genetic association observed in selected and unselected cancer cases. Out of four lists that we tested – autosomal dominant, autosomal recessive, tumor suppressor and Ras-Sos pathway genes, only the autosomal dominant model list shows enrichment in cancer cases (**Table 3.1**, **Figure 3.2A**). We looked into the frequency spectrum for the variants with MAC<=10 and observed that this association signal is driven almost entirely by singletons (**Figure 3.2B**).

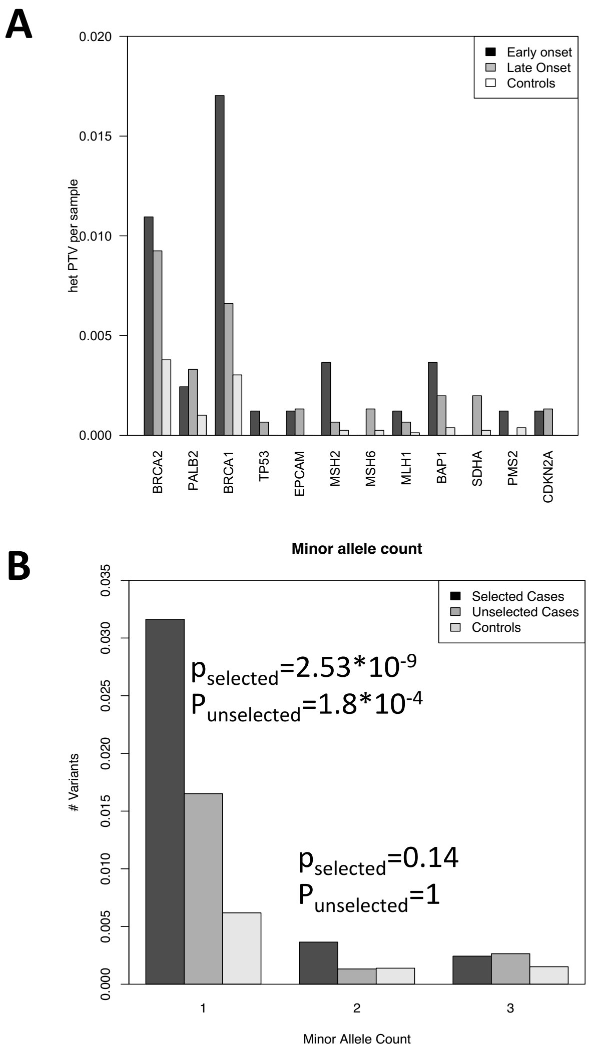
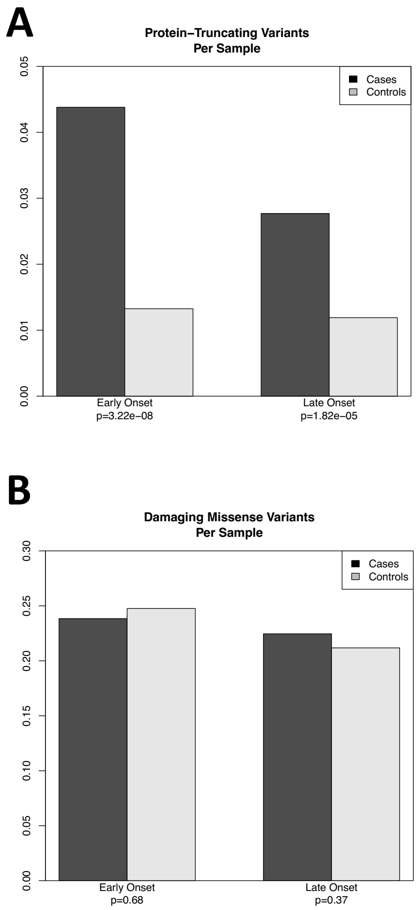


Figure 3.2. Autosomal dominant model analysis. (**A**) Genes with protein-truncating mutations show enrichment in cases. (**B**) Allele frequency spectrum of protein-truncating variants.

Table 3.1. Burden test of models of different inheritance.

Autosomal dominant. Protein-truncating variants					
Selected Cases (N=822)	Controls (N=7924)	P	Unselected Cases (N=1514)	Controls (N=7924)	P
40	105	**5.26x10^{-10}**	44	94	**6.41x10^{-6}**
Autosomal dominant. Damaging missense variants					
196	1962	0.68	340	1678	0.37
Autosomal recessive. Protein-truncating variants (homozygotes or double hets)					
0	0	1	0	0	1
Autosomal recessive. Damaging missense variants (homozygotes or double hets)					
3	14	0.21	3	13	0.73
Ras-Sos pathway genes. Protein-truncating variants					
2	16	0.68	2	15	1
Ras-Sos pathway genes. Damaging missense variants					
12	154	0.42	21	154	0.18
Tumor suppressor genes. Protein-truncating variants					
8	102	0.43	17	81	0.67
Tumor suppressor genes. Damaging missense variants					
225	2141	0.87	287	1762	0.02

Separate analysis of protein-truncating variants (nonsense, frameshift and essential splice site) and damaging missense (Supplementary Methods) was performed. Interestingly, unselected cases (p=6.41*10^{-6}; OR=2.45; OR CI=1.66-3.56) show similar significant enrichment to genetically enriched cases (p=5.26*10^{-10}; OR=3.67; OR CI=2.47-5.37) with rare (minor allele count less or equal to 10) protein-truncating variants only, while we observed no enrichment in damaging missense variation (p=0.68 and p=0.37 for selected and unselected respectively, **Figure 3.3**).



Cohort	Alt Cases	Total Cases	Alt Controls	Total Controls	Fisher P
Selected Cases	40	822	105	7924	3.22E-08
Unselected Cases	44	1514	94	7924	1.82E-05

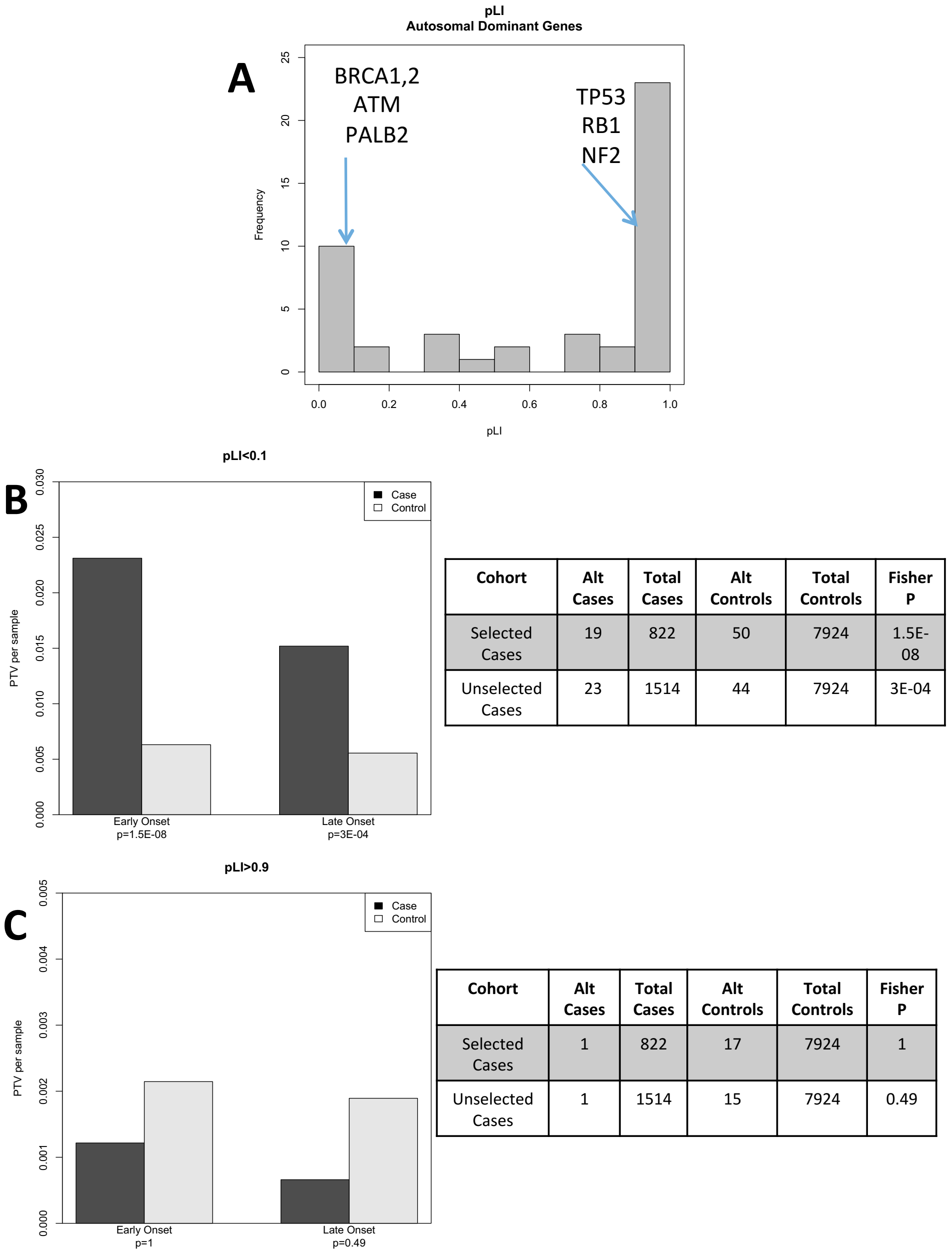
Cohort	Alt Cases	Total Cases	Alt Controls	Total Controls	Fisher P
Selected Cases	196	822	1962	7924	0.68
Unselected Cases	340	1514	1678	7924	0.37

Figure 3.3. Prevalence of alternative alleles in (**A**) Protein-truncating variants; (**B**) Damaging missense variants.

It is worth noting however that it is possible that genetically enriched cases have had in general had more genetic screening and that some diagnosed cases were removed before being entered in this study sample – in this circumstance our observed result for genetically enriched cases would be smaller than the true result. We tried multiple analyses for the recessive model, including counting of samples with more than 1 heterozygous genotype in the same gene and expanding the set of included variants up to minor allele frequency less than 1%, however the counts were still very low and inconclusive, thus the recessive model was ruled out.

We then asked whether there were any additional features characterizing which genes within the autosomal dominant list were driving the truncating variant association signal. Using a metric of genic tolerance to truncating variation (pLI) defined by the Exome Aggregation Consortium (ExAC), we separately estimated association in genes tolerant of loss-of-function mutations (pLI<0.1) and intolerant of such variation (pLI>0.9) (**Figure 3.4A**). While this list contains genes that carry either heritable risk of cancer or high-risk somatic mutations (or both) we observe high enrichment of protein-truncating variants in highly tolerant genes ($p=1.5*10^{-6}$ early and $p=3*10^{-4}$ late onset, **Figure 3.4B-C**), consistent with limited selective pressure from generally later adult onset of disease.

Using these identified properties of the known cancer susceptibility genes, we can infer what features we should expect to observe in novel germline candidate genes. We therefore targeted our search for mutations (primarily truncating) with autosomal dominant model of inheritance, in genes tolerant of loss-of-function mutations (as predicted by pLI score metric) and driven by a substantial burden of singletons (or independent variants) in both genetically enriched and unselected cases.



Cohort	Alt Cases	Total Cases	Alt Controls	Total Controls	Fisher P
Selected Cases	19	822	50	7924	1.5E-08
Unselected Cases	23	1514	44	7924	3E-04

Cohort	Alt Cases	Total Cases	Alt Controls	Total Controls	Fisher P
Selected Cases	1	822	17	7924	1
Unselected Cases	1	1514	15	7924	0.49

Figure 3.4. Analysis of the autosomal dominant model genes with pLI constraint metric. (**A**) pLI spectrum for autosomal dominant genes. (**B**) Prevalence of alternative alleles in cases in genes tolerant to loss-of-function variation (pLI<0.1). (**C**) Prevalence of alternative alleles in genes intolerant to loss-of-function variation (pLI>0.9).

Case-control analysis. We applied this search strategy to our complete dataset. We found 4021 and 6254 singleton protein-truncating variants in selected and unselected cases respectively. We kept only genes with pLI<0.1 for further analysis. Because of earlier demonstrated significant contribution of inherited risk in unselected cases we joined both case cohorts for further analysis. Considering the burden of singleton truncating variants, among the top 5 genes identified with our methodology, 3 are known cancer risk genes – *BRCA1, BRCA2* and *ATM*. While this serves as a good proof-of-concept and suggests that the exome sequencing and analysis approach has some degree of both sensitivity and specificity, there are no clearly significant novel candidates arising from this approach.

Individual Cohorts Analysis. We then performed analysis of the individual phenotypic cohorts for each cancer. Additional 3526 controls were matched to the unselected cases and were used as a replication set. In addition to our primary analysis focused on burden of protein-truncating variants, several other previously reported models for rare variant association studies (RVAS) were used for analysis[2] which added additional variants to the truncating variants: addition of the missense mutations (c-alpha, VT tests) and ultra-rare variation analysis (variants filtered for $MAF<10^{-5}$ in ExAC).

For each gene we record the best p-value out of 3 models and respectively adjusted the null-statistics by performing the same choice of the best p-value from 3 independent draws from uniform distributions of p-values.

For the analysis of the breast cancer patients, we eliminated all male samples from the dataset, resulting in comparison of 354 genetically enriched cases with 2190 matched controls. Despite the screening of the previously known *BRCA1* risk mutations in the breast cancer cohort we still observe genome-wide significant rare loss-of-function variants burden in this gene. Analysis was performed using three different but related statistical models and the minimal p-value was chosen for each gene. Genes with minimal p-value less than 10^{-4} were subjected to replication. *MKL2* was also included in the short-list of genes as it appears second only to *BRCA1* in the burden of protein-truncating variants. According to GTEx database[3] – *MKL2* is primarily expressed in adipose and mammary tissues. Two genes show evidence of replication – *BRCA1* and *HSD17B1*. Interestingly, four genes out of our short-list of candidates are known to be associated with worse outcome of breast cancer, once mutated or amplified in tumors – *BRCA1, HSD17B1, PCDHB15 and MED28*[4–7] (**Table 3.2**).

Similarly, *ATM* appears as a top gene in RVAS of the colon cancer cohort. Being a known predisposing gene for this phenotype it does not reach significance threshold due to statistical power limitations (**Table 3.3**). However, *OBP2A* is not expressed in colon tissues and *TMEM4C* does not have expression specificity to colon tissues.

Table 3.2. Early Onset Breast Cancer RVAS female cases and controls.

Target Set					Replication Set		
Gene Name	**Method**	**Cases (N=354)**	**Controls (N=2190)**	**P**	**Cases (N=504)**	**Controls (N=1870)**	**P**
ENPP5	VT	11	16	3.00×10^{-6}	3	14	1
BRCA1	PTV Burden	11	7	5.30×10^{-6}	5	3	0.014
PCDHB15	ExAC Burden	6	1	2.18×10^{-5}	0	0	1
USP35	VT	46	124	4.20×10^{-5}	15	53	0.76
CCDC9	VT	10	18	4.84×10^{-5}	5	12	0.224
COL5A2	C-alpha	23	72	7.38×10^{-5}	6	34	0.39
MED28	C-alpha	7	6	8.54×10^{-5}	0	6	1
HSD17B1	C-alpha	15	22	8.77×10^{-5}	3	5	0.039
MKL2	PTV Burden	5	2	8.85×10^{-4}	2	1	0.2

Table 3.3. Early onset colon cancer RVAS.

Target Set					Replication Set		
Gene Name	**Method**	**Cases (N=75)**	**Controls (N=7654)**	**P**	**Cases (N=190)**	**Controls (N=3526)**	**P**
ATM	PTV burden	4	12	1.66×10^{-5}	1	6	0.31
OBP2A	ExAC burden	5	43	4.81×10^{-5}	1	9	0.26
TMEM14C	C-alpha	2	12	9.83×10^{-5}	0	10	1

Discussion

Our study develops a systematic approach for search of the novel cancer risk genes through analysis of the rare variation in the known susceptibility genes. Moreover, we observe notable enrichment of the known inherited genetic risk factors in the unselected cancer cohort (TCGA). Despite common beliefs that sporadic cancer cases are mostly elucidated by aging and carcinogen exposure, genetic predisposition plays substantial role in the disease onset, comparable to genetically enriched cohort. List of the genes that we used as a training model includes well-known high-risk genes. Expectedly, we did not find potential candidate genes in our dataset with comparable effect size. At the same time, lower-risk genes require large cohorts providing enough statistical power. Individual cohort analysis provides alternative approach to the search for candidate genes. We identified three potential candidates – *HSD17B1, PCDHB15, MED28* with reported association to worse outcome for patients with somatic mutations in these genes. The *HSD17B1* gene produces an enzyme that catalyzes the conversion of estrone to estradiol and estrogen exposure influences risks of breast and endometrial cancer[5]. Instead of looking at the true cancer driver genes, inclusion of the missense variation allows to screen regulatory genes potentially altering functionality of the driver genes. Intriguingly, Lu et al. found that overexpression of full-length *MED28* in HEK293 human embryonic kidney cells or human breast cancer cell lines caused significantly increased *in vitro* cell proliferation[8]. Also, functional studies report importance of the *MED28* for breast cancer progression. *MED28* modulates cell growth through FOXO3a and NFkB in human breast cancer cells[7]. Not only is *MED28* involved in cellular migration and invasion but also in cell cycle progression in human breast cancer cells[7].

Analysis of the colon cancer cohort is strongly influenced by the small cohort size, although it still reveals a known predisposition gene – *ATM.* While proving the concept, a significantly larger genetically enriched cohort would be needed to facilitate discovery of the new candidates.

Materials and Methods

Patient cohorts. All patients provided written consent for this study and were enrolled at 4 sites- the Massachusetts General Hospital (MGH; cutaneous melanoma (CM), breast cancer (BC) and Li-Fraumeni syndrome patients), the A. Sygros Hospital in Athens, Greece (CM patients) and the Massachusetts Eye and Ear Infirmary (MEEI; ocular melanoma (OM) patients) in Boston, MA, the Memorial Sloan Kettering Cancer Center in New York, NY (MSKCC; BC and colon cancer (CC) patients) - in accordance with protocols approved at these institutions.

All probands were considered "genetically enriched" based on the following criteria.

1. **MGH**:
    a. **CM**: a histologically-proven CM AND at least one 1st degree affected relative OR ≥2 affected relatives on one side of the family regardless of degree of relationship (proband CM + relative with CM, "Familial CM/CM"; proband CM + relative with OM, "Familial CM/OM") OR ≥3 primary melanomas regardless of family history ("MPM CM-CM").

    b. **BC**: a histologically-proven BC AND at least one 1st degree affected relative (Familial BC) OR age at clinical diagnosis less than 35 years old (early onset BC).

    c. **Li-Fraumeni syndrome**: diagnosed with Li-Fraumeni syndrome at MGH cancer center AND age at clinical diagnosis less than 45.

2. **The A. Sygros Hospital**: a histologically-proven CM AND ≥1 affected relative on one side of the family (“Familial CM/CM”) OR ≥2 primary melanoma (“MPM CM-CM”).
3. **MEEI**: a histologically or clinically diagnosed OM AND ≥1 relative affected with either CM or OM (proband OM + relative with OM, “Familial OM/OM”; proband OM + relative with CM, “Familial OM/CM”) OR a second CM (“MPM OM-CM”).
4. **MSKCC**:
    a. **BC:** a histologically-proven BC AND at least one 1st degree affected relative (Familial BC) OR age at clinical diagnosis less than 35 years old (early onset BC).
    b. **CC:** a histologically-proven CC AND at least one 1st degree affected relative (Familial CC) OR age at clinical diagnosis less than 35 years old (early onset CC).

Exome sequencing, Variant processing and calling. Whole exome libraries were prepared using a modified version of Agilent's Exome Capture kit and protocol, automated on the Agilent Bravo and Hamilton Starlet. Libraries were then prepared for sequencing using a modified version of the manufacturer's suggested protocol, automated on the Agilent Bravo and Hamilton Starlet, followed by sequencing on the Illumina HiSeq 2000. We used an aggregated set of samples consented for joint variant calling resulting in 37,607 samples (germline from 292 cutaneous melanoma patients, 101 ocular melanoma patients, 397 TCGA cutaneous melanoma patients, 47 TCGA ocular melanoma patients, 355 breast cancer patients, 43 Li-Fraumeni syndrome with primary breast cancer patients, 75 colon cancer patients, 697 TCGA breast cancer patients, 250 TCGA colon cancer patients, 24,612 controls and 10,738 other individuals not used for association studies). All samples were sequenced using the same capture reagents at the Broad Institute and aligned on the reference genome with BWA[9] and the best-practices GATK/Picard Pipeline, followed by joint variant calling with all samples processed as a single batch using GATK v 3.1-144 Haplotype Caller[10–12]. The resulting dataset had 7,094,027 distinct variants. Haplotype Caller, which was used for the ExAC database[13], was also used to detect indels. Selected mutations in *CDKN2A, BRCA1* and *BAP1* were confirmed with Sanger sequencing.

We performed principal component analysis (PCA) on common (MAF>5%) autosomal independent SNPs to filter out all non-European samples with Eigenstrat[14]. Relatedness analysis among Europeans was conducted with PLINK[15,16] as suggested in the PLINK best practices. We used VEP[17] for functional annotation of the DNA variants. Common and rare variants analyses were conducted using PLINK/SEQ[18], which allows indexing of the large datasets. A burden test was used for rare protein truncating variants. In addition, the VT[19] and C-alpha[20] tests were chosen as an adaptive burden test and variance-component test, respectively, to complement each other and to boost the power of rare missense and protein truncating variation association detection[21]. See details in supplementary methods.

Statistical Methods. Gene-based association was performed using 3 distinct, but related, analytical frameworks. In the first analysis, a burden test was applied to all rare (MAF<1%) protein truncating variants (PTV) since the functional impact is presumed to be severe and most directly inferred. Then, to expand on all rare variants (missense and PTV), a second analysis using both the C-alpha and variable threshold (VT) tests was employed. A third analysis applied the burden test to examine "ultra-rare" (MAF<0.0001; ExAC database[13]) variants as these may represent the most highly penetrant alleles. In the case of a single-model association test – the null statistic was represented by the uniform distribution of p-values. Since four different test statistics (i.e. VT, C-alpha, burden of PTVs and burden of ExAC filtered variants) were applied and the lowest p-value was chosen, the null distribution was constructed by choosing the smallest p-value from 4 null single-statistic models (4 sets of uniform p-values). This process simulates the procedure of selecting the best p-value out of 4 different test statistics that was used for gene-association testing thus making it a more conservative approach. Genome-wide significance was determined by Bonferroni correction (0.05 /17,337 genes tested, i.e. $p<2.88x10^{-6}$).

Author contributions

Mykyta Artomov: designed and performed analysis, writing.

Vijai Joseph: collected clinical data, performed data quality check.

Grace Tiao, Kasmintan Schrader, Tinu Thomas: data analysis, quality check.

Robert Klein, Adam Kiezun:

Namrata Gupta, Lauren Margolin, Kristen Shannon: managed the samples and clinical data.

Alexander J. Stratigos, Ivana Kim, Leif W. Ellisen, Daniel Haber, Gad Getz, Hensin Tsao, Steven, Lipkin: provided DNA samples for the study.

Mark J. Daly, David Altshuler, Kenneth Offit: study designed and overall guidance.

Bibliography

1. Zhang, J. *et al.* Germline Mutations in Predisposition Genes in Pediatric Cancer. *N. Engl. J. Med.* **373,** 2336–2346 (2015).

2. Yu, H. *et al.* A role for genetic susceptibility in sporadic focal segmental glomerulosclerosis. *J. Clin. Invest.* **126,** (2016).

3. GTEx Consortium, T. Gte. The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45,** 580–5 (2013).

4. Zhang, C. *et al.* The Identification of Specific Methylation Patterns across Different Cancers. *PLoS One* **10,** e0120361 (2015).

5. Setiawan, V. W., Hankinson, S. E., Colditz, G. A., Hunter, D. J. & De Vivo, I. HSD17B1 gene polymorphisms and risk of endometrial and breast cancer. *Cancer Epidemiol. Biomarkers Prev.* **13,** 213–9 (2004).

6. Gunnarsson, C. *et al.* Amplification of HSD17B1 and ERBB2 in primary breast cancer. *Oncogene* **22,** 34–40 (2003).

7. Li, C.-I., Hsieh, N.-T., Huang, C.-Y., Chang, H.-C. & Lee, M.-F. MED28 Modulates Cell Cycle Progression in Human Breast Cancer Cells. *FASEB J.* **29,** (2015).

8. Lu, M. *et al.* The Novel Gene EG-1 Stimulates Cellular Proliferation. *Cancer Res.* **65,** 6159–6166 (2005).

9. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25,** 1754–1760 (2009).

10. Van der Auwera, G. A. *et al.* From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr. Protoc. Bioinforma.* **43,** 11.10.1-33 (2013).

11. DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43,** 491–8 (2011).

12. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20,** 1297–303 (2010).

13. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536,** 285–291 (2016).

14. Price, A. L. *et al.* Principal components analysis corrects for stratification in

genome-wide association studies. *Nat. Genet.* **38,** 904–909 (2006).

15. Purcell, S. PLINK.

16. Purcell, S. *et al.* PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *Am. J. Hum. Genet.* **81,** 559–575 (2007).

17. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol.* **17,** 122 (2016).

18. Https://atgu.mgh.harvard.edu/plinkseq/. PLINK/SEQ.

19. Price, A. L. *et al.* Pooled association tests for rare variants in exon-resequencing studies. *Am. J. Hum. Genet.* **86,** 832–8 (2010).

20. Neale, B. M. *et al.* Testing for an Unusual Distribution of Rare Variants. *PLoS Genet.* **7,** e1001322 (2011).

21. Lee, S., Abecasis, G. R., Boehnke, M. & Lin, X. Rare-variant association analysis: study designs and statistical tests. *Am. J. Hum. Genet.* **95,** 5–23 (2014).

Mosaic mutations in blood dna sequence are associated with solid tumor cancers

Abstract

Recent understanding of the causal role of blood-detectable somatic protein-truncating DNA variants in leukemia prompts questions about the generalizability of such observations across cancer types. We used TCGA exome sequencing (~8,000 samples) to compare 22 different cancer phenotypes with more than 6,000 controls using a case-control study design and demonstrate that mosaic protein truncating variants in these genes are also associated with solid-tumor cancers. The absence of these cancer-associated mosaic variants from the tumors themselves suggest these are not themselves tumor drivers.

Through analysis of different cancer phenotypes we observe gene-specificity for mosaic mutations. We confirm a specific link between *PPM1D* and ovarian cancer, consistent with previous reports linking *PPM1D* to breast and ovarian cancer. Additionally, glioblastoma, melanoma and lung cancers show gene specific burdens of mosaic protein truncating mutations. Taken together, these results extend existing observations and broadly link solid-tumor cancers to somatic blood DNA changes.

Introduction

Several recent studies[1,2] have reported associations of mosaic protein truncating variants (PTV) in *PPM1D, TET2, ASXL1* and *DNMT3A* with blood cancers. Intriguingly, such mosaic mutations in *PPM1D* have also been convincingly associated with breast and ovarian cancer[3] – however, since these mutations are somatic, rather than germline, their role in causation has not been clear. We sought to more fully explore the relationship of these somatic mutations, clearly causally linked to blood cancers, in solid tumor cancer using a large assembly of germline and somatic exome DNA sequences of 7,979 cancer cases from TCGA[4] and performed a large-scale case-control study with 6,177 population controls with no cancer phenotype reported.

Results

Using data available from dbGAP, we performed a large-scale joint variant calling of sequences generated from blood-derived germline DNA samples from cancer cases and controls – primarily from an assembly of TCGA samples (cases) compared with unselected population controls (with no known cancer status) from several studies (NHLBI-ESP, 1000 Genomes, ATVB, T2D, Ottawa Heart) appropriately consented for broad use as controls. Importantly, all cases and controls in this analysis have age at DNA sampling available.

Observations of the mosaic mutations might be affected by several parameters – both biological (age[5], clinical interventions[6,7]) and technical (depth of coverage, variant calling accuracy). To make the case-control comparison robust we first identified what adjustments to the model of association are needed.

We observed 348 PTVs (stop gain, essential splice site, frameshift mutations) in the four established somatic leukemia genes. Detection of somatic mutations with low non-reference allele balance is heavily dependent on sequencing depth. To ensure equal sensitivity in cases and controls we first compared coverage of these genes in cancer germline (average 33X coverage) and control (average 29X coverage) data. We next looked specifically at cases and controls carrying PTVs. For germline heterozygous sites the expected allele balance is 0.5, so we applied a binomial test to detect significantly low allele balance genotypes based on depth of coverage and number of alternative reads. Those with p<0.001 (i.e., heterozygotes with significantly less than 50% non-reference allele) and more than 20x coverage were determined to be mosaic and kept for further analysis (**Figure 3.5, 3.6**).

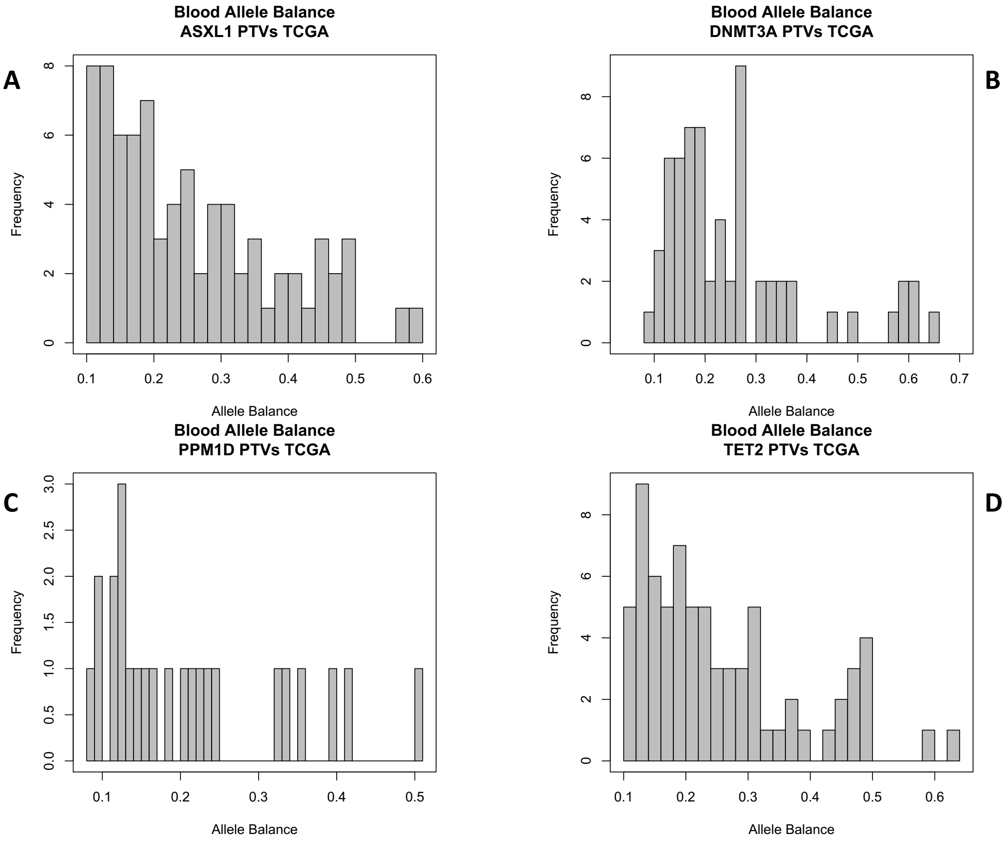


Figure 3.5. Allele balance for protein-truncating variants observed in blood DNA in TCGA cases in (**A**) *ASXL1*; (**B**) *DNMT3A*; (**C**) *PPM1D*; (**D**) *TET2*.

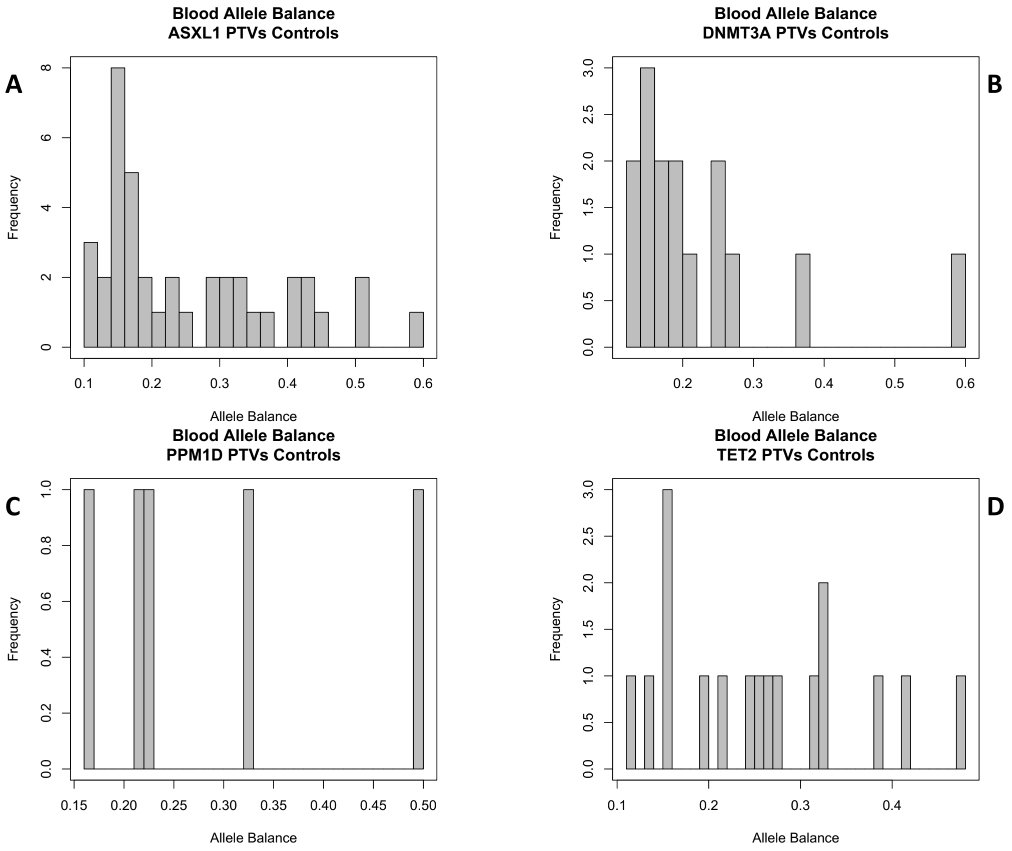


Figure 3.6. Allele balance for protein-truncating variants observed in blood DNA in controls in (**A**) *ASXL1*; (**B**) *DNMT3A*; (**C**) *PPM1D*; (**D**) *TET2*.

To further investigate any statistical bias due to sequencing coverage of cases and controls - we tested whether there is any statistical difference in coverage and ref/alt reads counts between cancer cases and controls that carry at least one PTV in the 4 candidate genes with generalized linear model testing (Supplementary Methods). Cancer status of a sample appears to be non-significant (p=0.279, p=0.898 if adjusted for age) parameter, confirming that called PTVs are adequately covered in both cases and controls and protein-truncating mosaic events have equal chances to be detected in both cohorts. We compared the probability of calling a protein truncating DNA variant in cases and controls with respect to coverage (**Figure 3.7**).

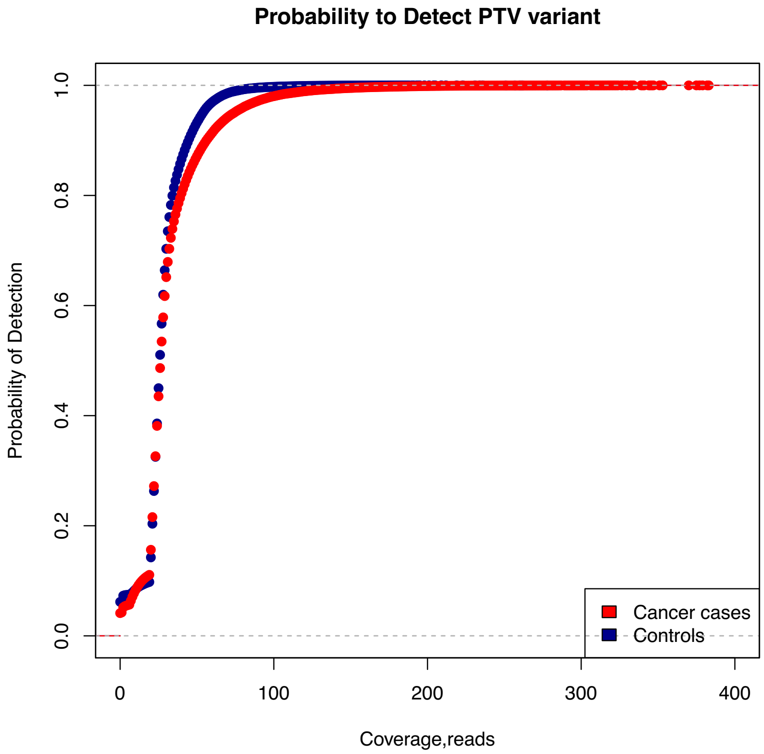


Figure 3.7. Probability of detection of a protein-truncating variant with respect to depth of coverage.

There is slightly higher sensitivity for the detection of DNA variants in cases, thus we adjusted further analysis for coverage differences. From these analyses, we conclude that all minor technical differences in sensitivity of mosaic variants search in cases and controls were accounted for – a pre-requisite for subsequent analyses.

We then investigated the effect of biological parameters on observation of the mosaic mutations – age and cancer therapy effects. Since our controls were, on average, roughly 10 years younger than the cancer cohort and age has been shown to be a strong predictor of the existence of somatic mosaic mutations, inclusion of age in the association model is critical. Older samples expectedly have higher probability of finding a mosaic variant (**Figure 3.8**)[5].

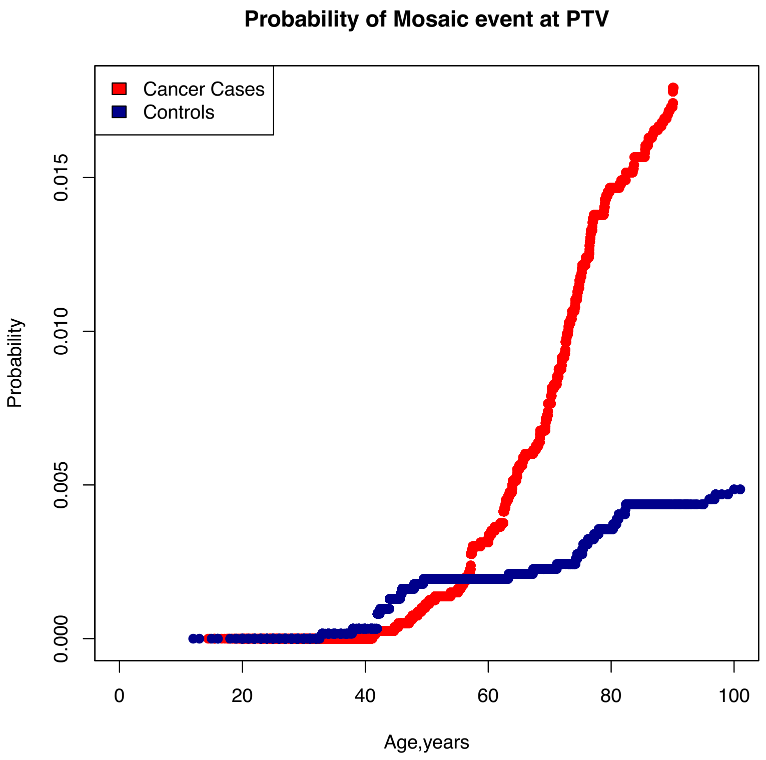


Figure 3.8. Probability of observing mosaic protein truncating variant with respect to age of DNA sampling.

Thus for case-control analysis we adjusted our model for age differences between cases and controls.

Another set of biological parameters to control for is clinical intervention. Specifically, chemotherapy and radiation treatment are of great importance and may alter somatic mutation rates. Within limited available clinical data in TCGA we saw no clear associations to treatment history - neoadjuvant treatment history (p=0.116), radiation therapy (p=0.348), pathologic tumor stage (p=0.354) or other outcome variables when adjusted for age and cancer subtype with mosaic PTV carrier status. This observation is consistent with previous reports of mosaicism in cancer case-control study[8]. Jacobs *et al.* reported no associations to smoking or cancer therapy using GWAS arrays, while confirmed associations to age and cancer status. Thus, we did not incorporate clinical parameters into further case-control model.

We then tested the association between mosaic PTV and cancer status by generating a data set consisting of 7,979 cancer cases and 6,177 controls. We applied a binomial generalized linear model considering age, coverage depth and mosaic PTV carrier status and found significant evidence of association with cancer status (P=0.00108, OR=1.26; OR CI=1.1-1.47). Since it was previously shown that *PPM1D* PTVs are associated with breast and ovarian cancers, we removed breast and ovarian cancer samples and repeated the analysis. It confirmed the observed association ($P=5.67x10^{-4}$, OR=1.3; OR CI=1.12-1.52) – suggesting that reported observations regarding *PPM1D* and breast and ovarian cancers are more general. We also adjusted our model for minor coverage differences between cases and controls.

It is known that PTVs in the last exon of *PPM1D* specifically that carry 'gain-of-function' effect are enriched in cases of breast and ovarian cancer[1]. We observed the same enrichment in our dataset – of 18 mosaic PTVs in *PPM1D*, 17 were in the last exon of the gene. We tested the 'gain-of-function' PTV hypothesis in other candidate genes as well (**Figure 3.9**).

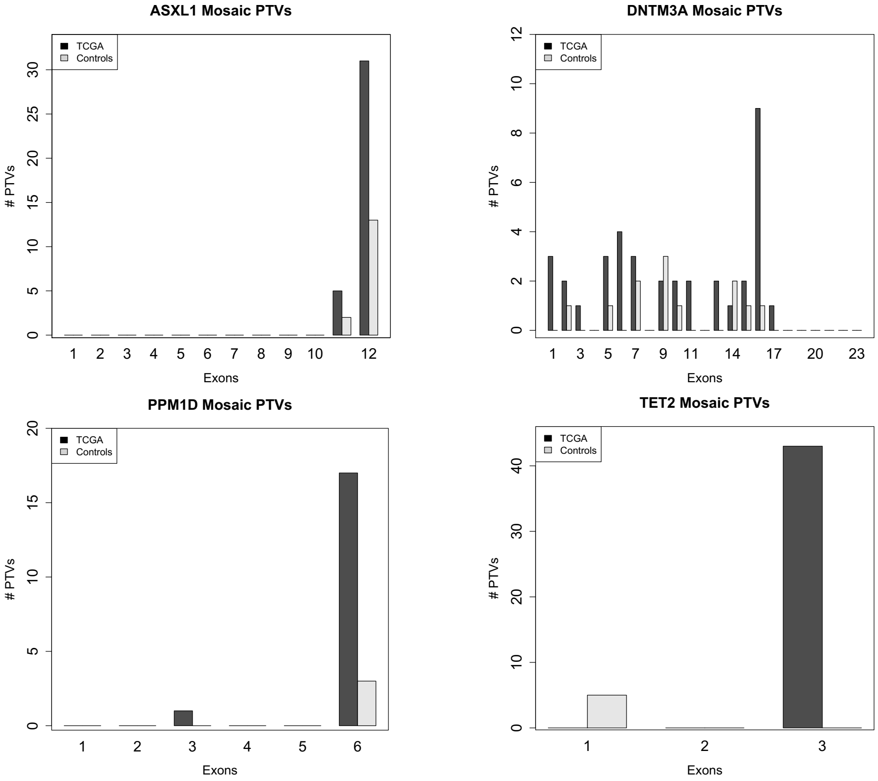


Figure 3.9. Exon specificity of mosaic protein truncating mutations.

ASXL1 follows the same pattern as *PPM1D* - 35 out of 40 PTVs in this gene are found in the last exon. *TET2* has strong enrichment of exon 3 – 44 out of 50 PTVs. This is intriguing because *TET2* transcript *ENST00000305737* has 3 exons and demonstrates enrichment of the last exon. Moreover, this transcript is mostly expressed in whole blood and EBV-transformed lymphocytes according to GTEx database. *DNTM3A* has no known pattern of mosaic PTVs distribution within the gene. Genovese et al, reported enrichment of the last exons of *DNMT3A* with mosaic missense mutations in leukemia cases. We observed similar enrichment in exons 17-23. However, no further studies are available to confirm whether missense mutations in this region also have ‘gain-of-function’ effect similar to the other candidate genes.

As previously demonstrated, mosaic PTVs in the list of candidate genes have been demonstrated to precede and predict the development of leukemia, indicating a causative role[1,2,9–11]. To determine the role of mosaic mutations in solid tumors we evaluated the quantity of mosaic PTVs between tumor and germline DNA in cancer samples. Mosaic PTVs in the candidate genes present in blood DNA were largely absent in tumor DNA from the same individual (**Figure 3.10**).

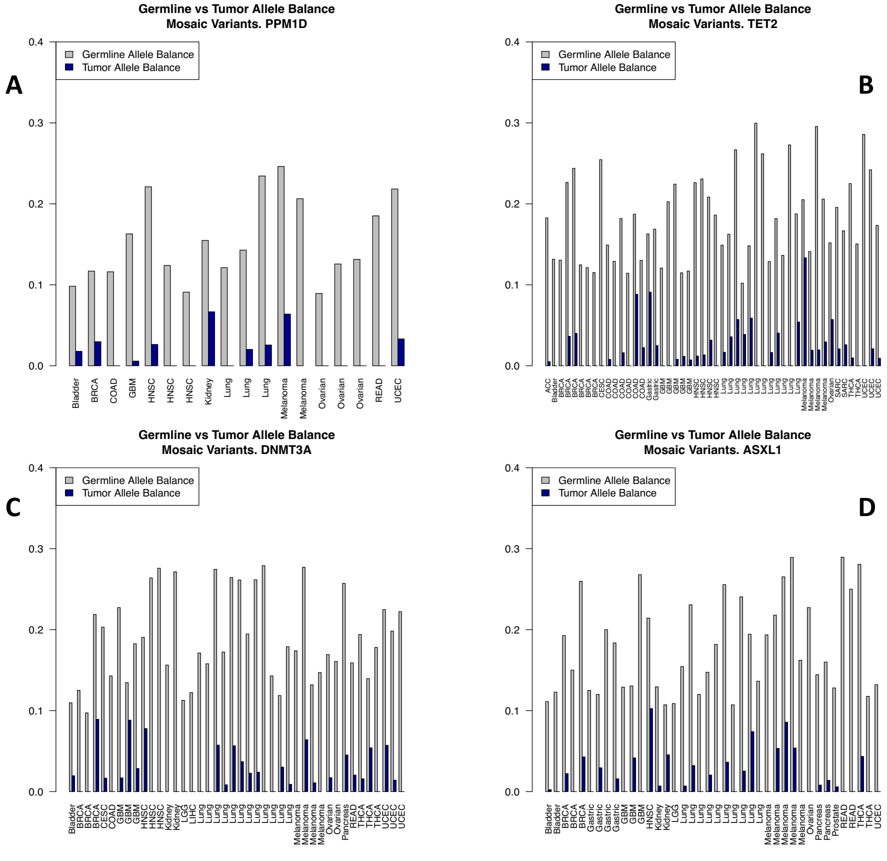


Figure 3.10. Allele balance of protein-truncating sites in blood DNA and in tumor DNA in (**A**) *PPM1D*; (**B**) *TET2*; (**C**) *DNMT3A*; (**D**) *TET2*.

Complete absence of these mutations in tumor sample is impossible due to ineluctable blood contamination of any tumor sample, however our data strongly indicates that these events in the blood did not represent residual evidence from driver mutations involved in tumor development (in which case we would have expected higher, or perhaps 100% of the mutated allele to be found). As before, we compared coverage in tumor and germline DNA samples and, consistent with the design of TCGA, that tumors have similar or better coverage indicating that the deficit of these mosaic events in tumors is not sensitivity based. This observation is consistent with the findings of mosaic *PPM1D* variants in breast/ovarian cancers[3].

We considered whether presence of mosaic PTVs showed any evidence of cancer specificity. Under the null hypothesis, mosaic PTVs are expected to be found in all candidate genes at the same rate in each of 20 cancer phenotypes. We first tested if any of the cancer phenotypes shows an unusual burden of mosaic PTVs. The empirical significance of observed mosaic PTV frequency deviation from null was assessed using the following scheme: For each cancer phenotype of N cases we drew random set of N samples from the pool of all cancer cases. Since age strongly affects the frequency of mosaic variants within cohort, only random sets with insignificant (as shown by Wilcoxon test) age difference from the target set were accepted. The empirical p-value was then calculated as the fraction of random sample sets with a mosaic PTVs frequency greater than the target set. Statistical significance threshold is given by multiple hypothesis testing correction considering 20 tested phenotypes – 0.05/20

(0.0025) (**Figure 3.11A**). Glioblastoma, melanoma and lung cancers demonstrate a significantly increased burden of mosaic PTVs compared to other cancers. We then examined the distribution of mosaic PTVs across the candidate genes in each cancer phenotype (**Figure 3.11B**).

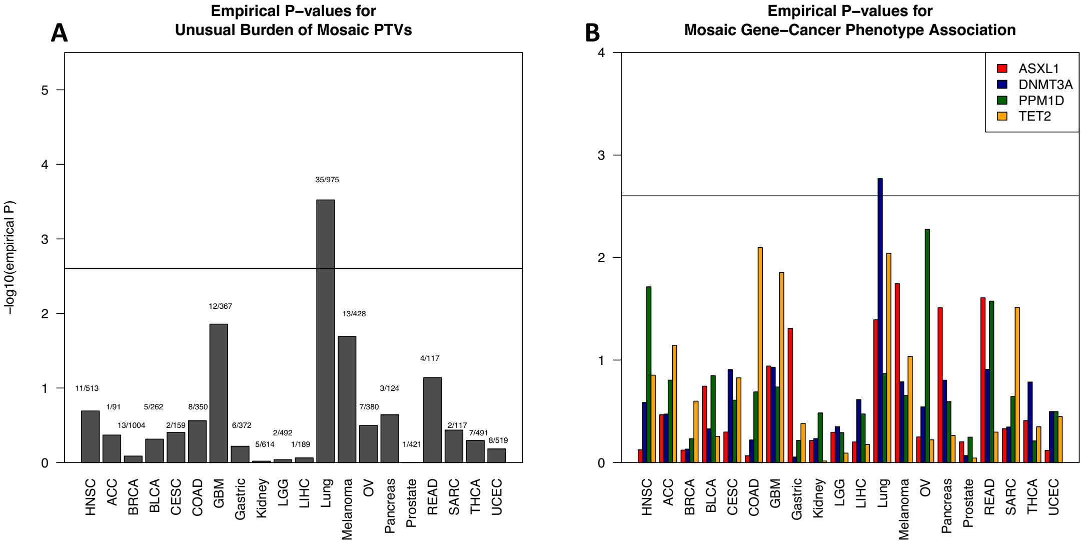


Figure 3.11. Testing for unusual burden of mosaic protein truncating variants. (**A**) Empirical significance of burden observed in all genes. (**B**) Empirical significance of burden observed in individual genes.

A similar approach was used as before: for each phenotype we estimated mosaic PTV frequencies in each of the candidate genes. Next, random sets of cancer cases with similar age distribution were generated. For each candidate gene the significance was estimated as fraction of random sets with greater mosaic PTV frequency in a gene of interest. The hypothesis of whether any gene has prevalent burden has been tested in 20 phenotypes, resulting in Bonferroni correction 0.05/20 for statistical significance threshold. It appears that several cancer types show a trend for accumulation of mosaic mutations in specific genes. Intriguingly, ovarian cancer is specifically associated with *PPM1D* mutations, which is supported by previous report[3]. We also observe associations of head and neck squamous cell carcinoma with *PPM1D*, colorectal adenocarcinoma and glioblastoma with *TET2*. Interestingly, cutaneous melanoma is associated with *ASXL1* mosaic mutations as *ASXL1* has protein-interaction with *BAP1*, a well-established risk factor for melanoma[12]. Lung cancer shows a burden of mosaic mutations that is distributed across several genes with *DNMT3A* being the most statistically significant. However, *ASXL1* and *TET2* show a nearly significant trend, suggesting no specificity in accumulation of the mosaic mutations.

We used the previously reported set of samples from Swedish national patient registers[1] to estimate the frequency of mosaic PTVs and associated solid-tumor cancer development in a population unselected for cancer. Accurate clinical records are available for this cohort so we sought to confirm our statistical approach for TCGA cohort.

We removed from analysis all samples that had an evidence of leukemia or lymphoma developed before the DNA collection as well as those samples that have mosaic missense mutations in *DNMT3A* to estimate the contribution of the PTVs only. The final dataset for this analysis consisted of (83 mosaic PTV carriers and 10867 non-carriers) samples. There were 11 individuals with pre-DNA collection record of the solid-tumor cancer in the cohort of mosaic PTV carriers and 1,105 samples with record of solid-tumor cancer among non-carriers. We tried using different thresholds for age of the samples to estimate significance of enrichment. However, due to a small incidence of the mosaic mutations in the population unselected for cancer, this test was inconclusive.

We added mosaic missense *DNMT3A* mutations carriers to the mosaic samples cohort and repeated population analysis. This resulted in a total of 153 mosaic mutation carriers. There were 26 individuals (~17%) with pre-DNA collection record of the solid-tumor cancer in the cohort of mosaic PTV carriers (1,104 – about 10% cancer records in 10,870 non-mosaic samples). Once corrected for age this enrichment appears to be insignificant, thus for samples unselected for cancer a much larger cohort is needed to reach a significant conclusion. However, we do observe a trend towards higher incidence of mosaic mutations in samples with cancer history. We analyzed effect of smoking among 4926 samples and saw no enrichment of smokers or former smokers in mosaic carriers (p=0.965 PTVs only, p=0.691 PTVs and mosaic missense in *DNMT3A*).

Analysis of larger clinical data should provide a clearer answer to whether mosaic mutations are precursors of cancer (and potentially play a causal role) or perhaps are non-causally associated as byproduct of previous therapy for an earlier cancer. Our analyses of these features are power-limited at this point and there is as yet no consensus surrounding this question. While genetic studies suggest that there is no correlation between cancer therapy and burden of mosaic mutations[8], clinical reports suggest that chemotherapy is one of the strong drivers of clonal expansion[6,7].

Discussion

Our study investigates the association of the mosaic protein-truncating variants in 4 genes previously associated with blood cancer risk in blood samples from patients with solid-tumor diagnoses.

We extend the previously observed strong association of mosaic PTVs with increased risk of leukemia to solid-tumor cancers. There are several possible explanations for such an observation. Recent findings in ovarian and breast cancer suggest a significant role of chemotherapy exposure in observed burden of mosaic PTVs in *PPM1D*[6,7]*.* Though our study lacks sufficiently detailed records of chemotherapy treatment to extend those observations, the breadth and robustness of the results here suggest that such an effect of treatment exposure may more generally apply to other candidate genes, cancer phenotypes and specific therapeutics. At the same time analysis of cancer case-control GWAS arrays did not report any association with cancer therapy regimens, or carcinogen exposure (smoking)[8]. While there is no unity in the field on this question, our observations of differences in PTV burden gene specificity according to cancer phenotype suggests that there could be some level of specificity of chemotherapy drugs to cause expansion/survival of certain mutated peripheral blood mononuclear cells clones. Importantly, however, such a link may provide a more general – and detectable – connection between early solid tumor diagnoses and enriched later incidence of leukemia.

There are other possible explanations for the observed association. First, there could be immune system changes in response to early pre-clinical stage of cancer. Our additional screening of early onset cancer cases (breast and ovarian cohort with cancer onset before 35, N=374) shows no enrichment in mosaic PTVs suggesting that this hypothesis is likely irrelevant and age of the samples plays important role (or serving as a trigger) for emergence of clonal expansion. Second, is a potential causal relationship. While a direct role as tumor drivers is ruled out by the absence of PTVs in tumors, we cannot completely eliminate the possibility that these represent a background cancer risk state but find no strong support for this hypothesis. Given fewer than 1% of the population carries a PTV in one of these candidate genes, a large-scale population study with a long-term pre- and post-cancer DNA collection and detailed treatment details will be needed to confidently answer the question whether blood mosaic PTVs are precursors or result of treatment for solid-tumor cancers.

Materials and Methods

Patient Cohorts

We used The Cancer Genome Atlas samples available at the Broad Institute (N=7979) and population controls without known cancer phenotype at the time of DNA collection (N=6177). All of the samples were sequenced at the Broad Institute. Libraries were then prepared for sequencing using a modified version of the manufacturer's suggested protocol, automated on the Agilent Bravo and Hamilton Starlet, followed by sequencing on the Illumina HiSeq 2000. Alignment and variant calling was performed using BWA/GATK/Picard pipeline where all of the samples were processed as a single batch. We removed sites with differential coverage to remove any potential bias between cases and controls.

A total of 12,380 Swedish research participants with psychiatric diagnoses were ascertained from the Swedish National Hospital Discharge Register, which captures all inpatient hospitalizations. Controls were randomly selected from population registers. We treated cases and controls as a single cohort for all analyses presented below, as none of the mutational variables analyzed below showed any relationship to psychiatric diagnosis after controlling for other factors such as age and smoking. Research participation and DNA sampling took place from 2005 to 2013. The 12,380 samples collected were sequenced in twelve separate waves. The first wave employed an earlier version of the hybrid-capture procedure (Agilent SureSelect Human All Exon Kit), which targets ~28 million base pairs of the human genome, partitioned in ~160,000 intervals, whereas the samples from the other waves used a newer version (Agilent SureSelect Human All Exon v.2 Kit), which targets ~32 million base pairs of the human genome, partitioned in ~190,000 intervals. The first wave was sequenced using Illumina GAII instruments and the remaining waves were sequenced using Illumina HiSeq 2000 and HiSeq 2500 instruments, with pair ended sequencing reads of 76 base pairs across all waves. Sequencing was performed at the Broad Institute of MIT and Harvard across the period of time from 2010 to 2013.

Dataset

Genotypes dataset was created by joint variant calling of cancer cases and non-cancer controls using HaplotypeCaller (GATK-3.0)[14–16] with Broad Institute calling pipeline. For functional annotation of variants we used Variant Effect Predictor by Ensembl[17].

PCA was performed to keep for analysis only samples of European ancestry to eliminate possible population effects. PCA was performed with EIGENSTRAT[18,19].

Resulting genotype file was used to create a PLINK/SEQ[20] project for further manipulations.

Clinical data

For testing relevance of the mosaic PTVs to medical treatment/outcome clinical data was downloaded from TCGA web-site https://tcga-data.nci.nih.gov/tcga/dataAccessMatrix.htm. All patients provided informed consent for research use of the collected data.

Generalized linear model and statistical tests

For all statistical tests we used R-3.0[21].

Defining mosaic genotypes: We kept only well-covered genotypes for further analysis (DP>=20). We used binomial test on the number of alternative reads for heterozygous sites to detect unusual distribution of reads suggesting a mosaic event. Under the null-hypothesis for heterozygous genotypes 50% of alternative reads are expected. So, using our lower boundary of 20x coverage, that provides minimal adequate statistics for binomial test we expect that only several alternative reads should evidence mosaic event rather than a true heterozygote. So, we required binomial $p<=0.001$ for true mosaics, which ensures matching with expectation (e.g. 3 alt reads vs 17 reference $p=0.002577$).

Controlling for coverage differences: To ensure that distribution of alternative and reference reads in PTV carriers does not depend on the cancer status we set up general linear model: (Ref reads, Alt reads) = $\beta_0 + \beta_1$*Cancer Status. With $p=0.279$ it appears that cancer cases and controls have similar distribution of the reads in the heterozygous genotypes at protein-truncating variants. Since age has strong impact on the mosaic status we adjusted the above model for age by performing it in two steps, fitting cancer status to age (Cancer Status = $\beta_0 + \beta_1$*Age) and then fitting the pairs of reference and alternative reads for each het genotype to the residues of the first fitting: (Ref reads, Alt reads) = $\beta'_0 + \beta'_1$*Resid(Cancer Status = $\beta_0 + \beta_1$*Age), adjusted for age $p=0.898$ confirms that there is no technical bias in reads distribution. We estimated probability of detecting a protein-truncating variant in cases and controls (Sup. Fig. 3), that is affected by quality of the DNA samples. Under the null we expect same probability of PTV detection with respect to coverage amongst cases and controls. Despite the difference between cases and controls is very tiny, we still adjusted association model for the mean coverage of the sample.

Controlling for biological parameters effects: We have set of parameters available from clinical data – age at the time of DNA sampling, neoadjuvant therapy and radiation therapy treatment, pathologic tumor stage. We used general linear model to assess significance of their contribution:

1) **Age**: We adjusted the model for coverage similarly to the previous model.

Mosaic Status = $\beta_0 + \beta_1$*Coverage

Age = $\beta'_0 + \beta'_1$*Resid(Mosaic Status = $\beta_0 + \beta_1$*Coverage)

2) **Clinical Intervention:** Model was adjusted for both coverage and age in order to ensure no bias is present.

Mosaic Status = $\beta_0 + \beta_1$*Coverage

Age = $\beta'_0 + \beta'_1$*Resid(Mosaic Status = $\beta_0 + \beta_1$*Coverage)

Resid(Age = $\beta'_0 + \beta'_1$*Resid(Mosaic Status = $\beta_0 + \beta_1$*Coverage)) = β''_0+ +β''_1*Neoadjuvant Therapy + β''_2*Radiation Treatment + β''_3*Tumor Stage

Case-Control association model: As we identified to ensure a robust comparison of cancer cases to controls model needs to be adjusted for age and coverage.

Mosaic Status = $\beta_0 + \beta_1$*Coverage

Age = $\beta'_0 + \beta'_1$*Resid(Mosaic Status = $\beta_0 + \beta_1$*Coverage)

Cancer Status = $\beta''_0 + \beta''_1$* Resid(Age = $\beta'_0 + \beta'_1$*Resid(Mosaic Status = β_0 + β_1*Coverage))

Unusual burden of the mosaic variants in cancer phenotype. Under the null model mosaic PTVs should have no specificity to cancer phenotype. This means, that frequency of the mosaic PTVs observation should be the same among each cancer cohort once accounted for age. Frequency of mosaic PTVs in a certain cancer phenotype cohort needs to be compared against large number of randomized sets of cancer samples that have similar age distribution. To generate the random sets of samples we ran a permutation scheme that ensures the age matching between the target and random cohorts. Example: Let the cancer phenotype A cohort have N=100 samples. We would randomly draw 100 samples out of all cancer samples in the dataset (N=7979) with age within two standard deviations of mean age in phenotype A cohort. We then ran Mann-Whitney test to confirm similarity of the age distributions between random and target sets of samples. If p<0.05 – we rejected this permutation and start over. For each permutation we recorded number of mosaic PTVs observed in random set of samples. Fraction of random sets with greater number of mosaic PTVs than in cohort with phenotype A is determined to be empirical p-value.

Mosaic gene specificity to cancer phenotype. This method is largely similar to the previous section. For each phenotype we estimated mosaic PTV frequencies in each of candidate genes. Next, random sets of cancer cases with similar age distribution were generated. For each candidate gene significance was estimated as fraction of random sets with greater mosaic PTV frequency in a gene of interest. Hypothesis of whether any gene has prevalent burden has been tested in 20 phenotypes, resulting in Bonferroni correction 0.05/20 for statistical significance threshold.

Author contributions

Mykyta Artomov: designed and performed analysis, writing

Manuel A. Rivas: designed analysis, overall guidance.

Giulio Genovese: provided Swedish Biobank genetic and clinical data.

Mark J. Daly: overall guidance, writing.

Bibliography

1. Genovese, G. *et al.* Clonal Hematopoiesis and Blood-Cancer Risk Inferred from Blood DNA Sequence. *N. Engl. J. Med.* **371,** 2477–2487 (2014).

2. Jaiswal, S. *et al.* Age-Related Clonal Hematopoiesis Associated with Adverse Outcomes. *N. Engl. J. Med.* **371,** 2488–2498 (2014).

3. Ruark, E. *et al.* Mosaic PPM1D mutations are associated with predisposition to breast and ovarian cancer. *Nature* **493,** 406–410 (2012).

4. The results published here are in whole or part based upon data generated by the TCGA Research Network. Available at: http://cancergenome.nih.gov/.

5. Xie, M. *et al.* Age-related mutations associated with clonal hematopoietic expansion and malignancies. *Nat. Med.* **20,** 1472–1478 (2014).

6. Pharoah, P. D. P. *et al. PPM1D* Mosaic Truncating Variants in Ovarian Cancer Cases May Be Treatment-Related Somatic Mutations. *J. Natl. Cancer Inst.* **108,** djv347 (2016).

7. Swisher, E. M. *et al.* Somatic Mosaic Mutations in *PPM1D* and *TP53* in the Blood of Women With Ovarian Carcinoma. *JAMA Oncol.* **2,** 370 (2016).

8. Jacobs, K. B. *et al.* Detectable clonal mosaicism and its relationship to aging and cancer. *Nat. Genet.* **44,** 651–658 (2012).

9. Ley, T. J. *et al. DNMT3A* Mutations in Acute Myeloid Leukemia. *N. Engl. J. Med.* **363,** 2424–2433 (2010).

10. Delhommeau, F. *et al.* Mutation in *TET2* in Myeloid Cancers. *N. Engl. J. Med.* **360,** 2289–2301 (2009).

11. Gelsi-Boyer, V. *et al.* Mutations of polycomb-associated gene *ASXL1* in myelodysplastic syndromes and chronic myelomonocytic leukaemia. *Br. J. Haematol.* **145,** 788–800 (2009).

12. Carbone, M. *et al.* BAP1 and cancer. *Nat. Rev. Cancer* **13,** 153–9 (2013).

13. Artomov, M., Rivas, M. A., Genovese, G. & Daly, M. J. Mosaic Mutations in Blood DNA Sequence Are Associated with Solid Tumor Cancers. *bioRxiv* (2016).

14. DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43,** 491–8 (2011).

15. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20,** 1297–303 (2010).

16. Van der Auwera, G. A. *et al.* From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr. Protoc. Bioinforma.* **43,** 11.10.1-33 (2013).

17. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol.* **17,** 122 (2016).

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

19. Patterson, N., Price, A. L. & Reich, D. Population Structure and Eigenanalysis. *PLoS Genet.* **2,** e190 (2006).

20. Https://atgu.mgh.harvard.edu/plinkseq/. PLINK/SEQ.

21. Team, R. C. A language and environment for statistical computing. (2013).

Rare variant, gene-based association study of hereditary melanoma using whole exome sequencing

Abstract

Extraordinary progress has been made in our understanding of common variants in many diseases, including melanoma. Since the contribution of rare coding variants is not as well characterized, we performed an exome-wide, gene-based association study of familial cutaneous melanoma (CM) and ocular melanoma (OM).

Using 11,990 jointly processed cases and controls, whole exome sequencing was performed followed by large-scale joint variant calling using GATK. Plink/SEQ was used for statistical analysis of genetic variation. Four models were used to estimate association among different types of variants. In vitro functional validation was performed using 3 human melanoma cell lines in 2D and 3D proliferation assays. In vivo tumor growth was assessed using xenografts of human melanoma A375 melanoma cells in nude mice (8 mice per arm). All statistical tests were two-sided.

Strong signals were detected for *CDKN2A* ($p=6.16x10^{-8}$) in the CM cohort (n=273) and *BAP1* ($p=3.83x10^{-6}$) in the OM (n=99) cohort. Eleven genes which exhibited borderline association ($p<10^{-4}$) were independently validated using the TCGA melanoma cohort (379 CM, 47 OM) and a matched set of 3,563 European controls with *CDKN2A* (p=0.009), *BAP1* (p=0.03) and *EBF3* ($p=4.75x10^{-4}$), a candidate risk locus, all showing evidence of replication. *EBF3* was then evaluated using germline data from a set of 132 familial melanoma cases and 4,769 controls of U.K. origin (joint $p=1.37x10^{-5}$). Somatically, loss of *EBF3* expression correlated with progression, poorer outcome and high MITF tumors. Functionally, induction of *EBF3* in melanoma cells reduced arrested cell growth in vitro, retarded tumor formation in vivo, and reduced MITF levels.

The results of this large rare variant germline association study further define the mutational landscape of hereditary melanoma and implicate *EBF3* as a possible CM predisposition gene.

Introduction

In 2016, an estimated 76,380 Americans will develop cutaneous melanoma (CM) and 10,130 will succumb to this disease making CM the fifth and seventh most common cancer among men and women, respectively[1]. Ocular melanoma (OM) is much rarer with only 2,500 estimated cases annually in the United States[2]. There is evidence for a strong genetic influence on melanoma risk. Ten percent of CM patients report a family history of melanoma[3,4], which confers a 2-fold risk of melanoma 1st degree relatives[5] and approximately 5-fold if two or more 1st-degree relatives are affected. Twin studies have estimated the heritability of melanoma to be 58%, which is statistically significantly higher than the 33% for cancers overall[6].

Prior to the advent of high-density human genomic maps, linkage efforts implicated familial melanoma loci on 1p36[7] and 9p21[8]; interval gene screening revealed deleterious germline alterations of *CDKN2A* in a subset of 9p21-linked families. Subsequent linkage efforts have produced additional candidate loci on chromosomes 1p36[7], 9q21[9], 5q31[10] and 1p22[11] without isolation of specific disease causing mutations. Rare germline mutations in *CDK4*, *BAP1*, *MITF*, *TERT* (promoter), *POT1*, *ACD* and *TERF2IP* have also been reported in both OM and CM families though, collectively, they account for <5% of all hereditary melanoma cases[12,13]. Finally, a rare functional polymorphism in MITF (pE318K) has been shown to double melanoma risk and alter MITF sumoylation[14].

Common variant association studies (i.e. genome-wide association studies (GWAS)) recently culminated in an analysis of 15,990 CM cases and 26,409 controls, which substantiated 20 genome-wide statistically significant loci[15]. The general synthesis from GWAS and other candidate association studies is that loci associated with pigmentation (*MC1R*, *TYR*, *ASIP, OCA2* and *SLC45A2*), nevus count (*CDKN2A-MTAP*, *PLA2G6* and *TERT*), DNA repair (*PARP1* and *ATM*) and telomere length (*TERT, OBFC1*) represent core drivers of a risk phenotype that has been delineated by epidemiologic studies[16–23].

The systematic pursuit of rare disease-causing variants is just emerging. Given the high cost of sequencing in the past, linkage analysis provided a robust method to leverage recombination for positional information. For relatively common endpoints, linkage has been largely unsuccessful as, beyond rare examples of single pedigrees that are sufficiently large and carry a near-fully penetrant mutation to generate statistically significant linkage, the polygenic nature of common disease precludes gene localization. To overcome the fact that rare variants are distributed across many genes, we and others have proposed methodologies for gene discovery through rare variant association studies (RVAS[24,25]). One common approach is to group the individual variants into sets (e.g. gene-based association) and compare the aggregate frequency distribution in cases vs. controls. Using this framework, we set out to comprehensively map the mutational landscape of cutaneous and ocular melanoma by performing whole exome sequencing followed by a gene-based association study of melanoma cases and PCA-matched European non-cancer controls.

Results

Cohort and overview

For the discovery set, individuals with familial CM/OM or multiple primary CM/OM were identified and their germline DNA were subjected to whole exome sequencing. A total of 273 CM (M/F 128/145; **Figure 3.12A**), 99 OM (M/F 46/53) and 7,629 (M/F 5451/2178) European non-cancer controls passed quality control and were included in the subsequent analysis (19 CM and 2 OM cases failed QC). The first replication cohort included 379 CM (18 samples were eliminated in case-control matching procedure) and 47 OM cases from TCGA and 3,563 European non-cancer controls. These additional samples were jointly processed through the same alignment and variant calling pipeline as the initial discovery set and subjected to the same quality control standards. To ensure ancestral matching, we performed principal component analysis (PCA; **Figure 3.12B**) between the cases and control cohorts in both the primary and TCGA replication cohorts.

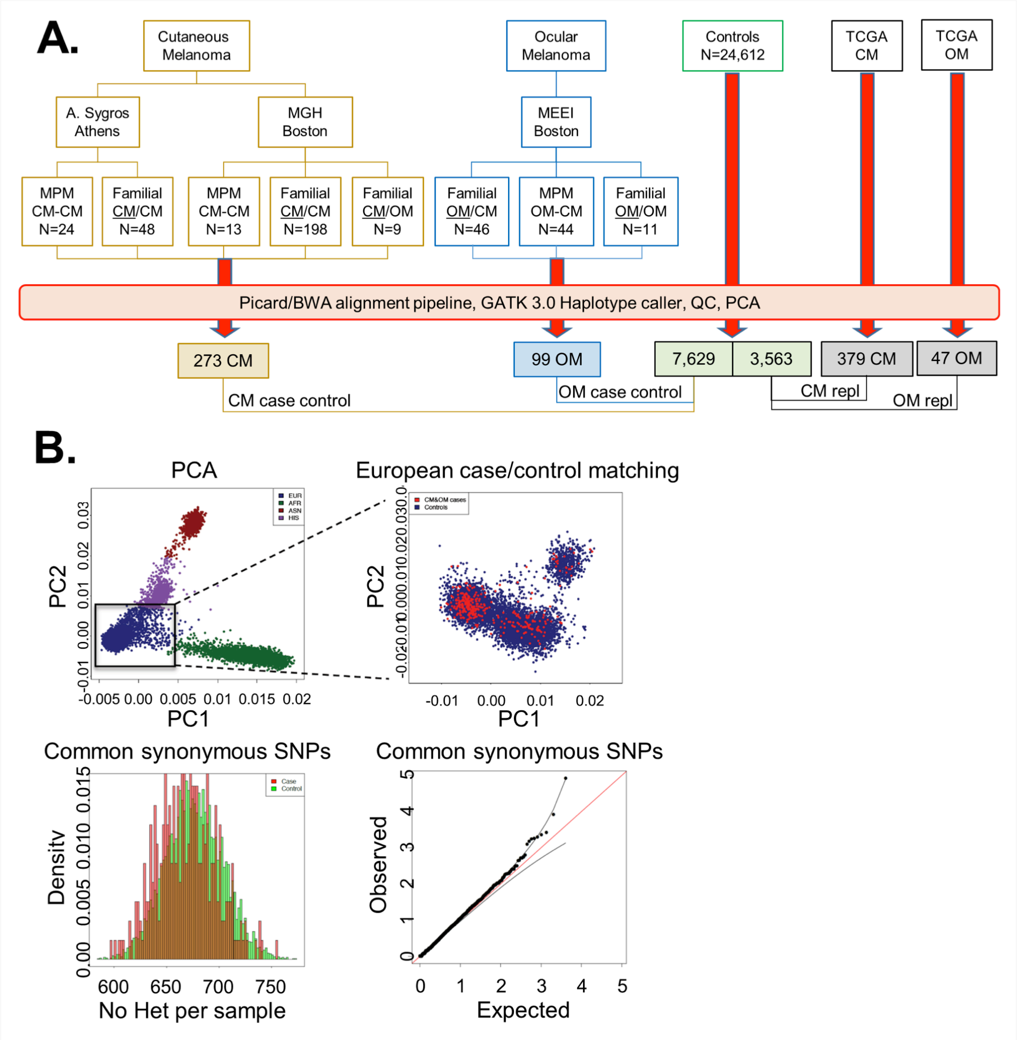


Figure 3.12. Study cohorts. (**A**) Cutaneous melanoma and ocular melanoma cases used in analysis. (**B**). Principal component analysis using the PCA module in PLINK; cases showing close matching with European controls. (**C**) Histogram of common synonymous SNPs between cases and controls and observed:expected ratios. MPM, multiple primary melanoma; CM, cutaneous melanoma; OM, ocular melanoma; PCA, principal component analysis; SNPs, single nucleotide polymorphisms; TCGA, The Cancer Genome Atlas; repl, replication study.

Our analysis was then restricted to European cluster of samples only. Examination of common synonymous variants (MAF>5%) revealed a null-distribution of the test statistic between cases and controls. There was an average of 25,633 SNPs per sample, which is within the expected range for a typical European germline exome[26]. In total, 11,990 samples were jointly processed and PCA-matched for analysis.

For *EBF3*, a second replication was performed using allele counts derived from 133 familial melanoma cases (i.e. 77 cases (66 families) from Leeds, U.K. and 56 cases (9 families) from Sydney, Australia) and 4,769 non-cancer controls from the UK10K population project; whole genotypes were not available for the U.K. replication cohort and thus were not matched by PCA.

Mutational landscape of cutaneous and ocular melanoma

We first interrogated our melanoma cases for rare PT mutations among known melanoma predisposition genes and their associated complexes or pathways (i.e. RB, telomerase/shelterin, BAP1). As expected, the strongest associations were for *CDKN2A* (p=6.16x10^{-8} statistically significant with CM) and *BAP1* (p=0.005 for CM and p=3.83x10^{-6} for OM). We detected 5 rare *POT1* mutations (p=0.002), including novel nonsense (p. Ser522X) and splice donor (chr7:124475332 C/T) mutations and a previously reported p.D224N variant. For *MC1R*, we examined red hair color (RHC) variants and observed 184 alternative alleles among CM (N=273) cohort and 2891 in controls (N=7,629) (p=5.00x10^{-12}; cumulative allele odds ratio=1.80, 95% CI: 1.53-2.10).

Figure 3.13A shows the P_{min} Manhattan plot for all 17,337 genes that were analyzed for association.

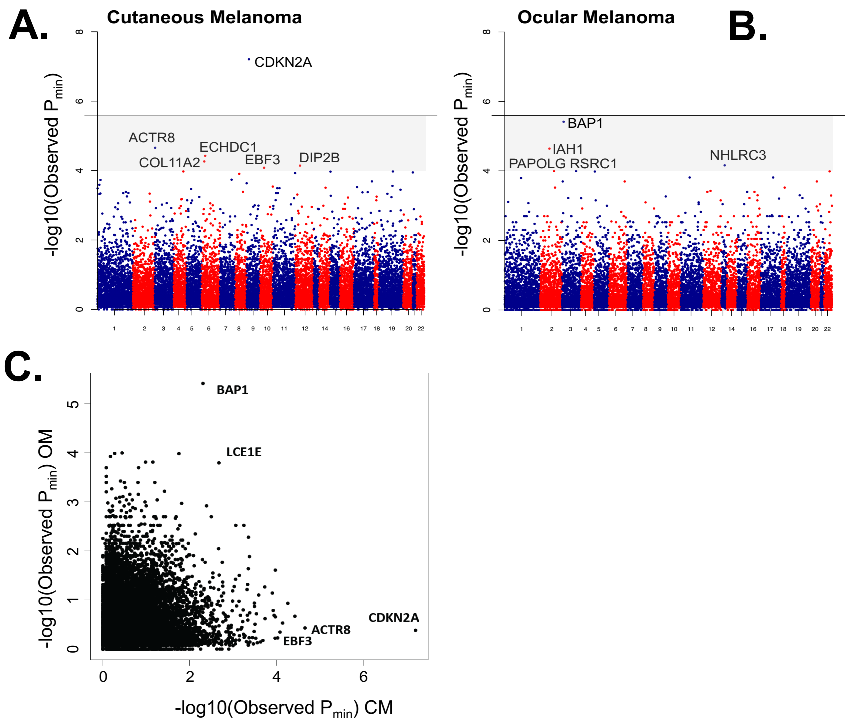


Figure 3.13. Mutational landscape of melanoma. (**A**). Manhattan plot showing gene-based associations across all loci. Genome-wide statistical significance is indicated by solid line. Genes which show near- statistically significant associations fall within the shaded region ($p<10^{-4}$). (**B**). Calculated $-\log_{10}$(P values) for genes in both CM and OM analyses. *BAP1* and *CDKN2A* clearly show strong preferential associations with OM and CM, respectively.

CDKN2A exhibited the strongest association ($P_{min} = 6.16 \times 10^{-8}$) and was the only locus to reach genome-wide statistical significance. Other candidates, which were near, but not reaching, genome-wide statistical significance were considered novel and which were subjected for further study included *ACTR8* ($P_{min}=2.18 \times 10^{-5}$, C-alpha), *ECHD1* ($P_{min}=3.73 \times 10^{-5}$, C-alpha), *COL11A2* ($P_{min}=5.42 \times 10^{-5}$, PTV burden) and *EBF3* ($P_{min}=8.22 \times 10^{-5}$, C-alpha).

In similar analyses for OM *BAP1* was the leading risk gene (P_{min}=3.83x10^{-6}; **Figure 3.13B**) though it did not quite reach genome-wide statistical significance. Other borderline candidates that were subjected to further analysis include *IAH1* (P_{min}=2.27x10^{-5}, C-alpha), *NHLRC3* (P_{min}=6.88x10^{-5}, C-alpha), *RSRC1* (P_{min}=1x10^{-4}, C-alpha), *PAPOLG* (P_{min}=1x10^{-4}, C-alpha).

All CM and OM loci with a association level of p<1.00x10^{-4} (6 CM genes, 5 OM genes) were then subjected to replication using the TCGA cohort as outlined above (**Table 3.4**). For CM, *CDKN2A* (p=9.31x10^{-3}; PTV burden) and *EBF3* (p=4.75x10^{-4}; C-alpha) both remained statistically significant while, for OM, *BAP1* also remained statistically significant (p=2.64x10^{-2}; PTV burden). As *EBF3* represents a potentially novel risk gene, we performed a second replication with this gene by interrogating whole exome and genome sequence data for 133 familial melanoma cases from 81 melanoma-prone kindreds. We identified a Leeds family with a p.N455S mutation which was detected in 1 of 2 affected members and a Sydney family with a p.G21S variant that was present in 2 of 6 affected members. Similar evaluation of 4,769 individuals from the UK10K cohort revealed 21 *EBF3* mutation carriers in this control set. In aggregate (**Table 3.4**), a joint burden test across all *EBF3* cohorts resulted in a statistically significant association between *EBF3* and cutaneous melanoma (9/784 cases and 42/15,961 controls; OR= 4.95 (95% CI 2.35-10.41), p=1.37x10^{-5}, Mantel-Haenszel chi-square test); to remove possible bias due to relatedness of carriers, we censored one of the two p.G21S carriers in the Sydney family while accounting for all non-carriers (i.e. 3 mutations out of 133 cases were identified in the U.K. replication but 2 carriers out of 132 individuals used for Mantel-Haenszel test).

Table 3.4. Melanoma Case/Control and Replication

Primary Cohort					TCGA Replication Cohort			U.K. *EBF3* Replication		Joint Odds Ratio¶ (CI 95%)
Gene	**Method***	**No. Variants CM CASES (N=273)**	**No. Variants CONTROL S (N=7629)**	**P$_{min}$**	**No. Variants TCGA CM (N=379)**	**No. Variants CONTROLS (N=3563)**	**P-value**	**No. Variants Sanger CM (N=132)‖**	**No. Variants UK10K (N=4769)**	
CDKN2A	PTV burden	5†	0	**6.16x10^{-8}**‡	2	0	**0.009** §	-	-	-
ACTR8	C-alpha	7	56	2.18x10^{-5}	5	26	0.14	-	-	-
ECHDC1	C-alpha	6	32	3.73x10^{-5}	1	14	0.83	-	-	-
COL11A2	PTV burden	4	3	5.42x10^{-5}	0	4	0.60	-	-	-
DIP2B	C-alpha	21	272	7.15x10^{-5}	4	78	0.20	-	-	-
EBF3	C-alpha	4	14	8.22x10^{-5}	3	7	**4.75x10^{-4}**	2	21	**4.95 (2.35-10.41)**
		No. Variants OM CASES (N=99)	**No. Variants CONTROL S (N=7629)**	**P$_{min}$**	**No. Variants TCGA OM (N=47)**	**No. Variants CONTROLS (N=3563)**	**P-value**			
BAP1	PTV burden	4	3	3.83x10^{-6}	1	1	**0.026**			
IAH1	C-alpha	6	68	2.27x10^{-5}	0	31	-			
NHLRC3	C-alpha	10	183	6.88x10^{-5}	0	84	-			
RSRC1	C-alpha	5	44	1.00x10^{-4}	0	18	-			
PAPOLG	C-alpha	6	47	1.00x10^{-4}	0	21	-			

* This column represents statistical test which was used to compute p-value, all listed tests are two-sided.

† Values in Table indicate number of rare (MAF < 0.01) variants Abbreviations: CI=confidence interval; CM, cutaneous melanoma; OM, ocular melanoma; TCGA, The Cancer Genome Atlas.

‡ genome-wide statistically significant.

§ Identical test used in both primary and TCGA replication cohorts for each gene (e.g. PTV burden for *CDKN2A* in both primary and TCGA replication cohorts)

‖ Melanoma kindreds of British ancestry were included to approximate the U.K.10K control collection. The U.K. replication included 77 cases (66 families) from Leeds, U.K. and 56 cases (9 families) from Sydney, Australia. A p.N455S mutation was found in one Leeds family (1/2 cases positive for mutation) and a p.G21S mutation was identified in one Sydney family (2/6 cases positive for mutation).

¶, Mantel-Haenszel chi-square test p value = 1.37x10^{-5}. In the joint burden analysis, we undertook a more conservative stance and censored the second carrier from the Sydney family while including all the non-carriers. Thus, 3 mutations were detected in 133 total cases but only 2 mutation carriers from 132 cases were used in the calculation.

Taken together, these results confirm the known contribution of *CDKN2A* and *BAP1* as strong risk loci for CM and OM, respectively, and nominate *EBF3* as a novel risk candidate for CM. In addition, the recovery of these known hereditary melanoma loci further verified the technical and methodological pipeline used in this RVAS.

Although the study was not designed to compare genes which selectively confer risk for either, or both, CM and OM, we did compare the P_{min}'s for each gene relative to their melanoma type (**Figure 3.13C**). There was some evidence that *LCE1E* can confer risk for both CM and OM (P_{min}=1.59x10^{-4} for OM and 2.56x10^{-3} for CM) though neither reached genome-wide statistical significance.

Functional validation of EBF3

Among the secondary candidates, none has been previously linked to cancer predisposition. Since *EBF3* ranked in 3 of 4 framework analyses, was replicated in the TCGA and European cohorts and has been reported to possess tumor suppressive activity in a number of non-melanoma cancers[27,28], we decided to examine *EBF3* in greater detail. Among melanoma cases, there were 4 mutations in our discovery set (one p.Q137R, one p.A409T and two p.N484S), 3 mutations in the TCGA cohort (one p.P594L and two p.G459C) and 2 variants in the U.K.-descent melanoma panel (p. N455S and p. G21S). Among the *EBF3* variants in our cases, 2 have been previously observed - p.N484S (MAF=3.00x10^{-5}; European/Non-Finnish) and p.G21S (MAF=2.21x10^{-4}; European/Non-Finnish) albeit at a much lower rate. The p.Q137R mutation falls within the DNA binding domain, the p.A409T mutation lies in the third helix of the HLHLH domain and the p.N484S alteration sits in and the C-terminal Pro/Ser/Thr-rich (PST) region. As *EBF3* have had minimal connection to melanoma biology, we set out to perform proof-of-concept validation for *EBF3* as a tumor suppressor using available somatic data and empirical cell-based assays.

Cancer susceptibility loci (e.g. *CDKN2A* and *BAP1*), often function as tumor suppressor genes and thus exhibit structural damage and/or expression loss during malignant progression. To this end, we surveyed the TCGA melanoma tumors for evidence of *EBF3* somatic copy-number loss. As shown in **Figure 3.14A**, *EBF3* harbored statistically significantly more deletions than other genes (p=2.00x10^{-151}, Wilcoxon test), which may, in some cases, reflect loss of the entire 10q arm. There is evidence of concurrent shallow deletions of *EBF3* and *PTEN* (also on 10q) though deep deletions of *EBF3* are uncommon and independent of *PTEN* while expression levels of the two genes appear to be also uncorrelated. On a mutation level, there were there were 28 missense alterations identified in the TCGA melanoma cohort with no loss-of-function variants.

With regards to expression, primary melanomas exhibited lower *EBF3* RNA levels when compared to either normal skin or benign nevi (**Figure 3.14B**; melanoma vs nevus, p=0.005, ANOVA) suggesting that loss of *EBF3* may contribute to melanoma progression. Diminished *EBF3* expression was also associated with heightened tumor aggressiveness. Using microarray data from a panel of 125 stage III melanoma tumor specimens obtained at Lund University and RNA-seq data from 470 TCGA tumor specimens, we found that lower *EBF3* levels showed a statistically significant correlation with poorer overall survival in the Lund data set (**Figure 3.14C**; p=0.02, log-rank test) and a marginal trend toward a worsened outcome among the TCGA tumors (p=0.10, log-rank test).

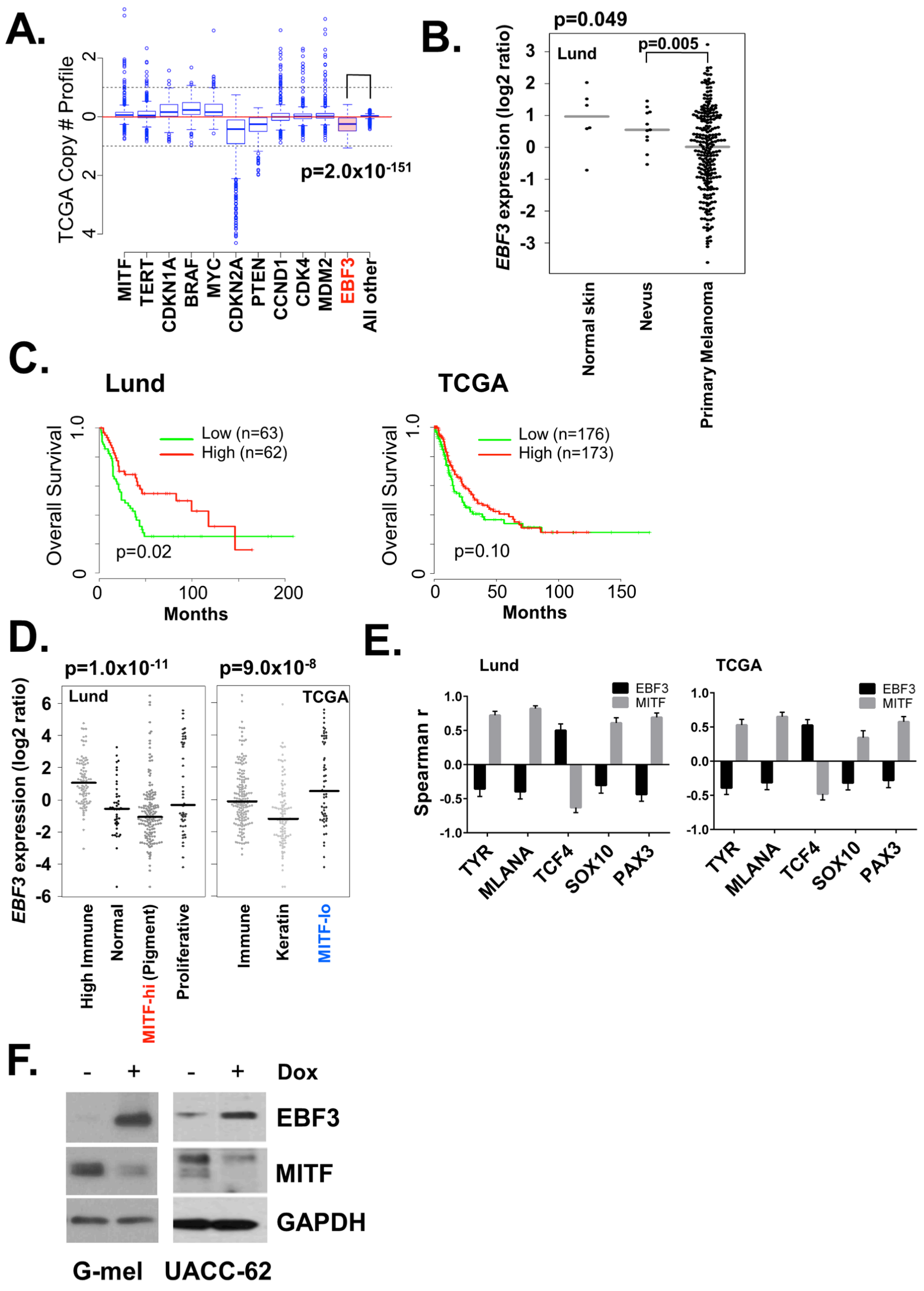


Figure 3.14. *EBF3* validation. (**A**). Copy number profile of *EBF3* (shaded red box) and various melanoma drivers among TCGA tumor specimens. Two-sided Wilcoxon test was used to compute p-values. (**B**). *EBF3* expression in normal skin, benign nevi and primary melanoma. P-values calculated using two-sided Wilcoxon test. (**C**). Lund University Medical Center and TCGA patients survival with metastatic stage III melanoma with high and low *EBF3* expression. Two-sided Cox proportional model was used to estimate significance. (**D**) *EBF3* expression in different molecular subtypes of melanoma (**E**) Spearman

Figure 3.14 (continuied) correlations for both *EBF3* and *MITF* and melanocyte lineage genes in the Lund and TCGA data sets; error bars indicate half-width of 95% confidence interval (**F**) Induction of *EBF3* in G-mel and UACC-62 cells using a Tet-responsive promoter. Dox, doxycycline, error bars represent mean +/- SD.

To better understand the molecular context of *EBF3* function, we examined *EBF3* levels across known molecular subtypes of melanoma. Strikingly, *EBF3* appears to be inversely correlated with the *MITF*-anchored classes, i.e. *EBF3* expression was lowest in the MITF-hi ("Pigmentation") subtype from the Lund cohort and highest in the "MITF-lo" subtype from the TCGA set (**Figure 3.14D**). Given this intriguing relationship, we searched for interactions between levels of *EBF3* and melanocyte lineage genes. There was a statistically significant inverse relationship between *EBF3* and *MITF* expression levels in both the Lund (Spearman r, -0.36, 95% CI: -0.47 to -0.24, p<0.0001) and TCGA (Spearman r, -0.30, 95% CI: -0.40 to -0.18, $p<10^{-4}$) tumor sets. Moreover, examination of several known *MITF* targets (**Figure 3.14E**; *TYR*, *MLANA*) and upstream regulators of *MITF* (*TCF4*, *SOX10*, *PAX3*) all revealed a consistent pattern whereby lineage genes which positively correlated with *MITF* (i.e. *TYR*, *MLANA*, *SOX10* and *PAX3*) were negatively correlated with *EBF3* while *TCF4*, a known negative regulator of *MITF*, was positively correlated with *EBF3*. These results reveal a reciprocal relationship between *EBF3* and *MITF* and raise the possibility that *EBF3* may antagonize MITF. To test this hypothesis, we induced *EBF3* in two MITF-expressing melanoma lines (G-mel and UACC-62) using a

Tet-responsive system and found a concomitant reduction of MITF in both cell lines (**Figure 3.14F**).

We next sought to phenotypically credential *EBF3* in cellular and animal experiments. As shown in **Figure 3.15A**, there was dose-dependent suppression of cellular proliferation with escalating levels of *EBF3* in 3 distinct melanoma cell lines; exposure of cells with the Tet-GFP control vector to doxycycline had no measurable effect on survival (data not shown). A375 and LOX cells were then subjected to a 3-D matrigel spheroid formation assay and, for both lines, there was a statistically significant reduction in colony volume with the introduction of *EBF3* (**Figure 3.15B**). Finally, overexpression of *EBF3* in A375 melanoma cells led to statistically significant suppression of tumor growth in nude mice (**Figure 3.15C**). Taken together, these results suggest that *EBF3* is a veritable tumor suppressor in melanoma, a function consistent with the role as a predisposition gene.

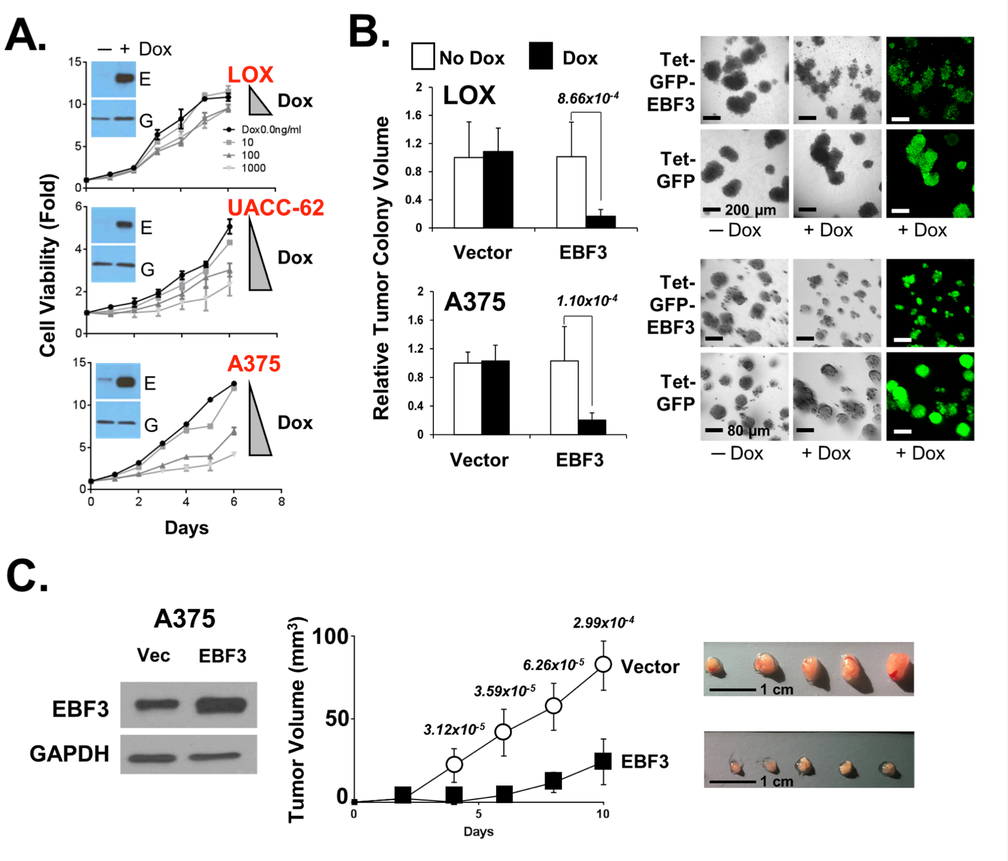


Figure 3.15. Functional accreditation of *EBF3*. (**A**). Cell viability assay demonstrating growth arrest of human melanoma cell lines with *EBF3* upregulation by dose-dependent induction with doxcycyline; experiments were performed in triplicates (relative mean viability +/- SD) and independently replicated using three different cell lines (LOX, UACC 62 and A375). (**B**) 3D spheroid formation is dramatically reduced by the induction of *EBF3* (relative mean tumor colony size +/- SD) (**C**). Constitutive overexpression of *EBF3* in A375 cells led to a profound suppression of tumor growth in *nu*/*nu* mice (n=8 in each arm; mean tumor volume +/- SD). P-values generated from two-sided T-tests are shown above each time point. Five representative dissected tumors from each arm are also shown. Vec, vector; Dox, doxycycline.

Discussion

The contribution of rare coding mutations to disease risk, such as melanoma, remains a bourgeoning but largely unexplored domain in human genomics. To the best of our knowledge, this is the first exome-based, rare variant association study in melanoma. Several technical advancements, which were introduced to permit the robust assembly of the large exome sequence dataset include joint variant calling and an advanced quality-check protocol.

Our results indicate that in the landscape of hereditary melanoma, *CDKN2A* and *BAP1* exhibit the strongest association with CM and OM, respectively. By design, we did not exclude these samples from analysis but blindly included the cases in the entire pipeline as “positive” controls; we were reassured that our methodology did recover these genes. There were additional risk loci which approached genome-wide statistical significance, one of which, *EBF3*, was further replicated in the TCGA and European melanoma cohorts.

We propose that *EBF3* has many of the features of melanoma predisposition gene based on multiple lines of investigation: (i) the enrichment for germline *EBF3* variants among individuals with melanoma across disparate cohorts, (ii) the presence of deletions at the *EBF3* locus in tumor specimens, (iii) the increased clinical aggressiveness associated with *EBF3* loss, (iv) the reciprocal interaction between *EBF3* and the known lineage oncogene *MITF* and (v) the direct inhibitory effects of *EBF3* on cellular growth and tumor formation. In two families where multiple affected cases were available for analysis, *EBF3* mutations were identified only in a fraction of the cases (1/2 and 2/6 affected members) suggesting that these mutations are moderate risk alleles, which is similar to the risk conferred by the MITF(E318K)[14] variant and consistent with *EBF3*’s calculated odds ratio of ~5.

EBF3 belongs to a family of transcription factors (*EBF*1-4) known to be involved in B cell differentiation and the pathogenesis of several tumor types[27,28]. Although speculative, it is conceivable that the role of EBF proteins in immune cell specification[27,29] could explain, in part, the observed association between EBF3 and the “high immune” melanoma subclass. The *EBF3* gene is located on the chromosome 10q26.3 and encodes a 596 amino acid protein with a conserved N-terminal DNA binding region, an IPT/TIG domain, an unusual helix-loop-helix-loop-helix (HLHLH) motif and a C-terminal PST domain[30]. Functionally, the EBF transcription factors bind to DNA with a consensus sequence of 5′-*CCCNNGGG*-3′ as homo- or heterodimers and can interact with p300[27]. Our mutations do not appear to cluster in any single domain though genotype/functional experiments are currently underway.

For melanoma, there is a single report of *EBF1* SNPs being correlated with survival among stage III and IV patients[31]. In our study, *EBF3* showed strong association in all tests except the LoF burden test. The gene is highly intolerant of LoF variation (ExAC; pLI=1.0) thereby suggesting that full loss of *EBF3* function may undergo strong negative selection. The precise mechanism of EBF3 action in melanoma remains to be elucidated though one recent report suggests that EBF3 might play a role in cell migration, and possibly proliferation, in a subset of melanoma lines[32]. There is precedence for transcription factors in cancer-predisposition. Perhaps the most relevant is a low-prevalence SNP in the *MITF* transcription factor (p.E318K) which alters a sumoylation site and which confers risk for both cutaneous melanoma and renal cell carcinoma[14,33].

Germline mutations in other transcription factors such as *TP53* and *RUNX1* also produce strong cancer phenotypes in Li-Fraumeni syndrome (OMIM #151623) and familial AML (OMIM #601399), respectively.

The design of our study incorporates several classic RVAS strategies[25] though there are also several limitations. Full exome sequencing is still more expensive compared to genotyping but the costs are converging. Although we enriched for genetic causation by focusing on rarer familial and multiple tumor cases, statistical power is still lower than those reported for common variant GWAS. Recognizing this risk, we subjected all available cases to our analytical pipeline including the *CDKN2A* and *BAP1* families and were reassured that statistically significant association signals were readily detected for these loci.

Despite limitations, the analyses in this report represent a major first step towards understanding the landscape of rare mutations in hereditary melanoma. The statistical methodologies which were blindly deployed in the case/control design unequivocally recovered several anticipated signals (i.e. *CDKN2A* and *BAP1*). Moreover, several sub-threshold loci have been nominated for future studies including a proof-of-concept validation of one such gene, *EBF3*.

Materials and Methods

Patient cohorts.

Primary discovery set. Cutaneous and ocular melanoma patients provided written consent for this study and were enrolled at 3 sites- the Massachusetts General Hospital (MGH; CM patients), the A. Sygros Hospital in Athens, Greece

(CM patients) and the Massachusetts Eye and Ear Infirmary (MEEI; OM patients) in Boston, MA- in accordance with protocols approved at these institutions.

All probands were considered "genetically enriched" based on the following criteria.

5. **MGH**: a histologically-proven CM *AND* at least one 1st degree affected relative *OR* ≥2 affected relatives on one side of the family regardless of degree of relationship (proband CM + relative with CM, "Familial CM/CM"; proband CM + relative with OM, "Familial CM/OM") *OR* ≥3 primary melanomas regardless of family history ("MPM CM-CM").
6. **The A. Sygros Hospital**: a histologically-proven CM *AND* ≥1 affected relative on one side of the family ("Familial CM/CM") *OR* ≥2 primary melanoma ("MPM CM-CM").
7. **MEEI**: a histologically or clinically diagnosed OM *AND* ≥1 relative affected with either CM or OM (proband OM + relative with OM, "Familial OM/OM"; proband OM + relative with CM, "Familial OM/CM") *OR* a second CM ("MPM OM-CM").

The dataset is registered in dbGAP under study number dbGaP Study Accession: phs000823.v1.p1.

Validation cases

1. **Sydney cases**: Individuals with a family history of melanoma were ascertained to the Genetic Epidemiology of Melanoma study at the Centre for Cancer Research, Westmead Institute of Medical Research; ultimately this

was as part of the international GenoMEL consortium (www.genomel.org), a multidisciplinary study of the genetic epidemiology of melanoma[34]. Briefly, multiple-case melanoma families have been ascertained from south eastern Australia since the 1980s through either: (i) a family member who attended the Sydney Melanoma Unit (the largest dedicated melanoma treatment service in the world, now Melanoma Institute Australia), the Victorian Melanoma Service, or other clinics, for treatment of melanoma, (ii) referral from health professionals such as clinical geneticists or dermatologists or occasionally, (iii) self-referral after media publicity of melanoma. Data on family structure, cancer history, illness characteristics, skin phenotype, other melanoma risk factors, and genotype are collected. Sequenced families had 6 or more cases of melanoma

2. **U.K. cases**. The cohort used for the whole exome analysis included families recruited to the U.K. Familial Melanoma Study (Section of Epidemiology and Biostatistics, University of Leeds (Leeds, UK)[35–38]. Inclusion for sequencing were families with 5 or more melanoma cases.

Blood samples were collected from eligible patients and DNA was extracted using routine commercial kits at the local institutions.

Exome sequencing, variant processing and calling. Whole exome libraries were prepared using a modified version of Agilent's Exome Capture kit and protocol, automated on the Agilent Bravo and Hamilton Starlet, followed by sequencing on the Illumina HiSeq-2000. We used an aggregated set of samples consented for joint variant calling resulting in 37,607 samples (germline from 292 CM patients, 101 OM patients, 397 TCGA CM patients, 47 TCGA OM patients, 24,612 controls and 12,158 other individuals included for joint variant calling only). All samples were aligned on the reference genome with BWA[39] and the best-practices GATK/Picard Pipeline, followed by joint variant calling with all samples processed as a single batch using GATK v 3.1-144 Haplotype Caller[26,40,41]. The resulting dataset had 7,094,027 distinct variants. Haplotype Caller, which was used for the ExAC database[42], was also used to detect indels. Selected mutations in *CDKN2A* and *BAP1* were confirmed with Sanger sequencing.

We performed principal component analysis (PCA) on common (MAF>5%) autosomal independent SNPs to filter out all non-European samples with Eigenstrat[43]. Relatedness analysis among Europeans was conducted with PLINK[44] as suggested in the PLINK best practices[45]. We used VEP[46] for functional annotation of the DNA variants. Common and rare variants analyses were conducted using PLINK/SEQ, which allows indexing of the large datasets. A burden test (Fisher's test with aggregated allele counts per gene) was used for rare protein truncating variants. Additionally, the VT and C-alpha tests were chosen as an adaptive burden test and variance-component test, respectively, to complement each other and to boost the power of rare missense and protein truncating variation association detection[47].

Statistical Methods. Gene-based association was performed using three distinct, but related, analytical frameworks. In the first analysis, a burden test was applied to all rare (MAF<1%) protein truncating (PT) variants since the functional impact is presumed to be severe and most directly inferred. Then, to expand on all rare variants (missense and PTV), a second analysis using both the C-alpha and variable threshold (VT) tests was employed. A third analysis applied the burden test to examine "ultra-rare" (MAF<0.0001; ExAC database http://exac.broadinstitute.org/gene/) variants as these may represent the most highly penetrant alleles. In the case of a single-model association test – the null statistic was represented by the uniform distribution of p-values. Since four different test statistics (i.e. VT, C-alpha, burden of PTVs, and burden of ExAC filtered variants) were applied and the lowest p-value was chosen, the null distribution was constructed by choosing the smallest p-value from four null single-statistic models (four sets of uniform p-values). This process simulates the procedure of selecting the best p-value out of four different test statistics that was used for gene-association testing thus making it a more conservative approach. Genome-wide statistical significance was adjusted by Bonferroni correction (i.e. 0.05 /17,337 genes tested, i.e. $p<2.88x10^{-6}$).

P-values in animal experiments were calculated using the Student T test. Burden of somatic deletions in *EBF3* was tested with Wilcoxon test. Survival correlation with *EBF3* expression level was assessed using log-rank test. All statistical tests were two-sided.

EBF3 validation

EBF3 and human melanoma specimens. TCGA RNAseq level 3 data (release 3.1.14.0, n=470) was quantile-normalized, and transformed as log2(data+1). Expression of *EBF3* was extracted from metastatic TCGA samples, divided into "low" and "high" groups by median expression. Survival analysis was performed using Kaplan Meier curves and log-rank test with TCGA follow-up data as of October 2015. To determine *EBF3* expression across gene expression subtypes, the subset of 329 TCGA samples with reported TCGA expression subtypes (Cancer Genome Atlas Consortium, 2015), was assigned to the Lund expression subtypes, as described previously[48]. From the TCGA copy number level 3 data (n=469) segmentation values were extracted for genes that have been reported to be lost or gained in melanoma, as well as for *EBF3*. For survival analysis in the Swedish data set, *EBF3* expression was extracted from the stage III patients of the Lund expression set[48], and survival analysis was performed as described for the TCGA cohort. To determine *EBF3* expression in normal skin, nevi and primary melanomas we used GSE57715 (GEO). Overall, we used gene expression data from 11 nevi and 237 primary melanomas to determine differences in EBF3 expression.

Correlations between *EBF3* and *MITF*, *TYR*, *MLANA*, *TCF4*, *SOX10* and *PAX3* and between *MITF* and the same panel of genes were generated using the mRNA Expression z-Scores (RNA Seq V2 RSEM) available through the cBioPortal website and calculated using Graphprism 6.0. Spearman correlations between *MITF* and all genes and *EBF3* and all genes were derived from the co-expression module of cBioPortal. Similar analyses were done using the Lund stage III melanoma set.

Cloning of EBF3 in Tet-On inducible expression vector. The pCMV6-Myc-DDK-hEBF3 cDNA clone was purchased from the OriGene Company (Rockville, MD) while PLVX-TRE3G-ZsGreen1 the tetracycline inducible, cDNA expression vector was purchased from Clontech Lab., (Mountain View, CA). Pseudo-typed lentiviruses were produced per manufacturer's protocols using the Lenti-X HTX Packaging Mix2 (Clontech) and Xfect polymer (Clontech) in HEK293FT (ATCC) producer cells. A day before transfection, 2.5 million HEK293FT cells were seeded on a 60 mm plate, in 5 ml of tetracycline free growth medium and incubated at 37ºC, 5% CO_2 incubator overnight. 48 hrs after transfection, the lentiviral supernatant was collected, centrifuged (500xg, 10 min), aliquoted and stored at –80°C until use. The virus production was estimated by using Lenti-X GoStix (Clontech).

LOX, UACC-62, G-mel and A375 melanoma cells were transduced with the lentiviral supernatant along with polybrene 8 µg/ml for 24 hrs. 24 hrs after infection, tetracycline-free growth medium containing Puromycin (4-6 µg/ml; Sigma-Aldrich) was added to the cells for antibiotic selection as described previously[49,50]. Stably transduced melanoma cells were utilized for further experiments.

Cell Proliferation Assay. Cell proliferation was measured CellTiter-Glo® (Promega). Briefly, viable cells were seeded at a density of 1,000 cells (in 150μl) per well in 96-well white plates (Corning, NY) with and without doxycycline in escalating doses (0, 10, 100, 1000 ng/ml). Cell viability was determined at regular 24-hour intervals for up to 6 days. Briefly, 30 μl of luminescence-based cell lysis/ATP reagent (CellTiter-Glo®) was added to each well and incubated on an orbital shaker for 15 minutes at room temperature. The luminescence was recorded on a spectrophotometer (SpectraMaxplus, Sunnyvale, CA) to measure cell viability. Background-subtracted luminescence values were plotted as fold change using GraphPad Prism.

RNA extraction for Real-Time PCR. Total RNA was isolated from 1x10^{7} cell pellet using RNeasy mini kit (Qiagen, Valencia, CA) according to the manufacturer's instructions. First-Strand cDNA Synthesis was performed by reverse transcription with high capacity RNA-to-cDNA Kit (Applied Biosystems, Waltham, MA) as instructed by the manufacturer. Levels of individual genes were quantified using a TaqMan Gene Expression Assays (Life Technologies): *MITF* (Hs01117294_m1), *EBF3* (Hs00406051_m1) and Human *GUSB* (4333767 T, Life Technologies, Grand Island, NY); *GUSB* was used as an endogenous control. Real-time QPCR was performed using the LightCycler480 (Roche, Indianapolis, IN) with denaturation at 95°C for 10 minutes followed by amplification at 95°C for 10 seconds, 55°C to 60°C for 10 seconds and 72°C for

45 cycles. The normalized, relative ratios of the genes between samples were expressed fold change.

Cell lysate preparation and Western Blotting. LOX, UACC-62, G-mel and A375 melanoma cells engineered with tetracycline-inducible *EBF3* were seeded in 10% tetracycline free fetal bovine serum (Atlanta Biologicals Flowery Branch, GA) 1x DMEM (Corning Manassas, VA) medium with and without doxycycline (Sigma Louis, MO) 100ng/ml for 2-4 days, and then washed with chilled phosphate-buffered saline and lysed with RIPA buffer supplemented with Halt protease inhibitor cocktail (Thermo Scientific, Rockford, IL). Equal amounts of protein (15 µg) were resolved onto 4–20% SDS polyacrylamide mini-gels (Bio-Rad, Hercules, CA) and transferred to nitrocellulose membranes. After blocking with nonfat 5% milk (Bio-Rad Hercules, California) in 1xTris-buffered saline–Tween20 for 1 hour, blots were incubated with primary antibodies against MITF 1:2000 (MS-771-P, Neo Markers Fremont, CA) and EBF3 1:2000 (SC-81999, Santa Cruz Biotechnology Dallas, Texas) for 2 hours, followed by horseradish peroxidase–conjugated secondary antibody (1:2,000) for 1h hour. Antigen–antibody complexes were detected by enhanced chemiluminescence (Bio-Rad Hercules, California).

3-D spheroid assay. Pre-chilled 96 well plates were coated with matrigel (BD Biosciences, San Jose, CA) (100 µl/well) and incubated for 30 min at 37°C. Both EBF3-inducible and control melanoma cells (A375 and LOX) were seeded on matrigel (500 cells in 100 µl media/well) and incubated for 30 min at 37°C. Then 100 µl media containing 10% matrigel was added to each well and incubated at 37°C for 24 hr. Then 100 ng/ml of Dox was added and incubated for 5 days to induce expression of EBF3. Images of colonies were taken using Olympus confocal microscope and the sizes of the colonies were measured and analyzed.

Animal studies. All experiments were performed in accordance with an approved Institutional Animal Care and Use Committee (IACUC) protocol. 6-week old nude mice (Gnotobiotic Mouse\Cox7 Core, MGH, Boston, MA) were injected subcutaneously with CD516B-2 (vector) or CD516B-2 – EBF3 transduced A375 cells. The cells were suspended in ice-cold HBSS and 1×10^6 cells were injected via 100 µl suspension in to the right hind leg of mice. There were 8 mice used per group. Animal body weights and tumor development were monitored and dimensions were measured by a Mitutoyo caliper (MSC, Melville, NY) every 48 hours. Tumor volume was calculated using mm^3= length × $width^2$ × 0.5. Animals were maintained in well-ventilated animal facility and tested in accordance with the MGH Animal Care and Use Committee guidelines. Data were expressed as mean ± S.D.; differences in tumor volume between vector- and *EBF3*-infected tumors were tested using the T-test at each time point.

Author contributions

Mykyta Artomov: sequencing data quality check and analysis, writing.

Alexander J. Stratigos: provided Greek cutaneous melanoma samples.

Ivana Kim, Evangelos S. Gragoudas, Anne Marie Lane: provided ocular melanoma samples.

Raj Kumar, Bobby Y. Reddy, Benchun Miao, Ching-Ni Njauw: *EBF3* functional accreditation.

Martin Lauss, Göran Jönsson: TCGA somatic mutations analysis and *EBF3* protein structural analysis.

Kristen Shannon: managed cutaneous melanoma samples.

Carla Daniela Robles-Espinoza, Aravind Sankar, Vivek Iyer, Julia A. Newton-Bishop, D. Timothy Bishop, Elizabeth A. Holland, Graham J. Mann, David J. Adams: provided UK replication set data.

Tarjinder Singh, Jeffrey Barrett: provided UK10K control data.

Mark J. Daly: study design, overall guidance, writing.

Hensin Tsao: study design, overall guidance, writing.

Bibliography

1. Siegel, R. L., Miller, K. D. & Jemal, A. Cancer statistics, 2016. *CA. Cancer J. Clin.* **66,** 7–30 (2016).

2. Jovanovic, P. *et al.* Ocular melanoma: an overview of the current status. *Int. J. Clin. Exp. Pathol.* **6,** 1230–44 (2013).

3. Greene, M. H. *et al.* High risk of malignant melanoma in melanoma-prone families with dysplastic nevi. *Ann. Intern. Med.* **102,** 458–65 (1985).

4. Kraemer, K. H. *et al.* Dysplastic naevi and cutaneous melanoma risk. *Lancet (London, England)* **2,** 1076–7 (1983).

5. Ford, D. *et al.* Risk of cutaneous melanoma associated with a family history of the disease. The International Melanoma Analysis Group (IMAGE). *Int. J. cancer* **62,** 377–81 (1995).

6. Mucci, L. A. *et al.* Familial Risk and Heritability of Cancer Among Twins in Nordic Countries. *JAMA* **315,** 68 (2016).

7. Bale, S. J. *et al.* Mapping the gene for hereditary cutaneous malignant melanoma-dysplastic nevus to chromosome 1p. *N. Engl. J. Med.* **320,** 1367–72 (1989).

8. Cannon-Albright, L. A. *et al.* Assignment of a locus for familial melanoma, MLM, to chromosome 9p13-p22. *Science* **258,** 1148–52 (1992).

9. Jönsson, G. *et al.* Mapping of a novel ocular and cutaneous malignant melanoma susceptibility locus to chromosome 9q21.32. *J. Natl. Cancer Inst.* **97,** 1377–82 (2005).

10. Falchi, M., Spector, T. D., Perks, U., Kato, B. S. & Bataille, V. Genome-wide search for nevus density shows linkage to two melanoma loci on chromosome 9 and identifies a new QTL on 5q31 in an adult twin cohort. *Hum. Mol. Genet.* **15,** 2975–9 (2006).

11. Gillanders, E. *et al.* Localization of a novel melanoma susceptibility locus to 1p22. *Am. J. Hum. Genet.* **73,** 301–13 (2003).

12. Soura, E., Eliades, P. J., Shannon, K., Stratigos, A. J. & Tsao, H. Hereditary melanoma: Update on syndromes and management. *J. Am. Acad. Dermatol.* **74,** 395–407 (2016).

13. Soura, E., Eliades, P. J., Shannon, K., Stratigos, A. J. & Tsao, H. Hereditary melanoma: Update on syndromes and management: Emerging melanoma cancer complexes and genetic counseling. *J. Am. Acad. Dermatol.* **74,** 411-20–2 (2016).

14. Yokoyama, S. *et al.* A novel recurrent mutation in MITF predisposes to familial and sporadic melanoma. *Nature* **480,** 99–103 (2011).

15. Law, M. H. *et al.* Genome-wide meta-analysis identifies five new susceptibility loci for cutaneous malignant melanoma. *Nat. Genet.* **47,** 987–95 (2015).

16. Elwood, J. M. *et al.* Pigmentation and skin reaction to sun as risk factors for cutaneous melanoma: Western Canada Melanoma Study. *Br. Med. J. (Clin. Res. Ed).* **288,** 99–102 (1984).

17. Holman, C. D. & Armstrong, B. K. Pigmentary traits, ethnic origin, benign nevi, and family history as risk factors for cutaneous malignant melanoma. *J. Natl. Cancer Inst.* **72,** 257–66 (1984).

18. MacKie, R. M., Freudenberger, T. & Aitchison, T. C. Personal risk-factor chart for cutaneous melanoma. *Lancet (London, England)* **2,** 487–90 (1989).

19. Gallagher, R. P. *et al.* Suntan, sunburn, and pigmentation factors and the frequency of acquired melanocytic nevi in children. Similarities to melanoma: the Vancouver Mole Study. *Arch. Dermatol.* **126,** 770–6 (1990).

20. Bliss, J. M. *et al.* Risk of cutaneous melanoma associated with pigmentation characteristics and freckling: systematic overview of 10 case-control studies. The International Melanoma Analysis Group (IMAGE). *Int. J. cancer* **62,** 367–76 (1995).

21. Holly, E. A., Aston, D. A., Cress, R. D., Ahn, D. K. & Kristiansen, J. J. Cutaneous melanoma in women. II. Phenotypic characteristics and other host-related factors. *Am. J. Epidemiol.* **141,** 934–42 (1995).

22. Naldi, L. *et al.* Sun exposure, phenotypic characteristics, and cutaneous malignant melanoma. An analysis according to different clinico-pathological variants and anatomic locations (Italy). *Cancer Causes Control* **16,** 893–9 (2005).

23. Iles, M. M. *et al.* The Effect on Melanoma Risk of Genes Previously Associated With Telomere Length. *JNCI J. Natl. Cancer Inst.* **106,** (2014).

24. Samocha, K. E. *et al.* A framework for the interpretation of de novo mutation in human disease. *Nat Genet* **46,** 944–950 (2014).

25. Zuk, O. *et al.* Searching for missing heritability: designing rare variant association studies. *Proc. Natl. Acad. Sci. U. S. A.* **111,** E455-64 (2014).

26. DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43,** 491–8 (2011).

27. Liao, D. Emerging roles of the EBF family of transcription factors in tumor suppression. *Mol. Cancer Res.* **7,** 1893–901 (2009).

28. Tao, Y.-F. *et al.* Early B-cell factor 3 (EBF3) is a novel tumor suppressor gene with promoter hypermethylation in pediatric acute myeloid leukemia. *J. Exp. Clin. Cancer Res.* **34,** 4 (2015).

29. Medina, K. L. *et al.* Assembling a gene regulatory network for specification of the B cell fate. *Dev. Cell* **7,** 607–17 (2004).

30. Siponen, M. I. *et al.* Structural determination of functional domains in early B-cell factor (EBF) family of transcription factors reveals similarities to Rel DNA-binding proteins and a novel dimerization motif. *J. Biol. Chem.* **285,** 25875–9 (2010).

31. Fang, S. *et al.* Association of Common Genetic Polymorphisms with Melanoma Patient IL-12p40 Blood Levels, Risk, and Outcomes. *J. Invest. Dermatol.* **135,**

2266–72 (2015).

32. Chatterjee, A. *et al.* Genome-wide methylation sequencing of paired primary and metastatic cell lines identifies common DNA methylation changes and a role for EBF3 as a candidate epigenetic driver of melanoma metastasis. *Oncotarget* **8,** 6085–6101 (2017).

33. Bertolotto, C. *et al.* A SUMOylation-defective MITF germline mutation predisposes to melanoma and renal carcinoma. *Nature* **480,** 94–98 (2011).

34. Holland, E. A., Schmid, H., Kefford, R. F. & Mann, G. J. CDKN2A (P16(INK4a)) and CDK4 mutation analysis in 131 Australian melanoma probands: effect of family history and multiple primary melanomas. *Genes. Chromosomes Cancer* **25,** 339–48 (1999).

35. Robles-Espinoza, C. D. *et al.* POT1 loss-of-function variants predispose to familial melanoma. *Nat. Genet.* **46,** 478–81 (2014).

36. Puntervoll, H. E. *et al.* Melanoma prone families with CDK4 germline mutation: phenotypic profile and associations with MC1R variants. *J. Med. Genet.* **50,** 264–70 (2013).

37. Harland, M. *et al.* A comparison of CDKN2A mutation detection within the Melanoma Genetics Consortium (GenoMEL). *Eur. J. Cancer* **44,** 1269–74 (2008).

38. Bishop, D. T. *et al.* Geographical variation in the penetrance of CDKN2A mutations for melanoma. *J. Natl. Cancer Inst.* **94,** 894–903 (2002).

39. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25,** 1754–1760 (2009).

40. Van der Auwera, G. A. *et al.* From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr. Protoc. Bioinforma.* **43,** 11.10.1-33 (2013).

41. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20,** 1297–303 (2010).

42. Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536,** 285–291 (2016).

43. Patterson, N., Price, A. L. & Reich, D. Population Structure and Eigenanalysis. *PLoS Genet.* **2,** e190 (2006).

44. Purcell, S. *et al.* PLINK: A Tool Set for Whole-Genome Association and

Population-Based Linkage Analyses. *Am. J. Hum. Genet.* **81,** 559–575 (2007).

45. PLINK best practices. Available at: http://pngu.mgh.harvard.edu/purcell/plink/.

46. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol.* **17,** 122 (2016).

47. Lee, S., Abecasis, G. R., Boehnke, M. & Lin, X. Rare-variant association analysis: study designs and statistical tests. *Am. J. Hum. Genet.* **95,** 5–23 (2014).

48. Cirenajwis, H. *et al.* Molecular stratification of metastatic melanoma using gene expression profiling: Prediction of survival outcome and benefit from molecular targeted therapy. *Oncotarget* **6,** 12297–309 (2015).

49. Kumar, R. *et al.* BAP1 has a survival role in cutaneous melanoma. *J. Invest. Dermatol.* **135,** 1089–97 (2015).

50. Ji, Z. *et al.* MITF Modulates Therapeutic Resistance through EGFR Signaling. *J. Invest. Dermatol.* **135,** 1863–72 (2015).

Insight into somatic variation in cancer: gender disparity and mutation burden in metastatic melanoma

Abstract

Gender differences in melanoma incidence and outcome have been consistently observed but remain biologically unexplained. We hypothesized that tumors are genetically distinct between men and women and analyzed the mutation spectra in 266 metastatic melanomas (102 women and 164 men) from The Cancer Genome Atlas (TCGA). We found a statistically significantly greater burden of missense mutations among men (male median 298 vs female median = 211.5; male-to-female ratio [M:F] = 1.85, 95% confidence interval [CI] = 1.44 to 2.39). We validated these initial findings using available data from a separate melanoma exome cohort (n = 95) and found a similar increase in missense mutations among men (male median 393 vs female median 259; M:F = 1.59, 95% CI = 1.12 to 2.27). In addition, we found improved survival with increasing log-transformed missense mutation count (univariate hazard ratio = 0.82, 95% CI = 0.69 to 0.98) for TCGA samples. Our analyses demonstrate for the first time a gender difference in mutation burden in cutaneous melanoma.

Introduction

Gender differences in both melanoma incidence and outcome are now well established. Last year, there were an estimated 43,890 cases of melanoma among men and only 32,210 cases among women[1]. Moreover, men accounted for 67% of the 9,710 melanoma-related deaths recorded in the United States in 2014[1]. This survival advantage among women has been confirmed in several additional studies[2–6]. While nonbiological explanations, including differences in clothing, sun-seeking behavior, and skin screening, have been hypothesized as sources of the disparity, little is known about intrinsic biological differences between melanomas from men and women. With the availability of several whole exome sequencing datasets for cutaneous melanoma, we set out to determine if genomic differences exist between male and female tumors.

Results

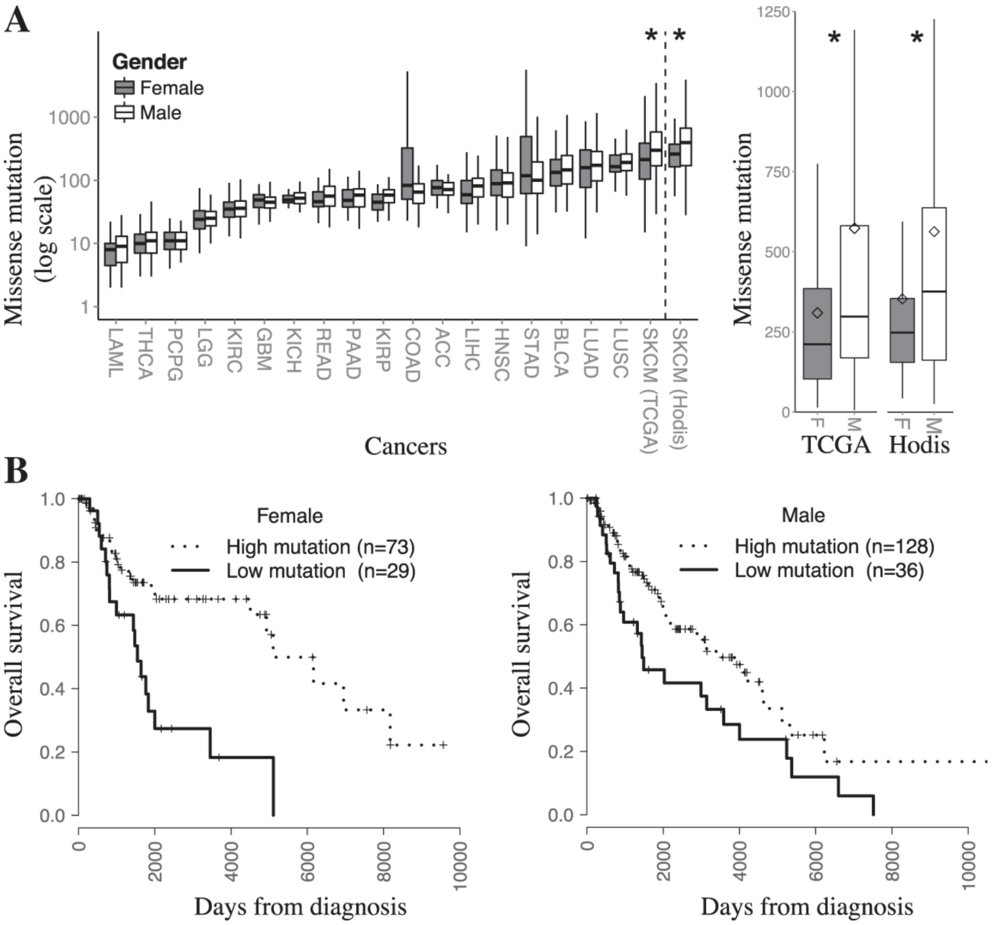
We analyzed the autosomal mutation spectra in 266 metastatic melanomas (102 women and 164 men) from The Cancer Genome Atlas (TCGA). Biospecimens from tumors were obtained from patients, with appropriate informed consent and institutional review board or ethics board approval facilitated by the National Cancer Institute (NCI) and National Human Genome Research Institute (NHGRI). A positive *NRAS* mutation status was determined if the tumor sample contained any of the following mutations: p.Q61H, p.Q61K, p.Q61L, p.Q61R, p.G12A, p.G12D, p.G12R, p.G13D, p.G13R, and p.61_62QE>HK. Positive *BRAF* mutation status was determined if the sample contained any of the following mutations: p.V600E, p.V600K, p.V600R, and p.600_601VK>E. Negative binomial regression predicting missense mutation counts in univariate and multivariable analyses were performed. Survival analyses were performed by fitting the Cox proportional hazard model. The effects of missense mutations are taken to be constant over time and implemented as 'stationary coefficients.' The proportional hazards assumption was tested by assessing correlation between survival times and Schoenfeld residuals. A *P* value of less than 0.05 was considered statistically significant, and all statistical tests were two-sided.

We found a statistically significantly greater burden of missense mutations among men (male median = 298 vs female median = 211.5; univariate male-to-female ratio [M:F] = 1.85, 95% confidence interval [CI] = 1.44 to 2.39), even after adjusting for age at diagnosis, primary tumor site, stage at diagnosis, site of sequenced tumor, history of neoadjuvant treatment, and *BRAF* and *NRAS* mutation status (multivariable M:F = 1.55, 95% CI = 1.19 to 2.02)[7]. Overall, there was also a greater burden of nonsense (M:F = 1.81, 95% CI = 1.39 to 2.36), stop loss (M:F = 2.83, 95% CI = 1.24 to 6.45), frameshift (M:F = 1.44, 95% CI = 1.18 to 1.77), and splice variant (M:F = 1.76, 95% CI = 1.38 to 2.25) mutations among men. Because the mutation imbalance was preserved for frameshift mutations and the number of "UV signature" tumors was not statistically significantly different between men and women (143/164 men vs 82/102 women, Fisher *P* = 0.16), it is unlikely that UV exposure differences would fully explain the gender mutation disparity[8].

To determine if the gender mutation differential was specific for melanoma, we also compared missense mutation load between men and women in 18 other TCGA cancers (**Figure 3.16A**). Only cutaneous melanoma demonstrated statistically significant gender differences in mutation burden among the 19 cancers after adjusting for multiple testing using the Holm-Bonferroni method. However, it is possible that tumors with smaller samples sizes and/or lower mutation counts were underpowered to detect a difference.

We subsequently validated our TCGA findings on an additional 95 melanoma exomes published by Hodis and colleagues (46 women and 49 men)[9]. There were statistically significant enrichments for missense (M:F = 1.59, 95% CI = 1.12 to 2.27), nonsense (M:F = 1.54, 95% CI = 1.09 to 2.17), and splice variant (M:F = 1.76, 95% CI = 1.26 to 3.02) mutations among men compared with women. Because the clinical information between the two cohorts was neither uniform nor consistently available, we performed a separate multivariable analysis of the validation set and found that gender mutation differences did not retain statistical significance. This may be due, in part, to the smaller validation sample size (n = 95), the statistically significant age disparity between men and women, and ascertainment differences between the TCGA and Hodis cohorts.

Missense mutation burden has been recently shown to be associated with survival in response to ipilimumab[10]. We next sought to determine if missense mutation burden influenced outcome of the TCGA patients and observed statistically significantly improved survival with increasing log-transformed missense mutation count (univariate hazard ratio [HR] = 0.82, 95% CI = 0.69 to 0.98; multivariable HR = 0.76, 95% CI = 0.63 to 0.91) independent of all other variables. The point of maximum statistical significance occurred with a missense mutation threshold of approximately 130 per tumor, a definition we used to plot survival curves (**Figure 3.16B**). We also tested for a gender interaction with missense mutation but observed no statistically significant interaction (*P* = 0.40).



Mutation	Time interval				
	0- 2000	2000- 4000	4000- 6000	6000- 8000	8000- 10000
High (n=73)	17/73	1/27	3/16	2/7	1/3
Low (n=29)	15/29	2/6	1/1	0/0	0/0

Mutation	Time interval				
	0- 2000	2000- 4000	4000- 6000	6000- 8000	8000- 10000
High (n=128)	36/128	11/52	6/20	1/4	0/1
Low (n=36)	17/36	4/11	2/6	2/3	0/0

Figure 3.16. Tumor Mutation Burden Associated with Gender and Survival. (**A**) To the left of the dotted line , boxplots depicting missense mutation distribution for 19 cancers on a log-scale are plotted in order of increasing median missense mutation count. To the right of the dotted line , boxplots depict missense mutation counts by gender from Hodis et al. (2012). TCGA and Hodis et al. (2012) samples are depicted on a normal scale to right of the main figure. Diamonds designate mean missense mutation count. Boxplot whiskers correspond to first and third quartiles of data. (**B**) While the positive relationship between mutation burden and survival is continuous, we determined that a definition of "high" vs "low" groups of missense mutation burden based on the threshold of 130 optimally captured survival differences for visualization purposes (univariate hazard ratio [HR] = 0.46, 95% confidence interval [CI] = 0.32 to 0.67; multivariable HR = 0.43, 95% CI = 0.28 to 0.64). These plots show survival

Figure 3.16 (continued) curves for high mutation tumors vs low mutation tumors among women on left and high mutation tumors vs low mutation tumors among men on right based on the 130 mutations threshold. Mutation counts will vary based on sequencing platform and bioinformatic pipeline and thus the '130' is not intended to represent an exact threshold of predictive discrimination for survival. A statistically significant survival advantage is observed for the women (univariate HR = 0.31, 95% CI = 0.17 to 0.58; multivariable HR = 0.32, 95% CI = 0.16 to 0.62) and men (univariate HR = 0.57, 95% CI = 0.36 to 0.92; multivariable HR = 0.49, 95% CI = 0.30 to 0.81). Among patients with "high mutation" tumors, the median survival was 167.9 months for women and 112.6 months for men. Events per patient at risk are indicated for 2000 days intervals below survival curves. Asterisks indicate a statistically significant difference (P < 0.05) in missense mutation burden by gender. P values were determined from the negative binomial regression model with use of two-sided Wald tests.

Our analysis indicates that male tumors harbor a higher mutation burden than female tumors. Yet, a higher mutation burden is also independently associated with better melanoma survival, which is supported by the recent ipilimumab findings[10]. It is worth noting that the mutation threshold (130), which optimizes survival differences in the TCGA metastatic samples (**Figure 3.16B**) is similar to the previously reported 100 threshold for ipilimumab responses[10]. This link between mutation burden and immune response may explain, in part, the female survival advantage observed clinically. A recent study found that female patients with melanoma had a statistically significantly higher frequency of tumor-associated, antigen-specific CD4+ T-cells than their male counterparts[11]. Furthermore, investigators have shown that anti-B16 melanoma immunity was better in B7-H1(-/-) female mice compared with syngeneic male mice as a result of reduced regulatory T-cell function[12]. Damian and colleagues have also shown that women are more resistant to UV-mediated immunosuppression compared with men[13]. Taken together, an "immune fitness" hypothesis suggests that men exhibit less effective antitumor immune surveillance, either because of innate or UV-mediated mechanisms, and are thus less able to clear the mutation-rich population of tumor cells compared with women; this may lead to a higher overall mutation burden in the tumor cell population. UV is an essential part of the model because the chronic DNA damage can lead to the generation of neo-epitopes and fuels immunogenicity. There are several corollaries to this hypothesis. First, melanomas on chronically irradiated skin (ie, head/neck) should harbor more mutations and show a higher male constituency—both of these appear to be supported by the TCGA data. Second, women should exhibit better survival than men, especially among the mutation-rich tumors. There appears to be a trend for improved survival among women with high-mutation (>130 mutations) tumors compared with men (median survival = 167.9 vs 112.6 months respectively, Wald test *P* = 0.13) (**Figure 3.16B**). Third, men should exhibit a higher risk of all skin cancers at chronically irradiated sites if in fact tumor immunity is fundamentally less fit; this male predominance has been in fact reported for other nonmelanoma skin cancer, including Merkel cell carcinoma, basal cell carcinoma, and squamous cell carcinoma[14–16].

This study has important limitations. First, the TCGA has delineated predominantly exomic sequences, and thus gender differences in the noncoding regions remain unscrutinized. Second, information about systemic, adjuvant therapy was only available for 36% (97) of the individuals and could not be incorporated into analyses. Lastly, accurate sun exposure history was not available to allow for gene-environment analyses. Despite these shortcomings, our results do begin to unravel, at a molecular level, the Gordian knot of gender disparity in melanoma.

Materials and Methods

All clinical and somatic mutation data for 19 cancers (SKCM, ACC, BLCA, COAD, GBM, HNSC, KICH, KIRC, KIRP, LAML, LGG, LIHC, LUAD, LUSC, PAAD, PCPG, READ, SKCM, STAD, THCA) in TCGA (http://cancergenome.nih.gov) were downloaded from the BROAD Firehose pipeline management system via the R package RTCGAToolbox (http://mksamur.github.io/RTCGAToolbox/; version 1.1.4), using the “20141017” run date[17]. The mutational data were restricted to autosomes. Duplicate somatic mutation calls for the same tumor sample, resulting from multiple comparisons between tumor sample sequences and multiple normal tissue sample sequences, were removed. We removed two individuals as extreme outliers from the collective dataset: TCGA-IB-7651 in PAAD and TCGA-AG-A002 in READ as these individuals were 412 and 249 interquartile ranges greater than the third quartiles respectively.

TCGA has predominantly metastatic melanoma tumor samples and a small set of primary tumor samples, and these analyses were restricted to tumor samples resected from non-primary sites, including regional cutaneous or subcutaneous tissue (includes satellite and in-transit metastasis), regional lymph nodes, and distant metastases. Data for two individuals, TCGA-D9-A4Z5 and TCGA-ER-A197, did not include tumor sample location. Pathology reports suggested these samples were sequenced primary tumors, and they were excluded from the analysis. Two individuals, TCGA-ER-A19T and TCGA-ER-A2NF, had multiple tumors collected from different sites and were likewise dropped from the analysis. Individual TCGA-ER-A2NC had both an NRAS and BRAF mutation and was removed out of concern for contamination. Seven individuals (TCGA-D3-A3C1, TCGA-D3-A3C3, TCGA-D3-A51G, TCGA-ER-A19O, TCGA-FR-A3YO, TCGA-RP-A695, TCGA-HR-A2OG) had no available time to event data and were excluded. In total, 266 tumor samples (102 females and 164 males) were included for final analysis.

The melanoma clinical dataset from Firehose did not include site of primary tumor, a variable that might be a statistically significant confounder, as there are known gender differences in incidence of melanoma by anatomic site, and anatomic sites are thought to have different amounts of sun exposure[18]. Therefore, anatomic site of primary tumor and pathology reports for each melanoma patient were downloaded directly from TCGA (https://tcga-data.nci.nih.gov/tcga/). Ambiguously coded sites such as “Extremities|Trunk” were recoded, after analyzing the available pathology report, to the most appropriate anatomic site or as not available (TCGA-EE-A2GR recoded as “Extremities”; TCGA-EE-A3JI recoded as not available; TCGA-D3-A2J9 recoded as not available; TCGA-D9-A4Z2 recoded as “Trunk”; TCGA-D3-A3CC recoded as not available; TCGA-ER-A194 recoded as “Trunk”; TCGA-ER-A198 recoded as not available). Only two individuals had a primary site other than extremities, trunk, or head and neck and were collapsed into an “other” category. American

Joint Committee on Cancer (AJCC) stage at initial diagnosis was collapsed to Stage 0, I, or II and Stage III and IV[7].

Validation Set

Gender status and mutation subtype counts for each individual included in the Hodis et al. 2012 literature were extracted from supplementary materials. This patient cohort was restricted to individuals with localization of primary melanoma to "skin" in order to exclude the uveal and mucosal melanomas in the dataset. AJCC stage at initial diagnosis was recoded as Stage 0, I, or II and Stage III and IV[7]. Anatomic sites of primary tumor were recoded as extremities, trunk, head and neck, or other.

Missing Data

For the pan-cancer analyses, only a small subset of clinical variables overlapped for all cancer types and analyses were restricted to univariate analyses comparing missense count and gender. There was no missing data for these variables, and thus complete data analysis was performed for pan-cancer analyses. For the melanoma cohorts, analyses were not restricted to univariate analyses. Notably, the variable with the most missing data is anatomic site of primary tumor (44 individuals; 16.5%) for the TCGA data and stage at initial diagnosis (21 individuals; 22.1%) for the Hodis validation set. For the melanoma mutation count analyses, we implemented multiple imputation using a bootstrapping-based EM algorithm with the Amelia II package (version 1.7.3) in R using the default and recommended parameters[19].

Statistical Methods

All analyses were performed using the statistical software R (R Development Core Team, 2010) version 3.1.1. Mutation subtype counts of missense, nonsense, frameshift, and splice variants among autosomes were tabulated in R. Negative binomial regression predicting missense mutation counts in univariate and multivariate analyses were performed. For the TCGA analysis, the multivariate model predicting mutation subtype counts included gender, age at diagnosis, anatomical site of primary, site of tumor sample sequenced, stage at diagnosis, history of neoadjuvant treatment, and *BRAF* and *NRAS* status as covariates. A matching multivariate analysis could not be performed on the Hodis validation set given the limited availability of clinical variables. However, a multivariate model including gender, age at diagnosis, anatomic site of primary, tumor stage at diagnosis, and BRAF/NRAS mutation status was performed. These analyses were implemented post-imputation with Amelia II using the 'zelig' package (http://gking.harvard.edu/zelig; version 3.5.4) in R[19]. For comparison across all 19 cancers, univariate analyses using negative binomial regression compared missense mutation burden by gender for each cancer. P-values were adjusted using the Holm-Bonferroni method for multiple testing. Regression coefficients and standard errors were converted to male-to-female mutation count ratios and confidence intervals. We also classified TCGA samples using a validated UV signature definition of tumors — C>T transitions at dipyrimidine sites for ≥ 60% of the total mutation burden or ≥ 5% CC>TT mutations[8].

For survival analysis in the TCGA, plotting of Kaplan-meier curves and calculating hazard ratios using the Cox proportional model using the R packages 'survival' (http://cran.r-project.org/web/packages/survival/; version 2.37-7) and 'zelig' (http://gking.harvard.edu/zelig; version 3.5.4). The effects of missense mutations are taken to be constant over time and implemented as 'stationary coefficients'. Missense mutation counts were log-transformed for survival analyses. Proportional hazards assumption was tested by assessing correlation between survival times and Schoenfeld residuals, followed by two-sided χ2 test, using the cox.zph() command in the R 'survival' package.

Threshold for Modeling High vs. Low mutation burden

We tested a range of threshold definitions, ranging from the first decile of missense mutation count and increasing by increments of 10 to the last decile of mutation count, using univariate analyses. We identified the threshold resulting in the maximum distance from 1 that had a statistically significant adjusted p-value (<0.05). P values were determined using the Cox proportional hazards model using two-sided Wald tests, adjusted for multiple testing using the Holm-Bonferroni method.

Author contributions

Sameer Gupta: study design, data assembling and analysis, writing.

Mykyta Artomov: statistical analysis, pipeline development.

William Goggins: consulting in statistical methods.

Mark J. Daly: overall guidance, manuscript comments.

Hensin Tsao: study design, overall guidance, writing.

Bibliography

1. Siegel, R., Ma, J., Zou, Z. & Jemal, A. Cancer statistics, 2014. *CA. Cancer J. Clin.* **64,** 9–29 (2014).

2. Joosse, A. *et al.* Gender Differences in Melanoma Survival: Female Patients Have a Decreased Risk of Metastasis. *J. Invest. Dermatol.* **131,** 719–726 (2011).

3. de Vries, E., Bray, F. I., Coebergh, J. W. W. & Parkin, D. M. Changing epidemiology of malignant cutaneous melanoma in Europe 1953-1997: rising trends in incidence and mortality but recent stabilizations in western Europe and decreases in Scandinavia. *Int. J. cancer* **107,** 119–26 (2003).

4. Robinson, J. K., Mallett, K. A., Turrisi, R. & Stapleton, J. Engaging Patients and Their Partners in Preventive Health Behaviors. *Arch. Dermatol.* **145,** 469–73 (2009).

5. Swetter, S. M. *et al.* Melanoma in Middle-aged and Older Men. *Arch. Dermatol.* **145,** 397–404 (2009).

6. Joosse, A. *et al.* Superior Outcome of Women With Stage I/II Cutaneous Melanoma: Pooled Analysis of Four European Organisation for Research and Treatment of Cancer Phase III Trials. *J. Clin. Oncol.* **30,** 2240–2247 (2012).

7. Balch, C. M. *et al.* Final Version of 2009 AJCC Melanoma Staging and Classification. *J. Clin. Oncol.* **27,** 6199–6206 (2009).

8. Brash, D. E. UV Signature Mutations. *Photochem. Photobiol.* **91,** 15–26 (2015).

9. Hodis, E. *et al.* A Landscape of Driver Mutations in Melanoma. *Cell* **150,** 251–263 (2012).

10. Snyder, A. *et al.* Genetic Basis for Clinical Response to CTLA-4 Blockade in Melanoma. *N. Engl. J. Med.* **371,** 2189–2199 (2014).

11. Wesa, A. K. *et al.* Circulating Type-1 Anti-Tumor CD4(+) T Cells are Preferentially Pro-Apoptotic in Cancer Patients. *Front. Oncol.* **4,** 266 (2014).

12. Lin, P.-Y. *et al.* B7-H1-Dependent Sex-Related Differences in Tumor Immunity and Immunotherapy Responses. *J. Immunol.* **185,** 2747–2753 (2010).

13. Damian, D. L. *et al.* UV Radiation-Induced Immunosuppression Is Greater in Men and Prevented by Topical Nicotinamide. *J. Invest. Dermatol.* **128,** 447–454 (2008).

14. Youlden, D. R., Soyer, H. P., Youl, P. H., Fritschi, L. & Baade, P. D. Incidence and Survival for Merkel Cell Carcinoma in Queensland, Australia, 1993-2010. *JAMA Dermatology* **150,** 864 (2014).

15. Bastiaens, M. T. *et al.* Differences in Age, Site Distribution, and Sex Between Nodular and Superficial Basal Cell Carcinomas Indicate Different Types of Tumors. *J. Invest. Dermatol.* **110,** 880–884 (1998).

16. Nguyen, K. D., Han, J., Li, T. & Qureshi, A. A. Invasive cutaneous squamous cell carcinoma incidence in US health care workers. *Arch. Dermatol. Res.* **306,** 555–560 (2014).

17. Samur, M. K. RTCGAToolbox: a new tool for exporting TCGA Firehose data. *PLoS One* **9,** e106397 (2014).

18. Whiteman, D. C. *et al.* Anatomic Site, Sun Exposure, and Risk of Cutaneous Melanoma. *J. Clin. Oncol.* **24,** 3172–3177 (2006).

19. James Honaker, Gary King, M. B. Amelia II: A Program for Missing Data. *J. Stat. Softw.* **7,** 1–47 (2011).

Chapter 4
Statistical framework for biological interpretation and improvement of genetic association studies

Abstract

Analysis of the genetic risk factors provides only unstructured pieces of information about the biology of a disorder. Generally, after identification of the associated loci massive follow-up studies are required to, first, prove the causal relationship, and, most importantly, understand the molecular mechanism of causality. Which locus should be prioritized for protein-level studies is currently determined based on empirical knowledge of protein function. Integration of the experimentally proven individual proteins functionality is then aimed to identify pathways affected by disease. Alternatively to this extensive approach, we developed a statistical framework that integrates genetic association data from multiple sources (GWAS, RVAS, etc.) and finds the protein-protein network returning the best cumulative association score. Using Bayesian model association results are then refined with evidence of the specific gene appearance in the best network. Our method provides a ranked list of genes prioritized based on both association strength and integration in the functional pathway. Such approach is essential for understanding biology of the disorders where it is impossible to build adequate animal model – autism, schizophrenia and other neuropsychiatric diseases.

Introduction

Since publication of the first draft of human genome in 2002[1] the ultimate aim for the computational biology was to aid therapy and drug discovery. Essential step in this is revealing disease biology by connecting genetic data to patient phenotype. Huntington's disease gene *HTT* (*HD*)[2] was the first in history gene mapped to a disease in 1983, followed by discoveries of cystic fibrosis causal *CFTR* gene in 1985[3]. These are examples of mendelian disorders which are caused by a specific type of alteration in patient's DNA. However, most of the common genetic disorders are complex (polygenic) and this makes interpretation of each individual risk locus significantly more challenging.

Methods like linkage analysis and GWAS were successful in identification risk loci but did not always point to specific genes and generally highlight not yet understood regulatory mechanisms. Fine-mapping[4] of GWAS hits has limited power and is mostly applicable for monogenic loci or in case when associated SNPs fall in coding sequence. Unfortunately, such ideal scenarios are quite rare. Lab techniques made limited progress in development of long-range DNA interactions maps. Thus interpretation of vast majority of the GWAS associations remains a challenge.

Additionally, many complex disorders have common risk factors (e.g. *MC1R* in cutaneous melanoma) and at the same time there are less frequently observed alterations in gens that are often left intact in affected samples (e.g. *EBF3* or *POT1* in cutaneous melanoma). Yet, such genes less frequently observed to be mutated in case cohorts provide valuable information about disease pathways. Often, they cannot be integrated together based on just genetics evidence and require additional reference from proteomics about functions of their downstream gene-products[5].

Phrase ‘pathway and network analysis’ would, thus, denote any analytic technique that benefits from biological pathway or molecular network information to gain insight into a biological system. The fundamental aim is to reduce data involving thousands of altered genes and proteins to a smaller and more interpretable set of altered processes. This process-oriented view helps generate testable hypotheses, identify drug targets, find cancer tumor subtypes with clinically distinct outcomes and identify disease-specific pathways[5].

Rare variation association studies have significant advantage[6] – they are focused on the coding variation and mapping of the association signal to a gene becomes straightforward. At the same time, complex disorders require large cohorts to gain enough statistical power for association detection, which cannot be achieved due to sequencing cost limitations. Integration of large GWAS statistical power and mapping simplicity of RVAS aligned on the map of protein interactions is a missing resource that will link genetic data to disease biology that can explain observed phenotype.

There are several conventional principles described in the literature for inferring molecular networks from different systems level data. These are matched to current experimental capabilities and will need revamping as technological leaps produce new types of data (e.g., more quantitative data and with real-time dynamics).

The key assumption used for network generation is that interaction between instances (e.g. proteins) generate statistical relations in the observed data. However, observed correlations in the data reflect only potential interactions. Robust reference set and statistical testing for significance of the observation are the base of the weighted approach.

Various statistical solutions have been successfully applied to network inference[7,8]. The commonality between the frameworks is that they model a target's behavior as a function of its regulators and search for the most predictive regulator set. More advanced model, using protein-protein interactions as a reference for the known pathways was developed by Dittrich *et al*[9], their algorithm utilizes Steiner decision tree to identify the best scoring network.

Protein-protein interactions databases have recently become important source of information that could serve as a basis for integration of genetic association results with functionality of the encoded proteins. Initial steps in this directions were made with DAPPLE software[10], which detected significantly higher connectivity within a set of proteins, encoded by genes identified in inflammatory bowel disease GWAS. Suggesting, that a truly causal genes should form a cluster of proteins with higher connectivity than random set. This observation is extended in breast cancer, where all of the genes that are currently screened in clinic for therapeutic needs form a highly interconnected cluster of proteins (mostly regulating DNA reparation pathway), which is never seen in random permutations.

Moreover, while first assemblies of the protein-protein interactions were unreliable and quite often had a substantial amount of false positives, current increased interest resulted in functionally validated databases (BioPlex[11], InWeb[12]). Recently released tissue specific PPI data could even further expand potential to use this data in interpretation of the genetic data and prioritization of the drug targets.

Results

Solution proposed by Dittrich *et al*[9,13] was implemented for microarray data. We sought to adapt this for the genetic associations data. Original approach uses association p-values to determine additive statistical weights for individual genes that are then used for search of the network maximizing cumulative score. Scores of individual genes are estimated using signal and noise decomposition of the p-values distribution. In this way genes with p-values falling into signal category receive positive scores and the noise ones receive negative scores (**Figure 4.1**).

Unlike in microarray data, in GWAS p-values of the significant associations could range between 10^{-8} and 10^{-100}. Original scoring method thus implies that genes with extremely small p-values would receive large positive scores. This makes the search of the most affected network very permissive and results in uninterpretable results. We restricted all of the genes with significant association p-values to receive the least positive score observed. In this way, all significant genes are treated equally inline with common interpretation of the association studies.

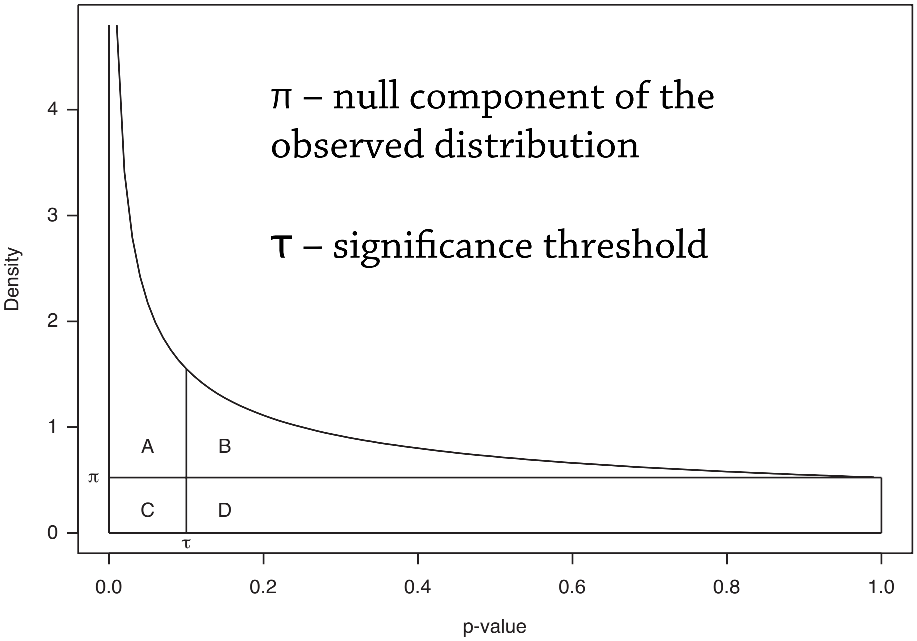


Figure 4.1. Betta-uniform fitting of the p-value distribution. A – true positive; B – false negative; C – false positive; D – true negative. Figure from Dittrich *et al*[9].

Significance threshold (τ) by default is given as Bonferroni derived correction for association test (0.05/number of genes tested). In case of a study that does not have any significant association observed significance threshold is determined from Betta-Uniform fitting for noise-signal decomposition[13].

To test non-randomness of the network we applied the scheme of connectivity-conserved permutations to generate random interactomes[10]. For each randomized interactome we repeated the search of the best scoring network and recorded its cumulative score and connectivity (number of edges per node).

Interpretation of the association studies

We tested this approach on the GWAS meta-analysis of coronary artery disease[14]. There are 48 previously known GWAS-identified loci associated with coronary artery disease reported and 8 new loci discovered in the meta-anlaysis. We used p-values for identified associated loci to build the most associated network. For polygenic loci the same p-value was assigned to each of the genes. Resulting network has significantly larger total score compared to the best network obtained from randomized interactome (p<0.001) and has significantly higher connectivity (1 edge per node, p=0.042), suggesting that clustering of these genes is non-random (**Figure 4.2**).

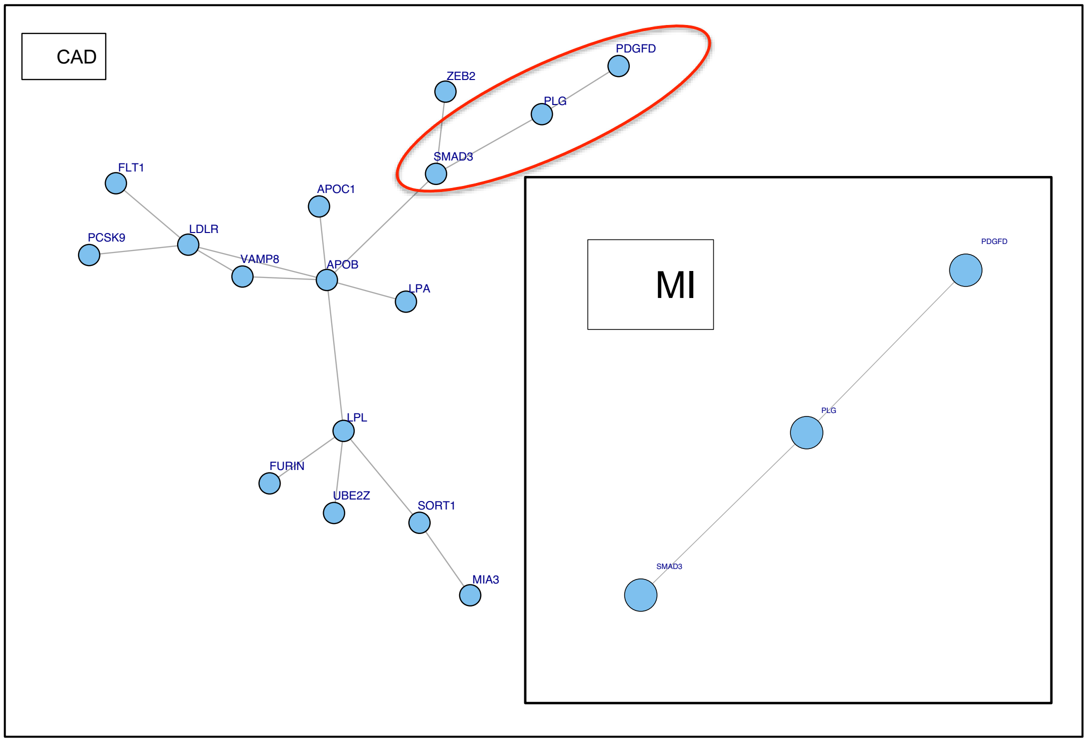


Figure 4.2. The best scoring networks from genetic association analysis of coronary artery disease patients (CAD) and patients with history of myocardial infarction (MI).

Two different functional clusters of proteins were identified within the network – lipid metabolism and inflammation proteins. Interestingly, we capture common drug targets for CAD – *Pcsk9* from lipid limb of the network for lowering blood LDL levels. We repeated analysis with association results obtained from cases with record of myocardial infarction. Three genes – *SMAD3, PLG* and *PDGFD* form the best scoring network for MI phenotype. Expectedly, these genes are also found in CAD network. These three genes are the core of the inflammation limb of the network. Suggesting, that inflammatory protein module has the strongest association in cases with history of MI.

Polygenic GWAS loci interpretation

Fine-mapping of the GWAS associations so far had limited success in identification of the signal driving genes[4]. It is common to find multiple genes within the same locus without clear evidence of any gene functionality to be relevant for the phenotype. In case of reasonably small number of genes (usually less than 5) it is possible to follow up with molecular biology studies to identify most likely relevant gene. However, quite often the number of genes within the locus is too large which makes wet-lab approach too expensive. Our default approach was to assign the same p-value for all genes within polygenic locus. However, based on functional relatedness evidence to the other candidate genes we are able to prioritize a specific gene from each locus. For example locus mapped to *SLC22A3-LPAL2-LPA* genes contributes only *LPA* to the most associated functional network (**Figure 4.2**). While this does not imply that *LPA* is the only candidate from this locus it does provide valuable information for prioritizing candidate genes.

Posterior association statistics

The fact that a gene appears in the best-scoring network provides extra information about its association signal. We sought to develop an empirical method to refine observed genetic association signal with protein-level data. There are two main factors contributing to chances of a gene to be selected for the best scoring network. First, the more positive is the score of a gene the more beneficial is inclusion of this gene to the network. In this way, genes with stronger association signals are more likely to be found in resulting network. Second, genes with greater connectivity despite their association signal strength are more likely to become a hub for positively scoring libs of the network. To test contribution of both factors we applied Bayesian approach for estimation of the posterior p-value of genetic association, given that a gene *A* with *N* connections is found in the best scoring network (**Equation 4.1**).

$$Pval|(gene\ A\ is\ in\ the\ best\ network)$$
$$= \frac{Prob(gene\ is\ in\ the\ best\ network|Pval) * Pval}{Prob(gene\ with\ N\ connections\ is\ in\ the\ best\ network)}\ (\boldsymbol{Equation\ 4.1})$$

Individual contributions of the association strength and connectivity were tested using two permutation schemes. First, to figure out prior probability of the gene with *N* connections to be included in the best network we randomly shuffled association p-values (*Pval*, **Equation 4.1**) for all genes. With this only effect of gene connectivity contributes to overall chances of gene selection. Ratio of the number of genes with the same number of connections found in resulting network to the total number of genes in the resulting network would return the probability of any gene with *N* connections to be found in the best scoring network.

Second, we estimated contribution of the association strength. We applied connectivity-conserved permutation scheme, similar to DAPPLE[10] and estimated ratio of the number of genes with association strength greater or equal than gene *A* to the total number of genes in the resulting network.

We used ranked list of genes from FSGS study[15] (Chapter 2) to find best-scoring PPI network. None of the genes originally reached significance threshold (0.05/1642 genes represented in reference interactome = $3x10^{-5}$). Resulting network consists of five genes – *WNK4, COL4A4, DLG5, KAT2B, UBC* (**Figure 4.3**).

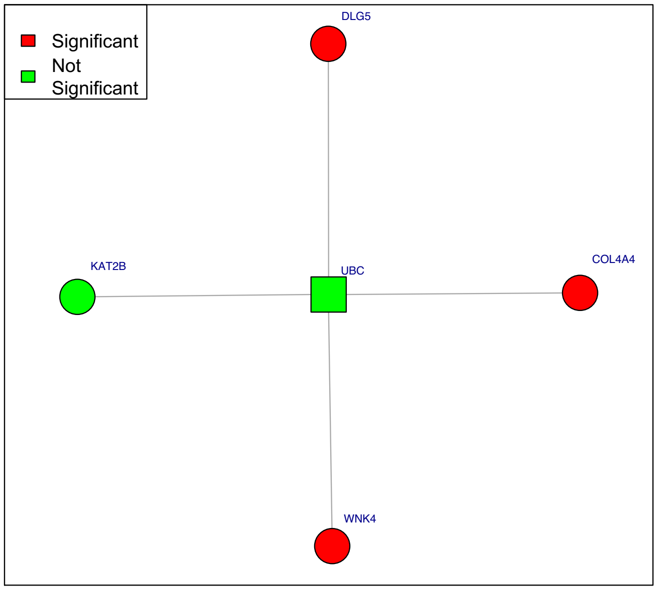


Figure 4.3. The best scoring network of FSGS rare-variant association study[15].

Then, posterior p-values of genetic association (probability of a gene being not associated to phenotype) were then calculated. 3 genes became significant – *COL4A4, DLG5, WNK4* (**Table 4.1**).

Table 4.1. Posterior association signal for FSGS rare-variant association study.

Gene	Original P-value	Posterior P-value
COL4A4	$6.76x10^{-5}$	$4.51x10^{-6}$
DLG5	$7.71x10^{-5}$	$1.54x10^{-5}$
KAT2B	$4.4x10^{-4}$	$8.73x10^{-4}$
UBC	$3.4x10^{-3}$	$1.4x10^{-4}$
WNK4	$3.1x10^{-4}$	$3.57x10^{-5}$

Interestingly, 2 genes had more than 10 fold improvement in association signal: *COL4A4* – is a known risk gene for FGSG; our mouse model[15] (**Chapter 2**) demonstrates that mice with non-functional *WNK4* develop proteinuria and FSGS, thus proving its causal role. *DLG5,* while gaining significance, has very modest improvement in posterior statistic – less than 3 fold enrichment, which is on the order of magnitude of the noise introduced by permutations to prior p-values. Our mouse model ruled out *DLG5* as a causal gene, as mice do not develop histologically proven FSGS with knockout of this gene.

One of the main challenges of our approach is *NP*-hardness of underlying combinatorial problem solved by FastHeinz algorithm[13]. While it is not a problem for the identification of the differential network, analysis of empirical significance requires tens of thousands runs to analyze random networks. We developed the R-based package ‘PPItools’ performing network analysis, beta-version of which is currently available at the Broad Institute computational cluster and undergoes testing by community. It uses parallel computations with UGER job scheduling environment[16] to perform individual random network analysis on a separate CPU. Source code could be found in the **Appendix**.

To overcome the computational challenges we replaced the permutation scheme for generation of the random interactomes with de-novo assembly of the interaction graph with pre-specified node connectivity. This solution has important benefit – it ensures uniform probability of each random graph to be generated. In this way a true estimate of the null distribution for the random networks could be achieved. This update is to appear in the second version of 'PPItools' package.

Discussion

Here, we developed a novel approach for interpreting and refining results of genetic association studies. Our methodology uses protein interactions data to find the most associated subset of connected genes. Interpretation of the biological mechanism underlying disorder and discovery of the therapeutic targets remains ultimate goal of genetic studies. Using examples of different disorders we show how our tool could be used for interpretation of polygenic GWAS loci, discovering network modules with different functions (lipid metabolism/inflammation in CAD) and statistical assessment of significance for individual genes within a network.

Currently, significant knowledge is accumulated about coding DNA regions, yet less is known about regulatory sequence. Interpreting of the eQTLs and non-coding variants (vast majority of known GWAS signals) still remains a great challenge. Composite methods taking into account long-range interactions within DNA locus, relevance of the protein product to a phenotype and its integration in the protein pathways are essential for integration of the genetic studies with molecular biology. Accordingly, our method is one of the modules of future integrated pipeline that interprets genetic studies to satisfy therapeutic demands.

Author contributions

Mykyta Artomov: designed and wrote original code and permutation schemes.

Alexey Sergushichev: software design consulting for the second release of the ‘PPItools’ package.

Maxim Artyomov: original idea and consulting.

Mark J. Daly: overall guidance, study design, statistics consulting.

Bibliography

1. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921 (2001).

2. Gusella, J. F. *et al.* A polymorphic DNA marker genetically linked to Huntington’s disease. *Nature* **306,** 234–8

3. Tsui, L. *et al.* Cystic fibrosis locus defined by a genetically linked polymorphic DNA marker. *Science (80-.).* **230,** (1985).

4. Altshuler, D., Daly, M. J. & Lander, E. S. Genetic Mapping in Human Disease. *Science (80-.).* **322,** 881–888 (2008).

5. Creixell, P. *et al.* Pathway and network analysis of cancer genomes. *Nat. Methods* **12,** 615–621 (2015).

6. Lee, S., Abecasis, G. R., Boehnke, M. & Lin, X. Rare-variant association analysis: study designs and statistical tests. *Am. J. Hum. Genet.* **95,** 5–23 (2014).

7. Friedman, N., Linial, M., Nachman, I. & Pe'er, D. Using Bayesian Networks to Analyze Expression Data. *J. Comput. Biol.* **7,** 601–620 (2000).

8. Bonneau, R. *et al.* A Predictive Model for Transcriptional Control of Physiology in a Free Living Cell. *Cell* **131,** 1354–1365 (2007).

9. Dittrich, M. T., Klau, G. W., Rosenwald, A., Dandekar, T. & Muller, T. Identifying functional modules in protein-protein interaction networks: an integrated exact approach. *Bioinformatics* **24,** i223–i231 (2008).

10. Rossin, E. J. *et al.* Proteins Encoded in Genomic Regions Associated with Immune-Mediated Disease Physically Interact and Suggest Underlying Biology. *PLoS Genet.* **7,** e1001273 (2011).

11. Huttlin, E. L. *et al.* The BioPlex Network: A Systematic Exploration of the Human Interactome. *Cell* **162,** 425–440 (2015).

12. Li, T. *et al.* A scored human protein–protein interaction network to catalyze genomic interpretation. *Nat. Methods* **14,** 61–64 (2016).

13. Beisser, D., Klau, G. W., Dandekar, T., Muller, T. & Dittrich, M. T. BioNet: an R-Package for the functional analysis of biological networks. *Bioinformatics* **26,** 1129–1130 (2010).

14. Nikpay, M. *et al.* A comprehensive 1000 Genomes–based genome-wide association meta-analysis of coronary artery disease. *Nat. Genet.* **47,** 1121–1130 (2015).

15. Yu, H. *et al.* A role for genetic susceptibility in sporadic focal segmental glomerulosclerosis. *J. Clin. Invest.* **126,** 1603–1603 (2016).

16. UNIVA. Univa Grid Engine. Available at: http://www.univa.com/resources/.

Chapter 5
Discussion

The main goal of this thesis was to develop a statistical method for interpretation of the genetic association data. First, we created a composite approach for rare variation tests that was used to extract a list of candidate genes from a modest-sized cohort of FSGS samples. Our statistical model predictions were validated in mouse knockout model that identified new susceptibility genes and proved genetic basis of sporadic FSGS. Second, on the larger scale case-control study our model identified a novel cutaneous melanoma gene – *EBF3*, through functional accreditation it was proven to have tumor suppressive properties. Finally, we built a software package for the interpretation of the genetic association studies results that finds differential network between cases and controls.

Summary of results

Composite model for rare DNA variation analysis

In this case-control study of sporadic and familial FSGS we used multiple gene-based association tests for analysis of rare coding DNA variants. While test statistics are related they have different power of capturing signal within specific effect size. Fisher's test is most suitable for ultra-low frequency high-impact protein-truncating variants, while VT and C-α tests are capturing cumulative effect of multiple moderate effect mutations that skew overall genotype distribution towards risk. Combination of all three models resulted in a short list of candidate genes – *WNK4, KAT2B, DLG5, ARGHEF17, KANK1* (**Table 2.2**)*.* All except *DLG5* were subjected for functional accreditation in the mouse model and found to cause proteinuria and development of FSGS phenotype (**Figure 2.5**).

Given the modest size of the case and control cohorts in this study our model was powerful enough to identify new susceptibility genes.

Large-scale analysis of cancer exome sequencing

We performed joint analysis of about 1,000 of early onset and about 1,500 unselected cancer cases with cutaneous and ocular melanoma, breast cancer and colon cancer versus about 8,000 matched controls. We identified common properties for the germline DNA variation in the list of known causal genes. Almost entirely statistical signal was driven by singleton protein-truncating variants in the genes tolerant to loss-of-function mutations that followed autosomal dominant inheritance pattern (**Figure 3.2**, **Table 3.1**, **Figure 3.3**). Analysis of pan-cancer dataset suggested that it is unlikely to find novel genes in this category with comparable effect size to the already known risk genes.

Thus, we concluded that there were two major directions of novel candidate genes search. First, identify genes that are intolerant of loss-of-function mutations but functionally are linked to the known cancer genes. We used earlier introduced composite model to analyze cutaneous melanoma cohort to identify novel susceptibility gene – *EBF3* (**Table 3.4**). Through functional accreditation overexpression of *EBF3* was found to correlate with suppression of *MITF* - a known cancer gene (**Figure 3.14**). We also sought to explain a known difference in survival rate among women and men with cutaneous melanoma. Analysis of somatic variation in metastatic tumors suggested, that a burden of the missense mutations in tumor have discriminative effect for gender disparity.

Second, we looked at another type of variation observed in blood DNA – somatic mutations. Mosaic mutations were known to cause leukemia and lymphoma. However, only several studies reported relevance to the solid-tumor cancers. We analyzed about 8,000 exomes from The Cancer Genome Atlas and identified significant burden of mosaic protein truncating variants in the leukemia genes (**Figure 3.8**). Interestingly, such mutations were absent in tumor (**Figure 3.10**). While, we only found statistical association, we were unable to conclude whether this is a predisposing factor or a consequence of the cancer phenotype.

Statistical framework for interpretation and improvement of association studies

We next used data-driven network inference algorithms to build a software module that identifies differential protein-protein interaction network most significantly perturbed in genetic association studies. Using generation of the random reference interactomes we assessed empirical significance of identified networks. With Bayesian statistics we designed a scheme for refinement of the association signal for individual genes using the evidence of inclusion to the best differential network. Our approach identified key functional differences between coronary artery disease patients with and without prior history of myocardial infarction (**Figure 4.2**). We also used data from analysis of FSGS cohort to predict which genes are most likely to be functionally relevant (**Figure 4.3**). Two of four genes validated with mouse model were predicted to be significant by our model.

Appendix

Below is a code for generation of single round of permutations for Bayesian analysis of the protein-protein network significance:

```
#!/bin/env Rscript

rm(list = ls(all = TRUE))

task.id <- Sys.getenv("SGE_TASK_ID")
#Run Parameters
n_perm<-1
args=commandArgs(trailingOnly=TRUE)
output_folder<-args[1]
input_file<-args[2]
GeneScoringMethod<-args[3]
prefix<-args[4]
reference_folder<-args[5]


substrRight<-function(x,n){
        substr(x, nchar(x)-n+1, nchar(x))
}
if (as.character(unlist(substrRight(output_folder,1)))=="/"){
        output_folder<-paste(output_folder,task.id,sep="")
}
system(paste("mkdir",output_folder,sep=" "))

library(BioNet)
library(igraph)


#Load references
network <- loadNetwork.tab(paste(reference_folder,"inweb_im_ppi.txt",sep=""),
header=FALSE,format="graphNEL", directed=FALSE)
network2 <- rmSelfLoops(network)
connectivity<-
read.table(paste(reference_folder,"inweb_im_gns_connections.txt",sep=""),
header=T,sep="\t")

#prepare inweb network
nodes_inWeb<-as.character(unlist(nodes(network)))
nodes_inWeb<-cbind(nodes_inWeb,nodes_inWeb)
colnames(nodes_inWeb)<-c("nodes_inWeb","V2")

connectivity_inWeb<-merge(nodes_inWeb,connectivity,by.x="nodes_inWeb",by.y="gns")
connectivity_inWeb<-connectivity_inWeb[,-2]
connectivity_rank<-rank(connectivity_inWeb[,2])
connectivity_rank<-.bincode(connectivity_rank,breaks=seq(0,17590,10))
connectivity_inWeb<-cbind(connectivity_inWeb,connectivity_rank)
connectivity_inWeb<-
connectivity_inWeb[match(nodes_inWeb[,1],connectivity_inWeb$nodes_inWeb),]
connectivity_inWeb<-cbind(connectivity_inWeb,connectivity_inWeb[,1])
colnames(connectivity_inWeb)[[1]]<-"Permuted_Nodes"
colnames(connectivity_inWeb)[[4]]<-"nodes_inWeb"

### Permute labels
permuteLabel<-function(x){
x[,1]<-sample(x[,1])
return(x)
}


#Dataset preparation
```

```
tt<-read.table(input_file)
colnames(tt)<-c("InWebID","pvals")
tt<-merge(tt,connectivity,by.x="InWebID",by.y="gns")
tt<-tt[!duplicated(tt),]
connectivity_rank<-rank(-as.numeric(as.character(unlist(tt$connections))))
tt<-cbind(tt,connectivity_rank)
#####define funcitons
#Gene scoring function:
scoreGenes<-function(x,method,FDR,GWSthreshold){
if (method=="B"){
        if(missing(GWSthreshold)){
                GWSthreshold<-0.05/length(nodes(network2))
        }
        maxSign<-max(x[which(x<GWSthreshold)],na.rm=T)
        x[which(x<GWSthreshold)]<-maxSign
        if(length(x)<length(nodes(network2))){
                tmpPvals<-c(x,runif(length(nodes(network2))-length(x)))
        }
        fb<-fitBumModel(tmpPvals,plot=FALSE)
        scoreEstimate<-function(y){
                return((fb$a-1)*(log(y,base=10)-log(GWSthreshold,base=10)))
        }
        scores<-mapply(scoreEstimate,y=x)

        return(scores)
}
if (method=="FDR"){
        if (missing(FDR)){
                FDR<-0.05
        }
        fb<-fitBumModel(x,plot=FALSE)
        subnet<-subNetwork(names(x),network2)
        scores <- scoreNodes(subnet, fb, fdr = FDR)
        return(scores)
}

}
#
#Network Permutations
NetworkPEstimate<-function(tt,network_interactome,connectivity_interactome,n_perm){
sim_scores<-NULL
nNodes<-NULL
nEdges<-NULL
EdgesPerNode<-NULL
number_of_gws_nodes<-NULL
allNodes<-NULL

        scatterRanks<-
split(connectivity_interactome,connectivity_interactome$connectivity_rank)
        scatterRanks<-lapply(scatterRanks,permuteLabel)
        connectivity_rank<-rank(connectivity_interactome[,2])
        gatherRanks<-gatherRanks<-do.call(rbind,scatterRanks)
        gatherRanks<-gatherRanks[match(nodes_inWeb[,1],gatherRanks$nodes_inWeb),]
        gatherRanks<-
gatherRanks[which(is.na(as.character(unlist(gatherRanks$Permuted_Nodes)))==FALSE),]
        gatherRanks<-gatherRanks[match(nodes_inWeb[,1],gatherRanks$nodes_inWeb),]


        nodes(network_interactome)<-as.character(unlist(gatherRanks$Permuted_Nodes))
        pvals<-tt$pvals
        names(pvals)<-tt$InWebID
        subnet<-subNetwork(names(pvals),network_interactome)
        scores<-scoreGenes(pvals,GeneScoringMethod)
        module <- runFastHeinz(subnet, scores)
        node_list<-unique(c(c(getEdgeList(module)[,1]),getEdgeList(module)[,2]))
        sim_scores<-
c(sim_scores,sum(scores[which(is.element(names(scores),node_list)==TRUE)]))
        gws_nodes<-subset(tt,pvals<0.05/length(nodes(network2)))
        number_of_gws_nodes<-
c(number_of_gws_nodes,length(which(is.element(nodes(module),gws_nodes[,1])==TRUE)))
        nNodes<-c(nNodes,length(nodes(module)))
        nEdges<-c(nEdges,length(unlist(edgeL(module)))/2)
        EdgesPerNode<-
c(EdgesPerNode,(length(unlist(edgeL(module)))/2)/length(nodes(module)))
        allNodes<-c(allNodes,nodes(module))

output<-cbind(sim_scores,number_of_gws_nodes,nNodes,nEdges,EdgesPerNode)
colnames(output)<-
c("Score","Number_of_gws_nodes","Number_of_Nodes","Number_of_Edges","Number_of_edges_p
er_node")
return(list(output,allNodes))
}
#
##############
#P(node in network) estimation
NodePEstimate<-function(tt,network_interactome,n_perm){
        allNodes<-NULL
        pvals<-tt$pvals
        names(pvals)<-sample(tt$InWebID)
        subnet<-subNetwork(names(pvals),network_interactome)
        scores<-scoreGenes(pvals,GeneScoringMethod)
        module<-runFastHeinz(subnet,scores)
        allNodes<-c(allNodes,nodes(module))

return(allNodes)
}
####

perm_scores<-replicate(n_perm,NetworkPEstimate(tt,network,connectivity_inWeb,1))
assembly_scores<-perm_scores[[1]]
assembly_nodes<-as.character(unlist(perm_scores[[2]]))
perm_node_list<-replicate(n_perm,NodePEstimate(tt,network,1))
perm_node_list<-as.character(unlist(perm_node_list))
#Save the scores of randomly drawn networks (connectivity-conserved permutations)
write.table(assembly_scores,paste(output_folder,"permuted_inweb_network_score.txt",sep
="/"),row.names=F,quote=F,sep="\t")
#Save the nodes that appeared in the randomly drawn networks (connectivity-conserved
permutations)
write.table(assembly_nodes,paste(output_folder,"permuted_inweb_network_nodes.txt",sep=
"/"),row.names=F,quote=F,col.names=F)
#Save the nodes that appeared in the randomly drawn networks (only p-values are
permuted)
write.table(perm_node_list,paste(output_folder,"permuted_pvals_network_nodes.txt",sep=
"/"),row.names=F,quote=F,col.names=F)
```

Results of multiple permutations are then assembled and Bayesian model is used to estimate posterior significance for genetic association signal:

```
#!/bin/env Rscript

rm(list = ls(all = TRUE))



library(BioNet)
library(igraph)
args=commandArgs(trailingOnly=TRUE)
input_file<-args[1]
GeneScoringMethod<-args[2]
prefix<-args[3]
reference_folder<-args[4]


substrRight<-function(x,n){
       substr(x, nchar(x)-n+1, nchar(x))
}


#Load references
network <- loadNetwork.tab(paste(reference_folder,"inweb_im_ppi.txt",sep=""),
header=FALSE,format="graphNEL", directed=FALSE)
network2 <- rmSelfLoops(network)
connectivity<-
read.table(paste(reference_folder,"inweb_im_gns_connections.txt",sep=""),
header=T,sep="\t")

#prepare inweb network
connectivity<-connectivity[!duplicated(connectivity[,1]),]
nodes_inWeb<-as.character(unlist(nodes(network2)))
nodes_inWeb<-cbind(nodes_inWeb,nodes_inWeb)
colnames(nodes_inWeb)<-c("nodes_inWeb","V2")

connectivity_inWeb<-merge(nodes_inWeb,connectivity,by.x="nodes_inWeb",by.y="gns")
connectivity_inWeb<-connectivity_inWeb[,-2]
connectivity_rank<-rank(connectivity_inWeb[,2])
connectivity_inWeb<-cbind(connectivity_inWeb,connectivity_rank)
connectivity_inWeb<-
connectivity_inWeb[match(nodes_inWeb[,1],connectivity_inWeb$nodes_inWeb),]
connectivity_inWeb<-cbind(connectivity_inWeb,connectivity_inWeb[,1])
colnames(connectivity_inWeb)[[1]]<-"Permuted_Nodes"
colnames(connectivity_inWeb)[[4]]<-"nodes_inWeb"


#Dataset preparation
tt<-read.table(input_file)
colnames(tt)<-c("InWebID","pvals")
tt<-merge(tt,connectivity,by.x="InWebID",by.y="gns")
tt<-tt[!duplicated(tt),]
connectivity_rank<-rank(-as.numeric(as.character(unlist(tt$connections))))
tt<-cbind(tt,connectivity_rank)

#####define funcitons
#Gene scoring function:
scoreGenes<-function(x,method,FDR,GWSthreshold){
if (method=="B"){
       if(missing(GWSthreshold)){
              GWSthreshold<-0.05/length(nodes(network2))
```

```
        }
        maxSign<-max(x[which(x<GWSthreshold)],na.rm=T)
        x[which(x<GWSthreshold)]<-maxSign
        if(length(x)<length(nodes(network2))){
                tmpPvals<-c(x,runif(length(nodes(network2))-length(x)))
        }
        fb<-fitBumModel(tmpPvals,plot=FALSE)
        scoreEstimate<-function(y){
                return((fb$a-1)*(log(y,base=10)-log(GWSthreshold,base=10)))
        }
        scores<-mapply(scoreEstimate,y=x)

        return(scores)
}
if (method=="FDR"){
        if (missing(FDR)){
                FDR<-0.05
        }
        fb<-fitBumModel(x,plot=FALSE)
        subnet<-subNetwork(names(x),network2)
        scores <- scoreNodes(subnet, fb, fdr = FDR)
        return(scores)
}

}

pvals<-tt$pvals
names(pvals)<-tt$InWebID
scores<-scoreGenes(pvals,GeneScoringMethod)

subnet<-subNetwork(names(pvals),network2)
subnet2 <- rmSelfLoops(subnet)

module <- runFastHeinz(subnet2, scores)

node_list<-nodes(module)
target_score<-sum(scores[which(is.element(names(scores),node_list)==TRUE)])
target_score<-
c(target_score,length(which(is.element(nodes(module),subset(tt,pvals<0.05/length(nodes
(network2)))[,1])==TRUE)),length(nodes(module)),length(unlist(edgeL(module)))/2,(lengt
h(unlist(edgeL(module)))/2)/length(nodes(module)))
names(target_score)<-
c("Score","Number_of_gws_nodes","Number_of_Nodes","Number_of_Edges","Number_of_edges_p
er_node")
assembly_scores<-read.table("permuted_networks_scores.txt",header=T,sep="\t")
assembly_nodes<-read.table("permuted_networks_nodes.txt",header=F)
assembly_nodes<-as.character(unlist(assembly_nodes))

perm_node_list<-read.table("permuted_pvals_nodes.txt",header=F)
perm_node_list<-as.character(unlist(perm_node_list))

output<-subset(tt,is.element(InWebID,node_list)==TRUE)
refined_pvals<-NULL

for(i in 1:length(output[,1])){
        worseGenes<-as.character(unlist(subset(tt,pvals<=output[i,2],select=InWebID)))
        rankSet<-
as.character(unlist(subset(tt,connectivity_rank==output[i,4],select=InWebID)))
        rankSet<-perm_node_list[which(is.element(perm_node_list,rankSet)==TRUE)]
        refined_pvals<-
c(refined_pvals,length(which(is.element(worseGenes,node_list)==TRUE))*output[i,2]/leng
th(which(is.element(perm_node_list,rankSet)==TRUE)))
}
output<-cbind(output,refined_pvals)
```

```
write.table(output,paste(prefix,"posterior_pvals.txt",sep="_"),row.names=F,sep="\t",qu
ote=F)
Pval_Score<-
length(which(assembly_scores[,1]>=target_score[1]))/length(assembly_scores[,1])
Pval_Connectivity<-
length(which(assembly_scores[,5]>=target_score[5]))/length(assembly_scores[,1])
target_score<-c(target_score,Pval_Score,Pval_Connectivity)
names(target_score)[c(6,7)]<-c("Pval_Score","Pval_Connectivity")
write.table(target_score,paste(prefix,"network_analysis.txt",sep="_"),col.names=F,sep=
"\t",quote=F)
system("rm permuted_networks_scores.txt permuted_pvals_nodes.txt
permuted_networks_nodes.txt")
module<-subNetwork(as.character(unlist(output[,1])),network2)
gws<-0.05/length(nodes(network2))
fchange<-NULL
for (i in 1:length(output[,1])){
        idx<-which(output[,1]==nodes(module)[1])
        if (output[idx,2]>gws & output[idx,5]<=gws){
                fchange<-c(fchange,0)
        }
        if (output[idx,2]<gws){
                fchange<-c(fchange,10)
        }
        if (output[idx,5]>gws){
                fchange<-c(fchange,-10)
        }
}
names(fchange)<-nodes(module)
pdf(paste(prefix,"network.pdf",sep="_"))
plotModule(module,diff.expr=fchange)
legend("topleft",c("Significant","Posteriory\nSignificant","Not
Significant"),fill=c("red","white","blue"),bty="n",pt.cex=1,cex=0.8)
dev.off()
```