University of Colorado
Anschutz Medical Campus

Exploring the role of gene interaction in idiopathic pulmonary fibrosis with exome sequencing

Adam J Richards[1], Tasha Fingerlin[2], Ivana V. Yang[1], James E. Loyd[3], Debbie A. Nickerson[4], Elizabeth Blue[4], Karynne Patterson[4], and David A. Schwartz[1]

[1]University of Colorado Denver, CO, USA
[2]Colorado School of Public Health, CO, USA
[3]Vanderbilt University, TN, USA
[4]University of Washington, WA, USA

Idiopathic pulmonary fibrosis (IPF)

IPF is characterized by thickening and scarring lung tissue and it is a fatal disease with no cure. A number of genetic loci have been associated with IPF and a major focus has been the discovery of novel markers associated with disease outcome. Despite a number of promising findings, like a MUC5B promoter variant, the known genetic underpinnings do not adequately explain disease risk. In this work we continue this search using exome sequencing data, but we further investigate the role of gene-gene interaction among specific genetic variants.

Background

- Cases where the cause is unknown are called idiopathic
- Survival is 3-5 years after diagnosis
- When more than one family member is affected it is called familial pulmonary fibrosis
- Mucin genes are associated with disease (i.e. MUC5B) [4]
- *PARN* and *RTEL1* have been associated [5]

Approach

- 286 cases
- variant association based on frequency differences
- variant interaction inferred by an iterative support vector machines approach

Variant calling

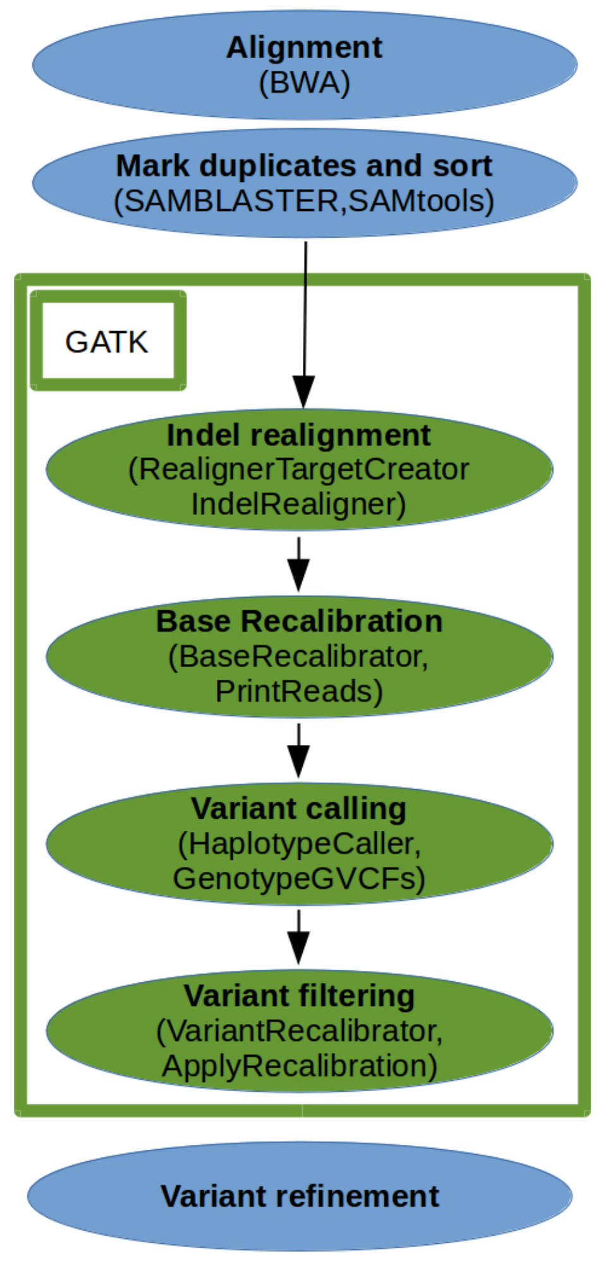

Figure 1: Variant calling pipeline based on BWA [2] and GATK [3]

Variant association

For each variant we subtract the case major allele frequency from the G1K reference population (EUR) to obtain frequency distributions.

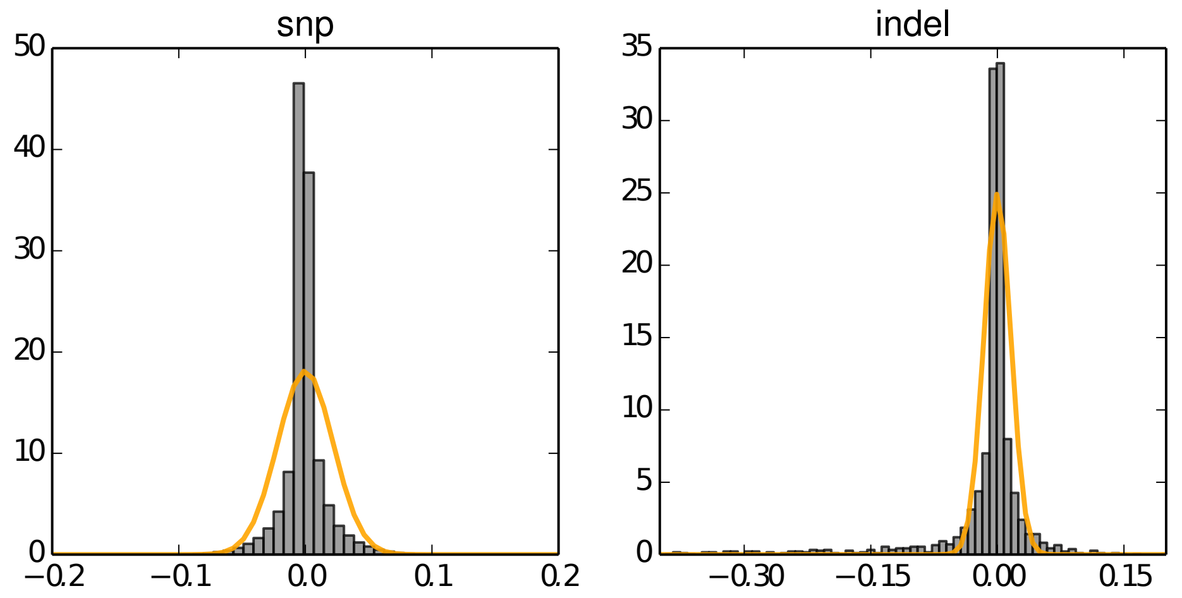

Figure 1: Frequency differences between called variants in the case cohort and variants in the G1K reference data.

The null hypothesis is that there is no difference in major allele frequency between the cases and controls. the frequency difference distributions were fit with Gaussians. The skewed indel distribution was fit by ignoring the tails. To test the significance of the ith frequency difference f we simply use a two-sided Z-test.

$$z_i = \frac{f_i}{\sigma} \qquad (1)$$

p-values were then obtained according to $2\Phi(-|z_i|)$, where Φ is the normal CDF. We controlled for multiple testing using the Benjamini and Hochberg FDR method [1].

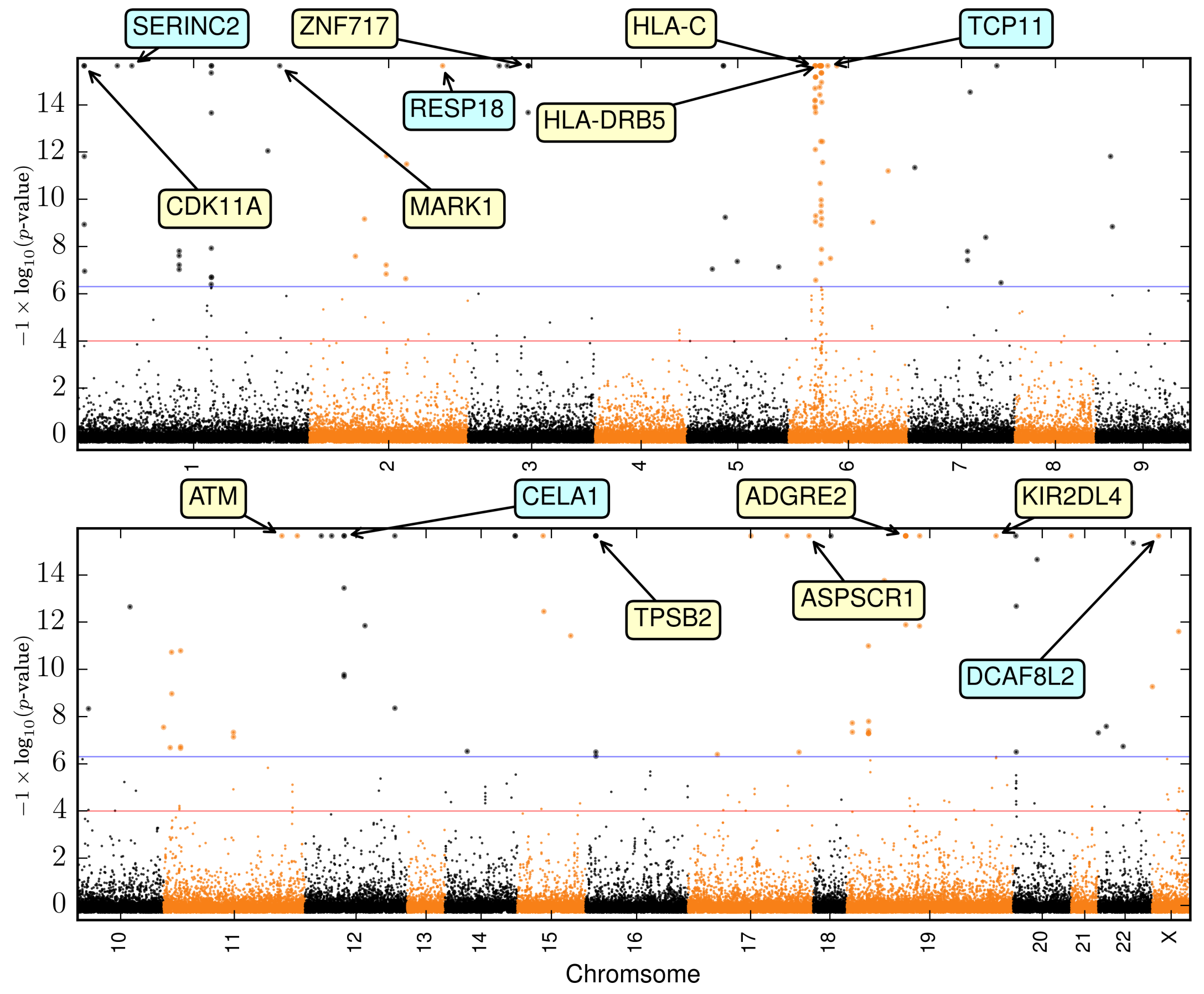

Figure 3: Manhattan plot of the p-values based on frequency differences. The top 15 variants are shown with the light oranges and blues indicating SNP and indels respectively. The two drawn thresholds correspond to 0.001 and $5 * 10e - 08$.

Variants from the HLA region were were highly associated with disease along with several Mucin genes (MUC2, MUC6, and MUC16)

Gene-Gene interaction

variant	0.001	$5 * 10e - 08$	B-H adjusted
SNP	413	142	274
indel	59	27	63

Table 1: Summary of candidate variants.

Imputed genotypes for both the cases and controls were transformed into $\{0, 1, 2\}$ space which represents reference, heterozygous-alternative and homozygous-alternative respectively.

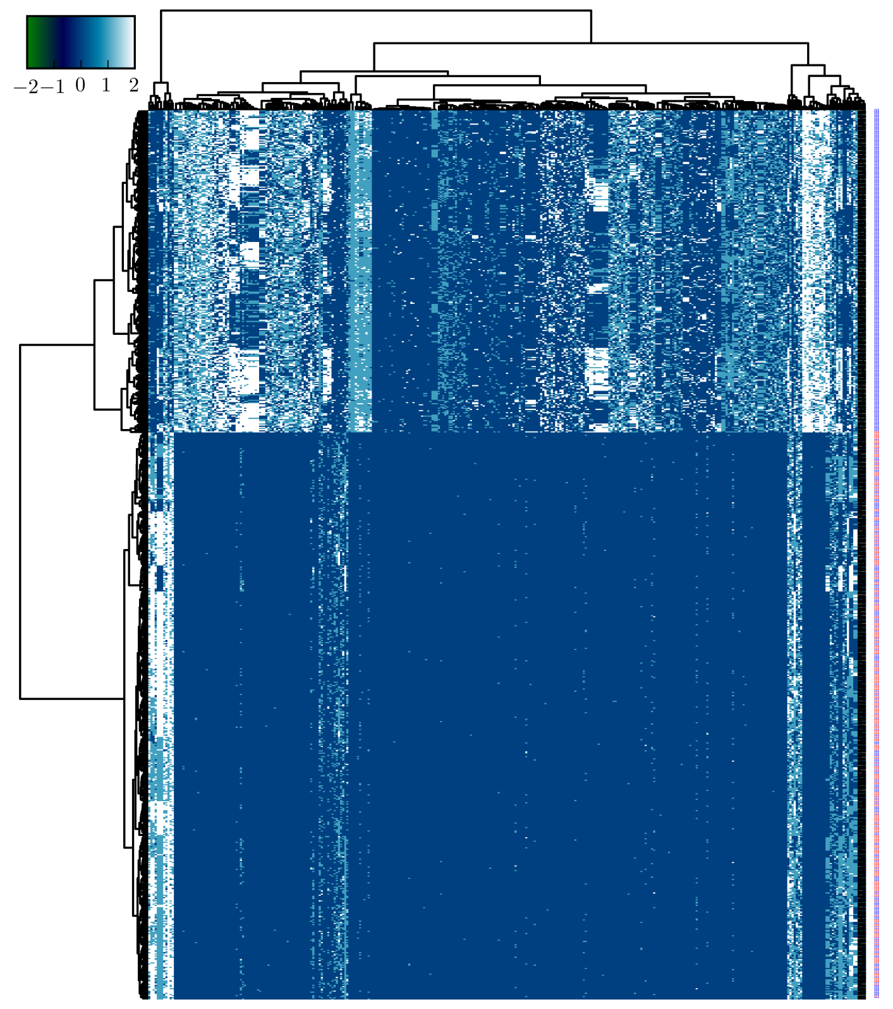

Figure 4: Imputed genotypes were discretized for input into machine learning algorithms. Here the values are shown with a clustering at the sample level (rows) and variant level (columns) The cases and reference population are readily distinguishable.

Of the 337 variants with significant p-values after adjustment we only considered the 279 with minor allele frequencies $\geq 5\%$.

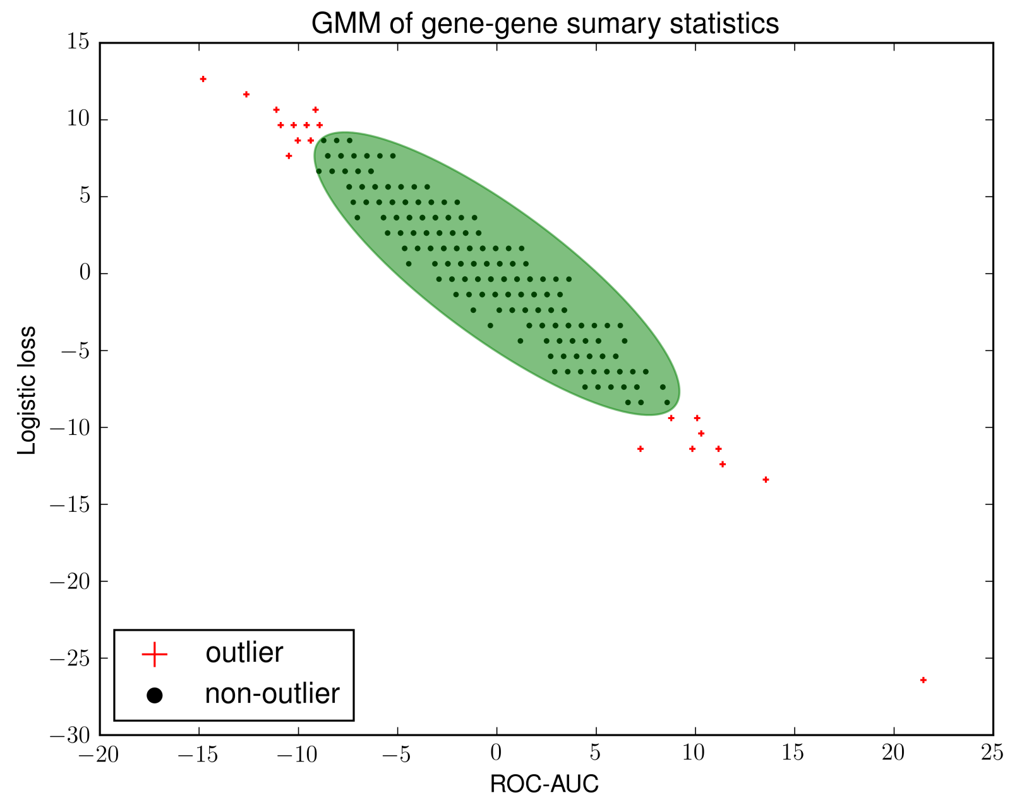

Figure 5: After running all pairwise combinations of variant-variant interactions summary statistics of predictive performance were plotted. The two statistics: the area under the ROC curve and a logistic loss function are shown in standardized space. A Gaussian mixture model (GMM) has been overlaid to determine outliers.

Variant1	Variant2	Gene 1	Gene 2	p-value	p-value 2
rs2074225	rs4245191	CD6	NLRX1	1.21e-05	7.72e-06
rs9284879	rs6120033	TOPAZ1	EFCAB8	5.96e-05	4.75e-05
rs3814355	rs12976493	CCNT2	ADGRE2	1.46e-07	2.22e-16
snv3311	rs1056286	CELA1	IL17RE	2.22e-16	0.0001
rs3814355	snv469	CCNT2	LIMD1	1.46e-07	2.22e-16
snv469	rs1136905	LIMD1	HLA-DRB5	2.22e-16	2.22e-16
rs11638215	rs1136905	SECISBP2L	HLA-DRB5	8.15e-05	2.22e-16
rs138579161	snv2825	RESP18	BPIFC	2.22e-16	1.84e-07
rs3814355	rs1136905	CCNT2	HLA-DRB5	1.46e-07	2.22e-16
rs147889095	rs1136905	ITPKB	HLA-DRB5	1.25e-06	2.22e-16
snv1454	rs10418767	MUC6	ADGRE2	2.07e-07	2.22e-16
rs1129152	snv469	INTS8	LIMD1	0.0001	2.22e-16
snv3311	rs9284879	CELA1	TOPAZ1	2.22e-16	5.98e-05
snv3311	rs11638215	CELA1	SECISBP2L	2.22e-16	8.15e-05
rs2308628	rs12976493	HLA-C	ADGRE2	2.22e-16	2.22e-16
rs3814355	rs11085765	CCNT2	MUC16	1.46e-07	1.73e-14
rs9284879	snv1454	TOPAZ1	MUC6	5.98e-05	2.07e-07
snv2825	rs192690014	BPIFC	HRNR	1.84e-07	2.22e-16
snv1454	snv2825	MUC6	BPIFC	2.07e-07	1.84e-07
rs9284879	rs1142888	TOPAZ1	GBP4	5.98e-05	6.04e-08
snv3311	rs11085765	CELA1	MUC16	2.22e-16	1.74e-14
rs12976493	snv469	ADGRE2	LIMD1	2.22e-16	2.22e-16
rs1056286	rs10418767	IL17RE	ADGRE2	0.0001	2.22e-16
rs72268642	rs6120033	CNTNAP2	EFCAB8	3.42e-07	4.75e-05
snv2825	rs3814355	BPIFC	CCNT2	1.84e-07	1.46e-07
rs192690014	rs3208105	HRNR	HLA-DQA1	2.22e-16	2.22e-16

Table 2: Gene-gene interactions prioritized by GMM $\log(p\text{−value})$

Discussion

The gene-gene interactions are based on a proxy of classification potential or more specifically how much that classification potential is perturbed when the pair of variants are removed from the analysis. Both the significant genes from these analyses as well as the significant interactions appear to be relevant to disease. Because these are preliminary results, both corroboration with the literature as well as independent experiments will help shed light on the reliability of these predictions. It is also important that these methods be carefully contrasted with count-based methods for determining individual variant significance.

References

[1] Y. Benjamini and Y. Hochberg, Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing Journal of the Royal Statistical Society. Series B (Methodological), Blackwell Publishing for the Royal Statistical Society, 57, 289-300, 1995

[2] H. Li and R Durbin, Fast and accurate short read alignment with Burrows-Wheeler transform Bioinformatics, 25, 1754-60, 2009

[3] A. McKenna, M. Hanna *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome research, 20, 1297-303, 2010

[4] M. Seibold, A. Wise *et al.* A common MUC5B promoter polymorphism and pulmonary fibrosis N Engl J Med., 364, 1503-12, 2011

[5] B. D. Stuart, J Choi *et al.* Exome sequencing links mutations in *PARN* and *RTEL1* with familial pulmonary fibrosis and telomere shortening Nature genetics, 47, 512-7, 2015

Acknowledgements

Support for this work is provided by the NIH (R01 HL097163 and P01 HL092870) and the VA Merit Review Program (1I01BX001534). Also, this work would not have been possible without the valuable contributions of several collaborating clinicians. The opinions, findings and recommendations expressed in this work are those of the authors and do not necessarily reflect the views of the UC Denver or other affiliates.