Distinguish between exomes of Crohn's disease patients and healthy individuals

Lipika R. Pal [1], Kunal Kundu [1, 2], Yizhou Yin [1, 2], John Moult [1, 3]*

[1] Institute for Bioscience and Biotechnology Research, University of Maryland, 9600 Gudelsky Drive, Rockville, MD 20850, [2] Computational Biology, Bioinformatics and Genomics, Biological Sciences Graduate Program, University of Maryland, College Park, MD 20742, USA, [3] Department of Cell Biology and Molecular Genetics, University of Maryland, College Park, MD 20742.

*Corresponding author
jmoult@umd.edu
Phone: (240) 314-6241
FAX: (240) 314-6255

All our Crohn's challenge predictions are based on the GWAS marker SNP genotypes present in each individual. One submission also includes a contribution from the rare missense variants present in each individual. Marker SNPs and the corresponding odds ratios were extracted from the GWAS catalog. For most GWAS loci, the markers do not fall within the exome sequences, and for these, we imputed marker genotypes using PLINK [1], ShapeIT [2] and IMPUTE2 [3], with 1000genome phase3 data as the reference data. Two sets of GWAS loci were used: a set of 90 we had previously compiled [4, 5] and a set of 138 loci from [6]. The exome sequence data were annotated using VARANT [7].

Our methods fall into three groups:

(a) Estimation of the relative risk of each individual being diagnosed with Crohn's based on Naïve Bayes models, using odds ratio information, marker genotypes, and genotype frequencies derived assuming Hardy Weinberg equilibrium. Two different methods were used. In one, ('odds ratio'), relative risks were approximated by the appropriate genotype odds ratio. In the other ('conditional probability') the relative probabilities of the appropriate genotype occurring in case versus control were used.

(b) Estimation of relative risk using standard machine learning, with marker genotypes as input data. We evaluated Naïve Bayes, Logistic regression, neural nets, and random forest methods using standard settings in WEKA [8]. WTCCC Crohn's microarray data were used for training, imputing marker genotypes where necessary as described above. Benchmarking was performed on WTCCC data not used for training, and with the CAGI 2013 Crohn's data. The neural net method performed badly and was not used. Results from the other methods were similar, with logistic regression apparently slightly more accurate.

(c) 'odds ratio' Naïve Bayes as in (a), with additional contributions from rare missense variants predicted to have a high impact on in vivo protein function. For this purpose we used a previously compiled set of mechanism genes for the set of 90 GWAS loci [4]. A missense variant in any of these genes was considered high impact if at least two of five methods assigned a pathogenic score. The methods are SNPs3D profile [9], SNPs3d stability [10], Polyphen [11], SIFT [12] and CADD [13]. For genes where we had previously assigned a high impact missense mechanism as underlying the association with disease [4], we used the same odds ratio contribution to Naïve Bayes as for the marker term. For other genes, it is unknown whether the mechanism underlying the disease association increases or decreases protein activity. We therefore used a weighted sum of the odds ratios implied by the two possibilities.

Submissions:

(1) Consensus prediction, averaging the predicted probabilities from the odds ratio based probability Naïve Bayes, Logistic regression, and Random Forest (1000 trees) for 138 loci.

(2) Odds ratio based Naïve Bayes, including rare variant contributions for 90 loci.

(3) Conditional probability based Naïve Bayes for 138 loci

(4) Logistic regression calculated using WEKA for 90 loci.

(5) Same Logistic regression as in (4) for 138 loci.

(6) Same consensus prediction as in (1) for 90 loci.

We had no data to base the standard deviations on, so arbitrarily assigned all of these as 0.2

Prediction of Early Age of Onset:

We checked presence of rare mutations in IL10RA gene as related to early onset of disease as mentioned in [14]. We gave arbitrary different standard deviations as we are not very certain for most of the predictions.

References:

1. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 2007;81:559–75. doi:10.1086/519795.

2. Delaneau O, Marchini J, Zagury J-F. A linear complexity phasing method for thousands of genomes. Nat Methods 2012;9:179–81. doi:10.1038/nmeth.1785.

3. Howie BN, Donnelly P, Marchini J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet 2009;5:e1000529. doi:10.1371/journal.pgen.1000529.

4. Pal, Lipika R., and Moult, J. (2015) Genetic basis of common human disease: Insight into the role of Missense SNPs from Genome Wide Association Studies. J. Mol. Biol., 427, 2271-2289. [http://dx.doi.org/10.1016/j.jmb.2015.04.014].

5. GWAS Catalog: http://www.genome.gov/gwastudies (accessed September, 2013).

6. Jostins L, Ripke S, Weersma RK, Duerr RH, McGovern DP, Hui KY, et al. Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. Nature 2012;491:119–24. doi:10.1038/nature11582.

7. VARANT: http://compbio.berkeley.edu/proj/varant/Home.html

8. Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, Ian H. Witten. The WEKA Data Mining Software: An Update; SIGKDD Explorations. 2009; 11(1).

9. Yue P, Melamud E, Moult J. SNPs3D: candidate gene and SNP selection for association studies. BMC Bioinformatics. 2006; 7:166

10. Yue P, Li Z, Moult J. Loss of protein structure stability as a major causative factor in monogenic disease. J Mol Biol 2005;353:459–73. doi:10.1016/j.jmb.2005.08.020.

11. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. A method and server for predicting damaging missense mutations. Nature Methods. 2010; 7:248–249.

12. Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat Protoc. 2009; 4(7):1073-81.

13. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet. 2014; 46:310-315.

14. Kotlarz D, Beier R, et al. Loss of Interleukin-10 signalling and infantile inflammatory bowel disease: implications for diagnosis and therapy. Gastroenterology 2012; 143:347-355.