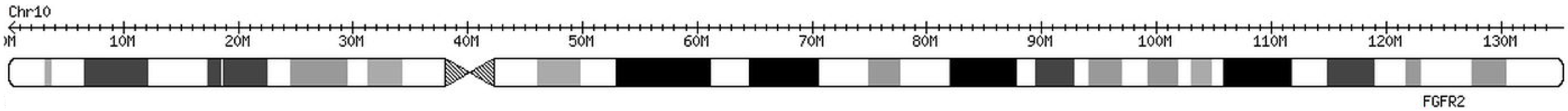
Beyond association: Bayesian analysis of a large-scale genome-wide association study using CAusal Variants Identification in Associated Regions (CAVIAR)

Jessica Shaw, MS Biostatistics student



MOTIVATION

Genome-wide association studies (GWAS) have provided a window into how small variations in the human genome impact risk for breast cancer. These studies are limited in their clinical value, however, by their inability to identify causal variants (Kisios and Zintzaras, 2009). Modest gains from GWAS come at a high expense, as do the laboratory techniques used to assess the functional implications of SNPs (Kisios and Zintzaras, 2009). Post-hoc analytical techniques offer one means of improving the ability of GWAS to identify causal genetic variants contributing to breast cancer risk.

OPPORTUNITY

Causal Variants Identification in Associated Regions (CAVIAR), developed by Farhad Hormozdiari et al. at the University of California, Los Angeles, is one analytical method with the potential to identify causal variants from GWAS before incurring the additional expense of laboratory-based investigation.

Whereas most statistical methods assume that each risk locus or gene contains only one risk modifying variant, CAVIAR allows for the possibility of many variants in the same gene, as seen in the BRCA2 gene. Using information about the correlation between SNPs from publicly available databases as a prior distribution, CAVIAR calculates a set of variants likely to contain a pre-determined percentage of the SNPs associated with a given phenotype. Set at 95%, the subset of SNPs is comparable to a confidence interval and can be used to better predict whether a variant or set of variants increases risk. Using data from a 2010 GWAS of 582,886 SNPs, I will use CAVIAR to generate a 95% causal set of SNPs in chromosome 2 (Turnbull et al., 2010).

OBJECTIVES

- Identify a causal set of SNPs that contains 95% of the causal SNPs in the region defined by the FGFR2 gene.
- Compare and contrast the inferences that can be drawn from GWAS under frequentist and Bayesian assumptions.

METHODS

1. Identify a public source of GWAS data.
2. Review published literature pertaining to the data available.
3. Identify a region of interest.
4. Isolate the region of interest within the GWAS data set, taking into account any changes to the reference genome since the time of the initial publication.
5. Extract the region data.
6. Clean the data to remove extraneous variables.
7. Calculate z-scores from the data set's reported -log base 10 p-values.
8. Transform data into Plink file formats.
9. Calculate linkage disequilibrium between all pairs of SNPs in the selected region.
10. Download C++ code from genetics.cs.ucla.edu/caviar.
11. Install all dependencies, including the GNU Scientific Library.
12. Ensure access to administrator privileges to access developer tools and programs.
13. Compile C++ code to make CAVIAR.cpp executable.
14. Execute CAVIAR, including your z statistic and linkage disequilibrium files in the arguments.
15. Compare causal set produced by CAVIAR to the SNPs reported as significant in the original publication.

THE PRESENT ANALYSIS

Nat Genet. 2007 Jul;39(7):870-4. Epub 2007 May 27.

A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer.

Hunter DJ[1], Kraft P, Jacobs KB, Cox DG, Yeager M, Hankinson SE, Wacholder S, Wang Z, Welch R, Hutchinson A, Wang J, Yu K, Chatterjee N, Orr N, Willett WC, Colditz GA, Ziegler RG, Berg CD, Buys SS, McCarty CA, Feigelson HS, Calle EE, Thun MJ, Hayes RB, Tucker M, Gerhard DS, Fraumeni JF Jr, Hoover RN, Thomas G, Chanock SJ.

Data source: GWAS Central

Publication:

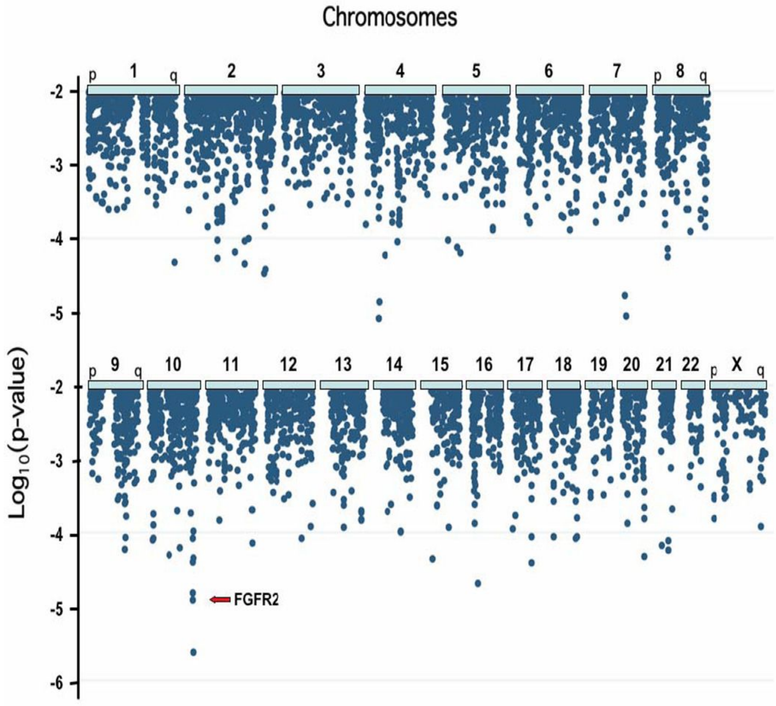
Hunter, et al., 2007. A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer.

Region:

FGFR2

Software:

R, Plink, CAVIAR, Mac Terminal via Homebrew and Xtools, Linux



Signal of association in the FGFR2 region (Hunter et al., 2007)

RESULTS

This analysis produced a set of SNPs likely to include 95% of the causal SNPs in the region.

LIMITATIONS

- Limited and inaccurate documentation
- High number of dependencies and permissions required