THE GENOMICS OF ORAL POLIOVIRUS VACCINE RESPONSE IN BANGLADESHI INFANTS

by

Genevieve L. Wojcik, MHS

A dissertation submitted to the Johns Hopkins University in conformity with the requirements for the degree of Doctor of Philosophy

Baltimore, Maryland, USA

October 2013

Abstract

The success of Oral Poliovirus Vaccine (OPV) in eradicating poliovirus has set an example for the immense potential of oral vaccines in preventing enteric infections. It is widely considered the standard for oral vaccines aiming to elicit a mucosal immune response. Despite being validated in diverse populations worldwide, there still remain some individuals that fail to mount an adequate response to vaccination with OPV. It has been hypothesized that this may be due to host genetics, as the heritability is estimated to be high (60%) and there have been ethnic differences in response. To address this question we conducted a genome-wide association study (GWAS) in 357 Bangladeshi children comparing individuals that fail to mount an immune response to high responders of OPV. Four different approaches were conducted to elucidate genetic risk loci: (1) a traditional GWAS analysis, (2) a correlation of the GWAS results with signatures of positive selection, (3) an application of gene-level methods to the GWAS results, and (4) an application of pathway-level methods to the GWAS results. Because there is no consensus as to the best gene- and pathway-level methods, a simulation experiment was conducted to systematically evaluate their relative performance.

The traditional GWAS assessed the association of 6.6 million single nucleotide polymorphisms (SNPs) across the human genome, adjusted for stunting (height-for-age Z-score (HAZ) < -2). While there were not any genome-wide significant results ($P<5\text{x}10^{-8}$), several suggestive associations were found on chromosomes 7 and 14 ($P<5\text{x}10^{-6}$). On chromosome 7, the top association was found at rs55906254 (OR=0.31, $P=3.5\text{x}10^{-6}$). Found

upstream of *SHH* (sonic hedgehog), the minor allele of this SNP conferred decreased odds of high seropositive status versus seronegative. On chromosome 14, the top association was downstream of *MAPK1IP1L* (mitogen-activated protein kinase 1 interacting protein 1-like) at rs113427985 (OR=0.22, P=2.9x10^{-6}). To measure regions under positive selection, the cross-population extended haplotype homozygosity (XP-EHH) was calculated. To correlate these with the GWAS results, a filter was used in which SNPs had to have a P-value from the GWAS less than 0.001 and a P-value from the selection scan below 0.01. A total of 32 SNPs reached this threshold, half of which were between *FAM86A* (family with sequence similarity 86, member A) and *RBFOX1* (RNA-binding protein, fox-1 homolog). The non-ancestral alleles of these SNPs were associated with high seropositive status. Therefore, it is likely that mutations arose in this region that were beneficial to either OPV immunity or another ancestral pathogen and were preserved.

Before the gene- and pathway-level methods were applied to the OPV GWAS, a simulation experiment was conducted to determine which methods were the best. These methods were developed to aggregate signals from the GWAS into gene- and pathway-level units, increasing the power to detect associations and offering biological interpretation. Using genotypic data from the Wellcome Trust Case Control Consortium (WTCCC), a phenotype was simulated assuming an additive polygenic model. A total of 12 gene-level methods and 10 pathway-level methods were systematically evaluated. The gene-level method with the best balance of sensitivity and specificity was VEGAS

using only the top 10% of the associated SNPs within the gene. MAGENTA and GSA-SNP had the best performance of all the pathway-level methods. These methods were then applied to the GWAS of OPV. The gene-level results highlighted the potential role of histone modifications as the top results included many histone marks within histone cluster 1 on chromosome 6. Pathway-level methods using the Gene Ontology Biological Processes showed enrichment in gene sets related to cyclic AMP as a second messenger and its relationship with G-protein signaling. Additional associations were found in neurological development.

Taken together, this dissertation seeks to elucidate the host genomics of immunity to OPV. The four different approaches were complementary to each other, highlighting different genes and pathways that may relate to the underlying mechanisms of the immunological response. The population-level results may be related to the individual response. Further investigation into the associations may reveal potential adjuvants and improved vaccines, not only for oral poliovirus vaccine but also for other mucosal vaccines for enteric infections.

Thesis Committee

Dr. Priya Duggal

Dr. W.H. Linda Kao

Dr. William Moss

Thesis Readers

Dr. David Sack (Committee Chair)

Dr. Priya Duggal (Thesis Advisor)

Dr. W.H. Linda Kao

Dr. Marsha Wills-Karp

Dr. William Moss (Alternate)

Dr. Neal Halsey (Alternate)

Dr. Alan F. Scott (Alternate)

Acknowledgements

I owe a great deal of thanks to a great deal of people for getting me through it all.

Everything would not have been possible without the support and guidance of my advisor, Dr. Priya Duggal. You are my role model, my second mother, and my sherpa through the wilderness of grad school. I cannot thank you enough for everything you have done for me. You have pushed me to become a better scientist and person. In short, the best advisor and mentor anyone could ask for.

I would like to thank my thesis readers: Dr. David Sack, Dr. Marsha Wills-Karp, Dr. Neal Halsey, and Dr. Alan Scott. Thank you for your time and feedback. Thank you to my thesis committee members, Dr. Linda Kao and Dr. Bill Moss, for your input and patience throughout this entire process. I would like to thank Dr. Bill Petri for always being so supportive of my involvement of this project. It has been an honor working with you.

I am indebted to Dr. Sandra Petersen for encouraging my fascination and love for genetics. You took me in as a child and let me run amok in your lab. I value the time you spent to foster my interest and give me hands-on experience. Even that entire summer I spent purifying RNA from rat livers and bleaching all my shirts while washing glassware.

The support I have received from my friends has been mind-blowing. Thank you to all of you who have made such a difference in my grad school experience. Much love to Kasia Fertala, my forever roommate who has been with me for nearly a decade. You have always been there to reassure me that there is a point to all this madness and to just remind myself: what would Tina do? A big thank you to Meg Parker for being the best work-wife a gal could ask for. True story. Thank you to Beth Linas, for suffering through numerous practice seminars and being a partner-in-crime for so many shows. Lastly, many thanks to my officemates past and present: Audrey Chu, Poojitha Balakrishnan, Jackie Bidinger and Stephanie Williams. You kept me sane and prevented me from ever taking myself too seriously.

Lastly, this entire dissertation is dedicated to my family. To Albert, who met me when the going got tough and still likes me. To my sister Kristen, who has been supportive of my years and years of studies. I wouldn’t share a Netflix account with anyone else. To my father, who has encouraged my scientific fanaticism since the beginning and has driven me back and forth from school while fast asleep too many times to count. To my mother, who has always let me know how proud she is. You have always pushed me to be better and I wouldn’t have made it this far without your Tiger Mom love.

I love you all.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

Oral poliovirus vaccine (OPV) has contributed to the global control of polio, with a 99% decrease in cases over the last twenty-five years. Both the safety and efficacy of the vaccine have been proven through the near eradication of polio with less than 300 cases in four countries in 2012.(WHO 2013) However, there still remain individuals who fail to elicit an immunological response to numerous doses of viable vaccine. By identifying the reasons why these individuals fail OPV, lessons may be learned to inform the development of other less well-characterized oral vaccines, such as those against cholera and rotavirus infections.

Upon vaccination with all four recommended doses of OPV, levels of systemic immunity can be measured by looking at the natural log titers of neutralizing antibodies. A large amount of variation has been observed in different populations around the world. (Richardson et al. 1995; Sabin et al. 1960; Habib et al. 2013; Reichler et al. 1997; World Health Organization Collaborative Study Group on Oral Poliovirus Vaccine 1995) While the CDC standard for seropositive status is having a log serum neutralizing antibody titer (LT) above 3 (serial dilution of less than 1:8), individuals can range from slightly above this cut-off to strong responders (LT > 7) adding another dimension to the systemic response to OPV. (World Health Organization Collaborative Study Group on Oral Poliovirus Vaccine 1995)

It has been hypothesized that an individual's response to vaccination may be in part due to host genetics because differential responses arise despite controlling for both host health-related factors, such as nutritional status, as well as vaccine-related factors, such as viral serotype concentrations. (Paul 2007; Newport et al. 2004) One way of investigating this hypothesis is through genome-wide association studies (GWAS), which test for the association of single nucleotide polymorphisms (SNPs) at various locations across the human genome. This method has been successful at elucidating risk loci for complex traits like asthma, hypertension, prostate cancer and age-related macular degeneration over the past 10 years, with 11,334 genome-wide significant variants identified. (Hindorff et al. 2009; Hindorff et al. 2013) GWAS of response vaccines, such as hepatitis B and smallpox, have identified significant associations with *HLA-DPB1* and *WDR92*, respectively. (Ovsyannikova et al. 2012; Kennedy et al. 2012; Png et al. 2011) The heritability, or proportion of phenotypic variability due to the host genetics, of the systemic response to OPV has been estimated to be high (60%), and is comparable to that of hepatitis B (77%).(Newport et al. 2004) This is also comparable with the heritability of human height (70%), and nearly double an estimated heritability of 30% for Type II Diabetes.(Zaitlen et al. 2013) GWAS of hepatitis B, human height, and type II diabetes have all previously found genome-wide significantly associated loci. This indicates that there is likely a genetic basis for the immune response to OPV.

To address this question, a GWAS of response to OPV was conducted among a birth cohort of Bangladeshi infants. These children had received four doses of vaccine by one year of age. This thesis aims to identify the host genetic factors that underlie the systemic immune response to oral poliovirus vaccine (OPV) in a cohort of Bangladeshi children using different genetic methods to elucidate genetic loci, genes, and pathways involved in this immune response. The specific aims are as follows:

Aim 1: *To identify genetic single nucleotide polymorphisms associated with the systemic immune response after four doses of oral poliovirus vaccine within a cohort of Bangladeshi children and correlate these signals with signatures of positive selection. (Chapter 3)*

Aim 2a: *To conduct a review and evaluation of gene-level methods for genome-wide association studies through simulation. (Chapters 4 and 5)*

Aim 2b: *To conduct a review and evaluation of pathway-level methods for genome-wide association studies through simulation. (Chapters 4 and 6)*

Aim 3: *To apply gene- and pathway-level methods to a genome-wide association study of oral poliovirus vaccine response in Bangladeshi children. (Chapter 7)*

Throughout human history, infectious pathogens have been strong agents of selective pressure on human populations. (Novembre and Han 2012; Fumagalli et al. 2011) The most well known example of this effect is malaria and sickle cell anemia. Malaria exerts selective pressure on individuals, as the illness could be fatal before an individual can reach reproductive age therefore discontinuing the further propagation of their genes. When beneficial mutations arose within the gene *HBB* (hemoglobin, beta), they were preserved within human populations by defending these individuals against the potentially fatal infection, leading to reproduction and the transmission of the protective alleles.(Jallow et al. 2009) The selective pressure of malaria was so great that these mutations persisted despite individuals with two copies developing sickle-cell anemia, a potentially fatal syndrome.(Gouagna et al. 2010) This phenomenon is not limited to malaria. Through the examination of global genetic adaptation, it has been suggested that many pathogens may be the main selective pressure throughout human evolution. (Fumagalli et al. 2011) Specifically, viruses have had a large influence on the innate immune system. (Zinkernagel, Hengartner, and Stitz 1985; Fumagalli et al. 2010) Among the top human genetic pathways that correlate with pathogen diversity within a human population, pathways involved in viral infection and subsequent replication are enriched when compared to bacterial or amoebic infection. (Fumagalli et al. 2011) Recently, measures of natural selection were estimated in Bangladeshi children and correlated with susceptibility with cholera, identifying risk loci in potassium channel genes and the NF-kB signaling pathways.(Karlsson et al. 2013) We will correlate measures of positive selection with our GWAS results to elucidate genomic regions that may have been beneficial to reduced morbidity and mortality with OPV or other disease with a similar mechanism and therefore preserved throughout multiple generations. (Aim 1)

Traditionally, genome-wide association studies require large sample sizes (>5,000) to identify an association using stringent significance thresholds (p-values) to correct for multiple comparisons for the 500,000-2.5 million SNPs being tested. SNPs that have low P-values ($0.001 \leq P \leq 5\text{x}10^{-8}$) but which do not reach this threshold are often ignored in the initial analysis. Gene- and pathway-level methods have been developed to look at SNPs that may be suggestive but not reach the stringent significance threshold. By combining signals from multiple SNPs within a gene, and subsequently in multiple genes in pathways, the enhancement of statistical signal in these regions can be determined. There is currently no consensus on the best method for this type of analysis, so a simulation will be conducted to evaluate gene- and pathway-level methods (Aims 2a and 2b). The best methods determined by this simulation will then be applied to the OPV GWAS data (Aim 3).

In the last fifty years, the efficacy of oral poliovirus vaccine has been proven by the eradication of wildtype poliovirus from many regions around the world. It is not well understood why some individuals fail to respond to OPV, a well-characterized and proven vaccine, while their peers with a seemingly similar health status respond robustly. As OPV can serve as a prototype for the future of oral vaccines, individuals who fail to respond to OPV may be likely to fail other oral vaccines. There are currently licensed oral vaccines for 5 pathogens: poliovirus, rotavirus, *Salmonella typhi* and two for *Vibrio cholera* infection, with varying efficacies.(Lycke 2012) By elucidating the genes and pathways that are involved with failure to respond to OPV, the underlying mechanisms inherent to oral vaccination may be better characterized and applied to the development of other oral vaccines.

References

Fumagalli, Matteo, Manuela Sironi, Uberto Pozzoli, Anna Ferrer-Admetlla, Anna Ferrer-Admettla, Linda Pattini, and Rasmus Nielsen. 2011. "Signatures of Environmental Genetic Adaptation Pinpoint Pathogens as the Main Selective Pressure Through Human Evolution.." *PLoS Genetics* 7 (11) (November): e1002355. doi:10.1371/journal.pgen.1002355.

Fumagalli, Matteo, Uberto Pozzoli, Rachele Cagliani, Giacomo P Comi, Nereo Bresolin, Mario Clerici, and Manuela Sironi. 2010. "Genome-Wide Identification of Susceptibility Alleles for Viral Infections Through a Population Genetics Approach.." *PLoS Genetics* 6 (2) (February): e1000849. doi:10.1371/journal.pgen.1000849.

Gouagna, Louis Clement, Germana Bancone, Frank Yao, Bienvenue Yameogo, Kounbobr Roch Dabiré, Carlo Costantini, Jacques Simporé, Jean-Bosco Ouedraogo, and David Modiano. 2010. "Genetic Variation in Human HBB Is Associated with Plasmodium Falciparum Transmission.." *Nature Publishing Group* 42 (4) (April): 328–331. doi:10.1038/ng.554.

Habib, M A, S Soofi, N Ali, R W Sutter, M Palansch, H Qureshi, T Akhtar, N A Molodecky, H Okayasu, and Zulfiqar A Bhutta. 2013. "A Study Evaluating Poliovirus Antibodies and Risk Factors Associated with Polio Seropositivity in Low Socioeconomic Areas of Pakistan." *Vaccine* 31 (15) (April 8): 1987–1993. doi:10.1016/j.vaccine.2013.02.003.

Hindorff, Lucia A, J Macarthur, J Morales, Heather A Junkins, P N Hall, A K Klemm, and Teri A Manolio, eds. 2013. *A Catalog of Published Genome-Wide Association Studies*. Accessed September 10. http://www.genome.gov/gwastudies.

Hindorff, Lucia A, Praveen Sethupathy, Heather A Junkins, Erin M Ramos, Jayashri P Mehta, Francis S Collins, and Teri A Manolio. 2009. "Potential Etiologic and Functional Implications of Genome-Wide Association Loci for Human Diseases and Traits.." *Pnas* 106 (23) (June 9): 9362–9367. doi:10.1073/pnas.0903103106.

Jallow, Muminatou, Yik-Ying Teo, Kerrin S Small, Kirk A Rockett, Panos Deloukas, Taane G Clark, Katja Kivinen, et al. 2009. "Genome-Wide and Fine-Resolution Association Analysis of Malaria in West Africa." *Nature Publishing Group* 41 (6) (May 24): 657–665. doi:10.1038/ng.388.

Karlsson, E K, J B Harris, S Tabrizi, A Rahman, I Shlyakhter, N Patterson, C O'Dushlaine, et al. 2013. "Natural Selection in a Bangladeshi Population From the Cholera-Endemic Ganges River Delta." *Science Translational Medicine* 5 (192) (July 3): 192ra86–192ra86. doi:10.1126/scitranslmed.3006338.

Kennedy, Richard B, Inna G Ovsyannikova, V Shane Pankratz, Iana H Haralambieva, Robert A Vierkant, and Gregory A Poland. 2012. "Genome-Wide Analysis of Polymorphisms Associated with Cytokine Responses in Smallpox Vaccine Recipients." *Human Genetics* 131 (9) (May 19): 1403–1421. doi:10.1007/s00439-012-1174-2.

Lycke, Nils. 2012. "Recent Progress in Mucosal Vaccinedevelopment: Potential and

Limitations." *Nature Reviews Immunology* 12 (8) (August 1): 592–605. doi:10.1038/nri3251.

Newport, M J, T Goetghebuer, H A Weiss, H Whittle, C-A Siegrist, and A Marchant. 2004. "Genetic Regulation of Immune Responses to Vaccines in Early Life." *Genes and Immunity* 5 (2) (January 22): 122–129. doi:10.1038/sj.gene.6364051.

Novembre, J, and E Han. 2012. "Human Population Structure and the Adaptive Response to Pathogen-Induced Selection Pressures." *Philosophical Transactions of the Royal Society B: Biological Sciences* 367 (1590) (February 6): 878–886. doi:10.1046/j.1469-1809.2001.6510001.x.

Ovsyannikova, Inna G, Richard B Kennedy, Megan O'Byrne, Robert M Jacobson, V Shane Pankratz, and Gregory A Poland. 2012. "Genome-Wide Association Study of Antibody Response to Smallpox Vaccine." *Vaccine* 30 (28) (June 13): 4182–4189. doi:10.1016/j.vaccine.2012.04.055.

Paul, Yash. 2007. "Role of Genetic Factors in Polio Eradication: New Challenge for Policy Makers." *Vaccine* 25 (50) (December): 8365–8371. doi:10.1016/j.vaccine.2007.09.068.

Png, E, A Thalamuthu, R T H Ong, H Snippe, G J Boland, and M Seielstad. 2011. "A Genome-Wide Association Study of Hepatitis B Vaccine Response in an Indonesian Population Reveals Multiple Independent Risk Variants in the HLA Region." *Human Molecular Genetics* 20 (19) (September 7): 3893–3898. doi:10.1093/hmg/ddr302.

Reichler, M R, S Kharabsheh, P Rhodes, H Otoum, S Bloch, M A Majid, M A Pallansch, P A Patriarca, and S L Cochi. 1997. "Increased Immunogenicity of Oral Poliovirus Vaccine Administered in Mass Vaccination Campaigns Compared with the Routine Vaccination Program in Jordan.." *The Journal of Infectious Diseases* 175 Suppl 1 (February): S198–204.

Richardson, G, R W Linkins, M A Eames, D J Wood, P J Campbell, E Ankers, M Deniel, A Kabbaj, D I Magrath, and P D Minor. 1995. "Immunogenicity of Oral Poliovirus Vaccine Administered in Mass Campaigns Versus Routine Immunization Programmes.." *Bulletin of the World Health Organization* 73 (6): 769–777.

Sabin, Albert B, Manuel Ramos-Alvarez, José Alvarez-Amezquita, William Pelon, Richard H Michaels, Ilya Spigland, Meinrad A Koch, Joan M Barnes, and Johng S Rhim. 1960. "Live, Orally Given Poliovirus Vaccine: EFfects of Rapid Mass Immunization on Population Under Conditions of Massive Enteric Infection with Other Viruses." *JAMA : the Journal of the American Medical Association* 173 (14): 1521–1526.

WHO. 2013. "WHO Polio Fact Sheet" (April 30): 1–3.

World Health Organization Collaborative Study Group on Oral Poliovirus Vaccine. 1995. "Factors Affecting the Immunogenicity of Oral Poliovirus Vaccine: a Prospective Evaluation in Brazil and the Gambia." *The Journal of Infectious Diseases* 171 (5) (May): 1097–1106.

Zaitlen, Noah, Peter Kraft, Nick Patterson, Bogdan Pasaniuc, Gaurav Bhatia, Samuela Pollack, and Alkes L Price. 2013. "Using Extended Genealogy to Estimate Components of Heritability for 23 Quantitative and Dichotomous Traits." Edited by Peter M Visscher. *PLoS Genetics* 9 (5) (May 30): e1003520.

doi:10.1371/journal.pgen.1003520.s011.

Zinkernagel, R M, H Hengartner, and L Stitz. 1985. “On the Role of Viruses in the Evolution of Immune Responses..” *British Medical Bulletin* 41 (1) (January): 92–97.

Chapter 2: Epidemiology of Polio and the Oral Poliovirus Vaccine

2.1: Poliovirus and Clinical Pathogenesis

Poliovirus (PV) was discovered to be the causative agent for poliomyelitis in 1909 by Karl Landsteiner and Erwin Popper. (De Jesus 2007) It is a small positive single-stranded RNA virus that is approximately 7400 nucleotides long. Poliovirus contains three serotypes: 1 [Mahoney], 2 [Lansing], and 3 [Leon]. PV uses the fecal-oral route for transmission, although the specific cell types that it replicates in immediately after entry are unknown. It is hypothesized that it likely replicates first in the lymphatic tissue of the gastrointestinal (GI) tract, such as in the tonsils, the Peyer's patches (PP) of the ileum, and the mesenteric lymph nodes (De Jesus 2007). Infection by PV is only in humans and other primates. It is highly infectious, most often in children under 5 years of age. (WHO 2013)

The majority of infected individuals (95%) are either completely asymptomatic, or experience a mild viremia. In these individuals, no neurological conditions occur. In the remaining 5% of infected individuals, the infection spreads to other sites of the reticuloendothelial system. In 4-8% of these individuals who develop a substantial primary viremia, a secondary major viremia develops which is described as the "minor, non-specific illness", or abortive poliomyelitis. These symptoms include an upper respiratory infection, GI illness, and an illness mimicking influenza. Of those who experience abortive poliomyelitis, only a very small percentage (<2%) go on to develop symptoms implicating the entry of PV into the central nervous system (CNS). This leads to either non-paralytic aseptic meningitis or paralytic poliomyelitis. Non-paralytic aseptic meningitis affects 1-2% of all PV infections, while paralytic poliomyelitis affects 0.1-1% of all infections. (Figure 2.1) This attack rate differs in virulence by the infecting serotype, with serotype 2 found less often in cases of paralytic poliomyelitis compared to the two other serotypes. (Ogra et al. 2011)

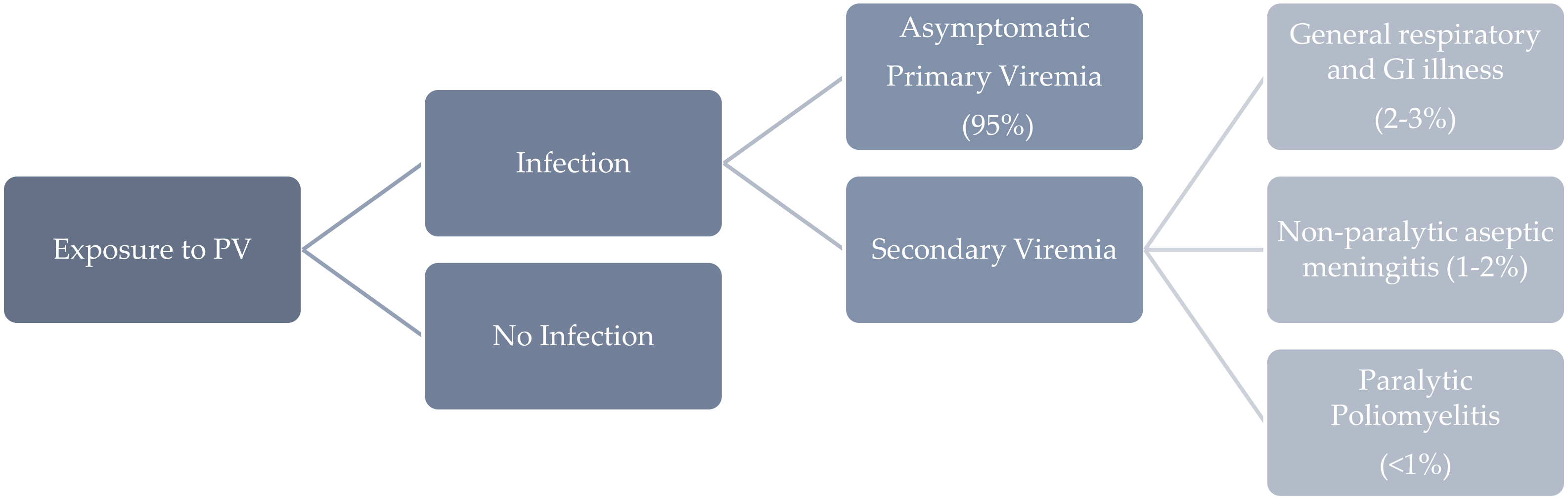

Exposure to PV
Infection
No Infection
Asymptomatic Primary Viremia (95%)
Secondary Viremia
General respiratory and GI illness (2-3%)
Non-paralytic aseptic meningitis (1-2%)
Paralytic Poliomyelitis (<1%)

Figure 2.1: Flowchart of Poliovirus Infection of Pathogenesis with Percentages of Terminal Symptoms within OPV-Naïve Infected Individuals. *Only the major outcomes are listed. Estimates are imprecise and are from numerous studies.*

Paralytic poliomyelitis’ with no effect on sensation or cognition are classified into three groups: Spinal poliomyelitis, bulbar poliomyelitis, and bulbospinal poliomyelitis. Spinal poliomyelitis causes acute flaccid paralysis because of the selective destruction of spinal motor neurons and the denervation of the associated skeletal muscles. Bulbar poliomyelitis results in the paralysis of respiratory muscles caused by infected neurons in the brainstem that regulate breathing. Bulbospinal poliomyelitis involves both the brain stem and the spinal cord. Paralytic poliomyelitis has a 2-5% case fatality rate in children, and a 15-30% fatality rate in adults with the highest rates in cases of bulbar poliomyelitis (De Jesus 2007).

The poliovirus is ingested and multiplies in the oropharyngeal and intestinal mucosa. The exact tissue that it colonizes first is unknown, partly due to the lack of an accurate mouse model for the natural pathogenesis of poliovirus. Despite this limitation, some advancement has been made. The cell receptor for poliovirus was discovered in the early 1990s to be CD155 (Koike et al. 1991) (Ren et al. 1990). CD155 is a glycoprotein that is part of the Immunoglobulin (Ig) superfamily (Racaniello 2006). It has 3 extracellular Ig-like domains used to bind poliovirus. The interaction of the V-type domain I of CD155 and the poliovirus capsid lead to a conformational change that releases the virus’ RNA genome into the cytoplasm for replication and translation. CD155 is also a recognition molecule for natural killer cells (NK), working with CD226 and CD96 to induce cytotoxic activity. (Racaniello 2006) Transgenic mice have been developed to express CD155, but it is not expressed on gut mucosal cells. The region surrounding CD155 was the target of ancient positive selection in simians. (Suzuki 2006) Because the receptor is deleterious to the fitness of an individual through polio infection, it is believed that it formed through the preferential binding of an unknown molecule. While CD155 defines the tropism of the initial infection, the route of invasion into the CNS is not known and the crossing of the blood-brain barrier is thought to be independent of the presence of cellular receptor CD155.(Racaniello 2006) Other popular theories are that the virus invades by retrograde axonal transport, or that it is imported by infected macrophages, deemed the "Trojan horse mechanism".

Because of the presence of intracellular RNA, it is hypothesized that TLR7/8 pathway is involved in poliovirus pathogenesis. In a subset of people, the virus spreads from the primary mucosal sites to the cervical and mesenteric lymph nodes, then to the blood. In 1-2% of poliovirus infections, the virus will then invade the central nervous system (CNS). It is hypothesized that because invasion of the CNS is unnecessary for the spread of the virus, it is an accidental diversion of the enteric stage. (Racaniello 2006) Tropism, or the tissues that poliovirus invades, is thought to be determined by IFNα/β in conjunction with CD155. In 99% of infections, this pathway limits the infection of poliovirus to the gastrointestinal tract. When poliovirus remains in the gastrointestinal tract, illness is restricted to milder non-fatal symptoms with little associated morbidity. It is when the virus crosses into the central nervous system that it may result in the most devastating effects of polio infection.

2.2: Poliovirus Vaccines

Jonas Salk developed the inactivated polio vaccine (IPV) in 1955 by exposing purified poliovirus to low concentrations of formaldehyde, therefore inactivating it. (De Jesus 2007; Nathanson and Kew 2011) Due to crosslinks in the external capsid proteins of the virus, it is unable to infect the patient; however, the formaldehyde leaves the antigenic epitopes capable of inducing neutralizing antibodies. IPV is administered intra-muscularly and provides systemic immunity. This is meant to prevent paralytic poliomyelitis by attacking the virus when it enters the bloodstream on the way to the central nervous system.(Nathanson and Kew 2011) IPV only provides low titers of mucosal immunity, and therefore allows colonization of the GI tract. Even with enhanced potency, IPV is less effective than OPV in inducing mucosal immunity to prevent and limit intestinal infection. (Belyakov and Ahlers 2009)

The oral polio vaccine (OPV) has been widely used since 1963. It was developed by Albert B. Sabin by successive tissue culture of virulent wild poliovirus and the isolation of individual clones. (Sabin et al. 1960; Belyakov and Ahlers 2009) Two key genetic properties of the virus segregated independently. This allowed the isolation of clones with attenuated neurovirulence that were still able to replicate in the GI tract. Administered in four doses, OPV produces both circulating and mucosal immunity. (Nathanson and Kew 2011) The mucosal immunity is essential for the prevention of poliovirus infection, a feature that the inactivated poliovirus vaccine (IPV) does not have. (Belyakov and Ahlers 2009) IPV provides strong systemic immunity, but unlike

OPV it does not provide strong intestinal immunity, and therefore does not prevent infection.

OPV does not offer the same level of protection to all three serotypes. In the original clinical trial in 1959 in Toluca, Mexico, at the end of 10 weeks after a single dose of trivalent OPV, seroconversion for type 1, 2, and 3, was found to be 68%, 82%, and 43%, respectively. (Sabin et al. 1960) Both monovalent OPV (mOPV), one for each serotype, and trivalent OPV (tOPV) are available but there are differences in seroconversion rates based on these vaccines. In studies from Leningrad in the 1970s, the seroconversion rates using tOPV were 82%, 80%, and 71% for serotypes 1, 2, and 3, respectively. For mOPV, seroconversion rates increased to 97%, 100%, and 96% for serotypes 1, 2, and 3, respectively. The reduced efficacy of tOPV is because the presence of all three serotypes introduces interference between the serotypes, however the efficacy of tOPV has been maximized by changing the proportions of virus for each serotype to minimize this interference since the original formulation. (Patriarca, Wright, and John 1991) The current vaccine has the proportions for serotypes 1, 2, and 3 as 10:1:3, which maximize efficacy for all three serotypes.(De Jesus 2007) Although the trivalent form of OPV was originally the most thermally labile vaccine in the World Health Organization’s Expanded Program of Immunization (WHO EPI), it has been chemically stabilized to minimize a loss of potency. (Patriarca, Wright, and John 1991) Because OPV is a live attenuated vaccine, the virus is capable of reversion to its virulent form. These circulating vaccine-derived polioviruses (cVDVP) can cause paralytic poliomyelitis.

However, in the United States it is estimated that there was one case of vaccine-associated paralytic poliomyelitis (VAPP) for every 2-3 million doses of OPV before the change to IPV. (De Jesus 2007)

The first priority in determining between the inactivated poliovirus vaccine (IPV) and the oral poliovirus vaccine (OPV) is the goal of the vaccination effort. If the goal is to stop transmission of wild-type poliovirus, OPV offers strong intestinal immunity, therefore preventing infection and subsequent shedding that propagates the virus. OPV is less expensive than IPV and is fast-acting.(Paul 2007) It is easily administered orally, while the administration of IPV is more invasive (injection) and requires trained personnel.(Nathanson and Kew 2011) OPV is a live attenuated vaccine and there is the probability of secondary spread to contacts of the vaccinated, protecting them against infection from wild type poliovirus. However, OPV can revert back to its virulent form, allowing cVDPV to infect unprotected children and cause vaccine-associated paralytic poliomyelitis (VAPP). (Paul 2007) If the goal of the vaccination campaign is instead to eliminate risk for paralytic poliomyelitis, then IPV offers strong circulating neutralizing immunity in the blood stream, preventing poliovirus from crossing the blood-brain barrier and causing paralysis. However, it does not induce adequate mucosal immunity allowing the infection and transmission of wildtype virus, leaving unvaccinated individuals susceptible to infection. IPV does not replicate and shed, offering no protection to the contacts of the vaccinated. (Paul 2007)

2.3: Polio Eradication Effort

In 1988, encouraged by the eradication of smallpox less than ten years earlier, the World Health Organization (WHO) launched a campaign to eradicate polio. (WHO 2013) This Global Polio Eradication Initiative (GPEI) was led by the WHO, Rotary International, the US Centers for Disease Control (CDC), and UNICEF. Its objectives were to interrupt the transmission of wild-type PV and to achieve certification of global polio eradication, while contributing to health systems development and strengthening routine immunization and surveillance in a systematic way. Because of the goal to interrupt transmission of the wild-type PV and a higher cost-effectiveness, OPV was the chosen vaccine. To achieve this, four strategies were adopted. First, infants were immunized with 4 doses within the first year of life at high rates. Second, supplementary doses of OPV would then be given to all kids under the age of 5 during Supplementary Immunization Activities (SIAs). Third, surveillance for wild poliovirus infection would be monitored through reporting and testing of all acute flaccid paralysis cases among children under the age of 15. Finally, there would be targeted "mop-up" campaigns once transmission was significantly decreased and limited to specific areas. (WHO 2013) To be certified as being polio-free, a region must meet three conditions: (i) they would have to be free of polio cases due to wild PV for at least 3 years, (ii) disease surveillance systems in the regional countries would need to meet international standards, and (iii) each country must demonstrate the ability to detect, report and respond to "imported" polio cases. As of 2010, this massive eradication effort has saved greater than 5 million people from getting paralytic poliomyelitis, and has immunized greater than 2 million children in SIAs. (WHO 2013) The annual incidence of poliomyelitis is now <1% of the pre-vaccination levels. The wild-type serotype 2 PV was eradicated globally in 1999. (Nathanson and Kew 2011) In 1994, the WHO Region of the Americas was certified as being polio-free, followed by the Western Pacific Region in 2000, and the European Region in 2002. (WHO 2013)

Due to large-scale vaccination efforts, the incidence of poliomyelitis has greatly decreased over the past 25 years. Since 1988 there has been > 99% decrease in cases, from 350,000 to 1,604 (WHO 2013). However, there still remain cases of poliomyelitis worldwide. During 2009-2010, 23 countries had imported cases, comprising a little over 75% of the annual incidence. As of 2013 only three countries had endemic wild-type PV transmission (Nigeria, Pakistan, and Afghanistan). Obstacles to eradication in these countries are the low efficacy of tOPV, as well as a failure to immunize a sufficient percentage of infants and toddlers. (Nathanson and Kew 2011)

2.4: Oral Poliovirus Vaccine Failure

The first reports of OPV failure were in the 1970s. (Patriarca, Wright, and John 1991) Developing countries showed low seroconversion for serotypes 1 and 3, while seroconversion reached 100% of recipients in developed countries. Reasons for failure have been cited to be both viral, as well as host-related. One potential issue is the vaccine's stability. Trivalent OPV is the most thermally labile in the WHO's EPI vaccination schedule. It requires a cold chain however it is chemically stabilized to minimize a loss of potency when exposed to higher temperatures. It has been shown to be resistant to numerous freezing and thawing cycles. There are differences in heat stability for the three serotypes, as well as interference from type 2. Despite these concerns, it has been shown that even with the proper handling of tOPV, there still exist failures. (Patriarca, Wright, and John 1991) Another variable is the vaccine's administration and schedule. The standard is 3 doses, with one supplemental dose at birth in countries that are endemic for poliomyelitis. In 1985, the Global Advisory Group suggested an accelerated immunization schedule, in which protection was provided at the youngest possible age. The first dose is less effective when administered at less than 4 weeks of age due to the interference of passively acquired maternal antibodies. Women in developing countries have a higher level of exposure to wild-type poliovirus, therefore they have higher circulating antibodies. This leads to infants passively-acquiring a higher concentration of antibodies and a higher level of interference with the first dose. (Patriarca, Wright, and John 1991) The median length of excretion of OPV was 21 days and continued excretion could interfere with subsequent doses. To minimize this potential interference, the EPI suggests 4-week intervals between the doses.(Table 2.1) Other vaccine factors include the vaccine potency, formulation, and dosage volume; however none of these have shown a high effect on seroconversion, especially after recent standardizations. Vaccine failure is cited as the major problem in the Indian provinces of Uttar Pradesh and Bihar. Vaccine efficacy against serotype 1 is 9% in Uttar Pradesh, 18% in Bihar, and 21% in the rest of India. Children in Uttar Pradesh also have similarly low seroconversion for serotype 3. (Paul 2007)

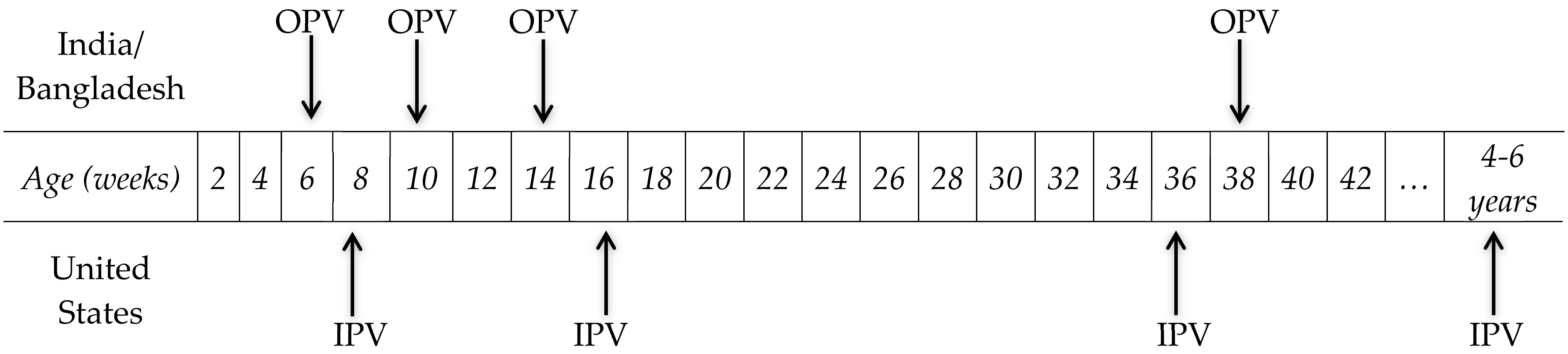


Figure 2.1: Timeline of Vaccinations for OPV (India/Bangladesh EPI Schedule) and IPV (United States Schedule). *IPV is administered in most developed countries, while OPV is still the recommended vaccine for the majority of countries where eradication has either not been achieved, or is recent.*

2.5: Epidemiological and Genetic Risk Factors for Vaccine Failure

2.5.1: Chronic Environmental Enteropathy and Immune Status

Host factors hypothesized to contribute to both mucosal and systemic vaccine failure include the interference of maternal antibodies, the nutritional status of the infant, as well as concurrent enteric infections. During the first few weeks of life, the newborn passively receives maternal antibodies through breast milk. These maternal antibodies can then attack the vaccine when it is administered, leading to vaccine failure. When the child stops receiving the maternal antibodies through breast milk, they will be unprotected without vaccination. This is not a major issue with older infants because of the lack of exclusive breast-feeding. (Patriarca, Wright, and John 1991) Concurrent enteric infections can produce lower rates of seroconversion in children. It is hypothesized that the diarrheal state with enteric infections alters the mucosal architecture, leading to more rapid gastrointestinal transit. (Patriarca, Wright, and John 1991) This leads to reduced colonization of the live attenuated virus, and a diminished antibody response to the vaccine. If this condition is ongoing, it is called chronic environmental enteropathy (CEE). Children in extreme poverty are highly susceptible to CEE because of poor sanitation, malnutrition, and intestinal flora overgrowth. (Czerkinsky and Holmgren 2009) This condition leads to histological changes through the inflammation and blunting of the small intestinal villi, leading to malabsorption of nutrients as well as vaccine antigens. (Korpe and Petri 2012) Strategies for improving vaccine response in children with CEE include a co-administration of the vaccine with agents that can improve the GI tract's integrity, such as zinc, vitamin A and probiotics. (Czerkinsky and Holmgren 2009) Other options include treatment for helminth infections before administration, as well as withdrawal of breast milk for a few hours before administration. It is hypothesized that CEE may contribute to the failure of oral vaccines due to the lack of gut integrity. (Korpe and Petri 2012; Guerrant et al. 2012)

2.5.2: Genetic Risk Factors

Genetic risk factors for OPV failure have not been extensively characterized, but there is evidence that genetic factors may play an important role. In 2004, Newport et al conducted a study of the genetics to OPV response, among other childhood vaccines (i.e. hepatitis B), in the Gambia. Using twins, they estimated that the heritability, the proportion of phenotypic variability due to human genetics, of antibody responses to OPV was 60% [CI:43-73%], using an additive genetic model with a unique environment. (Newport et al. 2004) Monozygotic (MZ) twins, who inherit identical genetic sequences, had a correlation of 64% in their serum-neutralizing antibodies titers for OPV. Dizygotic (DZ) twins, whom only share on average half of their genetics, had a 29% correlation in their titers. The variance between the twins due to environment is assumed to be the same. When twins that share a smaller proportion of their genetics (DZ) also have lower correlations in titers when compared to twins that share all of their genetics (MZ), it indicates that there may be a role for genetics with phenotype.

2.7: Conclusions

Within the past one hundred years, remarkable progress has been made to identify poliovirus, develop safe and efficient vaccines against it, and eradicate it from much of the world. Two major vaccines were developed: oral poliovirus vaccine (OPV) and inactivated poliovirus vaccine (IPV). OPV provides both mucosal and systemic immunity and is both easier and cheaper to administer, thus it has become the primary tool for the eradication of poliovirus. OPV has become the example that many oral vaccines developers (i.e. rotavirus) wish to emulate. However, despite the high efficacy of the vaccine, some individuals fail to mount an adequate response. This failed immune response remains after controlling for vaccine-related factors, such as potential variability in concentrations and attenuation, as well as host-related factors, such as general health status. One hypothesis has been that host genetic factors may play a role and this is supported by the high heritability of OPV response (60%) and distinct ethnic population failure of the vaccine. To-date, no large-scale genetic study to elucidate potential risk loci for the response to OPV has been performed. The aim is to identify genes and pathways that can inform future development and implementation of oral vaccines.

References

Belyakov, I M, and J D Ahlers. 2009. "What Role Does the Route of Immunization Play in the Generation of Protective Immunity Against Mucosal Pathogens?." *The Journal of Immunology* 183 (11) (November 18): 6883–6892. doi:10.4049/jimmunol.0901466.

Czerkinsky, C, and J Holmgren. 2009. "Enteric Vaccines for the Developing World: a Challenge for Mucosal Immunology." *Mucosal Immunology* 2 (4) (May 6): 284–287. doi:10.1038/mi.2009.22.
http://dx.doi.org/10.1038/mi.2009.22.

De Jesus, Nidia H. 2007. "Epidemics to Eradication: the Modern History of Poliomyelitis." *Virology Journal* 4 (1): 70. doi:10.1186/1743-422X-4-70.

Guerrant, Richard L, Mark D DeBoer, Sean R Moore, Rebecca J Scharf, and Aldo A M Lima. 2012. "The Impoverished Gut—a Triple Burdenof Diarrhoea, Stunting and Chronic Disease." *Nature Publishing Group* 10 (4) (December 11): 220–229. doi:10.1038/nrgastro.2012.239.

Koike, S, C Taya, T Kurata, S Abe, I Ise, H Yonekawa, and A Nomoto. 1991. "Transgenic Mice Susceptible to Poliovirus.." *Proceedings of the National Academy of Sciences* 88 (3) (February 1): 951–955.

Korpe, Poonum S, and William A Petri Jr. 2012. "Environmental Enteropathy: Critical Implications of a Poorly Understood Condition." *Trends in Molecular Medicine* 18 (6) (June 1): 328–336. doi:10.1016/j.molmed.2012.04.007.

Nathanson, Neal, and Olen M Kew. 2011. "Poliovirus Vaccines: Past, Present, and Future.." *Archives of Pediatrics & Adolescent Medicine* 165 (6) (June): 489–491. doi:10.1001/archpediatrics.2011.77.

Newport, M J, T Goetghebuer, H A Weiss, H Whittle, C-A Siegrist, and A Marchant. 2004. "Genetic Regulation of Immune Responses to Vaccines in Early Life." *Genes and Immunity* 5 (2) (January 22): 122–129. doi:10.1038/sj.gene.6364051.

Ogra, Pearay L, Hiromasa Okayasu, Cecil Czerkinsky, and Roland W Sutter. 2011. "Mucosal Immunity to Poliovirus." *Expert Review of Vaccines* 10 (10) (October): 1389–1392. doi:10.1586/erv.11.106.

Patriarca, P A, P F Wright, and T J John. 1991. "Factors Affecting the Immunogenicity of Oral Poliovirus Vaccine in Developing Countries: Review.." *Reviews of Infectious Diseases* 13 (5) (September): 926–939.

Paul, Yash. 2007. "Role of Genetic Factors in Polio Eradication: New Challenge for Policy Makers." *Vaccine* 25 (50) (December): 8365–8371. doi:10.1016/j.vaccine.2007.09.068.

Racaniello, Vincent R. 2006. "One Hundred Years of Poliovirus Pathogenesis." *Virology* 344 (1) (January): 9–16. doi:10.1016/j.virol.2005.09.015.

Ren, R B, F Costantini, E J Gorgacz, J J Lee, and V R Racaniello. 1990. "Transgenic Mice Expressing a Human Poliovirus Receptor: a New Model for Poliomyelitis.." *Cell* 63 (2) (October 19): 353–362.

Sabin, Albert B, Manuel Ramos-Alvarez, José Alvarez-Amezquita, William Pelon, Richard H Michaels, Ilya Spigland, Meinrad A Koch, Joan M Barnes, and Johng S Rhim. 1960. "Live, Orally Given Poliovirus Vaccine: Effects of Rapid Mass

Immunization on Population Under Conditions of Massive Enteric Infection with Other Viruses." *JAMA : the Journal of the American Medical Association* 173 (14): 1521–1526.

Suzuki, Yoshiyuki. 2006. "Ancient Positive Selection on CD155 as a Possible Cause for Susceptibility to Poliovirus Infection in Simians." *Gene* 373 (May): 16–22. doi:10.1016/j.gene.2005.12.016.

WHO. 2013. "WHO Polio Fact Sheet" (April 30): 1–3.

Chapter 3: Genome-wide association study of Oral Poliovirus Vaccine response and signatures of selection in Bangladeshi infants (Paper 1)

3.1: Abstract

Background: The Oral Poliovirus Vaccine (OPV) has been widely successful in the eradication effort of polio infection. However, it does not provide protection in some individuals despite multiple doses of viable vaccine. It was previously hypothesized that human genetics may be responsible for immune response failure to the oral vaccine. To examine the role human genetics may play, we performed a genome-wide association study (GWAS) of the response to OPV in 357 Bangladeshi infants. We also conducted a genome-wide scan for signatures of natural selection that may be relevant to poliovirus infection or immune response and may correlate with the GWAS results.

Methods: A genome-wide association study was performed using the log serum-neutralizing antibody titers (LTs) to OPV in 357 Bangladeshi children. The study compared seronegative (LT<3) to high seropositive (LT>7) individuals after four doses of OPV. Logistic regression was conducted on 6.5 million imputed SNPs across the human genome, adjusting for stunting (height-for-age Z-score <-2). A genome-wide scan of selection was conducted in the full cohort of 473 Bangladeshi children, calculating a standardized cross-population extended haplotype homozygosity (XP-EHH) score using the HapMap Nigerian Yoruba and Kenyan Luhya populations as a reference. Genetic locations were examined for overlap between the two genetic scans, GWAS (*P*<0.001) and selection (XP-EHH *P*<0.01).

Results: The GWAS did not identify any genome-wide significant (*P*<5x10^{-8}) variants, however two regions were suggestive of an association (*P*<5x10^{-6}). The top association was on chromosome 14 at SNP rs113427985 and showed a decreased odds of an adequate immune response for individuals with an LT > 7 compared to those with an LT < 3 (OR= 0.22, *P*=2.9x10-6). This SNP is located downstream of *MAPK1IP1L* and is in strong linkage disequilibrum with SNPs in *SOCS4*. An additional association was identified on chromosome 7 within the Sonic Hedgehog gene, *SHH*, and an *SHH* cis-regulatory element within a neighboring gene *LMBR1*. This SNP, rs55906254, also showed a decreased odds of OPV immune response for individuals with an LT > 7 compared to those with an LT < 3 (OR=0.31, *P*=3.6x10^{-6}). The selection scan identified significant regions under positive selection in this Bangladeshi population as compared to a Nigerian reference population (HapMap YRI). 32 SNPs had a both a GWAS P-value <0.001 and a selection P-value < 0.01, comprising 9 distinct regions. Half of these 32 SNPs were between the genes *FAM86A* and *RBFOX1* on Chromosome 16.

Conclusions: Genomic methods were used to identify loci associated with the immune response to OPV in a cohort of Bangladeshi children. The genome-wide association

study identified two regions associated with seronegative status after four doses of OPV, and when coupled with the selection scan additional suggestive regions were found. The derived (non-ancestral) alleles at this location were associated with a high seropositive status in response to OPV as well as strong positive selection, suggesting that beneficial mutations arose and were maintained in this genomic location that may have conferred protection against poliovirus. This study highlights the benefits of coupling a traditional GWAS with selection scans for immune or infectious traits like OPV response to identify novel host genetic regions that may warrant additional study.

3.2: Introduction

Poliovirus is the infectious agent responsible for poliomyelitis, a crippling infection that can result in flaccid paralysis. Over the past hundred years, vast leaps of progress have been made to identify this causative agent, develop two viable vaccines, and eradicate the virus from many regions of the world. In the past 25 years, there has been a 99% decrease in cases worldwide, with only 223 reported cases in four countries in 2012.(WHO 2013) An invaluable tool in this fight has been the oral poliovirus vaccine (OPV). Developed in 1960 by Albert B. Sabin, OPV is a live attenuated vaccine that contains all three serotypes (1-3).(Sabin et al. 1960) It is efficacious at eliciting both mucosal and systemic immune responses, with results replicated in diverse populations.(Ogra et al. 2011; John and Vashishtha 2013; Patriarca, Wright, and John 1991; Racaniello 2006)

The systemic immunity developed from OPV administration is measured as the log serum-neutralizing antibody titers (LTs). The World Health Organization (WHO) and Centers for Disease Control (CDC) standard cut-off for an adequate response is an LT > 3, with recognized variation occurring both within and between populations.(WHO 2013; World Health Organization Collaborative Study Group on Oral Poliovirus Vaccine 1995) Failure to mount an adequate systemic response to OPV may be due to numerous factors in both the vaccine and the host. Vaccine-related factors include the stability of the vaccine, relative concentrations of the three serotypes, as well as the timing of doses.(Sabin et al. 1960; Estívariz et al. 2013) Host-related factors include the child's nutritional status, whether or not the child is exclusively breast-fed, and any concurrent infections. (Habib et al. 2013)Even after controlling for these factors with identical viable vaccines and children from the same background, some individuals still fail to mount an immune response to the vaccine. It has been hypothesized that his may be due to differences in host genetics.(Paul 2007) The heritability, or percentage of phenotypic variability due to genetics, for the immune response to OPV has been estimated to be 60%. In a Gambian study of twins, the LTs of monozygotic twins had a higher correlation of titers (64%) than dizygotic twins (29%)(Newport et al. 2004) This increased correlation in monozygotic twins is expected if a disease has a higher heritability as monozygotic twins share identical genetic sequence and dizygotic twins, like other siblings, share only half of their genetic sequence on average. To identify host genes that may play a role in the immune response to OPV, we conducted a genome-wide association study in 357 children from Bangladesh who received four doses of OPV at one year of age and compared individuals at the extremes to OPV response; seronegative individuals (LT<3) to high seropositive individuals (LT>7).

To complement this study, we also conducted a population genetics scan of positive natural selection across the human genome in the same children. Throughout human history, it is thought that infectious pathogens have been responsible for the majority of selective pressure shaping the human genome.(Fumagalli et al. 2011) This is especially true for viruses, which have high mutation rates that allow them to adapt quickly to any changes in the human immune landscape. (Fumagalli et al. 2010; Zinkernagel, Hengartner, and Stitz 1985) Any mutations that are beneficial at preventing infection or limiting viral infections are likely to be preserved throughout successive generations, leading to positive selection. Positive selection can be detected by examining long runs of genotype homozygosity across the human genome. When the beneficial genetic variants are maintained in a population, the genetic sequence surrounding them is sometimes also preserved and can lead to long haplotypes of homozygosity. These extended regions of homozygosity can serve as markers harboring selected genetic variants. The cross population extended haplotype homozygosity (XP-EHH) is calculated by comparing these runs to a different reference population and standardizing across the genome. This method has been utilized successfully to identify verified signatures of positive selection in many global populations.(Pickrell et al. 2009) In this study, we identify regions of positive selection through the estimation of XP-EHH in a Bangladeshi population of children and then correlate these identified selection signals with loci associated with OPV LTs in the same Bangladeshi children. These overlapping regions may have been selected for in the development of immunity to poliovirus.

3.3: Materials and Methods

3.3.1: Study Population

Children were recruited at birth in Dhaka, Bangladesh and followed from birth until at least 2 years of age. All were recruited from Mirpur, an urban slum in Dhaka City. Mirpur, one of the 14 Thanas (subdistricts) of Dhaka, has a population density of one million people per 59 square kilometers. The average monthly expenditure in this population was 6000 BDT (Bangladesh Taka), which translates to roughly 77 US dollars.(Mondal et al. 2011) Despite being geographically closer to Nepal, the inhabitants of Dhaka are genetically closer to an Iranian-Indian-Afghan clade.(Roychoudhury and Nei 1985) The participants are visited bi-weekly in their homes, and in a clinical setting. Diarrheal episodes are recorded and stool samples collected. The stool samples are evaluated for the presence of numerous enteric infections, including *E. histolytica*, *Cryptosporidium*, rotavirus, and *E. coli*. Anthropometric measurements are available every few months, including height, age, and BMI, allowing the calculation of height-for-age Z-score (HAZ), weight-for-age Z-score (WAZ), and weight-for-height Z-scores (WHZ) standardized according to WHO guidelines.

For children completing at least one year of follow-up, serum-neutralizing antibody responses to the full 4-dose regimen were available for all three serotypes. Serum-neutralizing antibody titers were estimated at the CDC in triplicate according to the standard WHO procedure of a modified microneutralization technique in dilutions ranging from 1:4 to 1:1024 (LT of 2-10).(World Health Organization Collaborative Study Group on Oral Poliovirus Vaccine 1995) Of 448 children with OPV serum neutralizing antibody titers, 425 also had genotype data available. Vaccine failure was defined using the CDC standard cutoff of a $\log_2$ serum neutralizing antibody levels of 3, or a 1:8 dilution factor. Seroconversion rates were 93.41% for serotype 1, 96.47% for serotype 2, and 88.71% for serotype 3 (Figure 3.1). Due to the high rates of seroconversion for serotypes 1 and 2, only serotype 3 was examined.

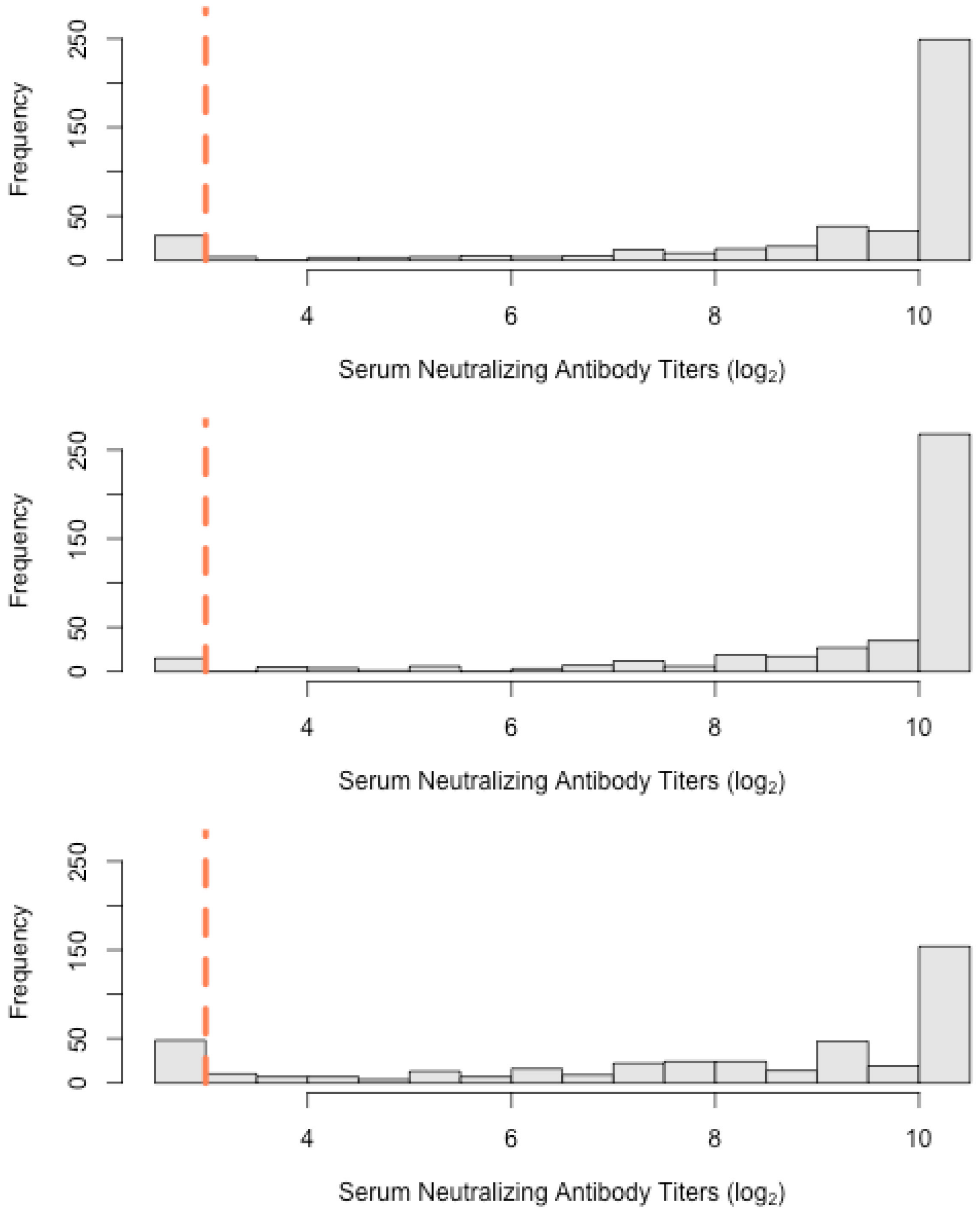


Figure 3.1: Serum neutralizing antibody titers for serotypes (A) 1, (B) 2, and (C) 3. *The red dashed line indicates an LT of 3 (1:8 dilution), the WHO/CDC cut-off for seropositive status.*

The titers were both right- and left-censored data (right at 10.5, left at 2.5) and do not represent a normal distribution, thus they could not be evaluated as a quantitative trait. Instead, the extremes were examined, with seronegative individuals classified as a titer below or equal to 3 (n=48), and a strong seropositive individual having a titer equal to or greater than 7 (n=309).

3.3.2: Genotype Data and Quality Control

DNA was extracted from whole blood at the ICDDR, B and shipped to the University of Virginia for genotyping. Two genome-wide arrays were used: 1M Illumina Duo and the 1M Illumina Quad. The overlap between these two Illumina arrays was 613,778 SNPs. The average call rate was 99.79%. Additional samples were genotyped using Illumina's 2.5M Quad array. To synchronize these three different genotyping arrays all samples were imputed to a 1000 Genomes reference data set using IMPUTE2.(Howie et al. 2012) SNPs were filtered for information content (>90%), minor allele frequency (>0.01) and a Hardy-Weinberg equilibrium (HWE) threshold of *P*<10E-5. The overall SNP and sample genotype missiningness was 5% or less. In addition, individuals with an excess or underrepresentation of heterozygosity were removed. Individuals were examined for identity-by-state clustering to identify duplicates and cryptic relatedness within the program Plink.(Purcell et al. 2007) This left 457 individuals and 6.5 million SNPs.

3.3.3: Analytical Methods

Association analysis was run using the program SNPTEST(Marchini et al. 2007) under an additive frequentist Expectation-Maximization (EM) model. The associations were adjusted for stunting, or a height-for-age Z-score (HAZ) below -2. SNPs were filtered by an information content of the test > 80%, and a minor allele frequency > 5%.

To identify regions of positive selection, cross population extended haplotype homozygosity (XP-EHH) was calculated.(Pickrell et al. 2009) Chromosomes were phased using the SHAPEIT(Delaneau, Zagury, and Marchini 2013) program using the 1000 Genomes phase 1 integrated data set, version 3 as a reference.(Delaneau, Zagury, and Marchini 2013) The genome was phased by using the genomic data and creating haplotypes. XP-EHH requires a reference population that is different from the study population for comparison, we used the 1000 Genomes African population (Yoruba (YRI) and Luhya (LWK)).

Both measures were standardized separately across all chromosomes. Because iHS is dependent upon allele frequency, it must be standardized within minor allele frequency bins genome-wide. We used bins with 5% frequency increments (5-10%, 10-15%, etc). From this standardized distribution a *P*-value was calculated under a normal distribution with a mean of 0 and standard deviation of 1.

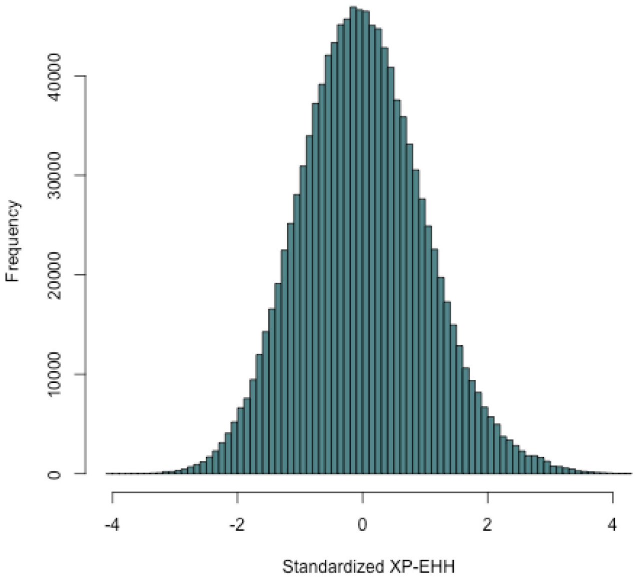


Figure 3.2: Distribution of Standardized XP-EHH (sXP-EHH). *After standardization, the XP-EHH estimates followed a normal distribution.*

Regions of interest were identified within three scenarios: the GWAS alone, the measures of selection alone (either sXP-EHH or stIHS), and the joint association between the two. The top associations for each scenario was investigated. To determine the joint association, regions with a GWAS p-value below 0.001 and an sXP-EHH p-value below 0.01 were used to filter for candidate regions. Fisher's combination test was used to combine the two p-values into an aggregate signal.

3.4: Results

3.4.1: Genome-wide association study

No genome-wide associated regions reached the threshold of significance (*P*<5x10^{-8}), but the top results were promising. (Figure 3.3, Table 3.1) The two main associations are on chromosomes 14 and 7. The top association on chromosome 14, rs113427985, was found 23 kilobases (kb) upstream of *MAPK1IP1L* (mitogen-activated protein kinase 1 interacting protein 1-like) (Figure 3.4). For each additional minor allele (T) an individual was less likely (OR=0.22) to be seropositive (*P*=2.9x10^{-6}) compared to those with the major allele (C; minor allele frequency = 0.07). Sixty kilobases away another association was identified on chromosome 14 was at rs112185488, within *SOCS4* (suppressor of cytokine signaling 4). A similar effect size was found with each additional minor allele (C) resulting in decreased odds of being seropositive (OR=0.21, *P*=5.8x10^{-6}). On chromosome 7 66kb upstream of *SHH*, or Sonic Hedgehog (Figure 3.5) at rs55906254 the minor allele was found to be associated with decreased odds of an adequate response (OR=0.31, P=3.6x10^{-6}).

Table 3.1: Top 20 Results from GWAS

SNP	*Chr*	*Position*	*A1*	*A2*	*All MAF*	*SP MAF*	*SN MAF*	*OR*	*P*	*Genes*
rs113427985	14	55560164	C	T	0.07	0.05	0.22	0.22	2.89E-06	MAPK1IP1L(dist=23252),LGALS3(dist=35771)
rs78866519	14	55561453	C	T	0.07	0.05	0.22	0.22	3.08E-06	MAPK1IP1L(dist=24541),LGALS3(dist=34482)
rs79358122	14	55562841	A	G	0.07	0.05	0.22	0.22	3.24E-06	MAPK1IP1L(dist=25929),LGALS3(dist=33094)
rs77273572	14	55563834	G	T	0.07	0.05	0.22	0.22	3.35E-06	MAPK1IP1L(dist=26922),LGALS3(dist=32101)
rs111628620	14	55566290	G	A	0.07	0.05	0.22	0.22	3.40E-06	MAPK1IP1L(dist=29378),LGALS3(dist=29645)
rs6541250	1	231173427	C	T	0.25	0.28	0.09	5.19	3.41E-06	FAM89A
rs55906254	7	155664061	C	T	0.50	0.47	0.70	0.31	3.61E-06	SHH(dist=59094),LOC285889(dist=566422)
rs79749285	11	84484264	G	A	0.11	0.08	0.25	0.23	4.13E-06	DLG2
rs112185488	14	55507179	T	C	0.07	0.05	0.21	0.22	5.82E-06	SOCS4
rs78575209	14	55505487	A	T	0.07	0.05	0.21	0.22	6.35E-06	SOCS4
rs6459953	7	155668247	A	T	0.46	0.49	0.25	0.34	6.60E-06	SHH(dist=63280),LOC285889(dist=562236)
rs112642967	14	55501802	C	T	0.07	0.05	0.21	0.22	6.68E-06	SOCS4
rs75495314	14	55502757	T	C	0.07	0.05	0.21	0.22	6.69E-06	SOCS4
rs111366012	14	55504297	C	T	0.07	0.05	0.21	0.22	6.71E-06	SOCS4
rs74364684	14	55500486	T	C	0.07	0.05	0.21	0.22	6.81E-06	SOCS4
rs76503733	14	55498451	A	G	0.07	0.05	0.21	0.22	7.24E-06	SOCS4
rs4716555	7	155665755	T	C	0.46	0.49	0.25	0.35	7.57E-06	SHH(dist=60788),LOC285889(dist=564728)
rs76518514	14	55495866	T	A	0.07	0.05	0.21	0.22	7.84E-06	SOCS4
rs12690728	7	155667439	A	T	0.39	0.43	0.19	2.96	7.98E-06	SHH(dist=62472),LOC285889(dist=563044)
rs112457757	14	55539891	G	A	0.07	0.05	0.20	0.21	8.10E-06	MAPK1IP1L(dist=2979),LGALS3(dist=56044)

*SNP= Single Nucleotide Polymorphism, Chr=Chromosome, MAF=Minor Allele Frequency, SP MAF= MAF in Seropositive Group, SN MAF=MAF in Seronegative Group, OR= Odds Ratio

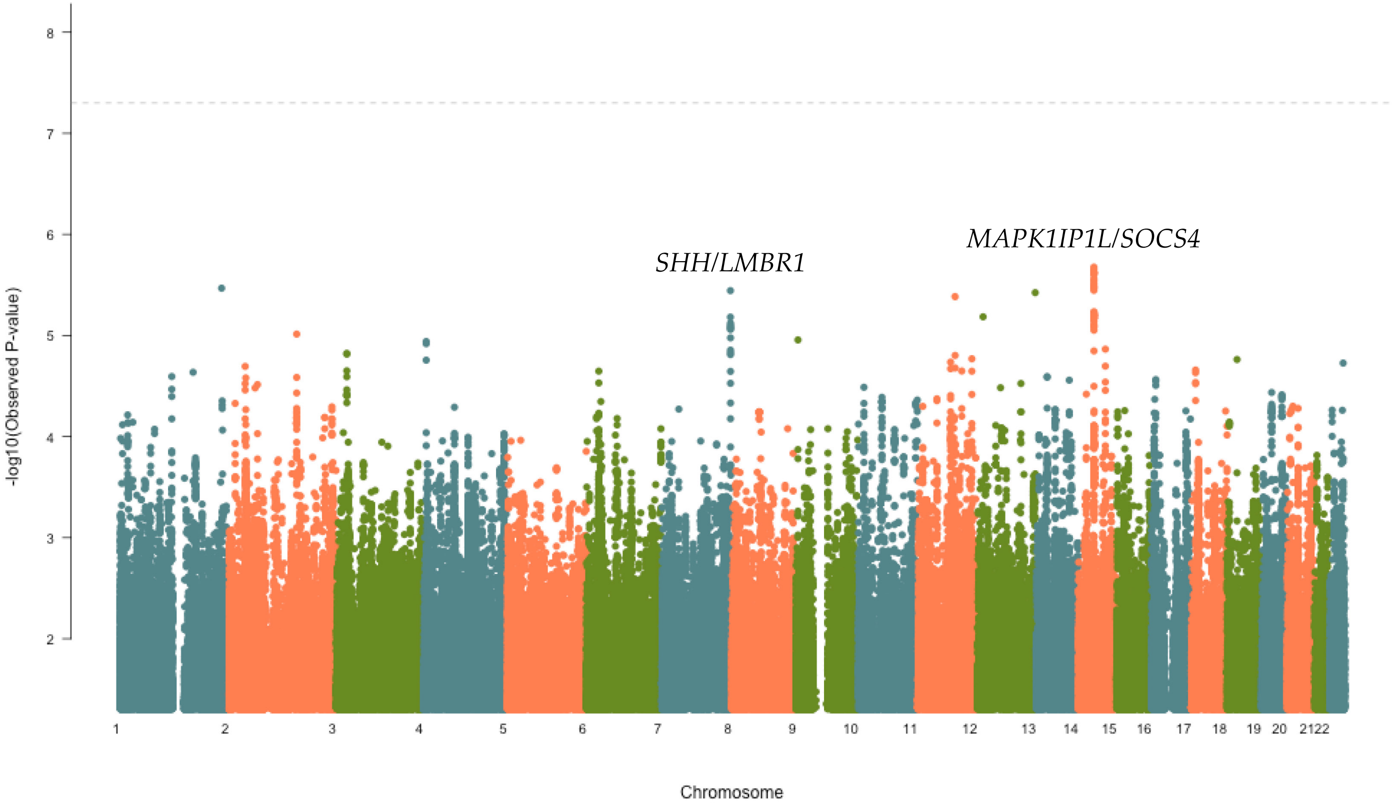


Figure 3.3: Manhattan Plot of GWAS Results for OPV Serotype 3 at 12 Months, Adjusted for Stunting. *The y-axis indicates significance in the form of –log*$_{10}$ *transformed P-values, and the x-axis is organized by chromosome (different colors) and position. The grey dashed line indicates genome-wide significance at 5x10*$^{-8}$.

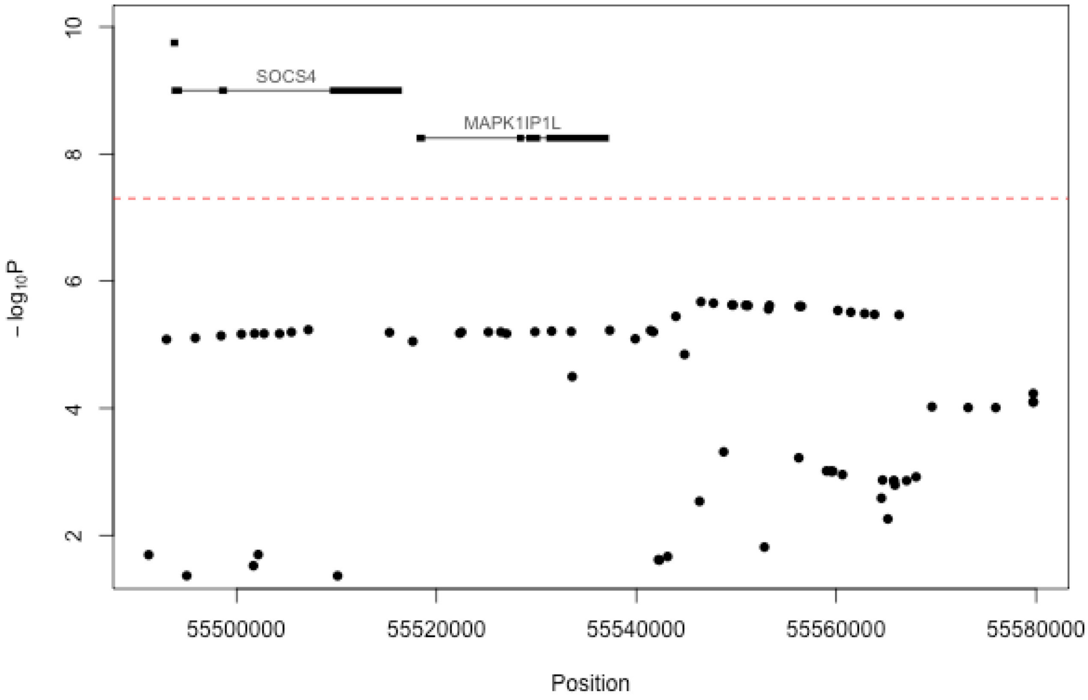


Figure 3.4: Association Results for Chromosome 14 Region. *The y-axis indicates the significance of the SNP-level P-values in terms of a –log$_{10}$ transformation, with the x-axis indicate position along chromosome 14. The red line indicates genome-wide significance of 5x10^{-8}. Genes are annotated above in black, with thicker lines symbolizing exons.*

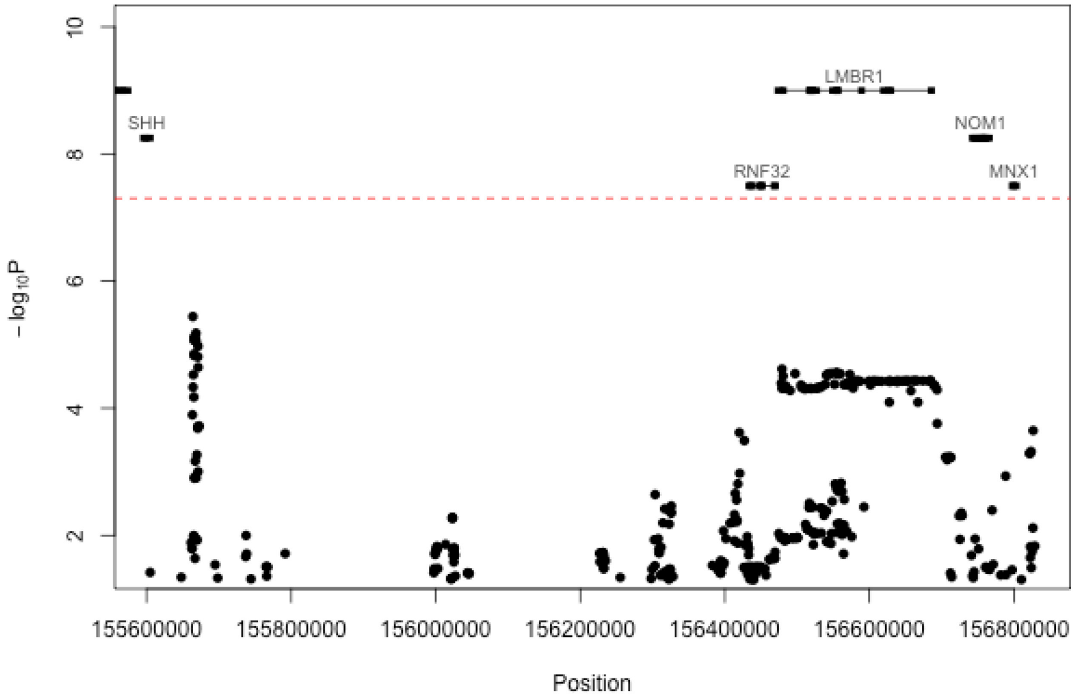


Figure 3.5: Association Results for Chromosome 7 Region. *The y-axis indicates the significance of the SNP-level P-values in terms of a –log*$_{10}$ *transformation, with the x-axis indicate position along chromosome 7. The red line indicates genome-wide significance of 5x10*$^{-8}$*. Genes are annotated above in black, with thicker lines symbolizing exons.*

3.4.2: Selection Scan

At 1,158,046 locations across the human genome, XP-EHH was calculated to detect signals of positive selection. Using two African populations from HapMap as a reference population for Bangladesh, the mean unstandardized XP-EHH was 0.57 with a standard deviation of 0.44. Since the mean genome-wide unstandardized XP-EHH was greater than 0 (0.57) it indicates that this Bangladeshi population has longer haplotype lengths than the Yoruba. This is expected as it has previously been noted that African populations have shorter haplotype blocks when compared to non-African populations due to their older age and decaying linkage disequilibrium.(Tishkoff and Williams 2002) For statistical evaluation, XP-EHH was standardized to the empirical distribution of statistics. From this standardized distribution a *P*-value was calculated assuming a normal distribution with a mean of 0 and standard deviation of 1.

The strongest signal was found within *BVES* at rs9391267 on chromosome 6 with an sXP-EHH of 4.29 (P=1.19x10^{-6}) (Table 3.2). *BVES*, or blood vessel epicardial substance, is also called *POPCD1* (popeye domain-containing protein 1). Another top region was on chromosome 1 within *EIF2C1*, now denoted *AGO1*, for argonaute RISC catalytic component 1. With an sXP-EHH of 4.23 (P=1.19x10^{-5}), haplotypes in this region are longer in this Bangladeshi population when compared to the Yoruba. In total, there were 9 different regions with an absolute value of sXP-EHH > 4. These results confirm prior findings for selection. (Tang, Thornton, and Stoneking 2007; Pickrell et al. 2009; Voight et al. 2006)

Table 3.2: Top 20 Selection Scan Results from Standardized XP-EHH

Chr	*Position*	*SNP*	*A1*	*A2*	*MAF*	*-log(HWE-P)*	*XP-EHH*	*XP-EHH,P-value*	*Region*	*Gene*
6	105565251	rs9391267	G	A	0.16	0.81	4.29	8.85E-06	intronic	BVES
6	105566414	rs9500032	G	A	0.16	0.67	4.28	9.19E-06	intronic	BVES
6	105562120	rs2001119	G	A	0.16	0.67	4.27	9.87E-06	intronic	BVES
6	105568685	rs9404601	G	A	0.15	1.00	4.26	1.00E-05	intronic	BVES
6	105569970	rs12523767	T	A	0.08	1.25	4.26	1.03E-05	intronic	BVES
6	105606018	rs768781	T	C	0.08	0.46	4.23	1.19E-05	ncRNA_UTR3	POPDC3
1	36363475	rs636832	A	G	0.18	0.12	4.23	1.19E-05	intronic	AGO1
6	105561560	rs9486037	C	A	0.15	1.00	4.22	1.22E-05	intronic	BVES
6	105583387	rs9500040	A	G	0.07	0.15	4.21	1.28E-05	intronic	BVES
6	105585511	rs9404605	G	A	0.16	0.52	4.21	1.30E-05	upstream	BVES,BVES-AS1
6	105559609	rs1018810	T	C	0.15	1.02	4.20	1.32E-05	intronic	BVES
3	96789865	rs7640007	A	G	0.08	0.30	4.20	1.35E-05	intronic	EPHA6
6	105558337	rs9322831	G	A	0.15	1.00	4.20	1.35E-05	intronic	BVES
1	36359669	rs2296470	G	A	0.14	0.25	4.19	1.37E-05	exonic	AGO1
6	105591282	rs6571219	G	A	0.07	0.15	4.19	1.38E-05	ncRNA_intronic	BVES-AS1
6	105596568	rs1933236	G	A	0.07	0.30	4.19	1.38E-05	ncRNA_intronic	BVES-AS1
6	105600322	rs6924620	C	T	0.08	1.25	4.19	1.42E-05	ncRNA_intronic	BVES-AS1
6	105599671	rs4626463	G	A	0.08	1.69	4.19	1.42E-05	ncRNA_intronic	BVES-AS1
3	96790746	rs9847081	G	T	0.08	0.30	4.18	1.43E-05	intronic	EPHA6
6	105595261	rs1190274	G	A	0.07	0.15	4.18	1.44E-05	ncRNA_intronic	BVES-AS1

*Chr=Chromosome, SNP= Single Nucleotide Polymorphism, A1=major allele, A2=minor allele, -log(HWE-P)= P-value associated with Hardy-Weingberg Equilibrium transformed by –log base 10.

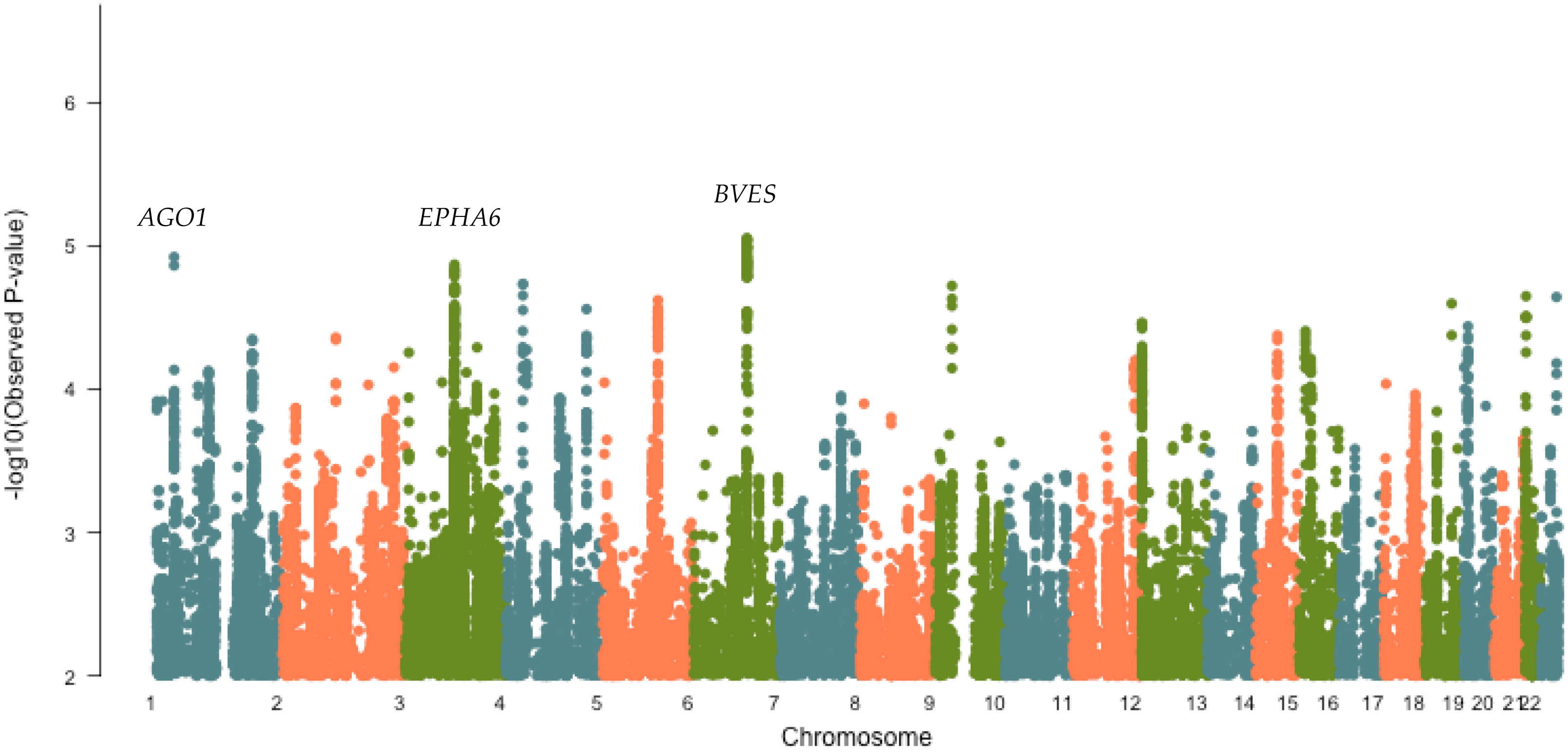


Figure 3.6: Selection Associations for sXP-EHH for the Bangladeshi population. *The Yoruba from Nigeria (HapMap YRI) were used as a reference population. The y-axis indicates significance through a –log₁₀ transformed P-value from the standardized XP-EHH. The x-axis indicates chromosome (by color) and position.*

3.4.3: Regions of Overlap between Selection Scan and GWAS for OPV Response

A total of 32 SNPs in 14 distinct regions overlapped between studies using a threshold of *P*<0.001 for the GWAS and *P*<0.01 for the selection scan (Table 3.3). Half of these SNPs (16/32) were found on chromosome 16 between *FAM86A* and *RBFOX1* at 16p13.3. Within this region, the SNP with the most significant *P*-value from the GWAS, rs11076928 (OR=2.62, P_{GWAS}=8x10^{-5}, $P_{sXP-EHH}$=1.6x10^{-3}), is within a retained intron of a non-coding transcript RP11-420N3.2. This SNP had a standardized XP-EHH of 2.94, indicating longer haplotype lengths when compared to the Yoruba. Each additional minor allele conferred 2.6 times the odds of having a high seropositive response to OPV versus being seronegative.

There were four other regions that had more than one signal within these 32 SNPs. Two signals were on 6q27 between *FRMD1* and *DACT2*. The top associated SNP in this region rs2054476 has a standardized XP-EHH of -2.49 (*P*=0.006), which indicates shorter haplotype lengths than the Yoruba. Located 22 kilobases upstream of *DACT2* (disheveled-binding antagonist of beta-catenin 2), the minor allele of this SNP (A) was associated with decreased odds of seropositivity (OR=0.39, P=1.7x10-4) or individuals were less likely to mount a strong immune response to OPV if they carried 1 or 2 copies of the A allele. The top dual GWAS and selection scan association was in *DOCK10* (dedicator of cytokinesis 10). The SNP rs9989765 had a standardized XP-EHH of -2.71 (P=3.4x10^{-3}) and a GWAS *P*-value of 1.7x10^{-5}. The Odds Ratio was large (4556) due to the minor allele frequency in the seronegative individuals being very small (<1%) while the high seropositive individuals reflected the general population with a minor allele frequency of 8%. Therefore, having the minor allele of this SNP (C) made an individual very likely to mount a high immune response to OPV.

Table 3.3: Cross-Method Associations between the Selection Scan and GWAS

Chr	Position	A 1	A 2	SNP	MAF	sXP- EHH	sXP-EHH P-Value	OR	GWAS P-Value	FCT P-Value	Gene
2	225850923	T	C	rs9989765	0.08	-2.71	3.37E-03	4556.45	1.72E-05	1.02E-06	DOCK10
16	5598818	G	A	rs11076928	0.28	2.94	1.64E-03	2.62	8.01E-05	2.21E-06	FAM86A(dist=451029), RBFOX1(dist=470314)
11	122458831	T	C	rs6589931	0.39	-2.38	8.63E-03	2.66	2.26E-05	3.21E-06	MIR100HG(dist=385061), UBASH3B(dist=67567)
12	52915172	G	T	rs89962	0.22	-2.35	9.32E-03	4.07	3.29E-05	4.91E-06	KRT5
16	5599065	T	C	rs4387604	0.17	2.84	2.27E-03	0.35	1.51E-04	5.44E-06	FAM86A(dist=451276), RBFOX1(dist=470067)
16	5587922	G	A	rs12930002	0.17	2.60	4.62E-03	0.34	1.08E-04	7.74E-06	FAM86A(dist=440133) ,RBFOX1(dist=481210)
16	5586594	T	C	rs3893314	0.17	2.57	5.04E-03	0.34	1.10E-04	8.51E-06	FAM86A(dist=438805), RBFOX1(dist=482538)
16	5590073	G	T	rs1486422	0.17	2.56	5.23E-03	0.34	1.09E-04	8.74E-06	FAM86A(dist=442284), RBFOX1(dist=479059)
16	5598466	G	C	rs11639793	0.17	2.60	4.71E-03	0.35	1.43E-04	1.02E-05	FAM86A(dist=450677), RBFOX1(dist=470666)
16	5594270	A	G	rs11076925	0.17	2.52	5.85E-03	0.35	1.21E-04	1.08E-05	FAM86A(dist=446481), RBFOX1(dist=474862)
16	5593455	A	G	rs11648316	0.17	2.57	5.08E-03	0.35	1.48E-04	1.13E-05	FAM86A(dist=445666), RBFOX1(dist=475677)
16	5604602	A	G	rs3927119	0.17	2.90	1.89E-03	0.37	3.98E-04	1.14E-05	FAM86A(dist=456813), RBFOX1(dist=464530)
16	5602219	C	G	rs8058741	0.17	2.86	2.15E-03	0.37	3.57E-04	1.16E-05	FAM86A(dist=454430), RBFOX1(dist=466913)
16	5607327	C	T	rs11645332	0.17	2.89	1.95E-03	0.37	3.97E-04	1.17E-05	FAM86A(dist=459538), RBFOX1(dist=461805)
16	5603855	A	C	rs11646049	0.17	2.89	1.95E-03	0.37	3.98E-04	1.17E-05	FAM86A(dist=456066), RBFOX1(dist=465277)

16	5602467	G	A	rs8057985	0.17	2.86	2.15E-03	0.37	3.77E-04	1.22E-05	FAM86A(dist=454678), RBFOX1(dist=466665)
16	5602716	C	T	rs8063667	0.17	2.87	2.04E-03	0.37	3.98E-04	1.22E-05	FAM86A(dist=454927), RBFOX1(dist=466416)
16	5598729	C	G	rs4442812	0.17	2.54	5.55E-03	0.35	1.47E-04	1.23E-05	FAM86A(dist=450940), RBFOX1(dist=470403)
16	5606961	A	G	rs11640260	0.17	2.81	2.49E-03	0.37	3.96E-04	1.46E-05	FAM86A(dist=459172), RBFOX1(dist=462171)
4	29909867	A	C	rs16882465	0.09	2.74	3.06E-03	10.30	3.54E-04	1.60E-05	MIR4275(dist=1088577), PCDH7(dist=812163)
6	168685559	G	A	rs2054476	0.40	-2.49	6.36E-03	0.39	1.74E-04	1.62E-05	FRMD1(dist=205720), DACT2(dist=22025)
9	124985785	G	A	rs10818652	0.45	-2.89	1.95E-03	2.24	5.98E-04	1.71E-05	LHX6
20	53497150	G	A	rs12329616	0.15	2.49	6.43E-03	0.36	2.00E-04	1.87E-05	DOK5(dist=229440), CBLN4(dist=1075263)
19	4571589	C	T	rs9304911	0.28	-2.51	5.97E-03	2.32	2.60E-04	2.23E-05	SEMA6B(dist=11818), TNFAIP8L1(dist=67938)
5	112152920	C	T	rs17164132	0.09	2.42	7.69E-03	11.94	2.04E-04	2.25E-05	APC
14	75729732	G	A	rs17183482	0.26	-2.72	3.25E-03	0.42	4.92E-04	2.29E-05	TMED10(dist=86383), FOS(dist=15749)
6	168685533	A	G	rs9346682	0.40	-2.33	9.95E-03	0.39	1.74E-04	2.47E-05	FRMD1(dist=205694), DACT2(dist=22051)
14	75709566	T	C	rs8013918	0.36	-2.63	4.33E-03	0.44	5.23E-04	3.17E-05	TMED10(dist=66217), FOS(dist=35915)
20	53493309	T	C	rs6023667	0.14	2.50	6.19E-03	0.39	6.61E-04	5.48E-05	DOK5(dist=225599), CBLN4(dist=1079104)
10	552355	C	T	rs11252842	0.05	2.47	6.73E-03	0.26	8.99E-04	7.87E-05	DIP2C
5	162031048	T	C	rs7708539	0.08	2.42	7.74E-03	0.30	1.00E-03	9.88E-05	GABRG2(dist=448503), CCNG1(dist=833529)
20	12926358	C	T	rs3903702	0.14	-2.38	8.76E-03	2.76	9.07E-04	1.01E-04	BTBD3(dist=1019115), SPTLC3(dist=63269)

3.5: Discussion

A GWAS of extreme responses to oral poliovirus vaccine in Bangladeshi children revealed two associations on 7q36.3 and 14q22-23. The chromosome 7 signal highlighted the role of *SHH* and a cis-regulatory element in *LMBR1*. Within intron 5 of *LMBR1* lies a long-range cis-regulatory sequence for *SHH.*(Lettice et al. 2002) and mutations within this intron are known to alter SHH expression.(Furniss et al. 2008) Previous studies of selection have found evidence of balancing selection within this regulatory region.(He et al. 2008) *SHH* is a gastric morphogen that drives epithelial cell differentiation. After acute injury, it helps to reconstruct the gastric epithelium.(Xiao et al. 2012) Other studies have shown that after infection with *Helicobacter pylori*, an enteric pathogen, the regeneration of the gastric epithelium is accompanied by the re-expression of *SHH*.(Nishizawa et al. 2007) A study done in mice found that a higher concentration of the SHH protein resulted in increased expression of the human poliovirus receptor, or CD155.(Solecki 2002) This direct link between the sonic hedgehog signaling pathway and poliovirus indicates that *SHH* may be important for the development of immunity against polio.

The other signal is near *SOCS4,* which is a negative regulator of cytokine activity, specifically STAT signaling. A study in biliary epithelial cells showed that infection with *Cryptosporidium parvum*, an enteric pathogen, resulted in an interaction between miRNAs (micro RNA) miR-98 and let-7 with *SOCS4* expression.(Hu et al. 2010) The let-7 family were the first microRNAs discovered, and are involved in the epithelial immune response.(Aalaei-andabili and Rezaei 2013) Despite the two top signals on chromosome 14 (rs113427985 and rs112185488) being over 50 kb away from each other and mapping to different genes (*MAPK1IP1L/LGALS3* and *SOCS4*, respectively), they are in high linkage disequilibrium, with an r^2 of 0.94 and a D′ of 0.98. Much of the association signals in this region exhibit high long-range linkage disequilibrium (Figure 3.4). In fact, a SNP (rs17128156, P=2.76x10^{-6}) located 20 kilobases downstream from *MAPK1IP1L* is an expression quantitative trait loci (eQTL) for *SOCS4*.(Zeller et al. 2010) The overall top GWAS association was rs113427985, which is located less than 7 kb away from this eQTL, indicating that it may also be involved in *SOCS4* expression.

The selection scan was performed on all the children in the study from Dhaka, Bangladesh regardless of OPV response outcome. Therefore, the genes under selection are not specific to an OPV response, but rather represent historic evolutionary pressures. The top associated region is on chromosome 6 within *BVES*—a highly conserved transmembrane protein that is expressed primarily in epithelial cells, such as the gut epithelium.(Osler, Smith, and Bader 2006) This region was previously identified under positive selection by looking at Continuous Regions of Tajima's D Reduction (CRTRs) within a European-descent population.(Carlson et al. 2005) Using the Composite of Multiple Signals (CMS), a measure of selection that incorporates both iHS and XP-EHH and other statistics, this region also exhibited signals of selection within a European population (CEU) with a CMS of 7.32 (CMS>3 is considered significant). (Grossman et al. 2013; Grossman et al. 2010; Karlsson et al. 2013) Selection was also high in Asian populations (Chinese and Japanese (CHB/JPT), CMS=9.88) and within an African population (YRI, CMS=6.25) using HapMap Phase II data. Genome-wide association studies have identified SNPs within *BVES* associated with age at menarche and human height, both of which are known to be under selective pressures. (Amato et al. 2011; Treloar and Martin 1990)

AGO1 located on chromosome 1 part of a cluster of closely related genes in this location including argonaute 3 and argonaute 4 that play a role in RNA interference. In our study, this region was under selection with a standardized XP-EHH of 4.19 (P=1.2x10^{-5}). Highly active immunologically, it is part of both the adaptive and innate immune systems. When compared to other studies, this region seems to be under selection in only non-African populations, such as Europeans and to a lesser extent Asian populations. Looking at CRTRs, enrichment was only found within the European populations.(Carlson et al. 2005) This is consistent when examining CMS for the three HapMap Phase II populations. Strong selection is found within CEU (CMS=11.89), and weaker selection in the CHB/JPT populations (CMS=3.53), while there isn't a CMS above 0 for this region within the YRI. This was consistent in a previous study looking at extended haplotype homozygosity (EHH) within a European population.(Tang, Thornton, and Stoneking 2007)

When it came to the overlap between the selection scan and the genome-wide association study, only 32 SNPs had a P_{GWAS}<0.001 and a $P_{sXP-EHH}$<0.01 within 14 distinct regions. Half of the associations were found in an intergenic region on chromosome 16 between *FAM86A* and *RBFOX1*. The region between *FAM86A* (family with sequence similarity 86, member A) and *RBFOX1* (RNA binding protein, fox-1 homolog 1), has previously been implicated in a genome-wide association study of visceral adipose tissue within women.(Fox et al. 2012) *RBFOX1* was also associated with weight, BMI, and fat mass in Hispanic children.(Comuzzie et al. 2012) An additional associated region was between *DACT2* and *FRMD1* on chromosome 6. *DACT2* is part of the TGF-beta receptor-signaling pathway. *FRMD1* (FERM domain containing 1) is associated with IL-2 secretion following smallpox vaccination (Kennedy et al. 2012)

The top region for the dual associations was in *DOCK10*. The minor (derived) allele for this SNP (rs9989765) was not found in any individuals seronegative for OPV antibodies, while it was found in 9.5% of individuals who were seropositive after four doses of OPV. This is consistent with European populations, in which the minor allele frequency (MAF) is 9%, while it is more rare in African populations (MAF=3%). The DOCK proteins are part of a family of Rho GTPase proteins.(Yelo et al. 2008) Inducible by IL-4, the mRNA transcripts of *DOCK10* are mainly expressed in peripheral blood leukocytes.(Yelo et al. 2008) IL-4 is essential for the development of adaptive immunity after vaccination of OPV indicating a potential link between *DOCK10* and the immune response to OPV.(Katrak et al. 1991)

By examining both the genetic polymorphisms that are associated with systemic immunity to OPV administration, as well as signatures of selection, we are able to elucidate genes involved in polio pathogenesis. Because the majority of poliovirus infections do not result in fatal sequelae such as flaccid paralysis, it is hard to justify that the positive selection found is due to poliovirus in its current form. By looking in simians, it was estimated that ancient positive selection acted on CD155, the poliovirus receptor.(Suzuki 2006) Positive selection refers to a beneficial mutation rising in frequency due to its increased fitness. Because positive selection is not likely to have arisen in response to an increased susceptibility to infection, it is likely that this selection was due to the ability to bind with another molecule.(Suzuki 2006) Therefore, the regions under selection and associated with response to OPV may be more universally relevant to the immune response to an enteric pathogen. By examining these regions we may better understand the biological mechanisms that are utilized to develop effective oral vaccines against enteric infection.

References

Aalaei-andabili, Seyed Hossein, and Nima Rezaei. 2013. “Toll Like Receptor (TLR)-Induced Differential Expression of microRNAs (MiRs) Promotes Proper Immune Response Against Infections: a Systematic Review.” *Journal of Infection* (July 26): 1–14. doi:10.1016/j.jinf.2013.07.016.

Amato, Roberto, Gennaro Miele, Antonella Monticelli, and Sergio Cocozza. 2011. “Signs of Selective Pressure on Genetic Variants Affecting Human Height.” Edited by Thomas Mailund. *PLoS ONE* 6 (11) (November 9): e27588. doi:10.1371/journal.pone.0027588.s002.

Carlson, Christopher S, Daryl J Thomas, Michael A Eberle, Johanna E Swanson, Robert J Livingston, Mark J Rieder, and Deborah A Nickerson. 2005. “Genomic Regions Exhibiting Positive Selection Identified From Dense Genotype Data..” *Genome Research* 15 (11) (November): 1553–1565. doi:10.1101/gr.4326505.

Comuzzie, Anthony G, Shelley A Cole, Sandra L Laston, V Saroja Voruganti, Karin Haack, Richard A Gibbs, and Nancy F Butte. 2012. “Novel Genetic Loci Identified for the Pathophysiology of Childhood Obesity in the Hispanic Population.” Edited by Dana C Crawford. *PLoS ONE* 7 (12) (December 14): e51954. doi:10.1371/journal.pone.0051954.s003.

Delaneau, Olivier, Jean-Francois Zagury, and Jonathan Marchini. 2013. “Correspondence.” *Nature Methods* 10 (1) (January 1): 5–6. doi:10.1038/nmeth.2307.

Estívariz, Concepción F, Mark A Pallansch, Abhijeet Anand, Steven GF Wassilak, Roland W Sutter, Jay D Wenger, and Walter A Orenstein. 2013. “Poliovirus Vaccination Options for Achieving Eradication and Securing the Endgame.” *Current Opinion in Virology* 3 (3) (June 1): 309–315. doi:10.1016/j.coviro.2013.05.007.

Fox, Caroline S, Yongmei Liu, Charles C White, Mary Feitosa, Albert V Smith, Nancy Heard-Costa, Kurt Lohman, et al. 2012. “Genome-Wide Association for Abdominal Subcutaneous and Visceral Adipose Reveals a Novel Locus for Visceral Fat in Women.” Edited by Molly Bray. *PLoS Genetics* 8 (5) (May 10): e1002695. doi:10.1371/journal.pgen.1002695.s005.

Fumagalli, Matteo, Manuela Sironi, Uberto Pozzoli, Anna Ferrer-Admetlla, Anna Ferrer-Admettla, Linda Pattini, and Rasmus Nielsen. 2011. “Signatures of Environmental Genetic Adaptation Pinpoint Pathogens as the Main Selective Pressure Through Human Evolution..” *PLoS Genetics* 7 (11) (November): e1002355. doi:10.1371/journal.pgen.1002355.

Fumagalli, Matteo, Uberto Pozzoli, Rachele Cagliani, Giacomo P Comi, Nereo Bresolin, Mario Clerici, and Manuela Sironi. 2010. “Genome-Wide Identification of Susceptibility Alleles for Viral Infections Through a Population Genetics Approach..” *PLoS Genetics* 6 (2) (February): e1000849. doi:10.1371/journal.pgen.1000849.

Furniss, D, L A Lettice, I B Taylor, P S Critchley, H Giele, R E Hill, and A O M Wilkie. 2008. “A Variant in the Sonic Hedgehog Regulatory Sequence (ZRS) Is Associated with Triphalangeal Thumb and Deregulates Expression in the Developing Limb.”

Human Molecular Genetics 17 (16) (May 7): 2417–2423. doi:10.1093/hmg/ddn141.

Grossman, Sharon R, Ilya Shlyakhter, Ilya Shylakhter, Elinor K Karlsson, Elizabeth H Byrne, Shannon Morales, Gabriel Frieden, et al. 2010. “A Composite of Multiple Signals Distinguishes Causal Variants in Regions of Positive Selection..” *Science* 327 (5967) (February 11): 883–886. doi:10.1126/science.1183863. http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=20056855&retmode=ref&cmd=prlinks.

Grossman, Sharon R, Kristian G Andersen, Ilya Shlyakhter, Shervin Tabrizi, Sarah Winnicki, Angela Yen, Daniel J Park, et al. 2013. “Identifying Recent Adaptations in Large-Scale Genomic Data.” *Cell* 152 (4) (February 14): 703–713. doi:10.1016/j.cell.2013.01.035.

Habib, M A, S Soofi, N Ali, R W Sutter, M Palansch, H Qureshi, T Akhtar, N A Molodecky, H Okayasu, and Zulfiqar A Bhutta. 2013. “A Study Evaluating Poliovirus Antibodies and Risk Factors Associated with Polio Seropositivity in Low Socioeconomic Areas of Pakistan.” *Vaccine* 31 (15) (April 8): 1987–1993. doi:10.1016/j.vaccine.2013.02.003.

He, Fang, Dong-Dong Wu, Qing-Peng Kong, and Ya-Ping Zhang. 2008. “Intriguing Balancing Selection on the Intron 5 Region of LMBR1 in Human Population.” Edited by Vincent Macaulay. *PLoS ONE* 3 (8) (August 13): e2948. doi:10.1371/journal.pone.0002948.g003.

Howie, Bryan, Christian Fuchsberger, Matthew Stephens, Jonathan Marchini, and Gon ccedil alo R Abecasis. 2012. “Fast and Accurate Genotype Imputation in Genome-Wide Association Studies Through Pre-Phasing.” *Nature Genetics* (July 22): 1–6. doi:10.1038/ng.2354.

Hu, Guoku, Rui Zhou, Jun Liu, Ai Yu Gong, and Xian Ming Chen. 2010. “MicroRNA-98 and Let-7Regulate Expression of Suppressor of Cytokine Signaling 4 in Biliary Epithelial Cells in Response to Cryptosporidium parvumInfection.” *The Journal of Infectious Diseases* 202 (1) (July): 125–135. doi:10.1086/653212.

John, T Jacob, and Vipin M Vashishtha. 2013. “Eradicating Poliomyelitis: India's Journey From Hyperendemic to Polio-Free Status.” *The Indian Journal of Medical Research* 137 (5): 881.

Karlsson, E K, J B Harris, S Tabrizi, A Rahman, I Shlyakhter, N Patterson, C O'Dushlaine, et al. 2013. “Natural Selection in a Bangladeshi Population From the Cholera-Endemic Ganges River Delta.” *Science Translational Medicine* 5 (192) (July 3): 192ra86–192ra86. doi:10.1126/scitranslmed.3006338.

Katrak, K, B P Mahon, P D Minor, and K H Mills. 1991. “Cellular and Humoral Immune Responses to Poliovirus in Mice: a Role for Helper T Cells in Heterotypic Immunity to Poliovirus..” *The Journal of General Virology* 72 (Pt 5) (May): 1093–1098.

Kennedy, Richard B, Inna G Ovsyannikova, V Shane Pankratz, Iana H Haralambieva, Robert A Vierkant, and Gregory A Poland. 2012. “Genome-Wide Analysis of Polymorphisms Associated with Cytokine Responses in Smallpox Vaccine Recipients.” *Human Genetics* 131 (9) (May 19): 1403–1421. doi:10.1007/s00439-012-1174-2.

Lettice, Laura A, Taizo Horikoshi, Simon J H Heaney, Marijke J van Baren, Herma C van der Linde, Guido J Breedveld, Marijke Joosse, et al. 2002. "Disruption of a Long-Range Cis-Acting Regulator for Shh Causes Preaxial Polydactyly.." *Proceedings of the National Academy of Sciences* 99 (11) (May 28): 7548–7553. doi:10.1073/pnas.112212199.

Marchini, Jonathan, Bryan Howie, Simon Myers, Gil McVean, and Peter Donnelly. 2007. "A New Multipoint Method for Genome-Wide Association Studies by Imputation of Genotypes." *Nature Genetics* 39 (7) (June 17): 906–913. doi:10.1038/ng2088.

Mondal, D, J Minak, M Alam, Y Liu, J Dai, P Korpe, L Liu, R Haque, and W A Petri. 2011. "Contribution of Enteric Infection, Altered Intestinal Barrier Function, and Maternal Malnutrition to Infant Malnutrition in Bangladesh." *Clinical Infectious Diseases : an Official Publication of the Infectious Diseases Society of America* 54 (2) (December 23): 185–192. doi:10.1093/cid/cir807.

Newport, M J, T Goetghebuer, H A Weiss, H Whittle, C-A Siegrist, and A Marchant. 2004. "Genetic Regulation of Immune Responses to Vaccines in Early Life." *Genes and Immunity* 5 (2) (January 22): 122–129. doi:10.1038/sj.gene.6364051.

Nishizawa, Toshihiro, Hidekazu Suzuki, Tatsuhiro Masaoka, Yuriko Minegishi, Eisuke Iwasahi, and Toshifumi Hibi. 2007. "Helicobacter Pylori Eradication Restored Sonic Hedgehog Expression in the Stomach.." *Hepato-Gastroenterology* 54 (75) (April): 697–700.

Ogra, Pearay L, Hiromasa Okayasu, Cecil Czerkinsky, and Roland W Sutter. 2011. "Mucosal Immunity to Poliovirus." *Expert Review of Vaccines* 10 (10) (October): 1389–1392. doi:10.1586/erv.11.106.

Osler, Megan E, Travis K Smith, and David M Bader. 2006. "Bves, a Member of thePopeye Domain-Containing Gene Family." *Developmental Dynamics* 235 (3) (March): 586–593. doi:10.1002/dvdy.20688.

Patriarca, P A, P F Wright, and T J John. 1991. "Factors Affecting the Immunogenicity of Oral Poliovirus Vaccine in Developing Countries: Review.." *Reviews of Infectious Diseases* 13 (5) (September): 926–939.

Paul, Yash. 2007. "Role of Genetic Factors in Polio Eradication: New Challenge for Policy Makers." *Vaccine* 25 (50) (December): 8365–8371. doi:10.1016/j.vaccine.2007.09.068.

Pickrell, J K, G Coop, J Novembre, S Kudaravalli, J Z Li, D Absher, B S Srinivasan, et al. 2009. "Signals of Recent Positive Selection in a Worldwide Sample of Human Populations." *Genome Research* 19 (5) (May 1): 826–837. doi:10.1101/gr.087577.108.

Purcell, Shaun, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel A R Ferreira, David Bender, Julian Maller, et al. 2007. "PLINK: a Tool Set for Whole-Genome Association and Population-Based Linkage Analyses." *The American Journal of Human Genetics* 81 (3) (September): 559–575. doi:10.1086/519795.

Racaniello, Vincent R. 2006. "One Hundred Years of Poliovirus Pathogenesis." *Virology* 344 (1) (January): 9–16. doi:10.1016/j.virol.2005.09.015.

Roychoudhury, A K, and M Nei. 1985. "Genetic Relationships Between Indians and Their Neighboring Populations." *Human Heredity*.

Sabin, Albert B, Manuel Ramos-Alvarez, José Alvarez-Amezquita, William Pelon, Richard H Michaels, Ilya Spigland, Meinrad A Koch, Joan M Barnes, and Johng S

Rhim. 1960. “Live, Orally Given Poliovirus Vaccine: EFfects of Rapid Mass Immunization on Population Under Conditions of Massive Enteric Infection with Other Viruses.” *JAMA : the Journal of the American Medical Association* 173 (14): 1521–1526.

Solecki, D J. 2002. “Expression of the Human Poliovirus Receptor/CD155 Gene Is Activated by Sonic Hedgehog.” *Journal of Biological Chemistry* 277 (28) (April 30): 25697–25702. doi:10.1074/jbc.M201378200.

Suzuki, Yoshiyuki. 2006. “Ancient Positive Selection on CD155 as a Possible Cause for Susceptibility to Poliovirus Infection in Simians.” *Gene* 373 (May): 16–22. doi:10.1016/j.gene.2005.12.016.

Tang, Kun, Kevin R Thornton, and Mark Stoneking. 2007. “A New Approach for Using Genome Scans to Detect Recent Positive Selection in the Human Genome.” *PLoS Biology* 5 (7): e171. doi:10.1371/journal.pbio.0050171.st013.

Tishkoff, Sarah A, and Scott M Williams. 2002. “Genetic Analysis of African Populations: Human Evolution and Complex Disease..” *Nature Reviews Genetics* 3 (8) (August): 611–621. doi:10.1038/nrg865.

Treloar, S A, and N G Martin. 1990. “Age at Menarche as a Fitness Trait: Nonadditive Genetic Variance Detected in a Large Twin Sample..” *American Journal of Human Genetics* 47 (1) (July): 137–148.

Voight, Benjamin F, Sridhar Kudaravalli, Xiaoquan Wen, and Jonathan K Pritchard. 2006. “A Map of Recent Positive Selection in the Human Genome.” *PLoS Biology* 4 (3): e72. doi:10.1371/journal.pbio.0040072.st004.

WHO. 2013. “WHO Polio Fact Sheet” (April 30): 1–3.

World Health Organization Collaborative Study Group on Oral Poliovirus Vaccine. 1995. “Factors Affecting the Immunogenicity of Oral Poliovirus Vaccine: a Prospective Evaluation in Brazil and the Gambia. .” *The Journal of Infectious Diseases* 171 (5) (May): 1097–1106.

Xiao, Chang, Rui Feng, Amy C Engevik, Jason R Martin, Julie A Tritschler, Michael Schumacher, Robert Koncar, et al. 2012. “Sonic Hedgehog Contributes to Gastric Mucosal Restitution After Injury.” *Laboratory Investigation* 93 (1) (October 22): 96–111. doi:10.1038/labinvest.2012.148.

Yelo, Estefanía, María Victoria Bernardo, Lourdes Gimeno, María José Alcaraz-García, María Juliana Majado, and Antonio Parrado. 2008. “Dock10, a Novel CZH Protein Selectively Induced by Interleukin-4 in Human B Lymphocytes.” *Molecular Immunology* 45 (12) (July): 3411–3418. doi:10.1016/j.molimm.2008.04.003.

Zeller, Tanja, Philipp Wild, Silke Szymczak, Maxime Rotival, Arne Schillert, Raphaele Castagne, Seraya Maouche, et al. 2010. “Genetics and Beyond – the Transcriptome of Human Monocytes and Disease Susceptibility.” Edited by Zoltan Bochdanovits. *PLoS ONE* 5 (5) (May 18): e10693. doi:10.1371/journal.pone.0010693.s012.

Zinkernagel, R M, H Hengartner, and L Stitz. 1985. “On the Role of Viruses in the Evolution of Immune Responses..” *British Medical Bulletin* 41 (1) (January): 92–97.

Chapter 4: Background and Review of Gene- and Pathway-Level Methods

4.1: The Success of Genome-wide Association Studies and Limitations

In less than a decade after their advent, genome-wide association studies (GWAS) have been remarkably successful in identifying risk loci for various complex diseases. As of September 2013, the National Human Genome Research Institute (NHGRI) Genome-wide Association Studies (GWAS) Catalog contained 1,673 publications and 11,194 SNP associations. (Hindorff et al. 2009; Hindorff et al. 2013) Working under the hypothesis of "common disease, common variants", GWAS has elucidated many loci that are moderately (Odds Ratio (OR)=1.2) to highly associated (OR>5) with complex phenotypes. However, there is still a large amount of "missing heritability". This missing heritability is the discrepancy between the low amount of within-population variation explained by GWAS results and the higher estimates of narrow-sense heritability, or proportion of phenotypic variance explained by additive genetics.(Vineis and Pearce 2010) One explanation for the missing heritability is that current studies are underpowered to identify variants that may be contributing to the overall heritability. Due to the large number of statistical tests, consideration of multiple comparisons requires conservative adjustment of the significance threshold (alpha) for the 1-2.5

million tests resulting in a threshold of ~5x10-8.(McCarthy et al. 2008) To counteract this limitation, larger sample sizes are needed to achieve adequate power.

Another potential reason for the missing heritability is that GWAS were not designed to uncover all types of associations, but to identify common variants. Under the hypothesis of “common disease, common variant”, the SNPs included in the current GWAS panels have minor allele frequencies (MAFs) on average > 1%. Therefore, rare variants (MAF<0.05) are underpowered for association. Standard analytical methods for GWAS cannot handle low allele counts in a stable manner. Better methods for handling these markers, such as collapsing methods used commonly in sequence analysis, must be developed and evaluated.

Many GWAS have been unable to replicate their findings. This can be due to numerous reasons, such as Type I error in the original analysis or unmeasured confounders in either the original discovery set or replication. It could also be due to allelic heterogeneity, in which different populations will have different alleles within the same locus or gene is associated with the outcome. Therefore, when SNPs are followed-up from the discovery set, they do not replicate even though the same gene may be involved in the pathogenesis of the outcome. The last factor that limits the performance of GWAS is alleles that only have a modest to small effect on the outcome of interest. The infinitesimal model states that there are many common variants of small effect, which contribute to the genetic variance of a phenotype (Figure 4.1).(Gibson 2012) GWAS are poorly equipped to handle these variants, as due to strict significance thresholds these variants would likely never be noticed without enormous sample sizes. Under this model, the heritability is not missing, but rather hidden. The truth is likely in between the CDCV and infinitesimal models, with the missing heritability being due to a finite number of smaller effect variants.(Bloom et al. 2013)

To address both of these limitations, a multitude of gene- and pathway-level analyses have been developed. These methods aggregate markers into biologically relevant units, such as a gene or pathway, and then analyze the effects within that unit. This method allows for allelic heterogeneity, as the exact alleles that are associated with the outcome are not important, only that there is an enrichment of signal in the unit of association. Also, by aggregating multiple signals, this may increase the power in weak or moderate associations. Another motivation to analyze variation at a gene or pathway level is that the analysis yields a biologically interpretable result in terms of the disease pathogenesis. Genes or pathways can be selected based on prior biological knowledge, or evaluated without prior biological information in a genome wide approach. While many of the issues surrounding these analytical methods are similar, the following review will discuss gene and pathway level separately.

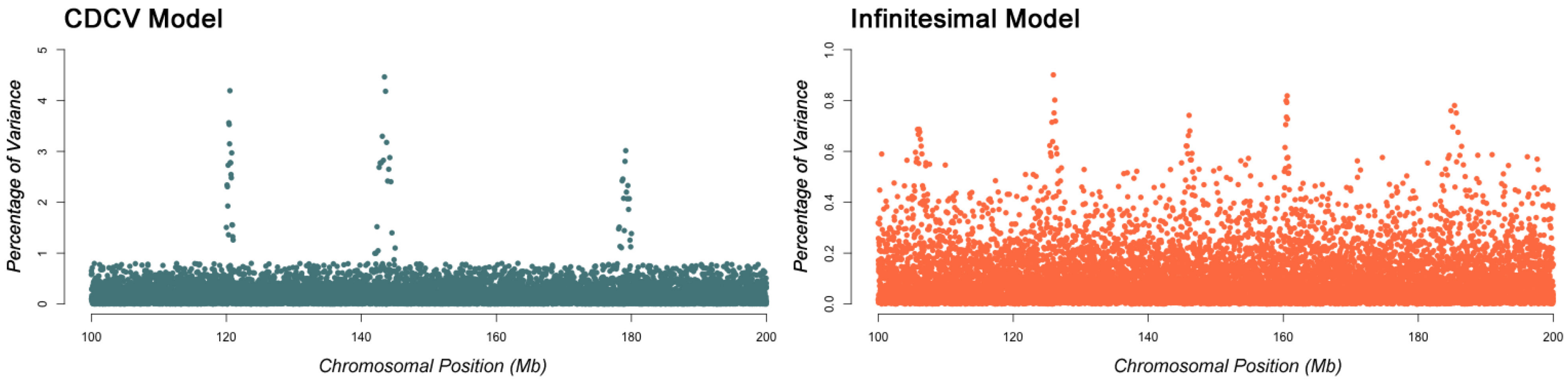


Figure 4.1: The Common Disease, Common Variant (CDCV) Model versus the Infinitesimal Model. *The CDCV model on the left indicates a few common variants being responsible for large proportion of the phenotypic variance (>1%), while the infinitesimal model on the right indicates that many variants (infinite) may be responsible for smaller percentages of the variance (<1%).*

4.2: Gene-Level Review

4.2.1: Methods

The goal of GWAS is to identify genetic variation associated with the phenotype, hopefully implicating a responsible gene. It is difficult to interpret when the significant variants lie in intergenic regions, even with the recent availability of the Encyclopedia of DNA Elements (ENCODE) data highlighting regulatory regions. Limiting markers to genic regions may ignore distant *cis*-regulatory elements or other functional regions associated with a gene, but it also reduces the potential of statistical noise clouding the interpretation of GWAS results. The set of SNPs assigned to a gene can be determined by either the physical location, or the functional variation.(la Cruz et al. 2010) Additionally, there are different methods to handle the correlation structure due to density of SNP coverage and linkage disequilibrium. These methods generally fall into three groups: classical methods, updates to classical methods, and newer methods that directly estimate the correlation structure.

4.2.1.1: SNP Classification

Publicly available databases such as RefSeq (NCBI) or Uniprot provide the physical location of the gene on the chromosomes. The SNPs that are included in these sets are determined by various criteria, such as exonic regions, translated regions, the entire genic region, or flanking regions ranging from 5-200 kilobases (kb). The flanking region size can be determined by the user's priorities. Previous eQTL mapping showed that most cis-regulatory SNPs are within 100kb of the transcribed region, while more than

93% of relevant functional nucleotides are found within 20kb of the transcribed region.(Huang et al. 2011) Because of this, a flanking region of 20 kb from the translated start and end sites is commonly used. Using these criteria, SNPs may contribute to more than one gene. This can be due to overlapping genes, or genes in close proximity having overlapping flanking regions. This will decrease the independence of the tests, and must be taken into consideration when interpreting results.

SNPs within a gene can also be categorized by their functional variation. This can include nonsynonymous SNPs (nsSNPs), variation around the transcription start and end sites, *cis* and *trans*-eQTLs, or variation only found in transcription factor binding sites. These classifications may be less interpretable than the physical location because only a fraction of information is available on known functional variation and the existing databases are not comprehensive.(la Cruz et al. 2010)

4.2.1.2: Classical Methods

1. ***Fisher's Combination Test (FCT)*** (Peng et al. 2009)

   All SNP p-values within the genes are combined, assuming independence. The resulting Z-score follows a X_{2K}^2 distribution, where K indicates the number of SNPs in the unit.

$$Z_F = -2 \sum_{i=1}^{k} \log P_i$$

2. ***Sidak's Combination Test (SCT)*** (Peng et al. 2009)

Only the best SNP in the gene (as determined by the lowest p-value) is used. This is also called minSNP, or Sidak's correction. The Z-score is distributed as follows to correct for the number of SNPs in the unit: $P(Z_B \leq w) = 1 - (1 - w)^K$.

3. ***Simes' Test (ST)*** (Peng et al. 2009)

The SNP p-values are ordered from least to most significant. For each of these p-values, the following adjusted p-value (P_s) is calculated: $k * \frac{P_i}{i}$, where (k) is the ordered position of the original p-value. The minimum P_s is the p-value for the gene.

4. ***False Discovery Rate (FDR)*** (Peng et al. 2009)

The SNP p-values mapped to the gene are ordered and a standard false discovery rate adjustment is applied to account for the number of SNPs within the gene. The minimum ordered false discovery rate is then assigned to the gene. The user must determine what the acceptable significance level is (α).

5. ***Logistic Regression (LR)***

In this standard model, each SNP is coded in the additive format of 0, 1, or 2 copies of the minor allele. The response variable is the case-control status. All SNPs in the gene are included as covariates in this logistic regression. The gene-level p-value is calculated using a likelihood ratio test comparing the full model with all the SNPs to a null model without any SNPs.

6. ***meanT*** (Lehne, Lewis, and Schlitt 2011)

The GWAS test statistics (χ^2) are aggregated over the genic region and the average test statistic is calculated over the entire genic region. Empirical p-values can be determined using multiple phenotype permutations and re-averaging the permuted genic test statistics.

**This method was not freely available, and was therefore not incorporated for further analysis.*

7. ***topQ*** (Lehne, Lewis, and Schlitt 2011)

Using the GWAS test statistics, only the top quartile of test statistics as determined by significance are considered. The mean test statistic of these top quartile SNP test statistics is calculated for the gene test statistic. Empirical p-values can be determined using phenotype permutations and recalculating the average test statistic in the top quartile.

**This method was not freely available, and was therefore not incorporated for further analysis.*

These methods were developed before GWAS and were not meant to handle correlated variables. Many aggregate single marker p-values into one test statistic (FCT, LR, meanT, topQ) that is tested for association against a null model and the markers are assumed to be independent. This assumption is violated with GWAS data due to the high density of markers, many of which are correlated or in linkage disequilibrium. This results in the inflation of test statistics, leading to increased type I error rates. Others only use the most significant SNP from the set, but may assume the SNPs within the gene represent a distribution (FDR). This still requires the lowest-level statistics to be independent of one another.

It should be noted that logistic regression is the only method that requires the raw genotype data for the classical methods. The other classical methods in their original form only require the statistics resulting from a GWAS, such as P-values or X^2 test statistics. However, to control the inflated type I error due to the presence of linkage disequilibrium violating the independence assumption, raw data may be used to run computationally intensive permutations.

4.2.1.3: Updates to the Classical Method

8. ***SLAT*** (la Cruz et al. 2010)

SLAT (Set-Level Association Testing), is related to Fisher's Combination Test. It employs two different basic modifications: truncation and weighting. The truncation consists of only including SNPs that reach a certain significance threshold in the original GWAS. The remaining SNPs are then weighted according to their linkage disequilibrium structure. To account for these two aspects, the Fisher's Combination Test becomes the following:

$$TS_{SLAT} = -\sum^{K} w_i \log(p_i)\, I_{p_i<\alpha_i}\,; w_i = weight\ given\ to\ the\ marker$$

The α used can be adaptive, or the same for all genes. The weights can be either LD, or possible functional relevance.

**This method was not freely available, and was therefore not incorporated for further analysis.*

9. ***GATES*** (M.-X. Li et al. 2011)

GATES is a Gene-based Association Test using Extended Simes procedure. The original Simes' test is detailed above. The altered p-value is as below:

$$P_{GATES} = \min\left(\frac{m_e p_{(j)}}{m_{e(j)}}\right)$$

The modification to the original Simes' test is that m_e is the effective number of independent p-values among the *m* SNPs, and $m_{e(j)}$ is the effective number of independent p-values among the top *j* SNPs. This is to account for the assumption inherent in the Simes' test, which requires the input to be the results from independent tests. The value of m_e is determined through a new approach using the following procedure of principal components analysis:

$$m_e = M - \sum_{i=1}^{M}[I(\lambda_i > 1)(\lambda_i - 1)]\lambda_i > 0$$

In this equation λ_i is the i^{th} eigenvalue of the p-value correlation coefficient matrix of the SNP-based statistic tests. With this procedure, negative eigenvalues are ignored by setting it as zero, which should be rare and only arises in the presence of missing data. If the SNPs are all independent, then the eigenvalues should all be 1 and they are all weighted equally.

10. ***aSUM*** (Han and Pan 2010)

This method combines the logistic regression, as well as the sum test, into an adaptive framework in five steps. The first step requires the original data, in which a marginal regression model is fit to each individual SNP, obtaining a $\beta_{M,j}$

and a $p_{M,j}$. The second step uses a pre-defined initial significance threshold, α_0, to reclassify SNPs. If $\beta_{M,j} < 0$ and $p_{M,j} < \alpha_0$, then the alleles are reclassified as the number of minor alleles -2. The other alleles are unchanged. In the third step, the new data is fitted on a common-effect model with a usual score statistic U, with its associated variance and p-value. The fourth step consists of permuting the disease variable, repeating steps 1-3. In the final step, the aSUM test statistic is calculated from the sample mean and variance from the permutations.

These methods have taken the classical methods described above, and altered them to account for the genetic architecture of the gene and the violation of the independence assumption found in GWAS data due to linkage disequilibrium. SLAT handles linkage disequilibrium by weighting SNPs based on their relative linkage disequilibrium, while GATES estimates the number of independent and representative SNPs

4.2.1.4: Methods that directly estimate correlation structures

11. ***Linear Combination Test (LCT)*** (Luo et al. 2010)

The LCT directly estimates the correlation matrix of the association statistics from the GWAS, and then transforms the association statistics by the inverse of the correlation matrix. This down-weights statistics that are highly correlated while up-weighting independent statistics. The equations for this are as follows:

$$e = (1,1,1,1,\dots,1)^T$$

$$R_g = correlation\ matrix\ of\ Z = \left(Corr(x_i - y_i, x_j - y_j)\right)_{k*k}$$

$$Z_i = \phi^{-1}(1 - P_i), Z = (Z_1, \dots, Z_k)^T \quad [\phi = standard\ normal\ distribution]$$

$$T_{LCT} = \frac{e'Z}{\sqrt{e'R_g e}}$$

After the transformation, the SNP-level test statistics are summed across the entire genic region.

**This method was not freely available, and was therefore not incorporated for further analysis.*

12. *Quadratic Test* (Luo et al. 2010)

The QT directly estimates the correlation matrix as well, but weights the test statistic matrix differently, instead applying a quadratic approach instead of the previous method's linear approach.

$$T_{QT} = Z^T R_g^{-1} Z$$

This method assumes that the test statistic is asymptotically distributed as a central χ_k^2 distribution. The quadratic approach consists of multiplying the test statistics by each other, instead of summing.

**This method was not freely available, and was therefore not incorporated for further analysis.*

13. *Decorrelation Test* (Luo et al. 2010)

The Decorrelation test (DT) directly transforms the dependent variables into independent variables. Once they are decorrelated, they can be combined using a traditional test, such as Fisher's Combination Test, or Sidak's Combination Test that was previously described. To decorrelate the variables, the following procedure is used:

$$R_g = CC^T; C = nonsingular\ matrix$$

$$W = C^{-1}Z = [W_1, W_2, \dots, W_k]^T$$

$$Cov(W, W) = C^{-1}Cov(Z, Z)(C^T)^{-1} = C^{-1}CC^T(C^T)^{-1}$$

$$W \sim N(0, 1)$$

Now each variable in W are independent and a new p-value can be determined from that distribution using FCT, ST, or any other methods that require independent signals.

**This method was not freely available, and was therefore not incorporated for further analysis.*

14. *VEGAS* (Liu et al. 2010)

VEGAS, or a Versatile Gene-Based Association Study, considers results from a variety of GWAS designs, taking the p-values from the *n* SNPs assigned to the gene and converting them to a series of χ^2_{1df} test statistics. These are then summed across the gene into a χ^2_{ndf} statistic. VEGAS accounts for the linkage disequilibrium present by using simulations from a multivariate normal distribution. A Monte Carlo approach cuts down on the computational resources required. VEGAS takes a gene with *n* SNPs and simulates an *n*-element multivariate normal with the covariance matrix (Σ) being an *n*x*n* matrix of pairwise LD (*r*) values. These variables are then multiplied by the Cholesky decomposition matrix of Σ. This new random vector will have a multivariate normal distribution, which is then transformed into a vector of uncorrelated χ^2_{1df} variable. The final test statistic is then the sum of these values. This is repeated a large number of times, and the empirical p-value is calculated as the proportion

> of these simulated test statistic that are greater than the observed original test statistic. This procedure is known as VEGAS-Sum. An alternative approach within VEGAS is known as VEGAS-Max. This procedure only considers the most significant SNP in the gene for the original test statistic. For each simulation, only the highest simulated test statistic from each run is used to create the empirical distribution. The best method between the VEGAS-Sum and VEGAS-Max tests will depend upon the genetic architecture of the gene.

These methods directly estimate the correlation structure of the SNPs assigned to the genic region. They then transform the association statistics from this region by the correlation structure, or linkage disequilibrium, seen with the markers. The resulting independent signals are combined for an aggregate test statistic.

4.2.2: Limitations

When evaluating these methods, various factors must be taken into consideration. One is the incorporation of potential confounders in the model. Methods that use the GWAS P-values as input can control for these variables by including variables in the original GWAS analysis, such as principal components to control for population substructure or known confounders for the outcome of interest. Methods (like SLAT) that require raw genotypes are unable to control for potential confounders, and therefore may be susceptible to bias in the same way as an unadjusted GWAS.

Another limitation is rare variants. GWAS genotyping panels and methods are not appropriate for rare variant detection and analysis. A simple model testing for the association of a marker with the outcome will be underpowered to detect an association with a rare variant (MAF<1%). Most often, these rare variants will be removed in standard GWAS quality control procedures before any analyses are done. Even if the rare variants are included in subsequent analyses, these methods do not account for the markers allele frequency as they are all weighted equally. An exception to this is aSUM, which was developed for both common and rare variants. While some methods may be able to manually handle weights determined by the user, they are not an inherent part of the method.

A last limitation of these methods is that they are highly dependent upon databases, which are continuously changing, being updated and improved on an irregular basis. Any results that are produced using these methods are therefore contingent on the build of the human genome, as well as the versions of the databases used. This may result in inconsistencies between studies done at different times.

4.2.3: Discussion

Previous literature has evaluated some of the programs described above. Lehne et al compared three basic methods: the most significant statistic from within the gene (Sidak), the mean test statistic of all SNPs (meanT) and the mean of the top quartile of test statistics (topQ).(Lehne, Lewis, and Schlitt 2011) In addition to these “uncontrolled” statistics, an empirical p-value was derived using permutations. They found that the maxT statistic, which only uses the strongest SNP P-value as the gene P-value, is subject to gene size bias. This is because large genes contain more SNPs and therefore are more likely to have a SNP be significant by chance. Because maxT only uses the top SNP, it does not account for this bias. The statistic meanT had the opposite problem, where the smaller genes were subject to extremes due to only have a small number of SNPs. Spurious associations will affect these smaller genes much more than when they may be averaged out with a larger number of SNPs in larger genes. The same problem occurred with topQ, in which smaller genes were found to be on the extremes more often than they would be by chance. All three of these methods performed similarly, with less than 2% difference between their Area Under the Curve (AUC) estimates. Lehne et al conclude that the performance is highly dependent upon the number of SNPs found in the gene, or genic region. When applied to real data, the different methods can rank genes very differently. For example, using a GWAS of Crohn's Disease, the known risk gene of *ZNF365* ranked 18th using maxT, 149th using meanT, and 67th using topQ. This gene is fairly large and had a total of 91 SNPs assigned to the region.

In a more recent study Bacanu and colleagues evaluated 6 different tests: VEGAS, GATES, Simes, aSUM, and a hybrid test that the author proposed.(Bacanu 2012) Using simulations, they determined that the different methods were optimized based on the number of variants, gene lengths. For multiple causal variants in smaller genes, aSUM had the best performance while Simes was the fastest and the best-performing method for single causal variant genes. For longer gene lengths, VEGAS performed better than the other methods. To optimize performance, the authors propose a two-step method, in which Simes is used as the first step to screen for suggestive signals. These genes are then followed up with more computationally intensive methods, aSUM or VEGAS depending on the gene length.

Further evaluation of gene-level methods is required to assess their relative performance in terms of sensitivity and specificity, as well as type I and II error. With nearly 20,000 genes currently cataloged with the National Center for Biotechnology Information (NCBI), multiple comparisons will remain an issue. Therefore, the ideal method would have low type I error to control false positives due to spurious associations. The balance between sensitivity (true positives) and specificity (true negatives) will depend on the priorities of the study. High sensitivity should be desired in the case of high-cost follow-up, in which there are heavier consequences for false positives. On the other hand, if the goal of the study is to generate hypotheses, a high specificity coupled with a lower sensitivity may be adequate.

Gene-level methods were developed to detect genes that were enriched for associations in GWAS. Signals that would otherwise be ignored by the traditional GWAS significance threshold are brought to the forefront allowing further examination. A thorough evaluation of these methods will provide insight into the relative performance of the programs, as well as the questions that could be answered with the application of gene-level methods to GWAS results.

Table 4.1: Review of Gene-Level Methods

Group	*Program/Method*	*Citation*	*Input*	*Output*	*Used in Aim 2a*
Classical	Fisher's Combination Test	(Peng et al. 2009)	SNP P-values	Chi-squared Test Statistic	X
	Sidak's Correction		SNP P-values	Minimum P-value	X
	Simes' Test		SNP P-values	Minimum Ranked P-value	X
	FDR		SNP P-values	Minimum False Discovery Rate	X
	Logistic Regression		Raw Genotype	Likelihood Ratio Test	X
	meanT	(Lehne, Lewis, and Schlitt 2011)	SNP P-values	Average P-value	
	topQ		SNP P-values	Average P-value (from top quartile)	
Updated Classical	SLAT	(la Cruz et al. 2010)	SNP P-values	Chi-squared Test Statistic	
	GATES	(M.-X. Li et al. 2011)	SNP P-values	P-value	X
	aSUM	(Han and Pan 2010)	Raw Genotypes	Empirical P-value	X
Direct Correlation Estimation	LCT	(Luo et al. 2010)	Raw Genotypes	T-Statistic	
	DCT		Raw Genotypes	Chi-squared Test Statistic	
	QT		Raw Genotypes	Normally-distributed Test Statistic	
	VEGAS	(Liu et al. 2010)	SNP P-values	Empirical P-value	X

4.3: Pathway-Level Review

A level higher than genes is grouping markers together within gene "sets". These methods are adapted from gene expression studies, in which gene sets were investigated for enrichment of signal within a ranked list of differential gene expression. The fundamental question of these approaches is different than in the gene-level analyses. Since these methods are typically "enrichment" analyses, they are a way of visualizing GWAS results on a pathway-level. They do not take into account multiple independent signals within a gene, and therefore may not increase power to identify multiple weaker signals. Instead, this approach will use the genes that your mid-level significance GWAS results represent, and summarize the results in an approachable format.

4.3.1: Databases

These gene sets are often genes found in known biological pathways, but can also be determined by protein-protein interaction (PPI) or other bioinformatics-informed networks. For this analysis, we will be focusing on biological pathways, as determined by canonical pathway databases such as the Kyoto Encyclopedia of Genes and Genomics (KEGG) or BioCarta. A brief description of each of these databases is below.

> *KEGG (Kyoto Encyclopedia of Genes and Genomes)*: KEGG is a database created from molecular-level information about understanding the functions and utilities of the biological system. Most of the large-scale molecular datasets were generated by genome sequencing and other high-throughput experimental technologies. Both the PATHWAY and BRITE aspects of KEGG are available.

KEGG PATHWAY details molecular interactions and reactions in manually drawn pathway maps. It uses datasets found in genomics, transcriptomics, proteomics, and metabolomics to inform these pathways. KEGG BRITE draws upon many other different types of relationships, such as bioinformatics and predicted networks. Pathways are classified by functional relevance, such as a particular product, as well as disease-specific pathways.

BioCarta: BioCarta is a commercial company that develops, supplies, and distributes reagents and assays for research. Their pathway database is open source with the academic community integrating emerging proteomic information. It currently has information about >120,000 genes in many different species. Pathways are classified by functional relevance, such as adhesion, apoptosis, and metabolism.

PANTHER (Protein Analysis Through Evolutionary Relationships): PANTHER classifies genes by their functional relevance. It draws upon scientific experimental evidence, and if not available, it uses evolutionary relationships to inform function. These genes are then classified by their families and subfamilies, Gene Ontology classes, PANTHER-specific protein classes, as well as known pathways. It is part of the Gene Ontology Reference Genome Project. It was developed for work with gene expression data. Some pathways are community-curated.

Reactome: Reactome is also an open source database that is manually curated and peer-reviewed. It is cross-referenced to many other databases, such as NCBI

Entrez Gene, Ensembl, and UniProt, as well as the UCSC and HapMap Genome Browsers. They also cross-reference with KEGG and ChEBI small molecule databases, PubMed and Gene Ontology. The focus for this database is the reaction, and therefore it mainly catalogs the small molecules involved in a specific reaction.

Gene Ontology (GO): The GO Consortium consists of a variety of collaborations, including Reactome and PANTHER. It is an effort to catalog and classify various bioinformatic information. It is separated into three groups: biological processes, molecular functions, and cellular components. GO is an ontology, meaning that these processes are not independent, but rather arranged in an hierarchical fashion. Cellular components are parts of a cell or the extracellular environment. Molecular functions detail the elemental activities of gene products. Lastly, a biological process is a set of events that has a start and an end, similar to a canonical biological process.

Molecular Signatures Database (MSigDB): This database is curated by the Broad Institute and includes 6 major collections: positional gene sets, curated gene sets, motif gene sets, computational gene sets, GO gene sets, and oncogenic signatures. MSigDB draws from numerous other databases into one central place. Originally developed to aid with gene expression data and GSEA, it can also be adapted for other uses. This site also hosts the original GSEA software.

4.3.2: Methods

Pathway-level methods differ in the treatment of gene-level associations, the handling of linkage disequilibrium, databases utilized, and the underlying hypotheses. These factors must be taken into account when considering the best, or most appropriate, program.

1. ***ALIGATOR*** (Holmans et al. 2009)

   This program exclusively tests for overrepresentation of association signals in Gene Ontology (GO) categories from a genome-wide association analysis. SNPs are mapped to the GO gene sets and filtered based on a pre-determined significance threshold. The genes that these SNPs represent are then determined to be significant, regardless of the number of SNPs in the gene. Further analyses are restricted to GO categories that have at least 2 significant genes. Replicate gene lists are simulated drawing the same number of SNPs as in the filtered GO categories from the original analysis. From these replicate gene lists, an empirical p-value is calculated. The simulations assume that the LD structure is identical between the different GO categories. A violation of these assumptions will lead to an overly conservative estimate in the presence of high LD. This method only requires the rsID of the SNP, as well as the associated p-value from the GWAS.

2. ***GenGen*** (Wang, Li, and Bucan 2007)

   The first incarnation of GenGen was developed in 2007 as a direct adaptation of the Gene Set Enrichment Analysis (GSEA) methods being used in gene expression analysis. SNPs are assigned to genes, in the coding regions, as well as a 500 kb region around the gene. For each gene, the most significant SNP test statistic is used as the gene test statistic. These gene scores are then sorted by strength of association. Using these rankings, the gene sets are analyzed using a "Kolmogorov-Smirnov-like running-sum statistic". This statistic tests for an overrepresentation of the genes in that set being highly ranked overall. The user provides the gene and pathway mapping, thus this method can be adapted to numerous pathway databases. Standard mapping files are available for some commercial arrays, as well as a composite of GO, BioCarta, and KEGG.

3. ***Gene Set-based Analysis of Polymorphisms (GeSBAP)*** (Medina et al. 2009)

GeSBAP is flexible with user input. It takes SNP-level p-values, gene-level p-values, or raw genotype data in Plink format.(Purcell et al. 2007) Gene Ontology, KEGG and Biocarta pathways are used for the analysis. SNPs are mapped to genes using a 5 kilobase flanking region on either side of the coding regions. The most significant SNP p-value is used as the gene-level p-value. These genes are then mapped to the pathways and ranked by significance. Fisher's Exact Test is then used to assess overrepresentation of functional categories in the top-ranked genes. P-values are FDR-corrected for multiple testing. GeSBAP is a web-server program.

**This method was not freely available, and was therefore not incorporated for further analysis.*

4. ***Gene Set Ridge Regression in Association Studies (GRASS)*** (Chen et al. 2010)

GRASS uses two steps for analysis. In the first step, the raw genotype data is aggregated into gene-level units, which are then decomposed into orthogonal components using Principal Components Analysis. The SNPs, which have the largest eigenvalues, are then called “nontrivial EigenSNPs”. These SNPs are considered individual signals within the gene. All of these SNPs are then considered predictors in the group ridge regression, which selects the representative SNPs associated with the outcome. The representative SNP beta estimates are then aggregated into a gene-level estimate. These statistics are evaluated for enrichment within a gene set, adjusting for gene size. Permutation is used to standardize the estimates.

5. ***GSA-SNP*** (Nam et al. 2010)

GSA-SNP is a stand-alone package that takes SNP p-values as input. SNPs are assigned to genes, including a 20 kilobase flanking region on either side of the coding region. The p-values are negative $\log_{10}$ transformed, and then the 2nd top SNP is selected. This was done to get the SNP most representative of the SNPs in the gene, not just the most significant by chance. Each gene-level p-value has a Benjamini-Hochberg multiple testing correction applied. These gene-level p-values may be evaluated at the pathway-level using three different analyses: Z-statistic, MAXMEAN, and iGSEA.

6. ***GSEA-SNP*** (Holden et al. 2008)

GSEA-SNP was developed as a direct adaptation of the original GSEA methods for gene expression data. SNP data is tested for association using an allele- or genotype-based statistic, such as the MAX-test. The MAX-test calculates three Cochrane-Armitage trend statistics according to the three different inheritance models (recessive, dominant, and additive). It uses the maximum of these three. A standard chi-squared model may also be used. The SNPs are ranked into a list according to significance and then compared to a gene set-specific list of SNPs. The gene sets are user-defined. Within each gene set, an enrichment score is calculated. This score shows if the SNPs in the gene set are overrepresented at the top of the original list including all SNPs ordered by significance. A running-sum statistic is used to determine overrepresentation. The phenotype is permuted to give the empirical P-value of the enrichment scores. A false discovery rate correction is applied to each SNP in the gene set. This program is available in R.

7. ***HYST*** (M.-X. Li, Kwan, and Sham 2012)

HYST was developed as a direct extension to GATES.(M.-X. Li et al. 2011) After performing GATES, an extended Simes procedure used for gene-level associations, HYST performs a scaled chi-squared test upon GATES output (SNP p-values). The procedure is similar to Fisher's Combination Test, but applied to gene-level p-values instead of SNP p-values. User-defined prior weights can be incorporated into the test statistic to account for functional significance of different members of a gene set.

8. ***i-GSEA4GWAS*** (Zhang et al. 2010)

This program is a web-server that performs a gene set enrichment specifically for GWAS. Given an input of SNPs and their p-values, i-GSEA4GWAS assigns SNPs to genes using various flanking regions, or the user can determine to only use functional SNPs. The maximum statistic or –log(P-value) within a gene is selected as the score for that gene. Permuting the SNP label normalizes these p-values. This corrects for gene variation, such as gene size or number of SNPs per gene. After this is done for all genes, they are ranked according to their scores. A Kolmogorov-Smirnov-like statistic is then calculated as the enrichment score for each gene. A significance proportion-based enrichment score (SPES) is calculated for a gene set, in which the number of significant genes in that set is divided by the number of significant genes in the entire dataset. A gene needs to have a SNP within the top 5% of SNPs to be considered significant. I-GSEA4GWAS draws upon pathways from MSigDB, which includes KEGG, BioCarta, and GO. The user may upload customizable gene sets.

**This method was not freely available with the current genomic build, and was therefore not incorporated for further analysis.*

9. *INRICH* (Lee et al. 2012)

INRICH is a unique method when compared to all the other methods in this review. Instead of taking input in the form of SNP-level test statistics or raw genotypes, it accepts genomic ranges that are found to be associated with outcome in the original GWAS. This can be done in Plink by scanning for all SNPs above a certain p-value threshold.(Purcell et al. 2007) The SNPs surrounding these index SNPs are then scanned for all SNPs below a less-stringent p-value threshold. After these intervals are estimated, INRICH calculates the number of intervals that overlap with a user-defined gene set. Permutations are conducted with intervals of the same length to assign empirical p-values to the gene set. An additional round of permutations using all gene sets is used to correct for multiple comparisons.

10. ***MAGENTA*** (Segrè et al. 2010)

MAGENTA uses gene set enrichment analysis (GSEA), adapted from gene expression studies, to evaluate the association of genetic data with pathways taken from public databases. These databases include KEGG, PANTHER, Reactome, BioCarta and Gene Ontology.(Segrè et al. 2010) It is a standalone package that runs on genome build 37 (hg19) or the older build 36 (hg18). MAGENTA's input is the SNP p-values, as well as their chromosomal positions. This can be from either a single GWAS, or a meta-analysis. MAGENTA maps the SNPs to genes using the UCSC genome browser coordinates from either hg18 or hg19. A gene is determined as the genic region, as well as user-defined flanking regions up and downstream of the transcribed start and end sites. In the second step, the minimum P-value from that gene is used to calculate a Z-score. The third step consists of correcting for possible confounders using a step-wise regression method. The six gene properties that are possibly corrected for are as follows: (1) physical gene size, (2) number of SNPs per kb, (3) number of independent SNPs per kb, (4) number of recombination hotspots per kb, (5) LD units per kb, and (6) genetic distance per kb. The adjusted gene p-value is then combined into gene sets, as determined by the databases previously mentioned.

Before an altered GSEA algorithm is applied to these sets, genes without any SNPs in the flanking regions are removed, as well as genes within a gene set that have the same most significant SNP to account for spurious associations. For each gene set, the proportion of genes with a corrected p-value below a certain cut-off is then calculated. This cut-off is predetermined as the 95th percentile of all the corrected gene-level p-values or the 75th percentile if a polygenic model is assumed. The GSEA p-value is then calculated using randomly sampled gene sets of the same size. A Bonferroni correction is applied to account for multiple testing.

11. ***PARIS (Pathway Analysis by Randomization Incorporating Structure)*** (Yaspan et al. 2011)

PARIS differs from other pathway-level methods in that it does not first assess significance at a gene-level, and then collapse it into a pathway, or gene-set. Instead it looks for independent "features" within the gene set. These features include LD blocks and individual SNPs in linkage equilibrium. LD blocks are defined using the HapMap CEU samples with the Gabriel et al method, and therefore may not be appropriate for GWAS of other ethnic groups. Any features that overlap with a gene's coding region is included in that gene's bin. PARIS then creates a "randomized feature collection" that has the same characteristics of the pathway's features from the rest of the genome. This is done to account for potential gene/pathway biases. An empirical p-value is then calculated comparing the enrichment of significance in the original pathway to the "randomized feature collection". This is done by calculating the number of significant features within the pathway, compared to the randomized set. Significance of at least one SNP with a p<0.05 within the feature.

**This method was not freely available for the server architecture used for analysis, and was therefore not incorporated for further analysis.*

12. ***PLINK Set-Based Test*** (Purcell et al. 2007)

PLINK's set-based test was designed originally to be for candidate gene studies, not GWAS due to its computational needs. The gene sets are user-defined. Within each gene set, the individual SNP association is conducted. Out of each gene set, the independent SNPs are extracted for further analyses. The mean of these independent SNPs' statistics is then calculated as the gene set statistic. The phenotype is then permuted for a user-specified number of times, repeating the same process. This maintains the LD structure found in the dataset. The empirical p-value for that gene set is then determined as the number of times the permuted set-statistic is greater than the original statistic for the set. While this corrects for the number of SNPs in the gene set, it does not correct for multiple testing on account of the number of gene sets. The r^2 threshold, p-value threshold, as well as the maximum number of independent SNPs selected per gene set can be user-specified.

13. ***RS-SNP*** (D'Addabbo et al. 2011)

RS-SNP is a Matlab package that can be used to assess if the significance found in a particular gene set is more than it should be by chance. In the first step, the association statistic is calculated for each SNP with five different models: general, dominant, recessive, multiplicative and additive risk models. After the individual SNP associations are computed, the enrichment of these associations in the user-defined gene sets is determined. This is done by using a hypergeometric distribution to calculate statistical significance under two null hypotheses simultaneously. The first null hypothesis is that there is no association between genotype and phenotype. The second null hypothesis is that the SNPs that are significant are not found in the gene set by chance. Significance is done by permutations in which the outcome status is permuted. For each permutation, the number of significant SNPs overall in the gene set is calculated using the mean and variance under the hypergeometric distribution. A false discovery rate and family wise error rate are computed to control for multiple testing.

**This method was not freely available, and was therefore not incorporated for further analysis.*

14. ***SNPtoGO*** (Schwarz et al. 2007)

SNPtoGO evaluates the enrichment of GO terms mapped to a set of SNPs. The input is a list of SNPs. SNPtoGO then maps the SNPs to GO terms, including a user-defined flanking region. A Fisher's exact test is used to determine if a GO term is overrepresented in a list of SNPs, compared to a random sample of SNPs. Because GO terms are hierarchical in structure, the *elim* algorithm is used {Alexa et al, 2006} to accommodate the tree structure and prevent there from being too many statistically relevant terms. A Bonferroni correction is applied to all results to account for multiple testing.

**This method was not freely available, and was therefore not incorporated for further analysis.*

15. ***SRT (SNP Ratio Test)*** (O'Dushlaine et al. 2009)

The SNP Ratio Test takes raw genotype files as an input, and computes the SNP-level association statistics as its first step. These SNPs are they aggregated into pathways using a user-defined database, ignoring the gene-level unit. The pathway-level units are evaluated by calculating the ratio of significant SNPs from a GWAS over a pre-determined threshold to the number of SNPs in the pathway unit. To assess significance, permutations are conducted using the raw genotype files given as input. The ratio of cases to controls is maintained throughout the outcome permutations. To prevent inflation, the same p-value threshold is not used as in the original analysis. Instead, the lowest M p-values are used from each pathway to create the new ratio. The empirical p-value is then calculated as the number of simulations that have a ratio larger than the original over the total number of simulations. Both the numerator and denominator have 1 added to them, to prevent a p-value of 0.

4.3.3: Limitations

The pathway-level methods have all of the same limitations as the gene-level methods. These include the inclusion of potential confounders, a lack of support for rare variants, and being dependent upon the databases used. In addition to these concerns, pathway-level analyses have their own issues. One of the fundamental differences between gene- and pathway-level analyses is that pathway-level analyses were not developed to find numerous additive effects within the same gene. Most of the programs only use the most significant SNP p-value as a surrogate for the overall gene p-value. This ignores all structure within the gene and all its information. Some programs ignore the gene structure all together. These programs directly map SNPs to their gene sets. While they may ignore this structure, the benefit is that they are much less computationally intensive without this extra step. Most of the programs take SNP p-values as their input, increasing the ease of computation.

An additional difference is the use of canonical pathway databases, such as GO and KEGG and the lack of directionality. While the program may indicate a pathway, it does not define a certain aspect of the pathway, nor the process that it may directly affect. The use of these canonical pathways may also limit the investigator's hypotheses. Other methods exist that only use the actual data to elucidate gene-gene interactions and potential networks of association through protein-protein interaction analyses (PPI).

4.3.4: Discussion

GWAS typically use a genome-wide significance threshold of $5x10^{-8}$. Associations with SNPs below this threshold are often ignored, at least in the first phase of analysis, leading to the loss of potential biologically relevant associations. These pathway methods were designed to look for enrichment of genes that are typically ignored within gene sets or pathways. All of these programs are highly dependent upon the databases.

Many are able to accept user-defined databases, which is especially helpful for disease-specific studies. The use of canonical pathways in GO, KEGG, and BioCarta contribute to a standardization of comparisons between various studies.

The interpretation of these programs should always be in the context of their methodology, as some programs rely upon the strength of associations for the genes within the gene sets. Others only rely upon the ranking of the genes, looking for enrichment within the top ranked genes regardless of their strength of association. Two of the methods (SNP Ratio Test and Plink Set Test) ignore gene structure altogether and only look at the SNPs in the gene set as a whole. It should be emphasized that pathway-level methods do not evaluate gene-gene or any other types of interactions. Results do not offer directionality or pinpoint the part of the pathway that is affected. To investigate these relationships, a different set of methods is required, such as protein-protein interactions or classical interaction analyses. The goal of pathway-level methods for GWAS is to visualize the data that is suggestive but not significant, looking for enrichment in some biological processes versus others. By evaluating enrichment of pathways, it offers the investigator the ability to see connections between the associated genes.

Table 4.2: Review of Pathway-Level Methods

Program/Method	**Citation**	**Input**	**Group**	**Pathways**	**Adjusted for Multiple Comparisons**	**Evaluated in Aim 2b**
ALIGATOR	(Holmans et al. 2009)	P (SNP)	C	GO		X
GenGen	(Wang, Li, and Bucan 2007)	Raw Genotype	C	User-defined		X
GeSBAP	(Medina et al. 2009)	P (SNP or Gene)	C	GO, KEGG, BioCarta	FDR	
GRASS	(Chen et al. 2010)	Raw Genotype	SC	User-defined		X
GSA-SNP	(Nam et al. 2010)	P (SNP)	C	GO	Benjamini-Hochberg	X
GSEA-SNP	(Holden et al. 2008)	Raw Genotype	C	User-defined		X
HYST	(M.-X. Li, Kwan, and Sham 2012)	P (SNP)	C	User-defined		X
i-GSEA4GWAS	(Zhang et al. 2010)	P (SNP)	C	GO, KEGG, BioCarta		
INRICH	(Lee et al. 2012)	Genomic Ranges	SC	User-defined	Permutations	X
MAGENTA	(Segrè et al. 2010)	P (SNP)	C	KEGG, PANTHER, Reactome, BioCarta, GO	Bonferroni, FDR	X
PARIS	(Yaspan et al. 2011)	P (SNP)	SC	User-defined		
PLINK Set Test	(Purcell et al. 2007)	Raw Genotype	SC	User-defined		X
RS-SNP	(D'Addabbo et al. 2011)	Raw Genotype	SC	User-defined		
SNPtoGO	(Schwarz et al. 2007)	SNP IDs	SC	GO	Bonferroni	
SRT	(O'Dushlaine et al. 2009)	Raw Genotype	C	User-defined		X

References

Bacanu, Silviu-Alin. 2012. "On Optimal Gene-Based Analysis of Genome Scans." *Genetic Epidemiology* 36 (4) (April 16): 333–339. doi:10.1002/gepi.21625.

Bloom, Joshua S, Ian M Ehrenreich, Wesley T Loo, Thúy-Lan Võ Lite, and Leonid Kruglyak. 2013. "Finding the Sources of Missing Heritability in a Yeast Cross." *Nature* (February 3): 1–6. doi:10.1038/nature11867.

Chen, Lin S, Carolyn M Hutter, John D Potter, Yan Liu, Ross L Prentice, Ulrike Peters, and Li Hsu. 2010. "Insights Into Colon Cancer Etiology via a Regularized Approach to Gene Set Analysis of GWAS Data." *The American Journal of Human Genetics* 86 (6) (June 11): 860–871. doi:10.1016/j.ajhg.2010.04.014.

D'Addabbo, Annarita, Orazio Palmieri, Anna Latiano, Vito Annese, Sayan Mukherjee, and Nicola Ancona. 2011. "RS-SNP: a Random-Set Method for Genome-Wide Association Studies." *BMC Genomics* 12 (1) (March 30): 166. doi:10.1186/1471-2164-12-166.

Gibson, Greg. 2012. "Rare and Common Variants: Twenty Arguments." *Nature Reviews Genetics* 13 (2) (February 1): 135–145. doi:10.1038/nrg3118.

Han, Fang, and Wei Pan. 2010. "A Data-Adaptive Sum Test for Disease Association with Multiple Common or Rare Variants." *Human Heredity* 70 (1): 42–54. doi:10.1159/000288704.

Hindorff, Lucia A, J Macarthur, J Morales, Heather A Junkins, P N Hall, A K Klemm, and Teri A Manolio, eds. 2013. *A Catalog of Published Genome-Wide Association Studies*. Accessed September 10. http://www.genome.gov/gwastudies.

Hindorff, Lucia A, Praveen Sethupathy, Heather A Junkins, Erin M Ramos, Jayashri P Mehta, Francis S Collins, and Teri A Manolio. 2009. "Potential Etiologic and Functional Implications of Genome-Wide Association Loci for Human Diseases and Traits.." *Pnas* 106 (23) (June 9): 9362–9367. doi:10.1073/pnas.0903103106.

Holden, M, S Deng, L Wojnowski, and B Kulle. 2008. "GSEA-SNP: Applying Gene Set Enrichment Analysis to SNP Data From Genome-Wide Association Studies." *Bioinformatics* 24 (23) (November 21): 2784–2785. doi:10.1093/bioinformatics/btn516.

Holmans, Peter, Elaine K Green, Jaspreet Singh Pahwa, Manuel A R Ferreira, Shaun M Purcell, Pamela Sklar, Michael J Owen, Michael C O Donovan, Nick Craddock, and The Wellcome Trust Case-Control Consortium9. 2009. "Gene Ontology Analysis of GWA Study Data Sets Provides Insights Into the Biology of Bipolar Disorder." *The American Journal of Human Genetics* 85 (1) (July 10): 13–24. doi:10.1016/j.ajhg.2009.05.011.

Huang, Hailiang, Pritam Chanda, Alvaro Alonso, Joel S Bader, and Dan E Arking. 2011. "Gene-Based Tests of Association.." *PLoS Genetics* 7 (7) (July): e1002177. doi:10.1371/journal.pgen.1002177.

la Cruz, De, Omar, Xiaoquan Wen, Baoguan Ke, Minsun Song, and Dan L Nicolae. 2010.

“Gene, Region and Pathway Level Analyses in Whole-Genome Studies..” *Genetic Epidemiology* 34 (3) (April): 222–231. doi:10.1002/gepi.20452.

Lee, P H, C O'Dushlaine, B Thomas, and S M Purcell. 2012. “INRICH: Interval-Based Enrichment Analysis for Genome-Wide Association Studies.” *Bioinformatics* 28 (13) (June 23): 1797–1799. doi:10.1093/bioinformatics/bts191.

Lehne, B, C M Lewis, and T Schlitt. 2011. “From SNPs to Genes: Disease Association at the Gene Level.” *PLoS ONE* 6 (6): e20133. doi:10.1371/journal.pone.0020133.t001.

Li, Miao-Xin, Hong-Sheng Gui, Johnny S H Kwan, and Pak C Sham. 2011. “GATES: a Rapid and Powerful Gene-Based Association Test Using Extended Simes Procedure.” *American Journal of Human Genetics* 88 (3) (March 11): 283–293. doi:10.1016/j.ajhg.2011.01.019.

Li, Miao-Xin, Johnny S H Kwan, and Pak C Sham. 2012. “HYST: a Hybrid Set-Based Test for Genome-Wide Association Studies, with Application to Protein-Protein Interaction-Based Association Analysis.” *American Journal of Human Genetics* 91 (3) (September 7): 478–488. doi:10.1016/j.ajhg.2012.08.004.

Liu, Jimmy Z, Allan F Mcrae, Dale R Nyholt, Sarah E Medland, Naomi R Wray, Kevin M Brown, AMFS Investigators, et al. 2010. “A Versatile Gene-Based Test for Genome-Wide Association Studies.” *American Journal of Human Genetics* 87 (1) (July 9): 139–145. doi:10.1016/j.ajhg.2010.06.009.

Luo, Li, Gang Peng, Yun Zhu, Hua Dong, Christopher I Amos, and Momiao Xiong. 2010. “Genome-Wide Gene and Pathway Analysis.” *European Journal of Human Genetics* 18 (9) (May 5): 1045–1053. doi:10.1038/ejhg.2010.62.

McCarthy, Mark I, GonCalo R Abecasis, Lon R Cardon, David B Goldstein, Julian Little, John P A Ioannidis, and Joel N Hirschhorn. 2008. “Genome-Wide Association Studies for Complex Traits: Consensus, Uncertainty and Challenges.” *Nature Reviews Genetics* 9 (5) (May): 356–369. doi:10.1038/nrg2344.

Medina, I, D Montaner, N Bonifaci, M A Pujana, J Carbonell, J Tarraga, F Al-Shahrour, and J Dopazo. 2009. “Gene Set-Based Analysis of Polymorphisms: Finding Pathways or Biological Processes Associated to Traits in Genome-Wide Association Studies.” *Nucleic Acids Research* 37 (Web Server) (June 29): W340–W344. doi:10.1093/nar/gkp481.

Nam, D, J Kim, S Y Kim, and S Kim. 2010. “GSA-SNP: a General Approach for Gene Set Analysis of Polymorphisms.” *Nucleic Acids Research* 38 (Web Server) (June 24): W749–W754. doi:10.1093/nar/gkq428.

O'Dushlaine, C, E Kenny, E A Heron, R Segurado, M Gill, D W Morris, and A Corvin. 2009. “The SNP Ratio Test: Pathway Analysis of Genome-Wide Association Datasets.” *Bioinformatics* 25 (20) (October 8): 2762–2763. doi:10.1093/bioinformatics/btp448.

Peng, Gang, Li Luo, Hoicheong Siu, Yun Zhu, Pengfei Hu, Shengjun Hong, Jinying Zhao, et al. 2009. “Gene and Pathway-Based Second-Wave Analysis of Genome-Wide Association Studies.” *European Journal of Human Genetics* 18 (1) (July 8): 111–117. doi:10.1038/ejhg.2009.115.

Purcell, Shaun, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel A R Ferreira, David Bender, Julian Maller, et al. 2007. "PLINK: a Tool Set for Whole-Genome Association and Population-Based Linkage Analyses." *The American Journal of Human Genetics* 81 (3) (September): 559–575. doi:10.1086/519795.

Schwarz, D F, O Hadicke, J Erdmann, A Ziegler, D Bayer, and S Moller. 2007. "SNPtoGO: Characterizing SNPs by Enriched GO Terms." *Bioinformatics* 24 (1) (December 19): 146–148. doi:10.1093/bioinformatics/btm551.

Segrè, Ayellet V, DIAGRAM Consortium, MAGIC investigators, Leif Groop, Vamsi K Mootha, Mark J Daly, and David Altshuler. 2010. "Common Inherited Variation in Mitochondrial Genes Is Not Enriched for Associations with Type 2 Diabetes or Related Glycemic Traits." Edited by Peter M Visscher. *PLoS Genetics* 6 (8) (August 12): e1001058. doi:10.1371/journal.pgen.1001058.t004.

Vineis, Paolo, and Neil Pearce. 2010. "Missing Heritability in Genome-Wide Association Study Research." *Nature Reviews Genetics* 11 (8) (August 1): 1–1. doi:10.1038/nrg2809-c2.

Wang, Kai, Mingyao Li, and Maja Bucan. 2007. "Pathway-Based Approaches for Analysis of Genomewide Association Studies." *The American Journal of Human Genetics* 81 (6) (December): 1278–1283. doi:10.1086/522374.

Yaspan, Brian L, William S Bush, Eric S Torstenson, Deqiong Ma, Margaret A Pericak-Vance, Marylyn D Ritchie, James S Sutcliffe, and Jonathan L Haines. 2011. "Genetic Analysis of Biological Pathway Data Through Genomic Randomization." *Human Genetics* 129 (5) (January 30): 563–571. doi:10.1007/s00439-011-0956-2.

Zhang, K, S Cui, S Chang, L Zhang, and J Wang. 2010. "I-GSEA4GWAS: a Web Server for Identification of Pathways/Gene Sets Associated with Traits by Applying an Improved Gene Set Enrichment Analysis to Genome-Wide Association Study." *Nucleic Acids Research* 38 (Web Server) (June 24): W90–W95. doi:10.1093/nar/gkq324.

Chapter 5: Evaluation of Gene-Level Methods (Paper 2)

5.1: Abstract

Background: Genome-wide association studies (GWAS) have successfully identified more than 10,000 SNPs associated with 840 traits. Despite this success, there still remains the problem of "missing heritability" for most traits. One contributing factor may be the result of examining single markers at a time as opposed to a group of markers that are biologically meaningful in aggregate. To address this problem, a variety of gene-level methods were developed to identify putative biologically relevant associations. A simulation was performed to systematically assess the performance of gene-level methods.

Methods: Using genetic data from the Wellcome Trust Case Control Consortium (WTCCC), we simulated case-control status based on an additive polygenic model where cases have more risk alleles than controls. A total of 20 gene sets and 226 genes were selected from Gene Ontology (GO). We evaluated 12 methods based on the sensitivity, specificity, as well as type I and type II error of each test. The influences of gene size, number of causal single nucleotide polymorphisms (SNPs) in each gene, and effect size were assessed. The effect of sample size was also examined using a traditionally underpowered (n=250 cases, 250 controls) and a larger (n=2,250 cases, 2,250 controls) set of cases and controls.

Results: Despite the low overall sensitivity (18-59%), across methods the specificity was high (89-100%) with low type I error (0.1-6%). Classical methods, not designed to handle linkage disequilibrium, had higher sensitivity, but also higher type I error. Newer methods that directly estimate correlation structures were underpowered to detect genes with smaller effect sizes, but type I error was low. All programs were significantly underpowered to detect signals in small sample sizes (n=500). Sensitivity was lowest for genes that had few causal SNPs, while they were increased if multiple independent signals were present.

Conclusions: The low type I error and high specificity found in most methods increase confidence in identified genes. Larger effect sizes and a higher number of causal SNPs increased accuracy in all programs. All methods were successful at identifying genes that would not been detected in a traditional GWAS.

5.2: Introduction

In less than a decade, genome-wide association studies (GWAS) have proven to be a useful tool in identifying risk loci for various complex diseases. As of August 2013, the NHGRI GWAS Catalog contained 1,666 publications and 11,082 associations.(Hindorff et al. 2009) Working under the hypothesis of "common disease, common variants", GWAS has elucidated many loci that are moderate to highly associated with complex phenotypes. However, there is still a large amount of "missing heritability". One example of this is human height. The heritability, or proportion of phenotypic variance due to genetics (as opposed to environmental influences), of human height has been estimated to be 80%.(Zaitlen et al. 2013) Through GWAS, 50 variants have been identified as being genome-wide significant, yet only 5% of the phenotypic variance has been explained. By including nearly 300,000 variants, 45% of the variance can be explained in a linear model.(Yang et al. 2010) The discrepancy between these two estimates is largely due to many of the variants not having a large enough effect size to be detected with stringent GWAS thresholds.

One method to detect these smaller effect sizes is through gene-level programs. These programs look for an enrichment of independent association signals within a gene. The underlying framework theorizes that genes that have multiple alleles associated with the outcome of interest (allelic heterogeneity) would not have any single nucleotide polymorphisms (SNPs) with a large enough effect size to be detected. This gene would be apparent when these SNPs are assessed for statistical significance in aggregate because many SNPs would have suggestive *P*-values. In recent years, a plentitude of methods have been developed to address this question. However, there has been no consensus as to the best method as determined by their sensitivity and specificity, as well as type I and II error rates. We seek to systematically evaluate gene-level programs for GWAS through a simulation that examines the effect of gene size, number of causal alleles, as well as sample size, on their relative performance.

5.3: Materials and Methods

5.3.1: Simulation Methods

5.3.1.1: Genotype Quality Control

The common control data was ascertained from the Wellcome Trust Case-Control Consortium following the appropriate IRB procedures. Data included the 1958 Birth Cohort (N=2,930) and the National Blood Service samples (n=2,737). These samples were collected to represent the overall population of the United Kingdom, regardless of health status. Within the original study, SNPs were filtered for having a Hardy-Weinberg Equilibrium p-value > 10^{-20}, an information content > 90%, genotype missingness <5 %, and a minor allele frequency (MAF) > 1%. This lead to a loss of 191,544 SNPs in the National Blood Service samples, and 192,375 SNPs in the 1958 Birth Cohort. The two studies were then combined for all further analyses. This resulted in a total of 926,604 SNPs, and 5,667 individuals. Quality control was repeated when the two studies were combined with the same criteria as within each original study. This resulted in a total of 913,763 SNPs in the same 5,667 individuals. Samples were then screened through individual-level quality control measures, including individual missingness < 5%, as well as heterozygosity outliers. For missingness, 5 people were dropped for having more than 5% of their SNPs missing. Heterozygosity was estimated and individuals more than 5 standard deviations away from the mean were dropped, leaving 5,627 individuals.

The data were converted from the probabilities in the Wellcome Trust format, to called genotypes in Plink format.(Purcell et al. 2007) Genotypes had to have a posterior probability of greater than 90%, otherwise they were annotated as missing. Data was again screened for genotype missingness (<5%, N=741), MAF (>1%, N=61), and individual missingness (<5%, N=19). Individuals were screened for excessive identity-by-descent (IBD). First-degree relatives were excluded (N=114).

To ascertain a relatively homogenous population, principal components analysis (PCA) was conducted. Non-autosomal markers were removed, as well as markers in known regions of population substructure. Markers were selected to be independent using Plink with a maximum r^2 cutoff of 0.05. A total of 42,913 SNPs were used in PCA analysis of 5,494 individuals. SNP weights in each analysis were examined for outliers, but none were found. Two successive rounds of PCA were conducted to remove outliers from the first two principal components. A total of 4,500 individuals remained in a homogeneous population.

Markers were then filtered for a Hardy-Weinberg Equilibrium p-value > 10^{-5}, MAF > 1%, and missingness < 5%. Marker coordinates were updated from hg18 to a more recent build of hg19 for further analyses. This was done using liftOver, a utility from the University of California, Santa Cruz (UCSC) Genome Browser.(Hinrichs 2006) SNPs that could not be mapped to the newer build were dropped from analysis (N=208). The cleaned data resulted in 4,500 individuals and 906,298 SNPs.

5.3.1.2: Pathway and Gene Selection

Pathways were downloaded from the Molecular Signatures Database (MSigDB) for Gene Ontology Biological Processes.(Subramanian et al. 2005) This database was chosen because the majority of the methods could use this database. The Gene Ontology (GO) Biological Processes (BP) are categorized as a series of events or molecular functions that have a beginning and an end. This is similar to a canonical pathway, which is found in KEGG and BioCarta, in that there is a process that begins and ends with an ultimate goal. There were 825 biological processes in this database, with a median size of 28 genes. A total of 20 pathways were randomly selected within two groups: 10 with over 28 genes (big) and 10 with under 28 genes (small). Features of these pathways are detailed in Table 5.1. The Entrez IDs from these pathways were then mapped using Ensembl and the BioMart package within Bioconductor in R under build hg19. The median gene size was 28.32 kilobases (kb) and the mean gene size was larger at 71.09 kb. This gene size includes introns. The median gene size is nearly on par with the average gene size estimated from the human genome of 27 kb.(Venter et al. 2001) However, the mean gene size is much larger. This is partly due to a few large genes that skew the distribution, but it is also due to a bias with well-characterized functional genes tending to be larger than those that are smaller, including pseudo-genes.(Venter et al. 2001) There was no difference in the distribution of gene size by the size of pathways (*P*<0.05).

From each class of pathways (big and small), a number of genes were selected to be "causal". Within each group, four pathways were selected to have only one associated gene, four pathways were selected to have 20% of their genes "causal", and two pathway were selected to have 50% of their genes "causal". Genes were removed that were found to be in numerous pathways to create relatively independent units of analysis. This lead to a total of 226 selected genes blind to the genes' various features such as size and SNP density. (Table 5.1)

Table 5.1: Pathway Characteristics

Group	*Biological Process*	*# Genes*	*Median Gene Size (kb)*	*Mean Gene Size (kb)*	*Percentage of Genes*	*# Simulated Genes*	*# Simulated Genes (Truth)*
SMALL	CDC42 Protein Signal Transduction	12	51.81	77.27	50%	6	4
	Defense Response to Virus	11	32.59	56.27	20%	3	3
	Establishment of Vesicle Localization	10	28.71	118.26	(1)	1	1
	G-Protein Signaling Adenylate Cyclase Activating Pathway	25	15.48	49.20	20%	5	4
	G1 Phase of Mitotic Cell Cycle	12	15.46	37.58	20%	3	3
	Morphogenesis of an Epithelium	17	31.64	53.37	20%	4	2
	Protein Complex Disassembly	15	23.55	98.05	(1)	1	1
	Protein Polyubiquitination	10	59.82	58.39	(1)	1	1
	Spindle Organization and Biogenesis	10	33.72	32.83	50%	5	5
	Ribonucleotide Metabolic Process	17	38.18	96.72	(1)	1	1
BIG	Anatomical Structure Morphogenesis	363	30.45	93.98	20%	73	70
	Cellular Defense Response	55	16.67	33.56	(1)	1	1
	Establishment and/or Maintenance of Chromatin Architecture	71	37.70	69.97	50%	36	36
	G-Protein Coupled Receptor Protein Signaling Pathway	332	14.78	59.50	20%	67	65
	Leukocyte Activation	65	20.87	59.77	(1)	1	0
	Lipid Transport	29	27.22	42.68	50%	15	13
	Membrane Lipid Metabolic Process	98	31.07	56.37	(1)	1	1
	Regulation of DNA Binding	44	25.79	53.89	(1)	1	1
	Response to Hypoxia	28	41.34	65.78	20%	6	6
	T-Cell Activation	41	26.30	42.52	20%	9	8

5.3.1.3: Phenotype Generation

SNPs that were within the genic region and a 20 kb flanking region on either side of the genomic coordinates were extracted from the GWAS genotype file. From each gene unit, tag SNPs were selected using Tagger and a cut-off of $r^2<0.2$ for "independent" SNPs.(de Bakker et al. 2005) Between the 226 genes, 75 (~1/3) genes had one SNP selected as causal, 76 (~1/3) had 2 SNPs selected as causal, and 75 (~1/3) had 5 SNPs selected as causal. This resulted in a total of 602 SNPs tagging 226 genes in 20 different pathways.

These causal SNPs were extracted from the genotype file and converted into an additive format, indicating the number of minor alleles per individual (0, 1, or 2). Genes were split into two groups, with effect size being assigned at random between an odds ratio (OR) of 1.2 and an OR of 2. The effect sizes were log transformed ($\log_2$) and multiplied by the individual's number of minor alleles to assume an additive model. This led to an individual per-marker score, with all SNPs in a gene having the same effect size. All 602 markers were then summed over an individual, leading to a liability score per person. Subtracting out the mean and dividing by the standard deviation of the overall distribution standardized the individual liability scores. To introduce a stochastic element into the phenotype assignment, scores had a random amount of variation added from a normal distribution. Individuals were then assigned a case or control status based on their underlying score using a binomial distribution. The resulting distribution can be seen in Figure 5.1. The study was evenly split between cases and controls, with 2,250 individuals in each group. The score distribution of cases and controls overlaps and was done intentionally to create a realistic additive polygenic model.

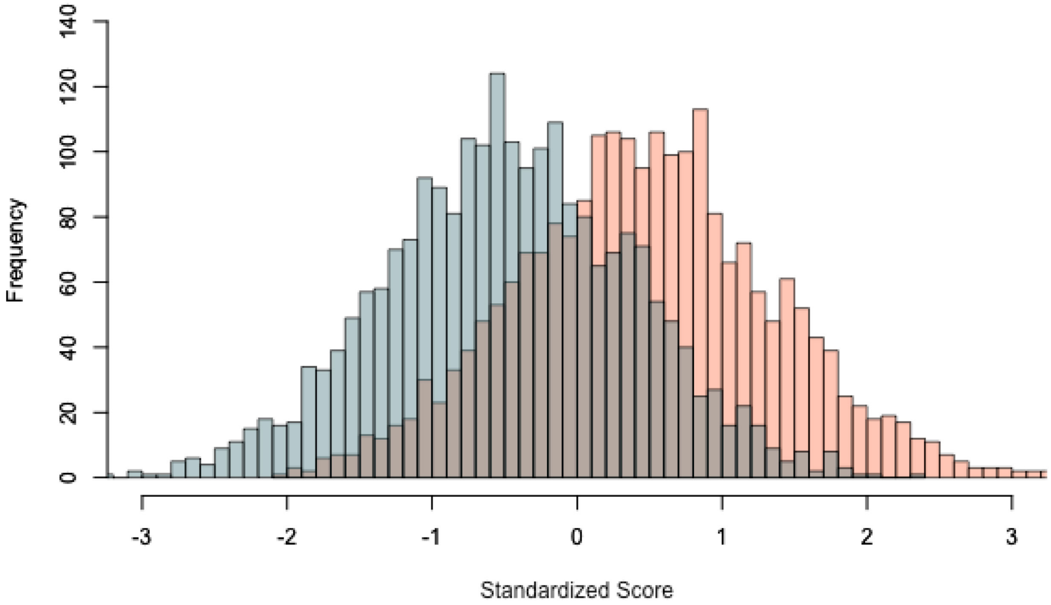


Figure 5.1: Frequencies of the standardized liability scores by simulated case (pink) and control (blue) status.

Two rounds of analyses were conducted with two datasets: one of a larger traditional GWAS sample size (N=4500), and another on a smaller sample size (N=500). The 500 individuals in the second analysis were randomly selected from the 4,500 individuals from the first analysis. This group consisted of 247 cases and 253 controls.

5.3.1.4: Genome-wide association study

Using the case-control phenotype assigned in the previous section, a genome-wide association study was conducted. Under an additive model, a logistic regression was performed for each marker. Genome-wide significance (P<5X10^{-07}) was reached for two regions: chromosomes 1 and 22 (Figure 5.2). No SNP with an effect size below 1.25 reached genome-wide significance (Figure 5.3).

To check the validity of the simulation, the correlation between an individual's SNP score and their case-control status was plotted versus that SNP's negative log p-value from the GWAS. They were separated out by the simulated effect sizes, 1.2 and 2 (Figure 5.4.1 and 5.4.2). It can be seen that the higher the correlation with the outcome, the more significant the association. This is more pronounced for the higher effect sizes, as expected, because of the increased power. While the effect sizes were split evenly between the genes, the more significant SNPs are highly skewed towards the larger effect size. However, this is consistent with many GWAS in which there is increased power for larger effect sizes. This will limit conclusions about the influence of effect sizes in later analyses for both gene- and pathway-level methods.

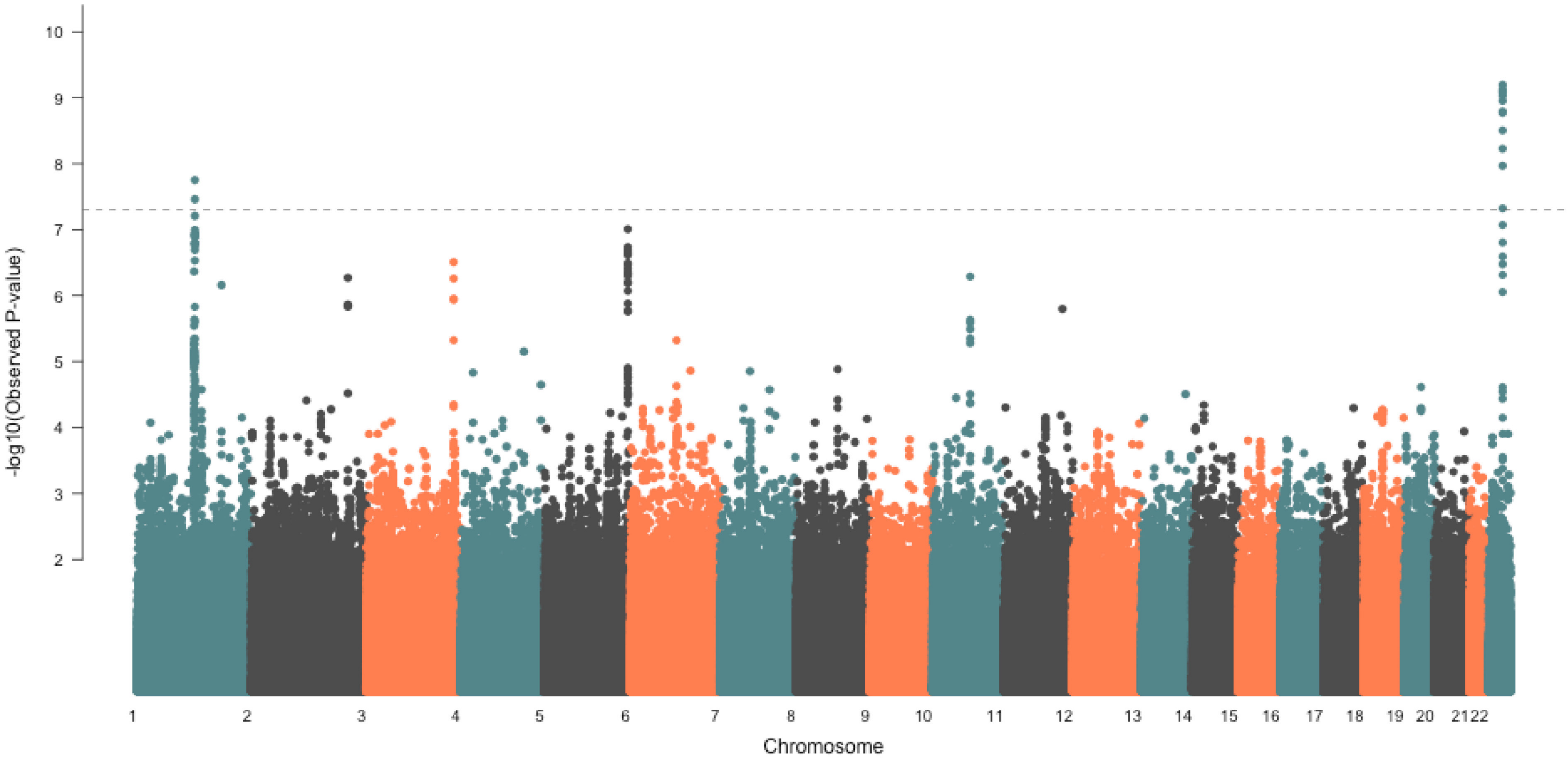


Figure 5.2: Manhattan Plot of genome-wide association by chromosome. *Significance is shown along the y-axis with the –log$_{10}$ transformation of the GWAS P-values. The grey line indicates genome-wide significance at 5x10^{-8}. SNPs are organized by chromosome (different colors) and position along the y-axis.*

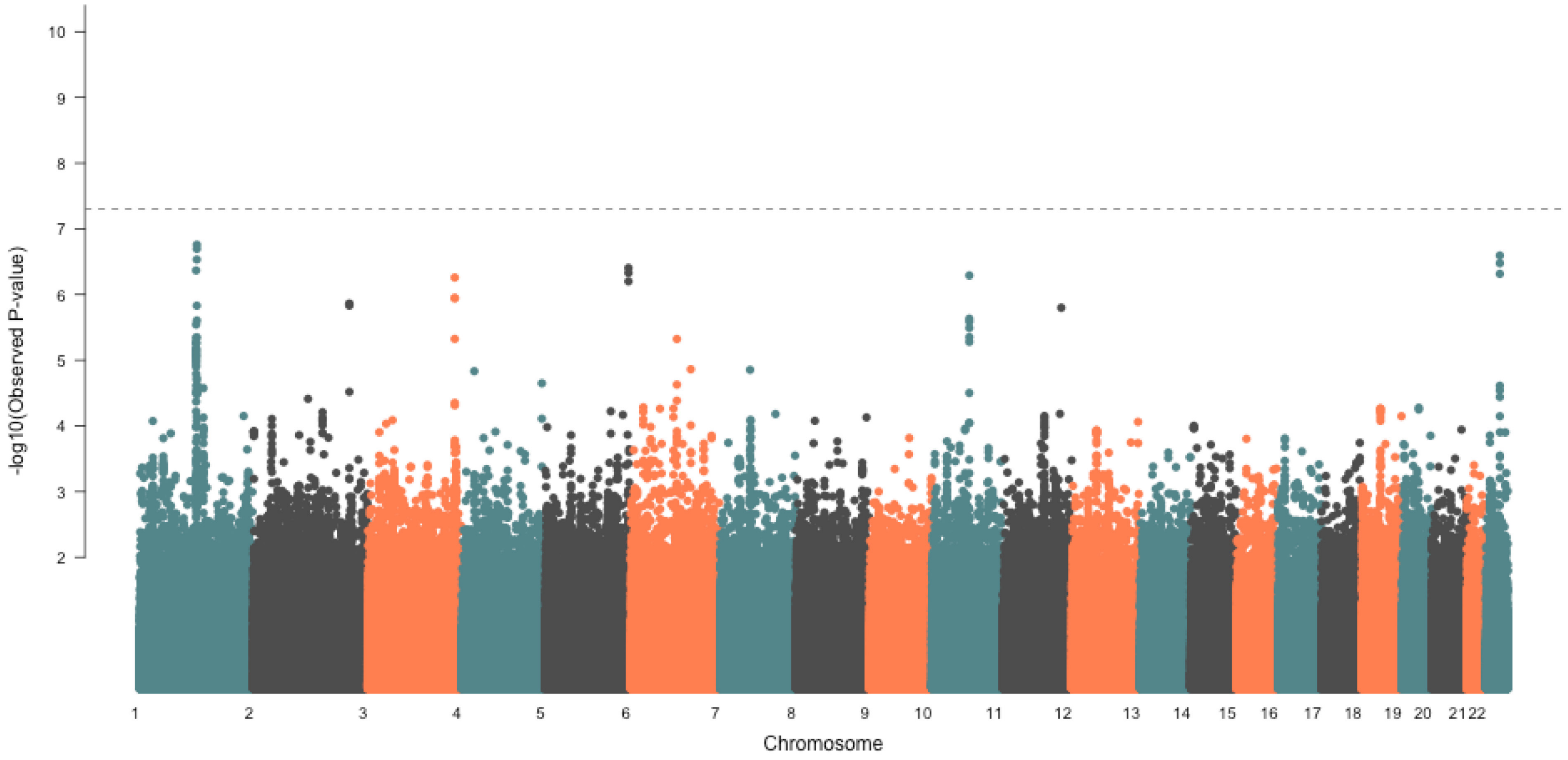


Figure 5.3: Manhattan Plot of SNPs with an effect size below 1.25 by chromosome. *Significance is shown along the y-axis with the –log_{10} transformation of the GWAS P-values. The grey line indicates genome-wide significance at 5x10^{-8}. SNPs are organized by chromosome (different colors) and position along the y-axis.*

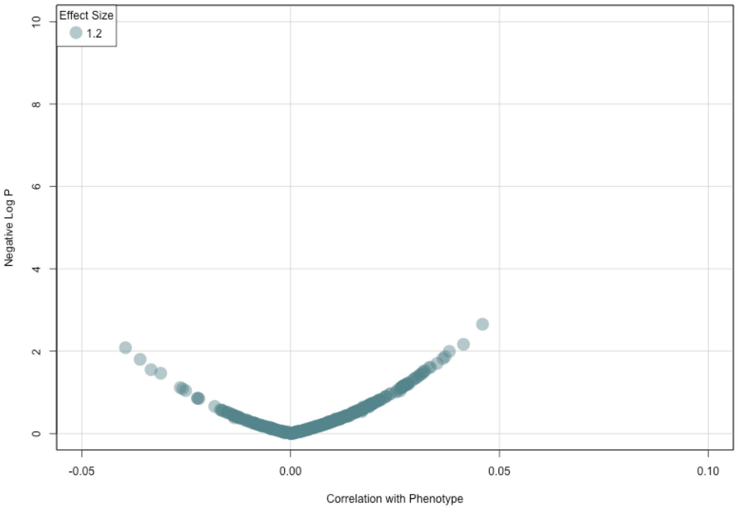


Figure 5.4.1: SNP score correlation with outcome (x-axis) versus significance for lower effect sizes (OR=1.2, y-axis).

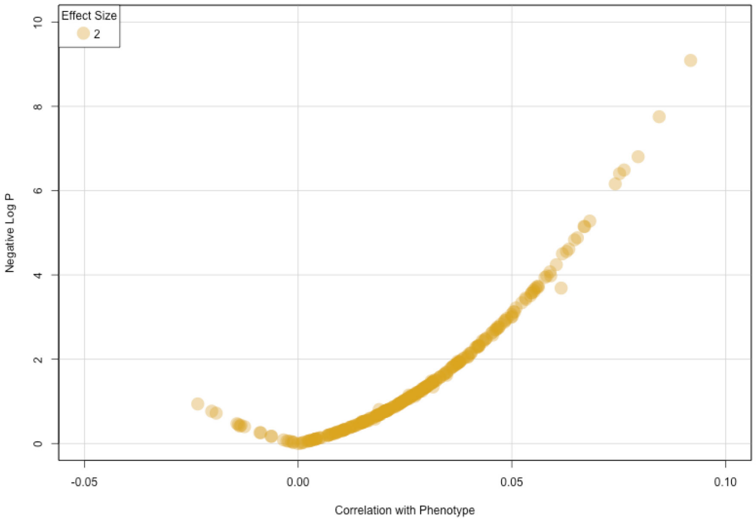


Figure 5.4.2: SNP score correlation with outcome (x-axis) versus significance for higher effect size (OR=2, y-axis).

5.3.2: Gene-Level Programs

A total of 12 programs were compared: VEGAS (all SNPs), VEGAS (Top 10% of SNPs), Fisher's Combination Test, Sidak's Combination Test, Simes' Test, False Discovery Rate (FDR), GATES, HYST, Weighed GATES, Weighted HYST, aSum, and the Score Test. While these methods were previously described in Chapter 4, they will be briefly summarized below.

1. **Fisher's Combination Test:** Fisher's combination test (FCT) takes the natural log of the SNP P-values, summing across all SNPs in the gene, and then multiplies by -2. The resulting chi-squared test statistic's degrees of freedom is determined by the number of SNPs in the gene.(Peng, Zhao, and Xue 2009)
2. **Sidak's Combination Test**: Sidak's Combination Test, also called Sidak's Correction, takes the minimum SNP from the gene and corrects for the number of SNPs.(Peng, Zhao, and Xue 2009)
3. **Simes' Test:** SNPs are ordered from the most to least significant, multiplied by the total number of SNPs, and divided by their rank. The minimum transformed P-value is then used as the gene-level P-value.(Peng, Zhao, and Xue 2009)
4. **False Discovery Rate (FDR):** The SNP P-values are ordered from most to least significant and are corrected for the False Discovery Rate. The minimum False Discovery Rate is then used as the gene-level output.(Peng, Zhao, and Xue 2009)

5. **GATES/Weighted GATES:** SNP P-values are assessed for correlations and independent representative SNPS are selected for each gene. The representative SNPs are then corrected using the Simes' procedure. The Weighted GATES methods incorporates weights for the SNPs depending on their functional relevance (intron, exon, nonsynonymous, etc).(Li et al. 2011)
6. **HYST/Weighted HYST:** HYST is part of the GATES package in which a modified hypergeometric test is used to determine a gene-level test statistic for enrichment. The weighted HYST procedure weights SNPs based on their functional relevance.(Li, Kwan, and Sham 2012)
7. **VEGAS (All/Top 10%):** VEGAS directly estimates the correlation structure of the genes by using a Cholesky decomposition. Permutations are conducted to determine an empirical P-value. All SNPs can be used within the gene, or just the top 10% of associated SNPs within each gene.(Liu et al. 2010)
8. **ASUM:** ASUM is an adaptive sum test that can be used for both rare and common variants. The effect size is first evaluated in a multivariate regression analysis for variants with a significant protective effect, which is then flipped. Then all variants are collapsed across the region and evaluated using the score test with logistic regression.(Han and Pan 2010)
9. **Score Test:** All variants are considered in a multivariate logistic regression using the score test with no transformations regarding effect size.(Han and Pan 2010)

All programs defined genic regions as the translated gene region plus 20 kilobases on either side. Because of the stochastic nature of the GWAS simulation, the determination of true positive and negative genes was dependent upon the GWAS results and the original framework. In order to be a "true positive", genes had to be one of the original list that the GWAS was simulated upon, as well as have at least one SNP with a *P*-value of less than 0.01. The "true negative" genes were then determined to be those that were not within 50 kilobases of either the start or stop of any of the original simulation genes. A total of 49 true positive genes and over 17,000 true negative genes were used to measure type I and type II error. To assess sensitivity and specificity, a subset of 50 true negative genes were randomly chosen to compare with the 49 true positive genes. Within the smaller sample size analyses, these sets of true negative and true positive genes were used, as well as an additional round in which a true positive gene had to have at least one SNP with a p-value of less than 0.01 within the smaller sample size GWAS results. This reduced the number of true positive genes to 23, instead of the previous 49 true positive genes from the larger analysis. A p-value threshold of 0.001 was used to determine statistical significance for all analyses. Due to the nearly 17,000 genes being evaluated, a Bonferonni correction would need a p-value threshold of $2.9\text{x}10^{-6}$ for α=0.05λ. However, this is a conservative estimate since many genes are in linkage disequilibrium. Bias was assessed for effect size, gene size, SNP density for the gene, and number of "causal" variants upon which the simulation was conducted.

A total of 10 different programs were compared for their sensitivity and specificity, as well as their type I and type II error rates (Table 5.2).

Table 5.2: Evaluation Methods

Measure	*Data*	*Assessment*
Sensitivity	50 true negative and 49 true positive genes	Ability to detect true positive
Specificity		Ability to detect true negative
Type I Error	Genome-wide (~17,000) true negative and 49 true positive genes	Incorrectly detecting false positives
Type II Error		Incorrectly detecting false negatives

Two of the programs evaluated (ASUM and Logistic Regression test) required individual raw genotype data making them computationally intensive. While this is not practical for a GWAS, a sub-analysis was performed using the 99 true positive and negative genes. Type I and type II error was not assessed due to only a subset of genes being run for these methods.

The role of potential biases was evaluated using the "gold standard" of true positive and negative genes. The accuracy of their prediction determined by accordance between the "truth" and statistical significance as determined by P<0.001. Correlation between ten of the programs (not aSum and Score test) were calculated for genes found in all the methods.

5.4: Results

5.4.1: Overall Results

Of the twelve programs evaluated, Fisher's Combination Test had the highest sensitivity. (Table 5.3) However, this statistical test also had the highest type I error (5.9%) and the lowest specificity. Sidak's Combination Test had the lowest sensitivity, despite having the lowest type I error rate (0.11%). Sidak's Combination Test only considers the most significantly associated SNP, ignoring any joint signals, leading to a conservative test.

Table 5.3: Performance Metrics of Gene-Level Methods

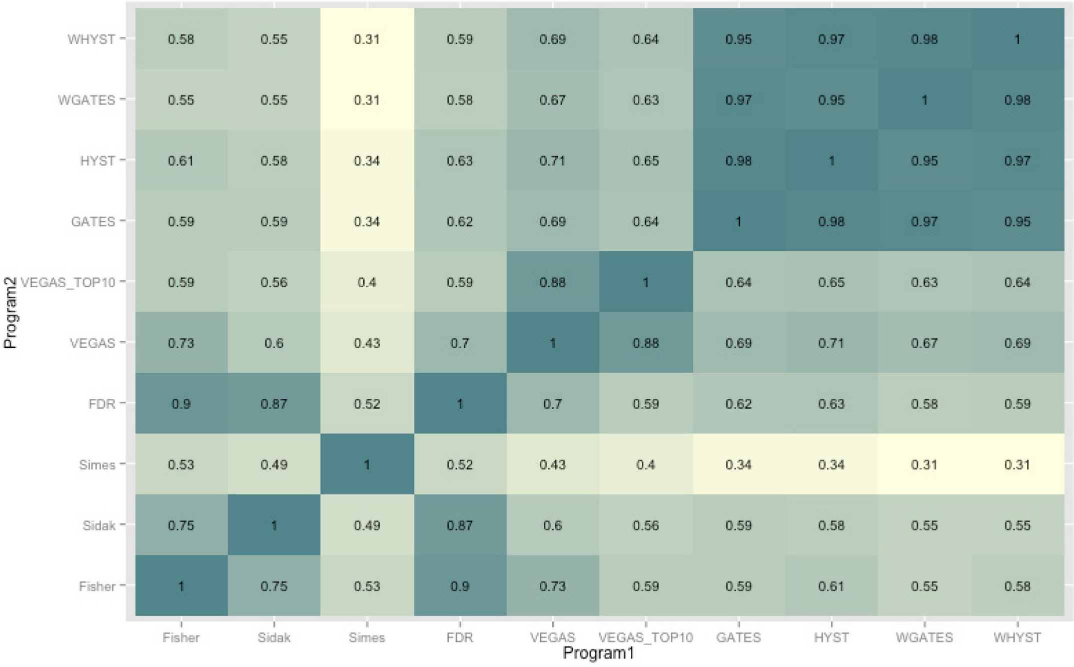
Group	*Method*	*Sensitivity*	*Specificity*	*Type I Error*	*Type II Error*
Classical	Fisher	59.18	88.64	5.89	40.82
	Sidak	18.37	97.73	0.11	81.63
	Simes	46.94	97.73	1.33	53.06
	FDR	24.49	97.73	0.13	75.51
Updated Classical	GATES	24.49	98.00	0.17	75.51
	WGATES	26.53	98.00	0.16	73.47
	HYST	24.49	98.00	0.16	75.51
	WHYST	24.49	98.00	0.16	75.51
Novel	VEGAS	20.41	100.0	0.16	79.59
	VEGAS (top10)	28.57	98.00	0.40	71.43
Regression	aSUM	24.49	100.00	-	-
	Score	18.37	100.00	-	-

**Type I and type II error rates were not estimated for aSUM and Score test due to them being computationally intensive.*

Newer methods all performed similarly. GATES and HYST were nearly identical in their predictions with sensitivity of 24.49%, specificity of 98%, and type I error rates of 0.17% and 0.16% respectively. VEGAS had similar performance with a sensitivity of

20.41% and 100% specificity. Type I error rate was 0.16%. With the exception of Fisher's and Simes' Test, all methods had a type I error rate below 1%.

Correlation was calculated using all genes from the 10 genome-wide programs (Fisher's, Sidak's, Simes', FDR, GATES, Weighted Gates, HYST, Weighted HYST, VEGAS, and VEGAS Top 10%). Correlation in the p-values ranged from 31-98% (Figure 5.5).

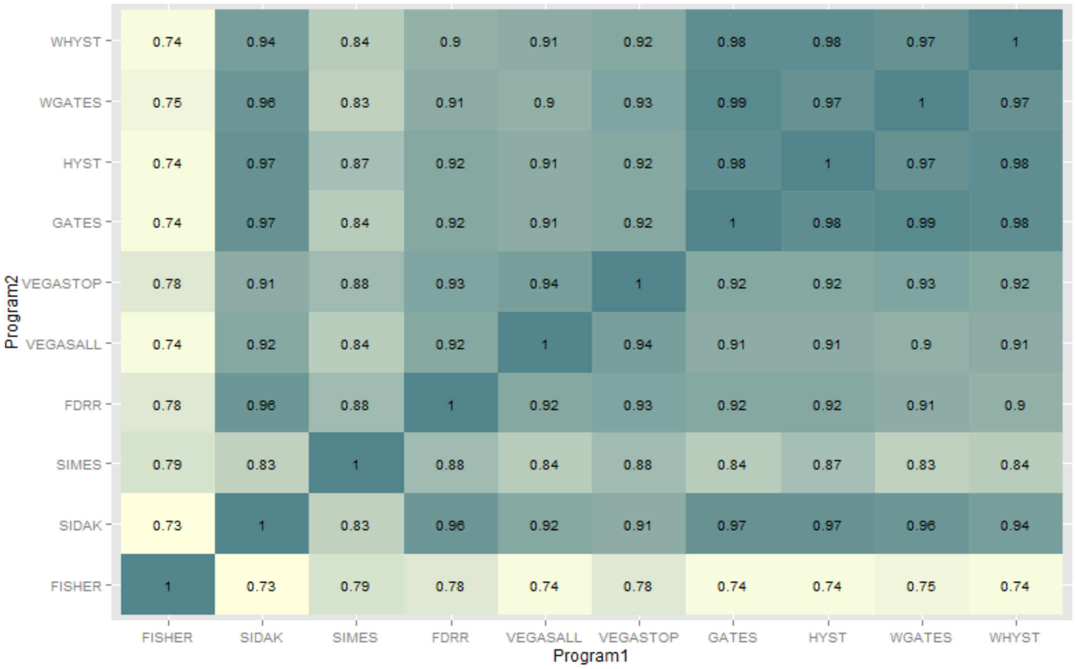


	Fisher	Sidak	Simes	FDR	VEGAS	VEGAS_TOP10	GATES	HYST	WGATES	WHYST
WHYST	0.58	0.55	0.31	0.59	0.69	0.64	0.95	0.97	0.98	1
WGATES	0.55	0.55	0.31	0.58	0.67	0.63	0.97	0.95	1	0.98
HYST	0.61	0.58	0.34	0.63	0.71	0.65	0.98	1	0.95	0.97
GATES	0.59	0.59	0.34	0.62	0.69	0.64	1	0.98	0.97	0.95
VEGAS_TOP10	0.59	0.56	0.4	0.59	0.88	1	0.64	0.65	0.63	0.64
VEGAS	0.73	0.6	0.43	0.7	1	0.88	0.69	0.71	0.67	0.69
FDR	0.9	0.87	0.52	1	0.7	0.59	0.62	0.63	0.58	0.59
Simes	0.53	0.49	1	0.52	0.43	0.4	0.34	0.34	0.31	0.31
Sidak	0.75	1	0.49	0.87	0.6	0.56	0.59	0.58	0.55	0.55
Fisher	1	0.75	0.53	0.9	0.73	0.59	0.59	0.61	0.55	0.58

Figure 5.5: Genome-wide Correlation in P-values for Gene-Level Methods

The highest correlation is found within the previously assigned groups (Classical, Updated Classical, Novel). The updated classical programs (GATES, Weighted GATES, HYST, and Weighted HYST) all had high correlation with each other (>95%). The two VEGAS programs (all and top 10%) had similarly high correlation in their p-values (88%). Surprisingly, the lowest correlation was found between the GATES-associated programs and Simes' (31-34%), considering that GATES is an extended Simes procedure.

Using a α=0.001, concordance was calculated between the 10 programs. Concordance was much higher than correlation, ranging from 93-100%. The high levels of concordance are more due to the large number of true negatives when compared to any other cell. When restricted to the subset of true negative and true positives, the concordances fell to 73-99% (Figure 5.6).



	FISHER	SIDAK	SIMES	FDRR	VEGASALL	VEGASTOP	GATES	HYST	WGATES	WHYST
WHYST	0.74	0.94	0.84	0.9	0.91	0.92	0.98	0.98	0.97	1
WGATES	0.75	0.96	0.83	0.91	0.9	0.93	0.99	0.97	1	0.97
HYST	0.74	0.97	0.87	0.92	0.91	0.92	0.98	1	0.97	0.98
GATES	0.74	0.97	0.84	0.92	0.91	0.92	1	0.98	0.99	0.98
VEGASTOP	0.78	0.91	0.88	0.93	0.94	1	0.92	0.92	0.93	0.92
VEGASALL	0.74	0.92	0.84	0.92	1	0.94	0.91	0.91	0.9	0.91
FDRR	0.78	0.96	0.88	1	0.92	0.93	0.92	0.92	0.91	0.9
SIMES	0.79	0.83	1	0.88	0.84	0.88	0.84	0.87	0.83	0.84
SIDAK	0.73	1	0.83	0.96	0.92	0.91	0.97	0.97	0.96	0.94
FISHER	1	0.73	0.79	0.78	0.74	0.78	0.74	0.74	0.75	0.74

Figure 5.6: Concordance for Significance for Gene-level Methods (α=0.001) Within Gold Standard Set of True Negative and True Positive Genes

The lowest concordance was found with Fisher's Combination Test and Simes' Test with any other method. This is likely due to these programs having the highest type I error. Therefore, they are more likely to call genes as significant that other programs do not call significant. As expected, the highest correlations were within related programs, such as the updated classical methods and the two versions of VEGAS.

5.4.2: Stratified Results

To examine the influence of effect size, sensitivities were estimated among genes that were simulated to have a strong effect size (OR=2) and a weaker effect size (OR=1.2). However, due to the underlying model, only 6 of the true positive genes were simulated based on a weaker effect size. The resulting sensitivities are found in Table 5.4 below.

Table 5.4: Stratified Sensitivities by Effect Size

Group	*Method*	*Sensitivity (OR*=2)*	*Sensitivity (OR*=1.2)*
Classical	Fisher	66%	17%
	Sidak	18%	33%
	Simes	50%	17%
	FDR	27%	17%
Updated Classical	GATES	25%	17%
	GATES [Weighted]	27%	17%
	HYST	25%	17%
	Weighted GATES/HYST	25%	17%
Novel	VEGAS	23%	17%
	VEGAS [Top 10%]	32%	17%

**OR=Odds Ratio*

Sensitivity was higher in the stronger effect sizes when compared to the weaker effect sizes, with the exception of Sidak's Combination Test. Additionally; the stratified sensitivity of strong signals (OR=2) was higher than the overall sensitivity from Table 4.3. This is expected as the genes that were simulated to have a stronger effect size will have lower p-values on a SNP-level which translates to the gene-level analyses.

Genes were also stratified based on the number of causal SNPs from the simulation. Out of the fifty total true positive genes, 8 were simulated using 1 causal SNP, 22 had 2 causal SNPs, and 20 had 5 causal SNPs.

Table 5.5: Stratified Sensitivities by Number of Causal SNPs

Group	*Method*	*Sensitivity (1 SNP)*	*Sensitivity (2 SNPs)*	*Sensitivity (5 SNPs)*
Classical	Fisher	50%	64%	60%
	Sidak	12%	18%	20%
	Simes	50%	50%	45%
	FDR	25%	27%	25%
Updated Classical	GATES	12%	18%	35%
	GATES [Weighted]	25%	18%	30%
	HYST	12%	18%	40%
	GATES/HYST [Weighted]	12%	18%	35%
Novel	VEGAS	0%	27%	25%
	VEGAS [Top 10%]	0%	32%	40%

Within the classical methods, the sensitivity estimates remain relatively consistent between the different number of causal SNPs. For the newer methods, sensitivity increased with the number of causal SNPs. This is consistent with their methodology, which is designed to combine independent signals for an enriched signal. The most extreme sample was in VEGAS [Top10%]. Neither version of VEGAS deemed genes with only one causal SNP as significant. Within genes with two causal SNPs, the sensitivity increased to 32% from the original overall 29%. When there were five causal SNPs, the sensitivity increased to 40%.

5.4.3: Smaller Sample Size Analysis

Within the smaller sample size analysis (n=500), measures of performance were recalculated. Using a significance threshold of *P*<0.001, type I error was found to be consistent from the larger analysis. Within the true negative and true positive genes from the original larger analysis, the majority of methods were unable to detect significant genes in the true positive categories (sensitivity=0%), with the exception of Fisher's Combination Test (sensitivity=12.24%) and Simes' Test (sensitivity=4.08%) (Table 5.6).

Table 5.6: Evaluation of Gene-Level Methods in Smaller Sample Size

Group	*Method/Program*	*Sensitivity*	*Specificity*	*Type I Error*	*Type II Error*
Classical	Fisher's	12.24	95.45	5.32	87.76
	Sidak's	0.00	100.00	0.03	100.00
	Simes'	4.08	100.00	0.98	95.92
	FDR	0.00	100.00	0.05	100.00
Updated	GATES	0.00	100.00	0.10	100.00
	Weighted GATES	0.00	100.00	0.13	100.00
	HYST	0.00	100.00	0.10	100.00
	Weighted HYST	0.00	100.00	0.12	100.00
Novel	VEGAS	0.00	100.00	0.10	100.00
	VEGAS, Top 10%	0.00	100.00	0.26	100.00

All specificity measures were above 95%, with only Fisher's Combination Test not reaching 100% specificity (specificity=95.45%). This is consistent with prior results showing the highest sensitivity and type I error within Fisher's Combination Test when compared to all other methods.

The generation of true positive and true negative genes was recalculated for the smaller analysis using the same steps used in the larger sample size analysis. This lead to only 23 true positive genes which had at least one SNP with a P-value <0.01, and the 50 original true negative genes. The programs were reevaluated with these updated gold standards. The only programs that were affected were Fisher's and Simes' Tests, with their sensitivities elevated to 47.83% and 13.04%, respectively.

If we lower the alpha value to adjust for the smaller sample size and reduced power to α=0.01 while using the updated gold standard of 23 true positive and 50 true negative genes, the performance increases for a few of the programs (Table 5.7).

Table 5.7: Evaluation of Gene-Level Methods in Smaller Sample Size, α=0.01

Group	*Method*	*Sensitivity*	*Specificity*	*Type I Error*	*Type II Error*
Classical	Fisher's	60.87	90.91	8.91	39.13
	Sidak's	8.70	100.00	0.54	91.30
	Simes'	100.00	93.18	8.28	0.00
	FDR	8.70	100.00	0.75	91.30
Updated	GATES	0.00	97.73	1.09	100.00
	Weighted GATES	0.00	97.73	1.11	100.00
	HYST	0.00	97.73	1.05	100.00
	Weighted HYST	4.35	97.73	1.01	95.65
Novel	VEGAS	0.00	100.00	0.92	100.00
	VEGAS, Top 10%	30.43	100.00	2.15	69.57

The most striking differences is seen in Simes' Test with the sensitivity increasing from 4% to 100% by decreasing α by an order of 10. This is likely due to the selection of true positive genes having at least one SNP with p<0.01, and Simes' Test weighting the most significant SNP. With a less stringent α, the type I error increased across the board, increasing by an order of 10 for the majority of the programs.

5.4.4: Potential Biases in Estimation

Gene-level methods for GWAS can be subject to a number of biases, such as gene size, SNP density, and the number of SNPs (both causal and all) considered within the gene. The effect of these variables was estimated using logistic regression. The mean gene size was 83.2 megabases (mb), with on average 176.1 SNPs, while the median gene size was 39.2 mb and 16 SNPs. Accuracy was determined as agreement between the "truth" and significance using α=0.001 for each of the program. Only 2 associations had a P<0.1. Fisher's Combination Test had a p-value of 0.08 showing that the accuracy of the method decreased with an increase in the number of SNPs within the gene. This is consistent with the method violating the inherent assumption of independent tests due to extensive linkage disequilibrium. The other association was between VEGAS using the top 10% of SNPs and the proportion of causal SNPs to total number of SNPs in the gene. Because this method only uses the top 10% of SNPs found in the gene, if the number of causal SNPs makes up a higher proportion of the SNPs, then the program is more accurate. This is consistent for there being enrichment for significance of independent signals in the top 10% of the genic SNPs.

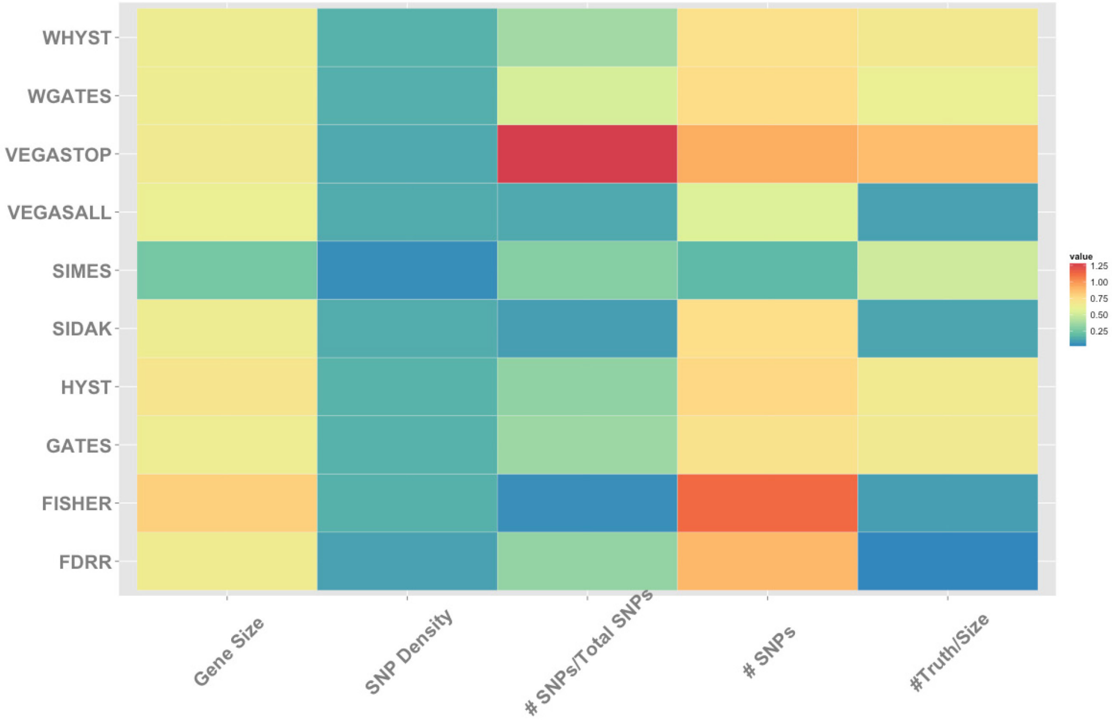


Figure 5.7: Heat map of the -log10 transformation of P-values from univariate logistic regression analyses for the effect of gene characteristics on accuracy. *Programs are organized alphabetically on the y-axis, with the variables on the x-axis. A lower P-value (more significant) is indicated in red.*

5.5: Discussion

The highest sensitivity was found using Fisher's Combination Test (59.18%), which was accompanied by the lowest specificity (88.64%) and the highest type I error (5.89%). This is expected, as Fisher's Combination Test is prone to test statistic inflation. FCT combines *P*-values which are assumed to be independent, but which are not because of linkage disequilibrium between genic SNPs on a GWAS panel. This generalized inflation leads to the highest sensitivity, paired with the highest type I error. The highest specificity was found with VEGAS, one of the more conservative approaches with a sensitivity of 20.41%. VEGAS adjusts for linkage disequilibrium with HapMap data from the CEPH population. This may be an overadjustment, as VEGAS is the most underpowered program, especially when it comes to smaller effect sizes. Within programs that have a type I error rate below 1%, the best balance between the two measures is likely VEGAS using the top 10% of SNPs with a sensitivity of 28.57% and specificity of 98%.

Both correlation and concordance between the programs clustered within related programs, such as GATES and the other updated classical methods (Weighted GATES, HYST, Weighted HYST), as well as the two VEGAS methods (All and Top 10%). The lowest correlation in p-values was found between Simes' Test and any other program (31-53%). This is likely due to Simes' Test only using the weighted most significant SNP, which is influenced by both the number of SNPs in the gene and the distribution of signals within the gene. Surprisingly, the lowest correlation is found between Simes'

Test and the GATES family, which is an extended Simes procedure. Using an α=0.001, concordance rates between the programs was much higher (73-99%). This may be due to the large number of "true negative" genes, which outweighs any other cell in the tabulation. Again, concordances were highest within related programs. The lowest concordance was between Fisher's Combination Test and the other programs (73-79%), most likely due to the highest type I error leading to the most false positives that are not found in other programs.

The stratified analyses reinforce the theory behind genome-wide association studies and a truly polygenic model. Within the simulation, the smaller effect sizes are underrepresented within SNPs with P<0.01, despite originally having equal weighting with the genes simulated upon higher effect sizes. Out of the 50 true positive genes, only 6 of them were originally simulated to have the smaller effect size (OR=1.2), despite that the original 226 genes were split evenly between the two effect sizes (OR=1.2 vs OR=2). This is consistent with larger effect sizes having increased power compared to smaller effect sizes. Sensitivity was increased for all programs within the stronger effect genes. The number of independent causal SNPs also had a large effect on the program's sensitivity. For most programs, sensitivity increased when the number of causal SNPs, and therefore independent signals, was increased. VEGAS, in either iteration, was unable to detect genes which only had one causal SNP while increasing the sensitivity within genes with 2 or 5 independent causal SNPs. If the underlying hypothesis is that there are multiple causal SNPs within a gene that could be contributing to the outcome as is the case with allelic heterogeneity, then this program will help to differentiate between genes that have multiple signals due to linkage disequilibrium or multiple independent biologically relevant signals.

A GWAS with a smaller sample size is woefully underpowered to detect signals, both in a traditional analysis as well as with these gene-level methods. Using the previously defined α=0.001, only Fisher's and Simes' Test detected any significant true positive genes. When α was increased to 0.01, sensitivity increased, however 4/10 programs still did not find any of the true positive genes to be significantly associated. There was also a large increase in type I error, leading to 7/10 programs having type I error above 1%, an unacceptable rate. Because of this large type I error, it is not recommended to lower the threshold for significance just because of sample size. On the other hand, if a gene is deemed significant with α=0.001 within smaller sample sizes, there is more confidence in the results.

All programs were relatively immune to theoretical gene size biases, however the absolute number of SNPs in the gene made more of a difference. Consistent with violating the underlying assumption of independence in Fisher's Combination Test, an increase in the number of SNPs resulted in a less accurate analysis. The proportion of causal SNPs to the total number of SNPs in the gene influenced the accuracy of VEGAS using the top 10% SNPs, increasing the accuracy with the higher proportion of causal SNPs.

When using gene-level methods to elucidate biological significance within GWAS results that fail to reach the genome-wide significance threshold, it is important to keep in mind the limitations of gene-level methods. Power to detect signals is limited, especially for smaller effect sizes. However, all programs identified genes that would have otherwise been ignored by a traditional GWAS. Fisher's Combination Test had the highest sensitivity, but also the highest type I error, therefore it should only be used if there is a low cost follow-up in place. VEGAS had the highest specificity, being the most conservative program with low type I error (0.16%). A good compromise would be to use VEGAS with the option of only using the top 10% of SNPs within a gene, with higher sensitivity (29%) and specificity (98%) coupled with low type I error (0.40%). Additionally, VEGAS was able to distinguish between genes with only one versus multiple causal variants. Gene-level methods can help to find genes that would previously have been ignored, but the programs are not all the same and they have individual caveats and limitations

5.6: Supplementary Methods

5.6.1: Code for Gene-Level Methods

Fisher's, Sidak's and Simes' Tests were performed within R using a user-created script. The major functions are shown below.

Fisher

```
for (i in 1:nrow(fish)){
  x=key[key$gene==fish[i,1],]
  y=merge(x, res, by.x="rsid", by.y="SNP")
  fish[i,2]=nrow(y)
  fish[i,3]=-2*sum(log(y$P))
  fish[i,4]=1-pchisq(as.numeric(fish[i,3]),
df=2*as.numeric(fish[i,2]))
}
```

Sidak

```
for (i in 1:nrow(sidak)){
  x=key[key$gene==sidak[i,1],]
  y=merge(x, res, by.x="rsid", by.y="SNP")
  sidak[i,2]=nrow(y)
  sidak[i,3]=min(y$P)
  sidak[i,4]=(1-(1-min(y$P))^nrow(y))
}
```

Simes

```
for (i in 1:nrow(simes)){
  x=key[key$gene==simes[i,1],]
  y=merge(x, res, by.x="rsid", by.y="SNP")
  y=y[order(-y$P),]
  simes [i,2]=nrow(y)
  if (nrow(y)>0) {
    y$rnk=1:nrow(y)
    y$simes=nrow(y)*y$P/y$rnk
    simes[i,3]=y[y$simes==min(y$simes),]$simes
  }
    print(i)
}
```

The False Discovery Rate (FDR) method utilized the p.adjust function from within R. (http://stat.ethz.ch/R-manual/R-devel/library/stats/html/p.adjust.html) The utilization of this package can be seen below.

FDR

```
for (i in 1:nrow(fish)){
  x=key[key$gene==fish[i,1],]
  y=merge(x, res, by.x="rsid", by.y="SNP")
  fish[i,2]=nrow(y)
  fish[i,3]=min(y$P)
  fish[i,4]=min(p.adjust(y$P, method="fdr"))
}
```

GATES and HYST were conducted within the Graphical User Interface (GUI) provided by the authors. (http://bioinfo.hku.hk:13080/kggweb/) Written within a java script, the program requires a user-defined reference dataset for LD estimation. While HapMap populations are available for download, the WTCCC data was used to build a genome for both analyses. Both a weighted GATES and HYST program were available, but they yielded the same results as their unweighted counterparts in this simulation.

VEGAS was run using a command-line interface. While a web-interface is available (http://gump.qimr.edu.au/VEGAS/) a command-line interface allows a script to be reproducible. A gene-list with correct build coordinates was created from Entrezgenes FTP data. Chromosomes were run separately, using HapMap's CEU data as LD references. The default test uses all SNPs within the gene. An additional option was run using the top 10% of associated SNPs within the genic region. The method was not used with custom LD estimation, due to its computationally intensive nature.

References

de Bakker, Paul I W, Roman Yelensky, Itsik Pe'er, Stacey B Gabriel, Mark J Daly, and David Altshuler. 2005. "Efficiency and Power in Genetic Association Studies." *Nature Genetics* 37 (11) (October 23): 1217–1223. doi:10.1038/ng1669.

Han, Fang, and Wei Pan. 2010. "A Data-Adaptive Sum Test for Disease Association with Multiple Common or Rare Variants." *Human Heredity* 70 (1): 42–54. doi:10.1159/000288704.

Hindorff, Lucia A, Praveen Sethupathy, Heather A Junkins, Erin M Ramos, Jayashri P Mehta, Francis S Collins, and Teri A Manolio. 2009. "Potential Etiologic and Functional Implications of Genome-Wide Association Loci for Human Diseases and Traits.." *Pnas* 106 (23) (June 9): 9362–9367. doi:10.1073/pnas.0903103106.

Hinrichs, A S. 2006. "The UCSC Genome Browser Database: Update 2006." *Nucleic Acids Research* 34 (90001) (January 1): D590–D598. doi:10.1093/nar/gkj144.

Li, Miao-Xin, Hong-Sheng Gui, Johnny S H Kwan, and Pak C Sham. 2011. "GATES: a Rapid and Powerful Gene-Based Association Test Using Extended Simes Procedure." *American Journal of Human Genetics* 88 (3) (March 11): 283–293. doi:10.1016/j.ajhg.2011.01.019.

Li, Miao-Xin, Johnny S H Kwan, and Pak C Sham. 2012. "HYST: a Hybrid Set-Based Test for Genome-Wide Association Studies, with Application to Protein-Protein Interaction-Based Association Analysis." *American Journal of Human Genetics* 91 (3) (September 7): 478–488. doi:10.1016/j.ajhg.2012.08.004.

Liu, Jimmy Z, Allan F Mcrae, Dale R Nyholt, Sarah E Medland, Naomi R Wray, Kevin M Brown, AMFS Investigators, et al. 2010. "A Versatile Gene-Based Test for Genome-Wide Association Studies." *American Journal of Human Genetics* 87 (1) (July 9): 139–145. doi:10.1016/j.ajhg.2010.06.009.

Peng, Qianqian, Jinghua Zhao, and Fuzhong Xue. 2009. "A Gene-Based Method for Detecting Gene-Gene Co-Association in a Case-Control Association Study." *European Journal of Human Genetics* 18 (5) (December 23): 582–587. doi:10.1038/ejhg.2009.223.

Purcell, Shaun, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel A R Ferreira, David Bender, Julian Maller, et al. 2007. "PLINK: a Tool Set for Whole-Genome Association and Population-Based Linkage Analyses." *The American Journal of Human Genetics* 81 (3) (September): 559–575. doi:10.1086/519795.

Subramanian, Aravind, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael A Gillette, Amanda Paulovich, et al. 2005. "Gene Set Enrichment Analysis: a Knowledge-Basedapproach for Interpreting Genome-Wideexpression Profiles." *Proceedings of the National Academy of Sciences* 102 (43) (October 25): 15545–15550. doi:10.1073/pnas.0506580102.

Venter, J C, M D Adams, E W Myers, P W Li, R J Mural, G G Sutton, H O Smith, et al.

2001. “The Sequence of the Human Genome..” *Science* 291 (5507) (February 16): 1304–1351. doi:10.1126/science.1058040.

Yang, Jian, Beben Benyamin, Brian P McEvoy, Scott Gordon, Anjali K Henders, Dale R Nyholt, Pamela A Madden, et al. 2010. “Common SNPs Explain a Large Proportion of the Heritability for Human Height.” *Nature Publishing Group* 42 (7) (June 20): 565–569. doi:10.1038/ng.608.

Zaitlen, Noah, Peter Kraft, Nick Patterson, Bogdan Pasaniuc, Gaurav Bhatia, Samuela Pollack, and Alkes L Price. 2013. “Using Extended Genealogy to Estimate Components of Heritability for 23 Quantitative and Dichotomous Traits.” Edited by Peter M Visscher. *PLoS Genetics* 9 (5) (May 30): e1003520. doi:10.1371/journal.pgen.1003520.s011.

Chapter 6: Evaluation and Application of Pathway-Level Methods for Genome-Wide Association Studies

6.1: Abstract

Background: In the past ten years, many investigators have used the genome-wide associated study (GWAS) design to identify risk loci for various phenotypes. While there has been success with this route, there still remains "missing heritability" for most traits. Many biologically relevant associations may have strong signals, but fail to meet the stringent genome-wide significance threshold (5x10^{-8}). To address this problem, a variety of pathway-level methods were developed to identify putative biologically relevant associations but they do not test for gene-gene interactions. There is currently no consensus as to the best method. A simulation was conducted to systematically assess the performance of pathway-level methods.

Methods: Using genetic data from the Wellcome Trust Case Control Consortium (WTCCC), a case-control status was simulated based on an additive polygenic model where cases have more risk alleles than controls using a traditional GWAS sample size (2,500 cases, 2,500 controls). A total of 20 pathways and 226 genes were selected from GO biological processes to create this simulated phenotype. We evaluated 10 different methods and examined the influence of pathway size and proportion of simulated "causal" genes. The simulation resulted in only 15 pathways having associated genes, consisting of 9-33% of the gene set’s total genes. Because of high computational burden, four of these programs were only run on the subset of 20 pathways (PST and GRASS), both self-contained tests. All competitive methods (ALIGATOR, gengen, MAGENTA, GSEA-SNP, SRT, and GSA-SNP) were run on the full GO biological processes (N=825).

Results: All methods were highly dependent upon the database used. INRICH is the most conservative approach and is unique among the methods for its use of linkage disequilibrium blocks instead of genes as the second level of analysis. The least conservative approach was using GRASS with an asymptotic distribution, which led to false positives especially in small pathways. By incorporating permutations, the false positives were decreased but not eliminated. Gengen, MAGENTA, and GSA-SNP (all competitive methods) clustered together in their performance, with lower P-values being associated with a higher proportion of “causal” genes.

Conclusions: Pathway-level methods should always be interpreted within the context of the database that is utilized. Competitive methods require the analysis of a large number of gene sets, as well as the entire genome-wide association data set. While the self-contained tests were less computationally intensive and only required candidate pathways, they were less accurate. These results support the underlying hypothesis of a polygenic model in elucidating biologically relevant genetic relationships in genome-wide association studies.

6.2: Introduction

When the genome-wide association study (GWAS) was designed, it relied upon the "common disease, common variant" (CDCV) hypothesis. This hypothesizes that common diseases, such as Type II Diabetes, are due to common genetic variants. These SNPs should be easily detectable in population samples through association. However, the results have fallen short of expectation. Many traits still have a large amount of missing heritability—the proportion of phenotypic variability due to genetics rather than environmental influences. It has been hypothesized that the missing heritability may be due to the truth being in between the CDCV model and the infinitesimal model, in which the phenotypic variance is explained by an "infinite" number of small effect variants.(Gibson 2012) are typically underpowered to detect smaller effect size, essentially ignoring suggestive associations with these smaller effect SNPs. To address this issue, a number of pathway-level analytical methods were developed for GWAS results.

Pathway-level methods for GWAS aim to examine if genetic associations within a GWAS are enriched within a set of genes, or pathway. This goal is different than the previously described gene-level method in that the gene-level programs aim to aggregate signals into a joint association test statistic. Pathway-level methods differ in that multiple association signals due to allelic heterogeneity are often ignored. These methods differ in their assessment of "enrichment", whether it is top ranked genes or an aggregate test statistic looking for joint association between the genes. With a wide variety of methods recently published, the field still lacks a consensus as to the best method. To address this knowledge gap, we evaluated 10 different programs using a simulation of real genotypic data from the Wellcome Trust Case Control Consortium on 20 pathways from the Gene Ontology Biological Processes.

6.3: Materials and Methods

6.3.1: Genotype Data

Genotype data was obtained from the Wellcome Trust Case-Control Consortium (WTCCC) following their release procedures. The Wellcome Trust Case-Control consortium genotype data included in this study was genome-wide SNP data off a custom Illumina 1.2M chip from the 1958 Birth Cohort (N=2,930) and the National Blood Service samples (N=2,737). Standard quality control measures were performed (previously described in Section 5.3.1.1). Principal components analysis was performed to evaluate ancestry and outliers were removed, reducing the sample size to 4,500 individuals of European ancestry. Additional filters were applied including minor allele frequency > 5%, Hardy-Weinberg Equilibrium ($p < 10^{-5}$) and genotype missing rate < 5% resulting in 906,298 genome-wide SNPs.

6.3.2: Simulation

Pathways were downloaded from the Molecular Signatures Database (MSigDB) for the Gene Ontology (GO) Biological Processes (BP).(Subramanian et al. 2005) This database was chosen for consistency since the majority of methods used this in their programs The GO BP are categorized as a series of events or molecular functions that have a beginning and an end. Out of all the GO terms, the biological processes most resemble a canonical pathway, such as those found in KEGG or BioCarta.

There are 825 total biological processes with a median number of 28 genes in the database. From these 20 pathways two groups were selected: 10 with greater than the median number of genes (large) and 10 with under the median number of genes (small). From each of these 20 pathways, a subset of genes were also selected. Within each size group (small/large), four pathways were selected to have only 1 gene, four pathways had 20% of their genes selected, and in two pathways 50% of the genes were selected. Any genes in numerous pathways were removed so that the pathways were "independent" of each other. This resulted in the inclusion of 226 genes (Table 5.1).

The selection of "causal" SNPs from within each of the genes is described in more detail in section 5.3.1.3 under "Phenotype Generation". In short, a number of "causal" independent SNPs were chosen from each gene. Each gene had one, two, or five causal SNPs included, and an odds ratio of either 1.2 or 2 assigned to them. An additive polygenic model was used to generate an underlying liability score, which was standardized to the mean. With some overlap generated by adding a random amount of variation to the liability score, individuals were stochastically assigned to case or control status according to their transformed liability score generating 2,250 cases and 2,250 controls. An unadjusted logistic regression was run on all the SNPS in PLINK for a traditional GWAS analysis. In addition, a standard case/control association using chi-squared statistics was also conducted on the same data in PLINK for methods that required a chi-squared statistic rather than the Z-score generated in logistic regression.

Since the simulation followed a stochastic process, the proportion of genes that were simulated to be associated with the outcome did not always result in a GWAS association. Additional details on the simulation are included in Chapter 5.

The results of the simulation for the pathways are detailed in Table 6.1. The "true positive" genes were annotated as such if they had at least one SNP within the genic region with P<0.01. Many of the pathways had fewer "true positive" genes with at least one SNP having a p-value < 0.01 than intended through the simulation. For example, the first pathway listed "Anatomical Structure Morphogenesis" was simulated to have 73 associated genes, which is approximately 20% of the total 363 genes. However, only 50 of these genes had at least one SNP with a p-value < 0.01 (14%). This was especially pronounced in the smaller pathways, with 5 pathways having no genes at all associated (P_{SNP}<0.01). However, if we include genes with a P-value below 0.05, then many pathways have more "associated" genes than the simulation intended. Thus, we used a cut-off of P<0.01 for a truly "associated" gene.

Table 6.1: Pathway Characteristics from Simulation

GOID	*Biological Process*	*# Genes*	*% of Genes*	*# Simulated Genes*	*# P<0.01*	*% P<0.01*
GO:0009653	Anatomical Structure Morphogenesis	363	20	73	50	14
GO:0008277	G-Protein Coupled Receptor Protein Signaling Pathway	332	20	67	40	12
GO:0006643	Membrane Lipid Metabolic Process	98	(1)	1	15	15
GO:1902275	Establishment and/or Maintenance of Chromatin Architecture	71	50	36	9	13
GO:0006869	Lipid Transport	29	50	15	8	28
GO:0045321	Leukocyte Activation	65	(1)	1	7	11
GO:0006968	Cellular Defense Response	55	(1)	1	6	11
GO:0032488	CDC42 Protein Signal Transduction	12	50	6	4	33
GO:0007189	G-Protein Signaling Adenylate Cyclase Activating Pathway	25	20	5	4	16
GO:0042110	T-Cell Activation	41	20	9	4	10
GO:0051101	Regulation of DNA Binding	44	(1)	1	4	9
GO:0043241	Protein Complex Disassembly	15	(1)	1	3	20
GO:0002009	Morphogenesis of an Epithelium	17	20	4	3	18
GO:0001666	Response to Hypoxia	28	20	6	3	11
GO:0051607	Defense Response to Virus	11	20	3	2	18
GO:0051650	Establishment of Vesicle Localization	10	(1)	1	0	0
GO:0000080	G1 Phase of Mitotic Cell Cycle	12	20	3	0	0
GO:0000209	Protein Polyubiquitination	10	(1)	1	0	0
GO:0009259	Ribonucleotide Metabolic Process	17	(1)	1	0	0
GO:0007051	Spindle Organization and Biogenesis	10	50	5	0	0

6.3.3: Programs

Using the simulated data, we evaluated the following pathway programs: Meta-Analysis Gene-set Enrichment of variaNT Associations (MAGENTA)(Segrè et al. 2010) Interval-based Enrichment Analysis Tool for Genome Wide Association Studies (INRICH)(Lee et al. 2012), Plink Set Test(Purcell et al. 2007), Gene Set Analysis for SNPs (GSA-SNP)(Nam et al. 2010) Gene Set Enrichment Analysis for SNP data (GSEA-SNP) (Holden et al. 2008) , Gene Set Ridge Regression in Association Studies (GRASS) (Chen et al. 2010), Association List Go AnnoTatOr (ALIGATOR)(Holmans et al. 2009), GenGen (Wang, Li, and Bucan 2007), Hybrid Set-Based Test for Genome-wide Association Studies (HYST)(M.-X. Li, Kwan, and Sham 2012) and SNP Ratio Test (SRT)(O'Dushlaine et al. 2009) .

These programs can be divided into two categories: competitive and self-contained. Competitive programs compute their test statistics depending on the distribution of all gene set test statistics. Therefore, the results are in comparison with the other gene sets that were used for this analysis. With these programs, it is important to do a genome-wide approach, instead of a candidate gene set approach, because of the dependence on a distribution of test statistics that is representative of largely the null. On the other hand, self-contained tests do not depend on other gene sets, so can be used on both genome-wide and candidate studies. These programs often use permutations to form a test statistic null distribution.

Table 6.2: Programs Evaluated by Category			
Category	*Program*	*Input*	*Citation*
Competitive	ALIGATOR	SNP P-values	(Holmans et al. 2009)
	GENGEN	SNP P-values	(Wang, Li, and Bucan 2007)
	GSA-SNP	SNP P-values	(Nam et al. 2010)
	GSEA-SNP	Raw Genotypes	(Holden et al. 2008)
	MAGENTA	SNP P-values	(Segrè et al. 2010)
	SRT	Raw Genotypes	(O'Dushlaine et al. 2009)
Self-Contained	GRASS	Raw Genotypes	(Chen et al. 2010)
	HYST	SNP P-values	(M.-X. Li, Kwan, and Sham 2012)
	INRICH	Genomic Coordinates	(Lee et al. 2012)
	PLINK	Raw Genotypes	(Purcell et al. 2007)

All programs allow the user to define the assignment of SNPs to genes. For consistency, SNPs were assigned to a gene if they were within the translated region and if they were within 20 kilobases of either end of the gene.

6.3.3.1: Competitive Methods

1. **ALIGATOR**(Holmans et al. 2009) : ALIGATOR is a method that looks for the enrichment of significant genes within Gene Ontology gene sets. The input is SNP p-values. ALIGATOR then filters by a pre-set P-value threshold ($p < 0.05$). Any gene that has at least one SNP below this P-value threshold is

annotated as being “significant”. Simulations are then conducted in which SNPs are randomly drawn from the GWAS and if they are in a gene that gene is added to the simulated gene list. This is repeated until the gene list is the same length as the original study’s significant gene list. This process is repeated to form 5,000 null gene lists. An empirical p-value is then calculated from the distribution of these gene lists in GO pathways. Because of this simulation procedure, this method is categorized as competitive. Multiple comparisons issues are controlled using a bootstrap procedure. This method is dependent upon all genes within a GO set having comparable linkage disequilibrium patterns. When a gene set has higher levels of linkage disequilibrium, the estimate tends to be overconservative. A total of 1,000 permutations were used in this analysis.

2. **GenGen**(Wang, Li, and Bucan 2007): GenGen is the oldest method available, using a modified GSEA which was originally developed for gene expression analyses. The most significant SNP is assigned as the gene’s overall P-value. Genes are then sorted by their significance from smallest to largest p-value. Using these rankings, a Weighted Komogorov-Smirnov-like running sum Enrichment Score (ES) is calculated to see the overrepresentation of highly ranked genes within the gene set. Phenotype permutation adjusts for gene size biases. The original ES is the normalized by the permutations’ enrichment scores to form a Normalized Enrichment Score (NES). A False Discovery Rate (FDR) or a Family-Wise Error Rate (FWER) can be used to

control for multiple comparisons. This method is also competitive. A thousand permutations are used to calculate the normalized enrichment scores.

3. **GSA-SNP**(Nam et al. 2010): GSA-SNP is an updated method adapted from gene expression studies. It uses the –log transformed SNP p-values as an input and the *k*th most significant SNP is selected as the gene-level P-value (default *k*=2). This is to minimize the effect of spurious associations for the top SNP (??) in the summarization of gene-level statistics. Three different methods are then offered within the package: (1) Z-score, (2) Restandardized-GSA, and (3) GSEA. The Z-score compares the average gene score within the gene set to an overall distribution. Both the Restandardized-GSA and GSEA use permutations to assess significance with pooled set scores. GSA-SNP is available as a graphical user interface (GUI).
4. **GSEA-SNP**(Holden et al. 2008) : A direct adaptation of the original GSEA algorithm(Subramanian et al. 2005), GSEA-SNP uses the raw genotypes as an input. Three inheritance models (recessive, dominant, and additive) are used and the most significant test statistic is calculated per SNP. These test statistics are then ranked genome-wide. Using a running sum statistic, an enrichment score is calculated to determine if a gene set's SNPs are overrepresented at the top of the genome-wide SNP list. This ES is tnormalized by the gene size to establish a Normalized Enrichment Score (NES). A false discovery rate is calculated to control for false positives. For

this project, GSEA-SNP was conducted as part of the SNPath package within R (http://linchen.fhcrc.org/grass.html).

5. **MAGENTA**(Segrè et al. 2010): MAGENTA requires SNP P-values as input, mapping SNPs to genes and using the most significant SNP P-value within that gene as the raw gene-level P-value. Gene p-values can be adjusted for multiple confounders, such as gene size, using regression and permutations. The adjusted gene-level P-values are ranked and "significant" genes are selected using a static cut-off, such as the 95th percentile. Gene sets are checked against this list of significant genes for over-enrichment, similar to a standard GSEA analysis. The rank can also be decreased if a polygenic model is hypothesized (i.e.75th percentile and up).
6. **SNP Ratio Test** (SRT)(O'Dushlaine et al. 2009): The SNP Ratio Test requires SNP P-values as input, as well as the SNP P-values from permutations calculated using Plink. Using a p-value threshold determined by the user, the ratio of significant SNPs to the number of all SNPs within a pathway is calculated. Gene-level classifications are ignored. Using permutations, an empirical p-value is calculated for the distribution of this ratio. The P-value threshold for SNP significance can be adjusted depending on the hypothesis. For example, a lower P-value threshold (0.01) would assume numerous smaller effects being important in contrast to a few large effects with a more stringent threshold (P=0.001). A total of 1,000 permutations were conducted in this simulation evaluation study.

6.3.3.2: Self-Contained Tests

1. **GRASS** (Chen et al. 2010): GRASS requires raw genotypes to directly estimate the genetic architecture of the genes involved in the evaluated gene sets/pathways. Within each gene, a Principal Components Analysis (PCA) is conducted to determine the SNPs that represent the unique linkage disequilibrium patterns. These "nontrivial" SNPs are then fed into a Group Ridge Regression with Lasso penalty to determine the "most representative eigenSNPs" in regards to their association with disease risk. A gene set association is then conducted by summarizing all of the effects from these "most representative eigenSNPs" across an entire gene set. Permutations are used to create a null distribution and calculate a P-value. For this analysis 1,000 permutations were used.
2. **HYST** (M.-X. Li, Kwan, and Sham 2012) **:** HYST is an extension to the gene-level method of GATES. (M.-X. Li et al. 2011) HYST uses the same graphical user interface (GUI) as GATES (KGG2.5). GATES is an extended Simes procedure to assess gene-level associations that directly accounts for linkage disequilibrium patterns by selecting "independent" SNPs. After performing GATES, HYST uses a scaled chi-square test to assess significance on the GATES P-values output., similar to the Fisher's Combination Test used in gene-level analyses. Prior weights can be incorporated into the blocks, or genes, if appropriate.

3. **INRICH** (Lee et al. 2012): INRICH uses associated genomic intervals as the input. These intervals are determined using a SNP p-value threshold (0.001) and included the surrounding SNPs in linkage disequilibrium around this index SNP. These intervals are estimated in a program such as Plink and then INRICH tests for the number of intervals that overlap with the target genes in any given gene set. Permutations for intervals of the same length are calculated to determine an empirical P-value separately for each gene set. A multiple comparisons correction is applied using additional permutations for the minimum empirical P-value across all sets analyzed.
4. **Plink Set Test**(Purcell et al. 2007) **:** The Plink Set Test assesses the joint significance of a set of SNPs, whether they be within a gene, or within a pathway. Using raw genotype data, the linkage disequilibrium patterns are estimated using all SNPs in a region. After single SNP-association testing, only SNPs below a certain P-value threshold are selected. Then, in decreasing order of significance, "independent" SNPs within that set are selected to be representative of the overall genetic variation in that region using the original LD patterns estimated from the raw genotype data. The average statistic within these "independent" SNPs is then used as the original set statistic. Permutation of the phenotype is conducted to determine an empirical p-value for the set.

6.3.4: Program Evaluation

In a subset of these programs, we included only the 20 simulated pathways because of the computational burden: Plink Set Test, GRASS, GSEA-SNP, and ALIGATOR. The rest of the programs (MAGENTA, INRICH, SRT, HYST, and GSA-SNP) were run on ***all*** Gene Ontology Biological Processes (N=824). However, GO processes are not independent from each other, and some genes may be involved in numerous processes. This is due to the hierarchical nature of Gene Ontology, and the pleiotropic effects of many genes. Because only a small number of pathways were evaluated in all programs, standard measures such as sensitivity and specificity as well as Type 1 and Type II errors could not be calculated for these programs. Instead, a qualitative assessment was conducted. Correlation between programs was calculated using Spearman's correlations within the R software package. Pathways below a p-value threshold of 0.001 were considered significant.

6.4: Results

6.4.1: Pathway-method level results

Of the 10 programs, 6 programs had at least one pathway that was below a threshold of 0.001 (gengen, PST, GSA-SNP, GRASS, GSEA-SNP, and HYST). The most consistently significant small pathway was the “Defense Response to Viruses” with 11 total genes. Half of the methods categorized this pathway as significant (P<0.001) and 70% of the methods had P-values below 0.01. Of the larger pathways, “Lipid Transport” consistently yielded lower P-values with 30% of the methods categorizing this as significant. INRICH had the least significant P-values (13/20 pathways, *P*-values=1), meaning none of the permutations had more extreme values than the original data. The method with the most significant P-values was HYST, with five pathways having *P*<0.001. Pathways in which there were no causal genes (all smaller pathways) did not have any significant results. No pathways were found to be significant that had less than 12% causal genes.

Table 6.3: Results from Pathway Analysis for Larger Pathways

Biological Process	*# Genes*	*# P<0.01*	*% P<0.01*	***Competitive Programs***						***Self-Contained Programs***			
				ALI.	*gengen*	*GSA*	*GSEA*	*MAG*	*SRT*	*GRASS*	*HYST*	*INR.*	*PST*
Lipid Transport	29	8	27.59%	0.244	0.005	*4.14E-04*	0.186	0.022	0.02	*<0.001*	*5.42E-08*	0.133	0.06
Membrane Lipid Metabolic Process	98	15	15.31%	0.198	0.143	0.056	0.057	0.018	0.127	0.014	0.02	0.169	*<0.001*
Anatomical Structure Morphogenesis	363	50	13.77%	0.457	0.055	*7.94E-06*	0.496	0.161	0.549	*<0.001*	0.1	0.893	1
Establishment and/or Maintenance of Chromatin Architecture	71	9	12.68%	0.983	0.036	0.113	0.896	0.033	0.002	0.116	*7.58E-09*	0.015	0.06
G-Protein Coupled Receptor Protein Signaling Pathway	332	40	12.05%	0.515	0.663	0.005	0.267	0.026	0.691	*<0.001*	0.19	0.51	0.99
Cellular Defense Response	55	6	10.91%	0.642	0.104	0.009	0.829	0.126	0.026	0.374	0.04	1	0.01
Leukocyte Activation	65	7	10.77%	0.996	0.761	0.534	0.955	0.944	0.246	0.146	0.74	0.804	0.45
Response to Hypoxia	28	3	10.71%	0.915	0.116	0.409	0.658	0.312	0.621	0.055	0.12	1	0.15
T-Cell Activation	41	4	9.76%	0.929	0.475	0.275	0.823	0.903	0.241	0.089	0.24	1	0.25
Regulation of DNA Binding	44	4	9.09%	0.962	0.838	0.949	0.918	0.93	0.287	0.907	0.18	1	0.87

**ALI= ALIGATOR, GSA=GSA-SNP, GSEA=GSEA-SNP, MAG=MAGENTA, SRT=SNP Ratio Test, INR=Inrich, PST=Plink Set Test*

Table 6.4: Results from Pathway Analysis for Smaller Pathways

Biological Process	*# Genes*	*# P<0.01*	*% P<0.01*	***Competitive Programs***						***Self-Contained Programs***			
				ALI.	*gengen*	*GSA*	*GSEA*	*MAG*	*SRT*	*GRASS*	*HYST*	*INR.*	*PST*
CDC42 Protein Signal Transduction	12	4	33.33%	0.996	0.357	0.097	0.942	0.227	0.026	0.583	0.24	1	1
Protein Complex Disassembly	15	3	20.00%	1	0.122	0.452	0.74	0.111	0.136	0.228	*3.30E-03*	1	1
Defense Response to Virus	11	2	18.18%	0.6	*<0.001*	*2.71E-09*	0.007	0.08	0.002	*<0.001*	*2.12E-05*	1	*<0.001*
Morphogenesis of an Epithelium	17	3	17.65%	0.109	0.042	0.032	0.016	0.016	0.221	0.002	0.04	1	0.07
G-Protein Signaling Adenylate Cyclase Activating Pathway	25	4	16.00%	0.47	0.167	0.026	0.724	0.053	0.95	0.345	*1.76E-04*	0.371	1
Establishment of Vesicle Localization	10	0	0.00%	0.551	0.89	0.851	0.595	0.689	1	0.137	0.75	1	0.78
G1 Phase of Mitotic Cell Cycle	12	0	0.00%	1	0.713	0.857	0.868	0.803	1	0.885	0.31	1	1
Protein Polyubiquitination	10	0	0.00%	1	0.811	0.934	0.716	0.764	1	0.986	0.79	1	1
Spindle Organization and Biogenesis	10	0	0.00%	0.388	0.702	0.888	0.543	0.471	1	0.008	0.15	1	0.28
Ribonucleotide Metabolic Process	17	0	0.00%	0.993	0.428	0.674	0.358	0.801	1	0.544	0.57	1	0.41

**ALI= ALIGATOR, GSA=GSA-SNP, GSEA=GSEA-SNP, MAG=MAGENTA, SRT=SNP Ratio Test, INR=INRICH, PST=Plink Set Test*

6.4.2: Competitive versus Self-Contained Methods

Pathway-level methods can be separated into two groups: competitive (ALIGATOR, GenGen, GSA-SNP, GSEA-SNP, MAGENTA, SNP Ratio Test) and self-contained (GRASS, HYST, INRICH, Plink Set Test). We evaluated these two groups using the results from the larger pathways. Self-contained tests had more significant findings than the competitive methods with the exception of INRICH. Within the competitive methods, only two gene sets were significant and only in GSA-SNP. However, within the five pathways with the most causal genes (12-28%), at least one self-contained method found them significant. INRICH, a self-contained approach, was an outlier for self-contained methods with no pathways being considered significant. This trend is exaggerated within the smaller pathways. Competitive methods only found one gene set to be significant ("Defense Response to Virus") while self-contained methods found three gene sets significant, but also many more gene sets with a *P*-value of 1. Because the smaller pathways had such few causal genes, they were not considered in further evaluation.

6.4.3: Rankings and the Influence of Proportion of "Causal" Genes

Many of the programs are competitive with their performance and depend on the distribution of the other gene set. We examined the rankings. Within each program the *P*-values for the sets were ranked from smallest (1) to largest (10). For each pathway, the mean ranking was calculated across the 10 programs. The correlation between the proportion of genes associated within the gene set and the mean ranking was -0.75, indicating that the larger the proportion of causal genes, the smaller the P-value. This is consistent with methodology and the goals of the program. Correlations between the programs' rankings and the proportion of associated genes ranged from -0.25 (Plink Set Test), and -0.65 (gengen). (Table 6.5) Correlations between the programs' rankings and the mean rankings ranged from 0.49 (SNP Ratio Test) to 0.83 (HYST), indicating the relative performance of the programs with each other varied.

Table 6.5: Correlations for Method Rankings

Group	***Program***	***Correlation (Proportion)***	***Correlation (Mean Ranking)***
Competitive	*ALIGATOR*	-0.58	0.76
	gengen	-0.65	0.79
	GSA-SNP	-0.59	0.78
	GSEA-SNP	-0.59	0.76
	MAGENTA	-0.61	0.93
	SRT	-0.44	0.49
Self-Contained	*GRASS*	-0.49	0.58
	HYST	-0.57	0.83
	INRICH	-0.59	0.68
	PST	-0.25	0.52

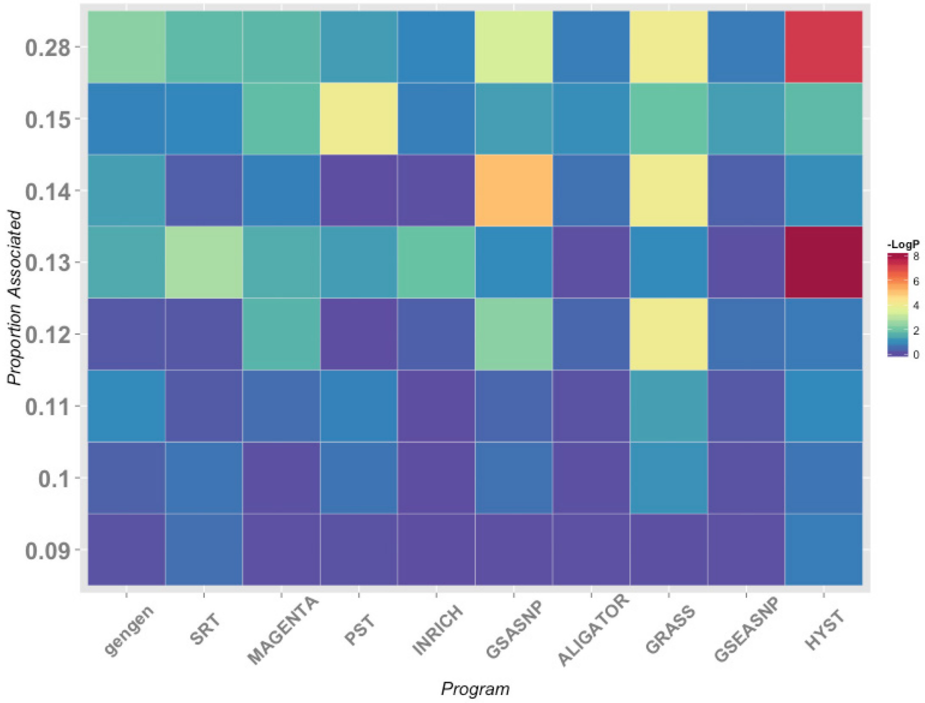


Figure 6.1: Association Results by Programs and Proportion of Genes Associated with a SNP with P<0.01.

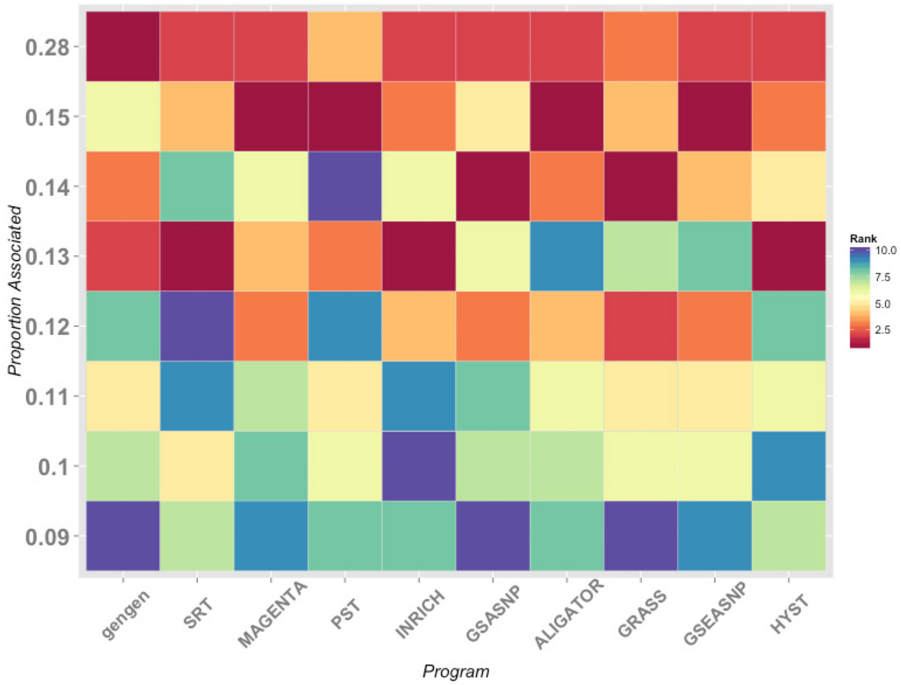


Figure 6.2: Ranking of Associations by Programs and Proportion of Genes Associated with a SNP with P<0.01.

Correlations between the results for the methods and the proportion of genes associated within the gene set varied from -0.29 (PST) to -0.63 (ALIGATOR). When the *P*-values were negative log transformed, the correlations ranged on a smaller scale, from 0.27 (PST) to 0.82 (MAGENTA). (Table 6.6)

Table 6.6: Correlations for Method Results with Proportion of Associated Genes

Group	*Program*	*Correlation (P)*	*Correlation (-logP)*
Competitive	*ALIGATOR*	-0.6346	0.6495
	gengen	-0.5130	0.8235
	GSA-SNP	-0.4767	0.6423
	GSEA-SNP	-0.6303	0.5517
	MAGENTA	-0.5034	0.6041
	SRT	-0.3476	0.3692
Self-Contained	*GRASS*	-0.411	0.627
	HYST	-0.3664	0.7009
	INRICH	-0.6266	0.4306
	PST	-0.293	0.2563

6.4.3: Relationships Between Programs

The correlation in P-values between the programs varied from -18% (SRT and GRASS) to 92% (ALIGATOR and GSEA-SNP). The SNP Ratio Test (SRT) had the lowest correlations with all the programs. It had negative correlation with ALIGATOR, GSEA-SNP, and GRASS. The only program with which the correlation was greater than 50% was with INRICH.

In a heatmap of the results from all pathways, organized from the gene sets with no genes within the pathway being associated to 33% of the genes being associated on the right hand side (Figure 6.4), three programs seem to cluster together: gengen, GSA-SNP, and MAGENTA. They exhibit a trend of less significant P-values with the smaller proportion-associated pathway, and stronger signals towards the pathways with more genes associated with outcome. The Plink Set Test and SNP Ratio Test clustered together. This may be because both methods treat the gene set as an aggregation of SNPs, instead of first creating a gene-level association. The Plink Set Test calculates the average test statistic within the set of SNPs as the gene set statistic, while the SNP ratio test calculates the ratio of significant SNPs to non-significant. Both methods test for the over-significance of associated SNPs within these regions.

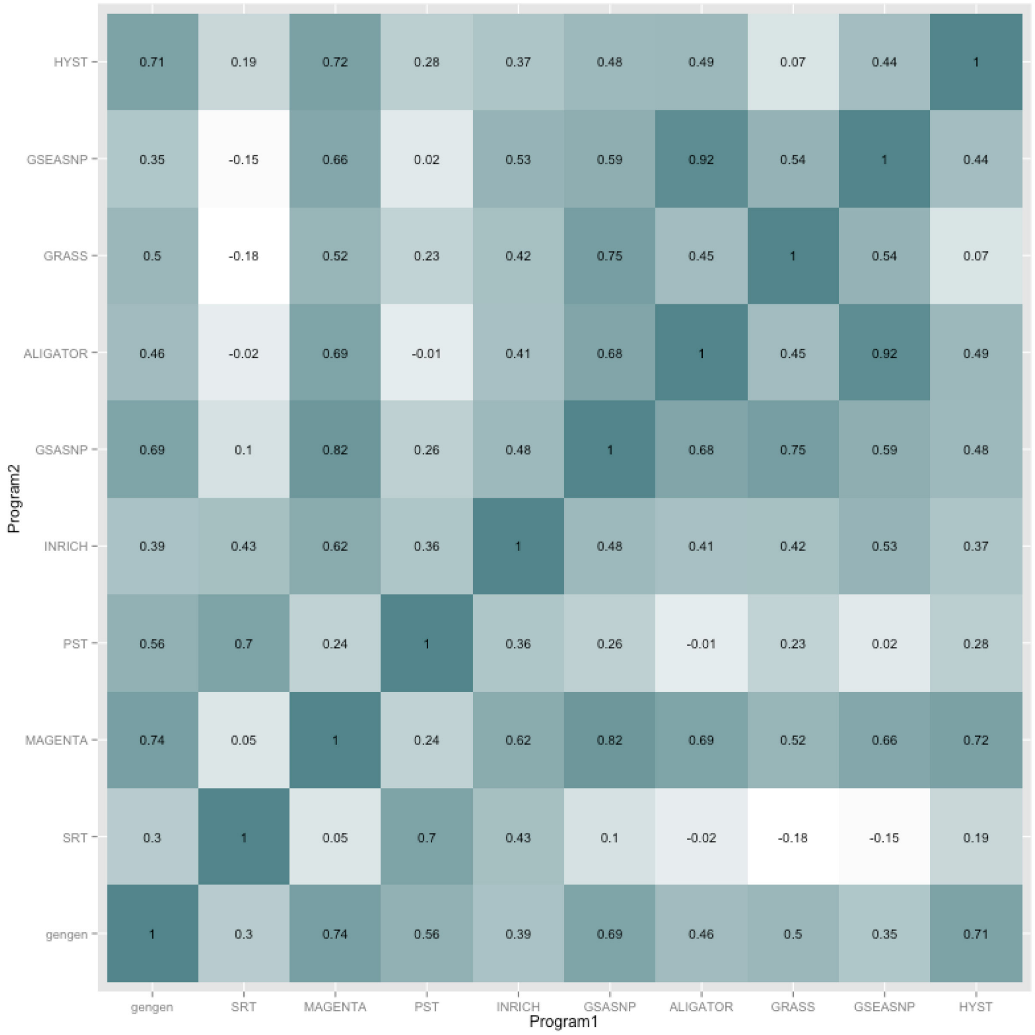


Figure 6.3: Correlation in Results Between Programs. *Correlation was calculated for the P-value results within only the larger pathways (# genes > 28).*

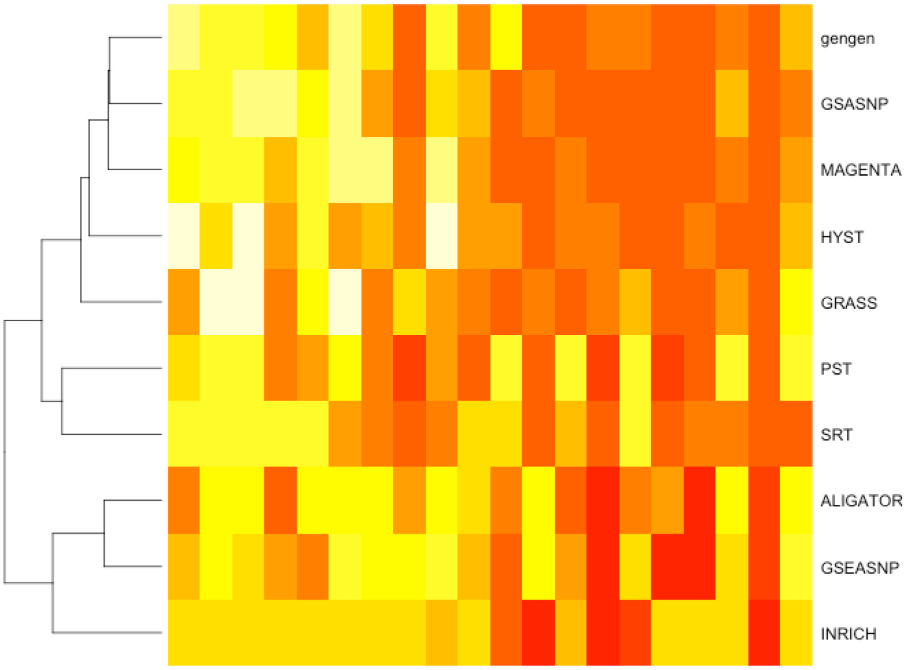


Figure 6.4: Heatmap of Results for Programs by the Proportion of Associated Genes within the Gene Sets. *The results were the P-values for all pathways using the programs for a complete assessment of performance. Pathways with similar performances will cluster together along the y-axis, as indicated by the dendrogram. Proportion of associated genes (at least one SNP with P<0.01) is indicated along the x-axis from left (0%) to right (33%).*

6.5: Discussion

The relative performance of 10 pathway-level programs for GWAS was evaluated through a simulation for 20 different gene sets from Gene Ontology (GO) Biological Processes. The underlying hypothesis for these methods states that there will be numerous genes that will be associated with the phenotype, a true polygenic model. Further, these genes will be clustered in certain sets of genes that will be related to the outcome of interest. Therefore, methods should find gene sets with a higher percentage of associated genes as more significant than gene sets with a lower percentage of associate genes. All of the methods evaluated here showed negative correlation between the proportion of associated, or causal, genes and the P-values, consistent with the underlying hypothesis.

The two methods with the lowest correlations supporting this hypothesis are the SNP Ratio Test and the Plink Set Test. These methods ignore gene architecture altogether, collapsing all SNPS within the genes into a massive gene set unit. Therefore, they are not looking for the enrichment of associated genes within a gene set, but rather an enrichment of SNP associations within genes that comprise the gene set. These methods may be susceptible to the gene size bias, in which a few large genes that contain a large number of associated SNPs exert influence through overrepresentation within the total number of SNPs. On the other hand, these are the only methods suited to handle allelic heterogeneity. Many of the methods assign the gene-level P-value from the minimum SNP P-value found in the genic region. This ignores the relevance of additional independent signals within this region.

The goal of this study was to determine the best-performing pathway-level method for GWAS through a simulation. If we consider the results (P-values) for all 20 pathways and cluster on their similarities between different programs, three methods cluster together: GSA-SNP, gengen, and MAGENTA. These methods show a decreased p-value with an increased proportion of associated genes. The correlation between the proportion of causal genes and the ranking within the program were the highest in these three methods. GSA-SNP showed a correlation of -0.56, MAGENTA had a correlation of -0.61, and GenGen had a correlation of -0.65. As these are all competitive methods, the rankings may be more important than the absolute P-value. This is because the results from a competitive method depend not upon a null model, but rather the enrichment of all gene sets evaluated. It is important to note that when interpreting results, users should not disregard results strictly based on a significance threshold.

Pathway-level methods for GWAS do not evaluate gene-gene interactions or pinpoint the downstream effects of polymorphisms in a gene. Instead, these methods offer a visualization of the data that did not reach genome-wide significance but may be suggestive and biologically relevant to the phenotype of interest. By determining which pathways are enriched for signal within a GWAS, candidate genes and regions may be generated and it may identify relationships between seemingly disparate phenotypes that may have a similar pathogenesis. The best performance was seen in three separate methods: GSA-SNP, MAGENTA, and GenGen. Pathway-level methods for GWAS remain useful tools for conceptualizing GWAS results beyond the traditional SNP-level results that require a strict significance threshold. By examining the relative importance of different gene sets with the results, researchers are allowed a more complete understanding of their genome-wide association study.

6.6: Supplementary Materials

6.6.1: GSA-SNP Options and Performance

A variety of different options were run within the GSA-SNP Software. The different options include using a Z-score estimation, both assuming an asymptotic distribution and using permutations, a GSEA approach using the MAXMEAN, and a traditional GSEA Enrichment Score. Performance between the Z-scores using the asymptotic distribution versus the permutations was nearly identical. The GSEA MAXMEAN method had test statistic inflation, with much smaller P-values across the board. The enrichment score was conservative with only 2 of the pathways reaching the significance threshold.

Supplemental Table 6.1: GSA-SNP Results from Simulation with Different Options

Size	*Biological Process*	#	% *P<0.01*	*Z* *(asym)*	*Z* *(perm)*	*GSEA* *maxmean*	*GSEA* *ES*
Large	Anatomical Structure Morphogenesis	363	13.77%	7.53E-06	7.94E-06	3.11E-03	0.01
	Cellular Defense Response	55	10.91%	0.01	0.01	6.73E-05	0.01
	Establishment and/or Maintenance of Chromatin Architecture	71	12.68%	0.12	0.11	1.78E-05	0.00
	G-Protein Coupled Receptor Protein Signaling Pathway	332	12.05%	0.01	0.01	1.87E-03	0.06
	Leukocyte Activation	65	10.77%	0.54	0.53	0.23	0.62
	Lipid Transport	29	27.59%	3.90E-04	4.14E-04	2.52E-06	3.92E-04
	Membrane Lipid Metabolic Process	98	15.31%	0.06	0.06	3.29E-03	0.02
	Regulation of DNA Binding	44	9.09%	0.95	0.95	0.88	0.85
	Response to Hypoxia	28	10.71%	0.40	0.41	0.15	0.14
	T-Cell Activation	41	9.76%	0.28	0.27	1.41E-04	0.22
Small	CDC42 Protein Signal Transduction	12	33.33%	0.09	0.10	0.09	0.10
	Defense Response to Virus	11	18.18%	8.08E-09	2.71E-09	9.17E-14	4.26E-03
	Establishment of Vesicle Localization	10	0.00%	0.86	0.85	0.65	0.40
	G-Protein Signaling Adenylate Cyclase Activating Pathway	25	16.00%	0.03	0.03	0.00	0.01
	G1 Phase of Mitotic Cell Cycle	12	0.00%	0.86	0.86	0.59	0.45
	Morphogenesis of an Epithelium	17	17.65%	0.03	0.03	0.01	0.04
	Protein Complex Disassembly	15	20.00%	0.45	0.45	0.04	0.06
	Protein Polyubiquitination	10	0.00%	0.94	0.93	0.78	0.82
	Spindle Organization and Biogenesis	10	0.00%	0.88	0.89	0.30	0.17
	Ribonucleotide Metabolic Process	17	0.00%	0.67	0.67	0.44	0.39

6.6.2: MAGENTA Options and Performance

The performance of the two cut-offs were evaluated. A cut-off of 95% is best for an oligogenic model, in which only a few genes are associated with outcome and therefore only the top 5% of genes will be relevant. On the other hand, if the underlying model is thought to be polygenic, in which many genes will play a role in the phenotypic variance then a cut-off of 75% should be used. Because the simulation was conducted under a polygenic model, results were only reported in the main chapter for the 75% cut-off threshold. The 95% cut-off was less conservative, with more significant P-values.

Supplemental Table 6.2: MAGENTA Results from Simulation with Different Cut-offs

Size	*Biological Process*	*# Genes*	*% P<0.01*	*0.95 Cut-off*	*0.75 Cut-off*
Large	Anatomical Structure Morphogenesis	363	13.77%	0.08	0.16
	Cellular Defense Response	55	10.91%	0.01	0.13
	Establishment and/or Maintenance of Chromatin Architecture	71	12.68%	9.00E-04	0.03
	G-Protein Coupled Receptor Protein Signaling Pathway	332	12.05%	0.12	0.05
	Leukocyte Activation	65	10.77%	0.62	0.94
	Lipid Transport	29	27.59%	1.60E-03	0.02
	Membrane Lipid Metabolic Process	98	15.31%	0.32	0.02
	Regulation of DNA Binding	44	9.09%	1.00	0.93
	Response to Hypoxia	28	10.71%	0.74	0.31
	T-Cell Activation	41	9.76%	0.32	0.90
Small	CDC42 Protein Signal Transduction	12	33.33%	0.40	0.23
	Defense Response to Virus	11	18.18%	0.09	0.08
	Establishment of Vesicle Localization	10	0.00%	1.00	0.69
	G-Protein Signaling Adenylate Cyclase Activating Pathway	25	16.00%	0.12	0.05
	G1 Phase of Mitotic Cell Cycle	12	0.00%	0.03	0.03
	Morphogenesis of an Epithelium	17	17.65%	0.54	0.02
	Protein Complex Disassembly	15	20.00%	0.16	0.11
	Protein Polyubiquitination	10	0.00%	1.00	0.76
	Spindle Organization and Biogenesis	10	0.00%	0.40	0.47
	Ribonucleotide Metabolic Process	17	0.00%	1.00	0.80

References

Chen, Lin S, Carolyn M Hutter, John D Potter, Yan Liu, Ross L Prentice, Ulrike Peters, and Li Hsu. 2010. "Insights Into Colon Cancer Etiology via a Regularized Approach to Gene Set Analysis of GWAS Data." *The American Journal of Human Genetics* 86 (6) (June 11): 860–871. doi:10.1016/j.ajhg.2010.04.014.

Gibson, Greg. 2012. "Rare and Common Variants: Twenty Arguments." *Nature Reviews Genetics* 13 (2) (February 1): 135–145. doi:10.1038/nrg3118.

Holden, M, S Deng, L Wojnowski, and B Kulle. 2008. "GSEA-SNP: Applying Gene Set Enrichment Analysis to SNP Data From Genome-Wide Association Studies." *Bioinformatics* 24 (23) (November 21): 2784–2785. doi:10.1093/bioinformatics/btn516.

Holmans, Peter, Elaine K Green, Jaspreet Singh Pahwa, Manuel A R Ferreira, Shaun M Purcell, Pamela Sklar, Michael J Owen, Michael C O Donovan, Nick Craddock, and The Wellcome Trust Case-Control Consortium9. 2009. "Gene Ontology Analysis of GWA Study Data Sets Provides Insights Into the Biology of Bipolar Disorder." *The American Journal of Human Genetics* 85 (1) (July 10): 13–24. doi:10.1016/j.ajhg.2009.05.011.

Lee, P H, C O'Dushlaine, B Thomas, and S M Purcell. 2012. "INRICH: Interval-Based Enrichment Analysis for Genome-Wide Association Studies." *Bioinformatics* 28 (13) (June 23): 1797–1799. doi:10.1093/bioinformatics/bts191.

Li, Miao-Xin, Hong-Sheng Gui, Johnny S H Kwan, and Pak C Sham. 2011. "GATES: a Rapid and Powerful Gene-Based Association Test Using Extended Simes Procedure." *American Journal of Human Genetics* 88 (3) (March 11): 283–293. doi:10.1016/j.ajhg.2011.01.019.

Li, Miao-Xin, Johnny S H Kwan, and Pak C Sham. 2012. "HYST: a Hybrid Set-Based Test for Genome-Wide Association Studies, with Application to Protein-Protein Interaction-Based Association Analysis." *American Journal of Human Genetics* 91 (3) (September 7): 478–488. doi:10.1016/j.ajhg.2012.08.004.

Nam, D, J Kim, S Y Kim, and S Kim. 2010. "GSA-SNP: a General Approach for Gene Set Analysis of Polymorphisms." *Nucleic Acids Research* 38 (Web Server) (June 24): W749–W754. doi:10.1093/nar/gkq428.

O'Dushlaine, C, E Kenny, E A Heron, R Segurado, M Gill, D W Morris, and A Corvin. 2009. "The SNP Ratio Test: Pathway Analysis of Genome-Wide Association Datasets." *Bioinformatics* 25 (20) (October 8): 2762–2763. doi:10.1093/bioinformatics/btp448.

Purcell, Shaun, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel A R Ferreira, David Bender, Julian Maller, et al. 2007. "PLINK: a Tool Set for Whole-Genome Association and Population-Based Linkage Analyses." *The American Journal of Human Genetics* 81 (3) (September): 559–575. doi:10.1086/519795.

Segrè, Ayellet V, DIAGRAM Consortium, MAGIC investigators, Leif Groop, Vamsi K Mootha, Mark J Daly, and David Altshuler. 2010. "Common Inherited Variation in Mitochondrial Genes Is Not Enriched for Associations with Type 2 Diabetes or

Related Glycemic Traits." Edited by Peter M Visscher. *PLoS Genetics* 6 (8) (August 12): e1001058. doi:10.1371/journal.pgen.1001058.t004.

Subramanian, Aravind, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael A Gillette, Amanda Paulovich, et al. 2005. "Gene Set Enrichment Analysis: a Knowledge-Based Approach for Interpreting Genome-Wide Expression Profiles.." *Proceedings of the National Academy of Sciences* 102 (43) (October 25): 15545–15550. doi:10.1073/pnas.0506580102.

Wang, Kai, Mingyao Li, and Maja Bucan. 2007. "Pathway-Based Approaches for Analysis of Genomewide Association Studies." *The American Journal of Human Genetics* 81 (6) (December): 1278–1283. doi:10.1086/522374.

Chapter 7: Application of Gene- and Pathway-Level Methods to a Genome-wide Association Study of OPV Response in Bangladeshi Children

7.1: Abstract

Background: The human infection of poliovirus has been eradicated in many parts of the world due to the successful use of the oral poliovirus vaccine (OPV). However, after numerous doses with viable vaccine, some individuals fail to mount an immunological response. It has previously been hypothesized that this may be due to underlying host genetics. To address this question, we conducted a genome-wide association study (GWAS) on Bangladeshi infants after four doses of OPV and one year of follow-up. To complement this GWAS, previously evaluated gene- and pathway-level methods of analysis were utilized. These methods collapse genetic variation into units on a gene-level, or a set of genes such as a canonical pathway. They aim to elucidate associations that are suggestive, but fail to meet the stringent GWAS significance threshold.

Methods: A genome-wide association study (GWAS) was conducted on 6.6 million imputed SNPs comparing extremes of the OPV immune response. High seropositive children had log serum neutralizing antibody titers above 7 and seronegative children had titers below 3 at one year of age. The GWAS results were analyzed for gene-level associations using VEGAS and the HapMap CEPH population as a reference. Pathway-level associations were also evaluated using the Gene Ontology Biological Processes and two programs: MAGENTA, and GSA-SNP.

Results: Gene-level results yielded suggestive signals in numerous histone variants within histone cluster 1 on chromosome 6 as well as representing the top associations of the GWAS. The pathway-level methods highlighted the role of cyclic AMP as a secondary messenger, especially when coupled with G-proteins. Numerous gene sets involved in the nervous system were also found to be suggestive.

Conclusions: The gene-level results suggest that variants in the host genome may affect histone modifications that will alter the immune response to OPV. Epigenetic studies are warranted to evaluate the role of histone modification in the immune response to OPV and other oral vaccines. The pathway-level results lack resolution, but suggest a role for cAMP that should be further investigated through functional studies. Both methods allow the dissection of genome-wide association studies beyond the traditional SNP-based testing and the stringent *P*-value threshold required for significance.

7.2: Introduction

Oral Poliovirus Vaccine (OPV) has had remarkable success over the past 50 years in the eradication of poliovirus and prevention of poliomyelitis. Since 1988, the number of cases has decreased over 99%. Wild circulating virus was only found in four countries in 2012: Afghanistan, Pakistan, Nigeria, and Chad.(WHO 2013) OPV has been shown to be effective at inducing both mucosal and systemic immunity in diverse populations, yet some individuals fail to mount an adequate response. One hypothesis has been that this is due to host genetics, as differential response is still seen in studies that use identical viable vaccines on a standardized schedule.(Paul 2007) The heritability of response to OPV has been estimated to be 60%, meaning that 60% of the variance in OPV neutralizing antibody titers is attributable to genetic variables.(Newport et al. 2004) Despite this high heritability estimate, there have been no genetic loci implicated so far.

To address this question, a genome-wide association study was conducted comparing extreme responders. The GWAS was underpowered and did not yield genome-wide significant association signals ($P<5\text{x}10^{-8}$). However, suggestive associations were found in *SHH* (sonic hedgehog) and *SOCS4* (suppressor of cytokine signaling 4) (*unpublished*, Chapter 3). A reason for the lack of genome-wide significant results may be the small sample size (N=357) and susceptibility loci with small effects. Due to these factors, there is limited power to detect associations. In recent years, a variety of gene- and pathway-level methods have been developed to increase power to detect genetic associations and find connections between suggestive findings that may inform the phenotype's pathogenesis. Gene-level methods aggregate single nucleotide polymorphisms (SNPs) to assess the joint associations. The underlying hypothesis with these methods is that in the presence of allelic heterogeneity, where different alleles within the same locus or gene are associated with outcome, no individual variant will be detected in a traditional GWAS because of limited power. However, the aggregate of their signals will have increased association. Results from this method would detect genes with elevated levels of association that would have typically been ignored using the stringent p-value threshold. Pathway-level methods are different in their underlying hypotheses and goals. Instead of looking to detect instances of allelic heterogeneity, they seek to determine if highly ranked genes are commonly linked by being disproportionately found in certain sets of genes, or a pathway. These gene sets may then be involved in the phenotype and inform further investigation. We evaluate both gene- and pathway-level methods using a GWAS of OPV titers.

7.3: Materials and Methods

7.3.1: Study Population

This study was conducted in 448 children from a Birth Cohort (DBC) established to evaluate enteric infections. Children were recruited from an urban slum in Mirpur, one of the 14 districts in Dhaka City, Bangladesh. Mirpur has a population density of one million people per 59 square kilometers. This cohort is followed from birth until at least 2 years of age. Children were visited multiple times per week, with data collected on numerous infectious diseases including enteric pathogens *E. histolytica*, *Cryptosporidium*, and rotavirus. Anthropometric measurements such as height-for-age Z-score (HAZ) and body mass index (BMI) are also available at multiple time points.

For children that had completed at least one year of follow-up, serum-neutralizing antibody responses to the full 4-dose regimen were available for all three poliovirus serotypes. A total of 448 children were assessed for OPV failure, of which only 425 were genotyped. Vaccine failure was defined using the CDC standard cutoff of a $\log_2$ serum neutralizing antibody levels of 3, or a 1:8 dilution factor. (World Health Organization Collaborative Study Group on Oral Poliovirus Vaccine 1995) Serotype 3 had a seroconversion rate of 88.71%. Because of the right- and left-censored data (left at 2.5, right at 10.5), the titers could not be assessed as a quantitative trait. Extremes were examined, with seronegative individuals classified as a titer below or equal to 3, and a strong seropositive individual having a titer equal to or greater than 7. This resulted in 48 seronegative and 309 seropositive children (previously described in Section 3.3.1).

7.3.2: Genotype Data and Quality Control

Whole blood was taken from the children at 6 months of age and DNA was extracted at the International Center for Diarrheal Disease Research, Bangladesh (ICDDR, B). The DNA samples were then shipped to the University of Virginia for genotyping. Two chips were used for the original round of genotyping: 1M Illumina Duo and the 1M Illumina Quad. Despite having over 1 million SNPs genotyped on each chip, there was only an overlap of 613,778 SNPs. For these two chips, the average call rate was 99.79% and no samples had a call rate of less than 96%. An additional round of samples was genotyped using Illumina's 2.5M Quad chip. To synchronize these different rounds, imputation was done using IMPUTE2 with a 1000 Genomes reference data set.(Howie et al. 2012) Standard quality control measures were applied to this data, as further detailed in section 3.3.2. After quality control, assessments there were 457 individuals with 6.5 million imputed SNPs.

7.3.3: Analysis Methods

Association analysis was run within SNPTEST(Marchini et al. 2007), using an additive frequentist EM model. The associations were adjusted for stunting, or a height-for-age Z-score (HAZ) below -2. SNPs were filtered by an information content of the test of more than 80%, as well as a minor allele frequency greater than 5%. No SNPs reached genome-wide significance, using a standard threshold of $5x10^{-7}$.

The gene-level method used was VEGAS(Liu et al. 2010). Two options were utilized: (1) using all SNPs and (2) using only the top 10% of associated SNPs within each gene. SNPs were assigned to genes according to hg19 coordinates, including 20 kilobase flanking regions on both sides. The adjusted GWAS *P*-values were used as input and the HapMap CEPH (CEU) population was used as a reference panel for the linkage disequilibrium estimates. Two different methods were used for the pathway-level analysis: MAGENTA(Segrè et al. 2010) and GSA-SNP(Nam et al. 2010). They both used the GWAS P-values as input and used the same Gene Ontology Biological Processes (N=825) downloaded from MSigDB.(Subramanian et al. 2005)

7.4: Results

7.4.1: Gene-Level Results

VEGAS was applied to 19,120 genes across the human genome. Genes with a *P*-value below 0.001 are detailed in Table 7.1. These top twenty genes consisted of two regions on chromosomes 6 and 12. The most significant gene was *HIST1H4J* (*P*=2.7x10^{-5}), a histone cluster 1 variant on chromosome 6 in the extended HLA region. With a total of 64 SNPs being assigned to the gene, the most significant SNP was rs183225 (*P*=2.95x10^{-5}). This SNP is located within an active promoter and CpG island. Seventeen out of the top twenty associations were in this region (Figure 7.1). Another top region is on chromosome 12 in the gene *TAS2R9* (P=9.2x10^{-5}), a taste receptor (type 2, member 9) that is a member of the G-protein coupled receptor superfamily.

VEGAS was also applied to the same genes using only the top 10% of associated SNPs within the genic regions. The results are largely consistent with the prior analysis using all of the SNPs with genes in histone cluster 1 remaining within the top ranked genes. The difference with this analysis was that results were more reflective of the original GWAS with *LMBR1* showing an association. Additional top genes within regions on chromosomes 7 and 14 reflect the top signals for the GWAS, such as *MAPK1IP1L* and *SOCS4* (previously described Chapter 3).

Table 7.1: Gene-Level Results using all SNPs

Chr	*Gene*	*# SNPs*	*Size*	*P-value*	*Best SNP*	*SNP P-value*
6	*HIST1H4J*	64	356	2.70E-05	rs183225	2.95E-05
6	*HIST1H4K*	61	354	2.90E-05	rs183225	2.95E-05
6	*HIST1H2BN*	63	449	5.10E-05	rs183225	2.95E-05
6	*HIST1H2AK*	64	460	5.60E-05	rs183225	2.95E-05
12	*TAS2R9*	132	1075	9.20E-05	rs11054019	1.03E-02
12	*TAS2R7*	136	1096	1.16E-04	rs11054019	1.03E-02
6	*HIST1H2AJ*	64	439	1.20E-04	rs183225	2.95E-05
6	*HIST1H2BM*	64	446	1.23E-04	rs183225	2.95E-05
12	*TAS2R8*	132	930	1.26E-04	rs11054019	1.03E-02
6	*HIST1H4L*	65	364	1.27E-04	rs188015	1.69E-04
6	*HIST1H3J*	73	478	1.30E-04	rs188015	1.69E-04
6	*HIST1H2BO*	73	467	1.41E-04	rs188015	1.69E-04
6	*HIST1H3I*	64	477	1.50E-04	rs188015	1.69E-04
6	*HIST1H2AM*	74	487	1.51E-04	rs188015	1.69E-04
6	*HIST1H3H*	65	473	1.58E-04	rs183225	2.95E-05
6	*HIST1H1B*	62	790	1.64E-04	rs200501	1.64E-04
6	*OR2B2*	72	1212	1.64E-04	rs188015	1.69E-04
6	*HIST1H2BL*	64	453	1.66E-04	rs183225	2.95E-05
6	*HIST1H2AL*	62	470	1.67E-04	rs200501	1.64E-04
6	*HIST1H2AI*	65	469	1.88E-04	rs183225	2.95E-05
11	*KCNE3*	113	12715	1.94E-04	rs686179	1.84E-05
10	*OGDHL*	98	27739	3.93E-04	rs1025742	3.29E-04
19	*ZNF709*	101	23635	5.77E-04	rs4804194	5.13E-04
19	*ZNF443*	86	11407	6.74E-04	rs4804194	5.13E-04
17	*ZSWIM7*	70	23132	7.46E-04	rs11869450	1.14E-04
19	*IGFL2*	53	13523	7.78E-04	rs11670023	1.84E-04
19	*IGFL3*	52	4604	8.14E-04	rs11670023	1.84E-04
17	*TTC19*	78	30030	8.32E-04	rs11869450	1.14E-04

** Chr=Chromosome, Best SNP= SNP with highest P-value in gene*

Table 7.2: Gene Results for using only the Top 10% of SNPs

Chr	*Gene*	*# SNPs*	*Size*	*Top 10% P*	*All SNPs P*	*Best SNP*	*SNP P*
7	*LMBR1*	250	212333	5.50E-05	1.74E-03	rs10242938	2.41E-05
6	*HIST1H2BL*	64	453	1.20E-04	1.60E-04	rs183225	2.95E-05
6	*HIST1H4K*	61	354	1.37E-04	4.10E-05	rs183225	2.95E-05
6	*HIST1H2AK*	64	460	1.51E-04	4.50E-05	rs183225	2.95E-05
6	*HIST1H2AJ*	64	439	1.52E-04	1.06E-04	rs183225	2.95E-05
6	*HIST1H4J*	64	356	1.62E-04	2.80E-05	rs183225	2.95E-05
6	*HIST1H2BM*	64	446	1.63E-04	9.80E-05	rs183225	2.95E-05
6	*HIST1H3H*	65	473	1.69E-04	1.61E-04	rs183225	2.95E-05
10	*CCDC3*	309	105080	1.70E-04	2.63E-03	rs10906260	3.25E-05
6	*HIST1H2AI*	65	469	1.82E-04	1.89E-04	rs183225	2.95E-05
6	*HIST1H2BN*	63	449	1.89E-04	4.20E-05	rs183225	2.95E-05
10	*SH2D4B*	232	108659	2.50E-04	3.33E-03	rs12360015	8.52E-05
6	*DACT2*	92	12819	2.70E-04	2.88E-03	rs9364424	1.23E-04
11	*KCNE3*	113	12715	2.77E-04	1.70E-04	rs686179	1.84E-05
7	*RNF32*	110	36468	3.00E-04	8.96E-03	rs10242938	2.41E-05
2	*ECEL1*	72	7996	3.26E-04	2.32E-03	rs746379	6.73E-05
19	*CCDC61*	80	23156	3.90E-04	6.50E-03	rs2302788	5.11E-05
1	*ACTL8*	152	71751	3.90E-04	9.49E-03	rs683259	6.12E-05
14	*SOCS4*	67	22363	4.30E-04	4.56E-02	rs17128156	2.76E-06
17	*ZSWIM7*	70	23132	4.61E-04	6.73E-04	rs11869450	1.14E-04
14	*LGALS3*	88	16214	4.70E-04	1.15E-02	rs17128156	2.76E-06
6	*HIST1H2AL*	62	470	5.10E-04	1.80E-04	rs200501	1.64E-04
2	*ALPI*	68	3910	5.20E-04	1.51E-02	rs746379	6.73E-05
17	*ADORA2B*	81	30980	5.30E-04	1.12E-03	rs11869450	1.14E-04
14	*MAPK1IP1L*	77	18551	5.50E-04	3.72E-02	rs17128156	2.76E-06
19	*IGFL4*	80	1269	5.58E-04	3.13E-03	rs2302788	5.11E-05
17	*TTC19*	78	30030	5.84E-04	8.53E-04	rs11869450	1.14E-04
19	*PGLYRP1*	79	4145	5.89E-04	3.61E-03	rs2302788	5.11E-05
6	*HIST1H1B*	62	790	6.48E-04	1.51E-04	rs200501	1.64E-04
1	*CTSE*	17	14646	6.50E-04	8.28E-03	rs28450935	3.87E-04
13	*ENOX1*	748	573451	6.53E-04	1.87E-03	rs9525777	2.59E-05
19	*NOVA2*	77	33887	6.90E-04	2.80E-02	rs2302788	5.11E-05
6	*HIST1H2BO*	73	467	7.50E-04	1.00E-04	rs188015	1.69E-04
17	*NCOR1*	141	185467	7.69E-04	2.60E-03	rs11869450	1.14E-04
6	*HIST1H3I*	64	477	7.77E-04	1.03E-04	rs188015	1.69E-04
6	*OR2B2*	72	1212	7.89E-04	1.57E-04	rs188015	1.69E-04
4	*MUC7*	115	52506	7.90E-04	1.26E-03	rs2130651	1.87E-04
6	*HIST1H3J*	73	478	8.07E-04	1.09E-04	rs188015	1.69E-04
10	*CHAT*	191	56010	8.23E-04	5.43E-03	rs1025742	3.29E-04

**Results are sorted by the P-value using only the top 10% of SNPs. The ALL SNPs P-value is the gene's corresponding P-value from the previous analysis using all SNPs within the gene.*

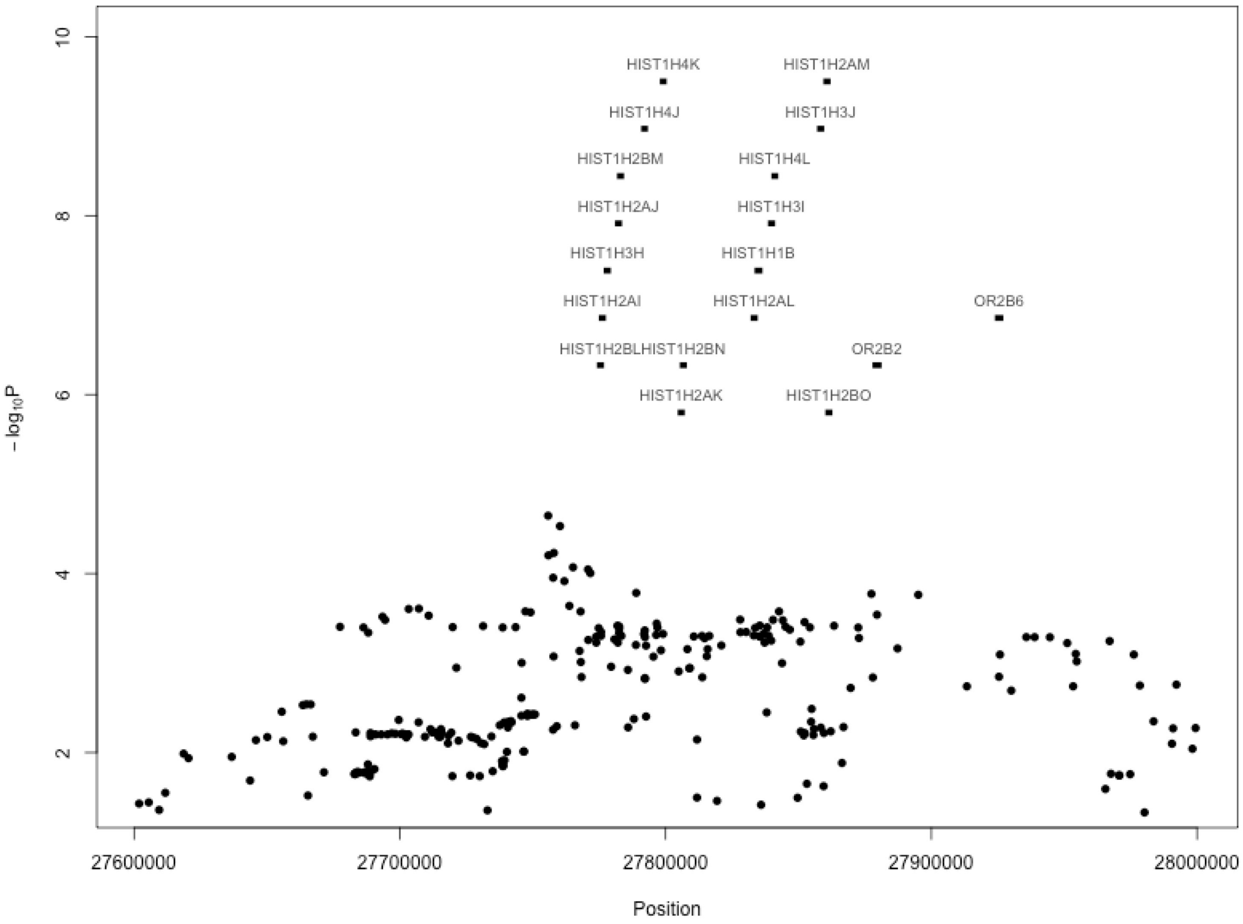


Figure 7.1: Chromosome 6 SNP Associations and Histone cluster 1. *Association is indicated along the y-axis with the –log*$_{10}$ *transformed P-values and chromosomal position is shown on the x-axis. Histone markers are labeled according to their hg19 coordinates.*

7.4.2: Pathway-Level Results

MAGENTA and GSA-SNP were applied to the whole GWAS of OPV Response using the Gene Ontology Biological Processes database. A P-value threshold of 0.01 for suggestive pathways was used for all programs. A polygenic model was assumed to use a 75th percentile cut-off within MAGENTA. A total of 16 pathways were suggestive (Table 7.3). The top pathway was "G-Protein Signaling Coupled to Camp Nucleotide Second Messenger" with 28 out of its 63 genes (44%) in the 75th percentile of all genes (P=5x10^{-4}, FDR-0.39). This pathway is now known as "Adenylate Cyclase-Modulating G-Protein Coupled Receptor Signaling Pathway" on Gene Ontology and affects the concentration of cyclic AMP (cAMP). Many of the top associated gene sets for MAGENTA were related to cAMP and the G-Protein signaling pathway with second messengers.

The top association for GSA-SNP was "Neurological System Process", a gene set that is an overarching organ system process carried out or involving any of the neurological system ($P_{corrected}$=2.8x10^{-4}) (Table 7.4). Other neurological gene sets were found to be highly associated such as "Neuron Differentiation", "Synaptic Transmission", "Generation of Neurons", "Nervous System Development", "Neurite Development", and "Transmission of Nerve Impulse".

MAGENTA and GSA-SNP overlapped greatly with 6 gene sets in common: "G-Protein Signaling Coupled to cAMP Nucleotide Second Messenger", "Cyclic Nucleotide Mediated Signaling", "cAMP Mediated Signaling", "G-Protein Signaling Coupled to

Cyclic Nucleotide Second Messenger”, “Neurological System Process”, and “Regulation of Developmental Process”. Four out of the six pathways were involved in cyclic nucleotide second messenger, specifically cAMP, signaling. Two gene sets were coupled additionally with G-protein signaling. The other two gene sets (“Neurological System Process” and “Regulation of Developmental Process”) were very large (N=336 and 440, respectively) and not as specific.

Table 7.3: MAGENTA Results for 75% Cut-off

Gene Set	***# Genes***	***P***	***FDR***	***Expected # Genes***	***Observed # Genes***
G-PROTEIN SIGNALING COUPLED TO CAMP NUCLEOTIDE SECOND MESSENGER	63	5.00E-04	0.39	16	28
CYCLIC NUCLEOTIDE MEDIATED SIGNALING	100	8.00E-04	0.24	25	39
CAMP MEDIATED SIGNALING	64	9.00E-04	0.26	16	28
G-PROTEIN SIGNALING COUPLED TO CYCLIC NUCLEOTIDE SECOND MESSENGER	98	1.70E-03	0.23	25	38
G-PROTEIN SIGNALING ADENYLATE CYCLASE ACTIVATING PATHWAY	24	2.00E-03	0.26	6	13
NEUROLOGICAL SYSTEM PROCESS	336	2.40E-03	0.20	84	106
REGULATION OF MAPKKK CASCADE	17	2.70E-03	0.20	4	10
G PROTEIN SIGNALING ADENYLATE CYCLASE INHIBITING PATHWAY	10	3.40E-03	0.24	3	7
AMINO ACID TRANSPORT	25	3.70E-03	0.19	6	13
ORGANIC ACID TRANSPORT	40	4.40E-03	0.20	10	18
AMINE TRANSPORT	37	4.60E-03	0.20	9	17
CARBOHYDRATE TRANSPORT	18	5.00E-03	0.22	5	10
SENSORY PERCEPTION	167	6.70E-03	0.32	42	56
REGULATION OF JNK CASCADE	11	7.50E-03	0.21	3	7
CARBOXYLIC ACID TRANSPORT	39	7.70E-03	0.30	10	17
REGULATION OF DEVELOPMENTAL PROCESS	400	9.70E-03	0.43	100	119

**Gene sets in common between MAGENTA and GSA-SNP are in italics.*

Table 7.4: Results from GSA-SNP

Gene Set	***# Genes***	***P***	***Corrected P***
NEUROLOGICAL SYSTEM PROCESS	379	3.57E-07	2.84E-04
CYCLIC NUCLEOTIDE MEDIATED SIGNALING	102	1.45E-06	5.76E-04
G PROTEIN SIGNALING COUPLED TO CYCLIC NUCLEOTIDE SECOND MESSENGER	100	1.85E-06	5.76E-04
NEURON DIFFERENTIATION	76	2.17E-06	5.76E-04
G PROTEIN SIGNALING COUPLED TO cAMP NUCLEOTIDE SECOND MESSENGER	64	8.72E-06	1.38E-03
SECOND MESSENGER MEDIATED SIGNALING	153	1.05E-05	1.39E-03
ANATOMICAL STRUCTURE MORPHOGENESIS	376	1.07E-05	1.39E-03
SYNAPTIC TRANSMISSION	174	1.58E-05	1.57E-03
cAMP MEDIATED SIGNALING	65	1.70E-05	1.57E-03
GENERATION OF NEURONS	83	3.97E-05	3.15E-03
ION TRANSPORT	185	4.11E-05	3.15E-03
G PROTEIN COUPLED RECEPTOR PROTEIN SIGNALING PATHWAY	342	4.22E-05	3.15E-03
NERVOUS SYSTEM DEVELOPMENT	385	5.85E-05	3.57E-03
REGULATION OF DEVELOPMENTAL PROCESS	440	6.14E-05	3.57E-03
NEURITE DEVELOPMENT	53	1.09E-04	5.80E-03
TRANSMISSION OF NERVE IMPULSE	189	1.10E-04	5.80E-03

*Gene sets in common between MAGENTA and GSA-SNP are in italics.

7.5: Discussion

The most significant gene-level result was found in Histone cluster 1 variation on chromosome 6 within the extended HLA region. These genes include many parts of the histone complex, including histones 1-4. Histones are responsible for the storage of DNA, both in coiling DNA around the octomer (H2-5), as well as forming the supercoils for the 30 nm nanofiber (H1).(Parseghian and Luhrs 2006) They are an essential part of epigenetics and genetic expression. It has previously been observed that bacterial toxins can alter the histone structure (histone modifications) through epigenetic imprinting.(Hamon et al. 2007) Specifically, in the early stages of *Listeria monocytogenes* infection, toxins were associated with the dephosphorylation of histone 3 and the deacetylation of histone 4— both core histones. A similar phenomenon was observed with *Clostridium perfringens* and *Streptococcus pneumonia* toxins. Epigenetic imprinting has also been observed in commensal probiotics, in which the expression of genes can be altered due to histone modifications induced by infection. A study of the linker histone (H1) in intestinal epithelial cells revealed a role in preventing microbial penetration into villous epithelial cells.(Rose et al. 1998) These studies highlight the potential role of histone marks with immunity to enteric infections.

An additional actor that may play a role in the response to OPV is environmental enteropathy, a sub-clinical syndrome in which a cycle of malnutrition and enteric infection leads to decreased gut integrity and response to oral vaccines. Because a newborn’s intestinal tract is originally free of a microbiome but is quickly populated, it can be hypothesized that epigenetic imprinting by this microbial population early in life may affect a child's mucosal immunity long-term.(Korpe and Petri 2012) Through this mechanism, host genetic differences may influence the way that gene expression is altered under these pressures, leading to differential systemic immunity to OPV. Future research should focus on epigenetic signatures in gut mucosa, as well as circulated serum, with response to vaccines. Gene expression studies may also pinpoint how these histone modifications alter the immune system's response to vaccination, or natural infection.

The pathway-level results highlight the role of cyclic AMP, as well as the nervous system. Cyclic AMP is a second messenger that is a negative regulator of T cell immune function.(Mosenden and Taskén 2011) Specifically, cAMP levels have been shown to correlate with suppressive capabilities of T regulatory cells. These cells suppress the immune system's response to foreign antigens. A disruption of this pathway could decrease vaccine efficiency. Some of the cAMP pathways were also coupled with G-protein signaling pathways. Numerous nervous system development and regulation pathways were also associated by both methods. This is not surprising, as poliomyelitis, the clinical presentation of poliovirus infection, results from poliovirus infecting the central nervous system (CNS). Polymorphisms in genes related to this system would then affect the ability of the virus to effectively invade and replicate within the CNS.

Pathway- and gene-level analyses are hypothesis-generating methods that do not offer a high level of resolution in their findings. They are methods that examine results from genome-wide association in aggregate. Gene-level methods will identify multiple signals within a gene that would otherwise have been undetected in a GWAS because they failed to reach the genome-wide significance threshold ($5x10^{-7}$). Results can inform further follow-up studies, such as sequencing, to identify risk loci. Pathway-level methods serve as a visualization tool for the genes that are enriched in your study. While many of these gene sets are broad and include many genes, they provide lists of candidate genes for follow-up. For this study, the gene-level method identified a potential role for genetic variations in histone cluster 1 that is densely packed with regulatory elements. It suggests a role for epigenetic research regarding immune responses to oral vaccines. The pathway results propose a role for cyclic AMP and G-protein coupled signaling in the response to OPV, as well as the involvement of the nervous system. Taken together they can inform future research not only in the response to OPV, but also in the response to other oral vaccines.

References

Hamon, Mélanie Anne, Eric Batsché, Béatrice Régnault, To Nam Tham, Stéphanie Seveau, Christian Muchardt, and Pascale Cossart. 2007. "Histone Modifications Induced by a Family of Bacterial Toxins.." *Proceedings of the National Academy of Sciences* 104 (33) (August 14): 13467–13472. doi:10.1073/pnas.0702729104.

Howie, Bryan, Christian Fuchsberger, Matthew Stephens, Jonathan Marchini, and Gon ccedil alo R Abecasis. 2012. "Fast and Accurate Genotype Imputation in Genome-Wide Association Studies Through Pre-Phasing." *Nature Genetics* (July 22): 1–6. doi:10.1038/ng.2354.

Korpe, Poonum S, and William A Petri Jr. 2012. "Environmental Enteropathy: Critical Implications of a Poorly Understood Condition." *Trends in Molecular Medicine* 18 (6) (June 1): 328–336. doi:10.1016/j.molmed.2012.04.007.

Liu, Jimmy Z, Allan F Mcrae, Dale R Nyholt, Sarah E Medland, Naomi R Wray, Kevin M Brown, AMFS Investigators, et al. 2010. "A Versatile Gene-Based Test for Genome-Wide Association Studies." *American Journal of Human Genetics* 87 (1) (July 9): 139–145. doi:10.1016/j.ajhg.2010.06.009.

Marchini, Jonathan, Bryan Howie, Simon Myers, Gil McVean, and Peter Donnelly. 2007. "A New Multipoint Method for Genome-Wide Association Studies by Imputation of Genotypes." *Nature Genetics* 39 (7) (June 17): 906–913. doi:10.1038/ng2088.

Mosenden, Randi, and Kjetil Taskén. 2011. "Cellular Signalling." *Cellular Signalling* 23 (6) (June 1): 1009–1016. doi:10.1016/j.cellsig.2010.11.018.

Nam, D, J Kim, S Y Kim, and S Kim. 2010. "GSA-SNP: a General Approach for Gene Set Analysis of Polymorphisms." *Nucleic Acids Research* 38 (Web Server) (June 24): W749–W754. doi:10.1093/nar/gkq428.

Newport, M J, T Goetghebuer, H A Weiss, H Whittle, C-A Siegrist, and A Marchant. 2004. "Genetic Regulation of Immune Responses to Vaccines in Early Life." *Genes and Immunity* 5 (2) (January 22): 122–129. doi:10.1038/sj.gene.6364051.

Parseghian, Missag H, and Keith A Luhrs. 2006. "Beyond the Walls of the Nucleus: the Role of Histones in Cellular Signaling and Innate immunityThis Paper Is One of a Selection of Papers Published in This Special Issue, Entitled 27th International West Coast Chromatin and Chromosome Conference, and Has Undergone the Journal's Usual Peer Review Process.." *Biochemistry and Cell Biology* 84 (4) (August): 589–595. doi:10.1139/o06-082.

Paul, Yash. 2007. "Role of Genetic Factors in Polio Eradication: New Challenge for Policy Makers." *Vaccine* 25 (50) (December): 8365–8371. doi:10.1016/j.vaccine.2007.09.068.

Rose, F R, K Bailey, J W Keyte, W C Chan, D Greenwood, and Y R Mahida. 1998. "Potential Role of Epithelial Cell-Derived Histone H1 Proteins in Innate Antimicrobial Defense in the Human Gastrointestinal Tract.." *Infection and Immunity* 66 (7) (July): 3255–3263.

Segrè, Ayellet V, DIAGRAM Consortium, MAGIC investigators, Leif Groop, Vamsi K Mootha, Mark J Daly, and David Altshuler. 2010. "Common Inherited Variation in Mitochondrial Genes Is Not Enriched for Associations with Type 2 Diabetes or Related Glycemic Traits." Edited by Peter M Visscher. *PLoS Genetics* 6 (8) (August 12): e1001058. doi:10.1371/journal.pgen.1001058.t004.

Subramanian, Aravind, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael A Gillette, Amanda Paulovich, et al. 2005. "Gene Set Enrichment Analysis: a Knowledge-Based Approach for Interpreting Genome-Wide Expression Profiles.." *Proceedings of the National Academy of Sciences* 102 (43) (October 25): 15545–15550. doi:10.1073/pnas.0506580102.

WHO. 2013. "WHO Polio Fact Sheet" (April 30): 1–3.

World Health Organization Collaborative Study Group on Oral Poliovirus Vaccine. 1995. "Factors Affecting the Immunogenicity of Oral Poliovirus Vaccine: a Prospective Evaluation in Brazil and the Gambia. ." *The Journal of Infectious Diseases* 171 (5) (May): 1097–1106.

Chapter 8: Conclusions

8.1: Research Questions and Goals

This dissertation had two major questions:

1. **What are the human genetics responsible for systemic immune response to oral poliovirus vaccine in children?** ***(Chapters 2, 3, and 7)***
2. **What are the best gene- and pathway-level methods for genome-wide association studies?** ***(Chapters 4, 5, and 6)***

Oral poliovirus vaccine (OPV) is known to be a model oral vaccine, largely responsible for decreasing global cases by over 99% in the last 25 years.(WHO 2013) It is the most well characterized oral vaccine and is efficient in diverse populations worldwide. Despite its success, there still remain some individuals who fail to respond adequately to the vaccine, leaving them susceptible to infection and the associated sequelae such as paralytic poliomyelitis. One hypothesis for this failure is host genetics.(Paul 2007) The heritability, or proportion of the phenotypic variation due to genetics, has been estimated to be high (60%) and ethnic differences in the efficacy of OPV have been observed.(Newport et al. 2004)

To address this question, we conducted a genome-wide association study (GWAS) in 357 Bangladeshi children to investigate the systemic immune response to OPV as determined by their log serum neutralizing antibody titers (LTs) comparing seronegative (LT<3) to high seropositive (LT>7) individuals. However, with a small sample size of 357 and a large number of comparisons (6.6 million), there was concern about limited power using this study design. Three secondary analyses were conducted on the GWAS results in addition to the original GWAS analysis to increase the overall power. The 1st approach was to correlate signatures of positive selection within the human genome with the GWAS results. This would highlight regions that contain a beneficial mutation that may be related to the response to OPV and therefore preserved throughout subsequent generations. We also applied gene- and pathway-level analytical methods for the GWAS results. There have been numerous methods developed in recent years to elucidate gene- and pathway-level associations. However, there has not been a consensus as to the most appropriate and accurate method. Therefore, we conducted a simulation experiment with an additive polygenic model simulated upon real genotypic data in which numerous genes and pathways were "causal". We then evaluated 22 programs (12 gene-level, 10 pathway-level) for their relative performance to determine the best methods. These best methods were applied to the GWAS; below we will describe the major findings for each of their components, organized according to these larger questions outlined above.

8.2: Major Findings

8.2.1: Genomics of the Response to Oral Poliovirus Vaccine

The genomics of the response to Oral Poliovirus Vaccine (OPV) were investigated using four different methods:

1. *A genome-wide association study (GWAS) comparing seronegative versus high seropositive responders to OPV in Bangladesh (Chapter 3)*
2. *A correlation of the GWAS with signatures of positive selection estimated within a larger sample of the same population (Chapter 3)*
3. *An application of gene-level methods to examine the joint association of single nucleotide polymorphisms (SNPs) from the GWAS (Chapter 7)*
4. *An application of pathway-level methods to examine the overrepresentation of highly ranked genes from the GWAS in biologically relevant gene sets (Chapter 7)*

Through these four methods we have highlighted different genes and pathways that may play a role in the pathogenesis of poliovirus and specifically the response to OPV.

The first approach was a traditional genome-wide association study design to identify risk loci involved in the systemic immune response to OPV as assessed by their log serum neutralizing antibody titers (LTs). Children were grouped into seronegative individuals (LT<3) and high seropositive individuals (LT>7). Association was assessed for 6.6 million single nucleotide polymorphisms (SNPs) across the human genome through logistic regression and adjusted for stunting (height-for-age Z score (HAZ) < -2), a previously established confounder for this relationship. While no markers reached genome-wide significance, several suggestive signals were found. The most significant finding was found on chromosome 14 at rs113427985 (OR=0.22, *P*=2.9x10-6) close to *MAPK1IP1L*. An additional association was found 50 kilobases away in *SOCS4* at rs112185488 (OR=0.22, *P*=5.8x10^{-6}). Although these two variants are far apart, they are in high linkage disequilibrium, indicating that the two signals could represent the same association. *SOCS4*, a suppressor of cytokine signaling, was previously implicated in enteric infections and the integrity of the gastric mucosa. An additional signal was found upstream of *SHH* at rs55906254 (OR=0.31, *P*=3.6x10^{-6}) on chromosome 7. A neighboring gene, *LMBR1*, also had numerous associations across the entire genic region. These two genes were previously shown to interact, with *LMBR1* containing a known cis-regulatory region for *SHH*.(Lettice et al. 2002) Sonic hedgehog is a gastric morphogen and has been shown to be associated with gut reconstruction following infection with enteric pathogens.(Xiao et al. 2012)

The second approach sought to correlate signatures of positive selection within the human genome with the GWAS results from the first approach. The cross-population extended haplotype homozygosity (XP-EHH) is a measure of positive selection. Through subsequent generations, a beneficial mutation will be conserved along with the surrounding variants due to linkage disequilibrium. XP-EHH detects these regions of the genome by looking at extended haplotypes and comparing them to a reference population. We used an African population as the reference population for the

Bangladeshi study subjects. We filtered genetic locations by a GWAS *P*-value < 0.001 and an XP-EHH *P*-value < 0.01. This resulted in 32 SNPs, half of which were found between *FAM86A* and *RBFOX1* on chromosome 16. Directionality for all these SNPs were consistent, showing positive selection (XP-EHH > 0) with longer haplotypes than the reference African population, and a protective effect with the derived (non-ancestral) allele (OR>1). This indicates that beneficial mutations in this region arose and were subsequently preserved in this population due to positive natural selection. We cannot determine if this was the result of poliovirus, but suspect that it may have an ancestral virus exerting selective pressure. This demonstrates the benefit in coupling the GWAS results with signatures of positive selection, especially when looking at infectious pathogens that have historically had a large effect on human populations.

The third approach was to aggregate the SNP-level associations using gene-level methods to test for joint associations within a gene that is likely due to allelic heterogeneity. Using only the top 10% of associated SNPs within the gene, including 20 kilobase flanking regions on either side of the gene, associations were reinforced for the top GWAS association signals in genes such as *LMBR1* (*P*=5.5x10^{-5}), *SOCS4* (*P*=4.3x10^{-4}), and *MAPK1IP1L* (*P*=5.5x10^{-4}). Additional associations were found for numerous histone marks in histone cluster 1 on chromosome 6. The highest association of these was *HIST1H2BL* (*P*=1.2x10^{-4}). These results suggest a role for epigenetic influences through histone modifications. It has previously been hypothesized that epigenetic imprinting may play a role in the response to OPV. Infections with both commensal probiotics as well as different enteric infections have previously been shown to cause epigenetic imprinting. (Ghadimi et al. 2012; Hamon et al. 2007) It is possible that epigenetic modifications may influence the gut integrity and mucosal immunity, influencing the immunological response to OPV.

The last approach sought to detect pathway-level associations within the GWAS. Two different programs were used for this analysis: MAGENTA and GSA-SNP. Their results were largely consistent, with 6 pathways in common with *P*<0.001. Four of these pathways were involved with cyclic AMP (cAMP) as a secondary messenger and G-protein signaling. Cyclic AMP is a known negative regulator of T cell immune function, influencing the suppressive abilities of T regulatory cells.(Mosenden and Taskén 2011) This includes the immune system's ability to respond to foreign antigens. A disruption in these signaling pathways could damage the ability to respond to the live attenuated poliovirus found in OPV, resulting in an inadequate systemic response. The other two pathways that were in common included neurological system processes and the regulation of the developmental process. Both pathways were very large and non-specific, however neurological system processes may play a role as the most serious sequelae of poliovirus infection is paralytic poliomyelitis which can occur when the virus cross the blood-brain barrier into the central nervous system.

8.2.2: Gene- and Pathway-Level Methods for Genome-wide Association Studies

We sought to conduct a systematic evaluation of the relative performance for gene- and pathway-level methods for genome-wide association studies. A simulation was conducted using real genotypic data from the Wellcome Trust Case-Control Consortium (WTCCC) and assuming an additive polygenic model. A total of 22 methods were evaluated: 12 gene-level and 10 pathway-level programs. Gene-level programs included: Fisher's Combination Test, Sidak's Combination Test, Simes' Test, False Discovery Rate Correction, Score Test, aSUM (Adaptive sum test), GATES, Weighted GATES, HYST, Weighted HYST, VEGAS, and VEGAS using only the top 10% of associated SNPs. Programs were evaluated based on their sensitivity and specificity, as well as type I and type II error. The highest sensitivity was found using Fisher's Combination Test (59.2%), which also had the lowest specificity (88.6%). Fisher's Combination Test also had the highest type I error rate (5.9%). The lowest sensitivity was found using Sidak's Combination Test (18.37%), with a specificity of 88.6% and type I error of 0.11%. Sensitivity was decreased for all methods when the analysis was limited to only genes with small effect size under the simulation (OR=1.2 vs. OR=2). When stratified by the number of causal SNPs within the gene, the highest sensitivities were found in genes having 5 causal SNPs versus 1 or 2 causal SNPs. This is consistent with the underlying hypothesis of gene-level methods, which aim to identify genes with multiple independent association signals. Out of the 12 methods, only VEGAS did not identify any genes with only one causal SNP. This is important as it means VEGAS is able to discern between genes with multiple causal SNPs and genes with only one association that may be due to high levels of linkage disequilibrium. The best balance of sensitivity, specificity, and type I error was present in VEGAS using only the top 10% of the associated SNPs. This method has a sensitivity of 28.6%, a specificity of 98%, and type I error rate of 0.4%.

A total of 10 programs were evaluated for pathway-level methods: ALIGATOR, gengen, GSA-SNP, GSEA-SNP, GRASS, HYST, INRICH, MAGENTA, Plink Set Test, and the SNP Ratio Test. These methods were divided into self-contained and competitive tests. Self-contained tests do not depend on the distribution of the other gene sets being tested while the significance of competitive tests does depend on the distribution of other genes. Because only 20 gene sets from the Gene Ontology Biological Processes were part of the simulation, a quantitative analysis of the programs was not possible. Instead a qualitative comparison of their results was evaluated. All programs had negative correlations between the proportion of associated genes within the gene set and the P-value. This supports the underlying hypothesis of pathway-level methods that a phenotype follows a polygenic model in which the higher proportion of genes that are associated within a gene set, the more important the gene set. This relationship was the clearest for GSA-SNP, gengen, and MAGENTA, all competitive methods. These methods had the advantage of having more stable estimates as well as strong correlation with the proportion of associated genes. The disadvantage of these methods is that they are dependent upon the gene sets being calculated and therefore results may not be reproducible across different pathway databases or releases of the same databases. Pathway-level methods do not evaluate gene-gene interactions or implicate a certain aspect of the gene set. Additional methods are required to ask these questions. Instead, pathways-level methods for GWAS provide an opportunity for researchers to conceptualize their GWAS results beyond the top associations.

8.3: Strengths and Limitations

The first component of this dissertation seeks to elucidate the genetic loci underlying the immune response to OPV. Traditional GWAS typically require large sample sizes and our study is no exception. With only 357 study participants, we have limited power to detect associations unless they have a very large effect size. This is reflected with the lack of genome-wide significant results. However, three different approaches were applied to this GWAS dataset to improve associations that may have been underpowered in the original analysis. By correlating signatures of positive selection with the phenotype of interest, results are put into their evolutionary context. Gene-level methods aim to increase power to detect association and pathway-level methods help researchers link suggestive signals and further explore these relationships. Another limitation of the study of OPV was that the associations were only adjusted for stunting. Prior publications suggest that exclusive breastfeeding and specifically breast-feeding at the time of vaccination may play a role in the decreased efficacy of OPV. Further studies should examine the potentially confounding role of breast-feeding with these associations. Additional confounders are the presence of symptomatic enteric infections leading to diarrhea and the presence of sub-clinical tropical enteropathy. Unfortunately, there is no consensus as to the best measurement of tropical enteropathy so adjustment is not possible. However, the presence of diarrhea or burden of enteric infections may be incorporated in subsequent analyses.

The simulation experiment offered a standardized evaluation of gene- and pathway-level methods. The systematic generation of phenotype with a large number of true negative and true positives allows for reliable and realistic estimates of sensitivity and specificity, as well as type I and type II error. For the gene-level methods, a limitation was the underrepresentation of smaller effect variants in the GWAS results. This prevented stable estimates of sensitivity within the smaller effect group due to a low number of smaller effect true positive genes. However, this is consistent with the infinitesimal model in which the majority of variance is found in small amounts at many small effect variants.(Gibson 2012) The majority of hidden heritability is expected to reside in these underpowered variants. Despite this limitation, the methods were still able to identify some smaller effect genes that would have otherwise been ignored by a traditional GWAS. A limitation for the pathway-level method comparison was the small number of pathways upon which the phenotype was simulated. This prevented a quantitative analysis for measures of accuracy. Despite this limitation, the simulation represents a realistic GWAS in that it is unlikely that there will be a large number (>20) of truly associated pathways. To answer this question it would be more ideal to simulate numerous phenotypes and assess their ability to identify the associated pathways across the different GWAS, instead of numerous pathways within one GWAS. However, the qualitative assessment of these programs does offer insight into their methodology, strengths, and limitations.

8.4: Future Directions

The four-pronged approach to identify the host genetics underlying the response to OPV has yielded numerous candidate genes that warrant follow-up. The first step would be to validate the GWAS findings in the *SHH*/*LMBR1* and *MAPK1IP1L*/*SOCS4* regions in a separate population. Recruitment is ongoing for the Exploration of the Biological Basis for Underperformance of Oral Polio and Rotavirus Vaccines in India (PROVIDE), a clinical trial for the efficacy of rotavirus vaccine and OPV ongoing in Bangladesh and India. These children will be genotyped for these candidate regions as they are similar to the cohort examined in this dissertation, allowing an opportunity to confirm and replicate our findings. Additionally, the GWAS and gene-level results could be followed-up with targeted resequencing to identify variants on a finer scale. The gene-level method identified an association within histone cluster 1 on chromosome 6, implicating a role for epigenetics in the immune response to OPV. Histone modifications could be examined in a longitudinal sample from birth to one year of age to see how different factors may influence the histones, as well as how the histone modifications influence different phenotypes such as gut integrity and response to OPV. Because of the tissue-specific nature of epigenetics, it will be important to choose the correct timing and sample to measure these modifications. Overall, the inquiries into the genetics of OPV response have generated a few candidates that are biologically plausible. Through targeted sequencing and alternative measures, such as histone modifications, the genomics of the immune response to OPV deserves a closer look.

8.5: Public Health Significance

The evaluation of these gene- and pathway-level methods will assist investigators as they evaluate their own associations. Traditional GWAS methodology requires stringent significance thresholds to handle multiple comparisons, essentially "tabling" all signals that fall below $5x10^{-8}$. Gene- and pathway-level methods for GWAS seek to formalize a test for multiple associations within a biologically relevant unit. Our results will inform future researchers as to the best method for their project so that all of the associated variation in GWAS may be elucidated.

The efficacy of OPV has been validated in diverse populations around the world. It has been highly successful through mass immunizations, which is largely due to the easy administration of the oral vaccine.(Pasetti et al. 2011) Other vaccines have been modeled after OPV to elicit mucosal immunity. The most notable is against rotavirus with two licensed vaccines: Rotareq and Rotarix. The latter was created through serial passage in tissue culture, similar to Sabin's OPV strain.(Pasetti et al. 2011) Both rotavirus vaccines and OPV show decreased effectiveness in developing versus developed countries. This may be due to biological factors within children in developing countries, such as the presence of tropical enteropathy leading to poor gut integrity and an inability to mount an adequate response to enteric pathogens.(Korpe and Petri 2012)

However, a study of Brazilian children estimated that the heritability of early childhood diarrhea was 54%, suggesting that the extent of tropical enteropathy may be partly genetic.(Pinkerton et al. 2011) By understanding the genetic risk factors for the response to OPV, it informs the general mechanisms of oral vaccines that aim to target the mucosal immunity.

The response to OPV has high levels of variability both within and between populations. Even with the same vaccine, children response differently. Human genetics may play a role in this variability, with individuals carrying mutations that confer stronger immunological responses. Genetic epidemiology seeks to detect these mutations on a population-level scale, which can then be related to the individual response. By examining the underlying human genomics of these diverse responses, not only do we better understand the mechanisms of the immune response to OPV but may lead to potential adjuvants and improved vaccines. This is a public health issue that can be addressed as we move genetic knowledge forward.

References

Ghadimi, D, U Helwig, J Schrezenmeir, K J Heller, and M de Vrese. 2012. "Epigenetic Imprinting by Commensal Probiotics Inhibits the IL-23/IL-17 Axis in an in Vitro Model of the Intestinal Mucosal Immune System." *Journal of Leukocyte Biology* 92 (4) (October 1): 895–911. doi:10.1189/jlb.0611286.

Gibson, Greg. 2012. "Rare and Common Variants: Twenty Arguments." *Nature Reviews Genetics* 13 (2) (February 1): 135–145. doi:10.1038/nrg3118.

Hamon, Mélanie Anne, Eric Batsché, Béatrice Régnault, To Nam Tham, Stéphanie Seveau, Christian Muchardt, and Pascale Cossart. 2007. "Histone Modifications Induced by a Family of Bacterial Toxins.." *Proceedings of the National Academy of Sciences* 104 (33) (August 14): 13467–13472. doi:10.1073/pnas.0702729104.

Korpe, Poonum S, and William A Petri Jr. 2012. "Environmental Enteropathy: Critical Implications of a Poorly Understood Condition." *Trends in Molecular Medicine* 18 (6) (June 1): 328–336. doi:10.1016/j.molmed.2012.04.007.

Lettice, Laura A, Taizo Horikoshi, Simon J H Heaney, Marijke J van Baren, Herma C van der Linde, Guido J Breedveld, Marijke Joosse, et al. 2002. "Disruption of a Long-Range Cis-Acting Regulator for Shh Causes Preaxial Polydactyly.." *Proceedings of the National Academy of Sciences* 99 (11) (May 28): 7548–7553. doi:10.1073/pnas.112212199.

Mosenden, Randi, and Kjetil Taskén. 2011. "Cellular Signalling." *Cellular Signalling* 23 (6) (June 1): 1009–1016. doi:10.1016/j.cellsig.2010.11.018.

Newport, M J, T Goetghebuer, H A Weiss, H Whittle, C-A Siegrist, and A Marchant. 2004. "Genetic Regulation of Immune Responses to Vaccines in Early Life." *Genes and Immunity* 5 (2) (January 22): 122–129. doi:10.1038/sj.gene.6364051.

Pasetti, Marcela F, Jakub K Simon, Marcelo B Sztein, and Myron M Levine. 2011. "Immunology of Gut Mucosal Vaccines.." *Immunological Reviews* 239 (1) (January): 125–148. doi:10.1111/j.1600-065X.2010.00970.x.

Paul, Yash. 2007. "Role of Genetic Factors in Polio Eradication: New Challenge for Policy Makers." *Vaccine* 25 (50) (December): 8365–8371. doi:10.1016/j.vaccine.2007.09.068.

Pinkerton, R C, R B Oria, J W Kent, A Kohli, C Abreu, O Bushen, A A M Lima, J Blangero, S Williams-Blangero, and R L Guerrant. 2011. "Evidence for Genetic Susceptibility to Developing Early Childhood Diarrhea Among Shantytown Children Living in Northeastern Brazil." *American Journal of Tropical Medicine and Hygiene* 85 (5) (November 2): 893–896. doi:10.4269/ajtmh.2011.11-0159.

WHO. 2013. "WHO Polio Fact Sheet" (April 30): 1–3.

Xiao, Chang, Rui Feng, Amy C Engevik, Jason R Martin, Julie A Tritschler, Michael Schumacher, Robert Koncar, et al. 2012. "Sonic Hedgehog Contributes to Gastric Mucosal Restitution After Injury." *Laboratory Investigation* 93 (1) (October 22): 96–111. doi:10.1038/labinvest.2012.148.

CURRICULUM VITAE

Genevieve L. Wojcik

PERSONAL DATA

615 N. Wolfe St. W6517
Baltimore, MD 21205
E-mail: gwojcik@jhsph.edu
Phone: 413.530.4338

EDUCATION

Johns Hopkins Bloomberg School of Public Health, Baltimore MD

PhD, Epidemiology — December 2013
Area of Concentration: Genetic Epidemiology
MHS, Human Genetics/Genetic Epidemiology — 2010
Certificate in Vaccine Science and Policy — 2010

Cornell University, Ithaca NY

BA, Biology — 2008
Area of Concentration: Genetics and Development, French

RESEARCH EXPERIENCE

Graduate Research Assistant — Spring 2009-Present
Johns Hopkins Bloomberg School of Public Health, Baltimore MD
Laboratory of Dr. Priya Duggal, Department of Epidemiology

Student Researcher — Fall 2007-Spring 2008
Cornell University, Ithaca NY
Laboratory of Dr. Andrew Clark, Department of Molecular Biology and Genetics

Student Researcher — Summer 2007
Cornell University, Ithaca NY
Laboratory of Dr. Paula Cohen, Department of Biomedical Sciences, Genetics

Research Assistant Summer 2005

Baystate Medical Center, Springfield MA

Department of Clinical Research

Student Intern Summer 2003, 2006

University of Massachusetts, Amherst MA

Laboratory of Dr. Sandra Petersen, Department of Veterinary and Animal Sciences

HONORS AND AWARDS

The Charlotte Silverman Award Recipient 2013

Johns Hopkins Bloomberg School of Public Health, Baltimore MD

Department of Epidemiology

The Johns Hopkins Vaccine Initiative, Vaccine Day Poster Winner 2012

Johns Hopkins Bloomberg School of Public Health, Baltimore MD

"Genome-wide Association Study of Response to Oral Poliovirus Vaccine in Bangladeshi Children"

The Charlotte Ferencz Fellowship 2012

Johns Hopkins Bloomberg School of Public Health, Baltimore MD

Department of Epidemiology

Mary Meyers Scholar, recognizing the top two doctoral students in the department 2010-2012

Johns Hopkins Bloomberg School of Public Health, Baltimore MD

Department of Epidemiology

Dean's List 2008

Cornell University, Ithaca NY

PUBLICATIONS

1. Jaffe A*, **Wojcik G***, Chu A, Golozar A, Maroo A, Duggal P, Klein AP. Identification of functional genetic variation in exome sequence analysis *BMC Proc* 2011, 5(9): S13 *[*shared authorship]*
2. Arav-Boger R, **Wojcik GL**, Duggal P, Ingersoll RG, Beaty T, Pass RF, Yolken RH. Polymorphisms in Toll-like receptor genes influence antibody responses to cytomegalovirus glycoprotein B vaccine *BMC Res Notes* 2012, 5:140.
3. Duggal P, Thio C, **Wojcik GL**, Goedert JJ, Mangia A, Latanich R, Kim AY, Lauer GM, Chung RT, Peters MG, Kirk GD, Mehta SH, Cox AL, Khakoo SI, Alric L, Cramp ME, Donfield SM, Edlin BR, Tobler LH, Busch MP, Alexander G, Rosen HR, Gao X, Abdel-Hamid M, Apps R, Carrington M, Thomas DL. Genome-wide association study of spontaneous resolution of hepatitis C virus infection *Annals of Internal Medicine* 2013, 158(4): 235-245.
4. Kim Y, Tilley MK, Parker MM, **Wojcik GL**, Maroo A, Klein AP, Duggal P. A Comparison of the accuracy of protein prediction methods to classify human genetic variation *PLoS One, in review*
5. **Wojcik GL**, Mosbruger T, Latanich R, Astemborski J, Kirk GL, Kim A, Seaberg EC, Busch M, Thomas DL, Duggal P, Thio CL. Genetics variants in HAVCR1 gene region as a partial explanation for high hepatitis C persistence in African-Americans, *Journal of Infectious Diseases, in press [Advance Access)]*
6. **Wojcik GL**, Thomas DL, Thio C, Duggal P, HCV Consortium. Admixture Analysis of Spontaneous Hepatitis C Viral Clearance in Individuals of African-Descent, *in preparation*
7. **Wojcik GL**, Duggal P. Review and Evaluation of Gene-Level Methods for Genome-wide Association Studies, *in preparation*
8. **Wojcik GL**, Mondal D, Alam M, Mychaleckyj J, Rich S, Concannon P, Haque R, Pallansch M, Petri WA, Duggal P. Signatures of Selection and a Genome-wide Association Study of Response to Oral Poliovirus Vaccine in Bangladeshi Children, *in preparation*

TEACHING EXPERIENCE

Johns Hopkins Bloomberg School of Public Health

Infectious Disease Dynamics, 4th Term	2013
Principles of Genetic Epidemiology, 1st Term	2012
Summer Institute Population Genetics	2011, 2012
Methods for Linkage Analysis in Genetic Epidemiology, 4th Term	2010, 2011,2012
Epidemiological Methods III, 3rd Term [Lead TA]	2012
Epidemiological Methods III, 3rd Term	2010,2011
Principles of Epidemiology, Summer Term	2009
Introduction to Population Genetics, 2nd Term	2009
Introduction to Genetic Epidemiology, 1st Term	2009

Cornell University

BioG112: Current Topics in Biology and Society	2008
BioGD281: Genetics	2007,2008
Biology and Genetics Tutor	2007,2008
BioBM330: Biochemistry	2006

ABSTRACTS

Platform Presentations

- **Wojcik GL**, Mondal D, Alam M, Mychaleckyj J, Rich S, Concannon P, Haque R, Petri WA, Duggal P. "Age-dependent genetic associations with Cryptosporidium infection in Bangladeshi children" ASTMH 61st Annual Meeting, Atlanta GA, 2012
- Thomas DL, **HCV Consortium**. "Genome-wide study of spontaneous hepatitis C virus infection" CROI, Seattle WA, 2012

Posters

- **Wojcik GL**, Mondal D, Alam M, Mychaleckyj J, Rich S, Concannong P, Haque R, Pallansch M, Kirkpatrick BD, Petri WA, Duggal P. "Host Genetic Regions under Natural Selection Associated with Oral Poliovirus Vaccine Response in Bangladeshi Children" ASTMH 62nd Annual Meeting, Washington DC, 2013
- Zignego AL, **Wojcik GL**, Cacoub P, Visentini M, Fiorilli M, Terrier B, Mangia A, Latanich R, Charles E, Khakoo SI, Busch MP, Dustin LB, Thomas DL, Duggal P. "Genome-wide association study of hepatitis C virus- and cryoglobulin-related vasculitis" The Liver Meeting, AASLD, Washington DC, 2013
- **Wojcik GL**, Kao W-HL, Duggal P. **"**Relative performance and application of gene- and pathway-level methods for genome-wide association studies" 63rd Annual American Society of Human Genetics Meeting, Boston MA, 2013
- **Wojcik GL**, Kao W-HL, Duggal P. "A systematic evaluation of gene- and pathway-level methods for genome-wide association studies through simulations" IGES, Washington DC, 2013
- **Wojcik GL**, Duggal P, HCV Consortium. "Admixture Analysis of Spontaneous Hepatitis C Virus Clearance among Individuals of African-Descent" Delta Omega Poster Competition. Johns Hopkins Bloomberg School of Public Health, Baltimore MD. 2013
- **Wojcik GL**, Mondal D, Alam M, Mychaleckyj J, Rich S, Concannon P, Haque R, Pallansch M, Petri WA, Duggal P. "Genome-wide Association Study of Response to Oral Poliovirus Vaccine in Bangladeshi Children" Vaccine Day Poster Competition. Johns Hopkins Bloomberg School of Public Health, Baltimore MD, 2012
- **Wojcik GL**, Mondal D, Alam M, Mychaleckyj J, Rich S, Concannon P, Haque R, Petri WA, Duggal P. "Age-dependent genetic associations with Cryptosporidium infection in Bangladeshi children" Young Investigators Competition ASTMH 61st Annual Meeting, Atlanta GA, 2012
- **Wojcik GL**, Duggal P, HCV Consortium. "Admixture Analysis of Spontaneous Hepatitis C Virus Clearance among Individuals of African-Descent" 62nd Annual American Society of Human Genetics Meeting, San Francisco CA, 2012
- Duggal P, **Wojcik GL**, HCV Consortium. "Genome-wide association study of spontaneous resolution of HCV virus infection" Genomics of Common Disease, Potomac MD, 2012
- **Wojcik G**, Thomas D, Duggal P. "Evaluating Associations and interactions of spontaneous clearance of Hepatitis C infection using logic regression" IGES, Boston MA, 2010

SERVICE

President, Epidemiology Student Organization 2011-12
Johns Hopkins Bloomberg School of Public Health, Baltimore MD

Chair, STARS (Students Teaching and Reaching Students) 2009-10, 2011-12
Johns Hopkins Bloomberg School of Public Health, Baltimore MD

Volunteer, STARS (Students Teaching and Reaching Students) 2008-12
Johns Hopkins Bloomberg School of Public Health, Baltimore MD

Volunteer EMT-Basic 2007-08
Dryden Ambulance, Dryden NY