

**UCSF**

**UC San Francisco Electronic Theses and Dissertations**

**Title**

Investigating host genes involved in HIV control by a novel computational method to combine GWAS with eQTL

**Permalink**

<https://escholarship.org/uc/item/12x6k23n>

**Author**

Song, Yi

**Publication Date**

2012

Peer reviewed|Thesis/dissertation

Investigating Host Genes Involved in HIV Control by a Novel  
Computational Method to Combine GWAS with eQTL

by

Yi Song

THESIS

Submitted in partial satisfaction of the requirements for the degree of

MASTER OF SCIENCE

in

Biological and Medical Informatics

in the

GRADUATE DIVISION

Copyright (2012)

by

Yi Song

## **Acknowledgement**

First and foremost, I would like to thank my advisor Professor Hao Li, without whom this thesis would not have been possible. I am very grateful that Professor Li lead me into the field of human genomics and gave me the opportunity to pursue this interesting study in his laboratory. Besides the wealth of knowledge and invaluable insights that he offered in every meeting we had, Professor Li is one of the most approachable faculties I have met. I truly appreciate his patient guidance and his enthusiastic supervision throughout my master's career.

I am sincerely thankful to Professor Patricia Babbitt, the Associate Director of the Biomedical Informatics program at UCSF. Over my two years at UCSF, she has always been there to offer her help when I was faced with difficulties. I would also like to thank both Professor Babbitt and Professor Nevan Krogan for investing their valuable time in evaluating my work.

I take immense pleasure in thanking my co-workers Dr. Xin He and Christopher Fuller. It has been a true enjoyment to discuss science with Dr. He, whose enthusiasm is a great inspiration to me. I also appreciate his careful editing of my thesis. Christopher Fuller, a PhD candidate in the Biomedical Informatics program, has provided great help for me on technical problems. I also take this opportunity to sincerely thank my classmates, especially to Javona Whitebears and Natalia Khuri, who have selflessly taught me knowledge in programming, without which I could not have finished this work.

Last but not least, I place on record my sense of gratitude to one and all who, directly or indirectly, have lent their helping hand in this venture.

## Abstract

Acquired immunodeficiency syndrome (AIDS) is one of the most deadly diseases worldwide. AIDS was first reported in 1981, with its disease causing virus discovered and isolated two years later (Gottlieb et al., 1981; Barré-Sinoussi et al., 1983; Gallo et al., 1983). Since then, the three decades of research has seen huge progress on many aspects, especially on lengthening the life span of HIV infected patients. Yet today, there are still more than 34 million people living with HIV/AIDS, ([http://www.amfar.org/About\\_HIV\\_and\\_AIDS/Facts\\_and\\_Stats/Statistics\\_\\_Worldwide/](http://www.amfar.org/About_HIV_and_AIDS/Facts_and_Stats/Statistics__Worldwide/)) and we are no where close to thoroughly understanding the pathogenesis of this virus and to finding an ultimate cure for the disease. Despite the enormous amount of studies, the enigma of HIV infection and how it progresses to AIDS remains elusive.

The progression of HIV infection varies greatly among individuals. Since HIV uses the cellular machinery to replicate, many researchers have been focusing on identifying the host factors that determine the resistance to HIV progression. Numerous genome wide association studies (GWAS) have been conducted to unveil these determining genetic variation and to infer causal genes, but there has been little success. Most GWA studies agree on the significant roles of HLA genes (mainly HLA-B and HLA-C), which are challenging candidates due to their complexity.

In this study, I adopt a novel computational method to identify candidate genes by leveraging the information in GWAS and expression quantitative trait loci (eQTL) data. The combination of GWAS and eQTL reveal several new genes, including MED28, CD151, A4GALT, and ANAPC2, that have never been implicated in previous GWA

studies. Substantial literature evidence support the potential roles of these genes. Hypergeometric test between HIV interactome data (Jager et al., 2011), RNAi screens (Brass et al., 2008; Konig et al., 2008; Zhou et al., 2008; Yeung et al., 2009) and my result shows significant overlap.

## Table of Content

Chapter One: Introduction .....	1
§1.1 Genome Wide Association Study (GWAS) .....	1
§1.2 New Approach to Interpret GWAS Results .....	2
§1.3 The Control of HIV-1 .....	4
§1.4 Previous works .....	6
§1.4.1 GWA Studies .....	7
§1.4.2 RNAi Screens .....	8
§1.4.3 Interactome Studies .....	9
§1.4.4 CCR5 and CXCR5 .....	10
§1.5 Objectives .....	11
Chapter Two: Analysis and Results .....	12
§2.1 Careful Choosing of the GWAS and eQTL Datasets .....	12
§2.2 Producing the List of Target Genes .....	13
§2.3 Brief Summary of the Top Ten Genes Produced in 2.2 .....	14
§2.4 Pruning the GWAS Datasets and Producing a Second List of Target Genes .....	16
§2.5 Brief Summary of the Top Ten Genes Produced in 2.4 .....	17
§2.6 Overlap Studies .....	20



§2.6.1 Significance of Overlap .....	20
§2.6.2 Rank-sum Test of the Overlap Genes .....	21
§2.6.3 Overlap Between the Interactome, the RNAi and My Result .....	22
Chapter Three: Discussion and Self Assessment .....	24
§3.1 New Ways to Interpret GWAS May Lead to the Discovery of New Genes .....	24
§3.2 Multiple GWAS Datasets and eQTL Datasets Will Improve the Accuracy .....	25
§3.3 Critiques About This Work and Future Direction .....	26
Bibliography .....	29
Appendix .....	38
Publishing Agreement .....	47

## List of Tables

Table 1: Program Parameters .....	38
Table 2: Top Ten Candidates Before Pruning .....	38
Table 3: SNPs Support for Four of the Top Ten Candidates .....	39
Table 4: Top Ten Candidates After Pruning .....	40
Table 5: SNPs Support for Top Ten Candidates After Pruning .....	41
Table 6: SNPs Support for CXCR4 and APOBEC3G .....	44

## List of Figures

Figure 1: Illustration of the Algorithm's Idea .....	45
Figure 2: Counts of the Overlapping Genes with Respect to Rank .....	46

## **Chapter One: Introduction**

In the first half of this chapter, I introduce the concept of GWAS and the new approach I adopt to understand GWAS data. The second part presents the biological question I try to answer by laying out the background of HIV infection progression and previous works on HIV control.

### **1.1 Genome Wide Association Study (GWAS)**

Mendel's laws of inheritance is the earliest model to explain the transmission of phenotypes in diploid organisms (Altshuler et al., 2008). Mendel Genetics can be elegantly summarized into the law of segregation and the law of independent assortment. Yet the advancement of molecular biology and molecular genetics has shown that most human phenotypes, such as body height, disease risk, and disease resistance, are not determined by single alleles. Naturally occurring phenotypes cannot be explained by Mendel Genetics. Instead, they are results of extremely complex interplay between genes, epigenetic and environmental factors.

The idea of genetic association studies emerged in the 1980s. In the 1990s, Hill et al. reported case-control studies that showed the association between alleles in Human Leukocyte Antigen (HLA) genes and the resistance to severe malaria (Hill et al., 1991; Hill et al., 1992). These early association studies were based on hypotheses about the functions of the target genes, but prior knowledge might not be available in most search for the targets. It was much desired to develop a systematic method to uncover true causal

genes with minimal prior knowledge. This did not become practical until the advent of GWAS.

GWAS examines common genetic variation in a population to see if the variation is associated with certain traits. Typically, GWAS focuses on the association between single-nucleotide polymorphisms (SNPs) and traits such as the resistance or susceptibility to certain disease, or the response to drugs and therapies. The study breaks test subjects into two groups. One group has the trait (case) and the other does not (control). By comparing the DNAs between the two groups, GWAS returns a list of SNPs, each followed by a p-value indicating the significance of the association. A small p-value indicates a high confidence about the association.

Most classical GWA studies focus on the genes that contain or in close proximity to the most significant signals. For instance, in 2005, the GWAS published by Klein et al. greatly supported the hypothesis that polymorphisms in Human Complement Factor H (CFH) can lead to Age-related Macular Degeneration (AMD). The strongest SNP signal rs380390 is contained within CFH gene (Klein et al., 2005). In another famous triumph of GWAS on Crohn's Disease, Duerr et al. identified IL23R as associated with the disease. IL23R was revealed due to multiple strong signals within the IL23R region (Duerr et al., 2006).

## **1.2 New Approach to Interpret GWAS Results**

The shortcoming of the approach mentioned above is that GWAS identifies loci but not genes per se. One reason for the successes in the cases of AMD and Crohn's

Disease is that the strongest signals all happen to reside within the genes, but this is not always the case. It did not take long for the researchers to realize that most loci identified do not map to amino acid changes in proteins. In fact, many of the loci do not map to the open reading frames of any recognizable protein, but intergenic regions, causing huge ambiguities in the mapping from SNPs to causal genes (Hardy and Singleton, 2009).

The speculation is that the loci alter transcriptional or translational efficiency (Hardy and Singleton, 2009). Studies have shown that complex-trait associated variants overlap with eQTL, a variant of GWAS that associate SNPs with the expression level of a certain gene (Emilsson et al., 2008; Nica et al., 2010; Stranger et al., 2011). A recent study also reported that on a global scale, the trait-associated SNPs are highly likely to be eQTL SNPs (eSNPs) (Nicolae et al., 2010). Thus, many scientists in the field have envisioned the prospect of using eQTL data to prioritize the GWAS SNPs and interpret GWAS results (Emilsson et al., 2008; Cookson et al., 2009; Mackay et al., 2009). Recently, there have been numerous attempts to combine GWAS data with eQTL data in order to leverage the information in these datasets. Many have produced promising results (Levy et al., 2009; Hsu et al., 2010; Speliotes et al., 2010; Sille et al., 2012). However, none of these studies provided a systematic method to relate eQTL to GWAS. As a result, much useful information could have been missed by these studies regardless of the costly and time-consuming production of GWAS data.

My colleagues in Dr Hao Li's lab, Dr. Xin He and a PhD candidate Christopher Fuller, have developed a Bayesian statistical method that matches GWAS signals with eQTL signals. Figure 1 (page 45) illustrates the basic idea of this method. The model is

built upon the notion that a genetic variation affecting the expression of a disease-causing gene must also affect the disease. Therefore, the set of SNPs of a disease-causing gene should overlap with that of the disease itself. It assesses the p-values of an overlapping SNP between GWAS and eQTL datasets, and returns a score (a Bayesian Factor) for each SNP. The score of a gene is the sum of the scores for each overlapping SNP. This method can also be considered as an alignment. For every gene, one can perform an alignment between the GWAS signals and the eQTL signals along the genome. The more the strong signals match up, the more likely it is a causal gene. The profiling can technically be done for every gene in the genome.

In addition to directly assessing the likelihood of a causal gene, this model comes with a great advantage to use the information in both the *cis*- and the *trans*- SNPs. *Trans*-SNP signals are thought to be weaker than *cis*- signals (Dixon et al., 2007). However, due to the large quantity of *trans*-SNPs, the combined effect might be remarkable. He and Fuller have performed analysis on Crohn's Disease, and have identified several novel candidate causal genes, most of which supported by existing literature evidence (manuscript in preparation). We believe that there is rich information embedded in the *trans*-SNPs, and that our method is one step ahead towards fully understanding the GWAS results.

### **1.3 The Control of HIV-1**

In my work, I based all of my analysis on the previous works on HIV-1. For the rest of my thesis, I refer to HIV-1 as HIV.

HIV affects human health by hijacking the immune system. The major sites of HIV infection includes lymphocytes, in particular CD4<sup>+</sup> T cells, macrophages and dendritic cells (Embretson et al., 1993). Once infected, the patient's immune system is doomed with destruction, followed by the onset of AIDS. In the initial stage of HIV infection, there is a short but intense period of viral replication (Tindall and Cooper, 1991). Possible symptoms of the this stage were reported to resemble that of an acute infectious-mononucleosis-like illness, such as fevers, sweats, malaise, pharyngitis, and so on (Cooper et al., 1985; Deeks and Walker, 2007). The outburst of HIV particles in the periphery blood cells is then followed by a prolonged period of clinical latency (Pantaleo et al., 1993). The length of the latency varies depending on the individual and many other factors. The short extreme could be < 2 years and the long extreme could be > 15 years (van Manen et al., 2011). Despite the apparently low viral activity and the steady-state viral count, HIV undergoes active replication during this time (Pantaleo et al., 1993) in the potential HIV reservoirs including lymphoid tissue, bone marrow and brain (Chun and Fauci, 1999). By the end of the infection course, the patient's immune system starts to fail. The CD4<sup>+</sup> T cell count drops precipitously and the viral count increases drastically.

To uncover the factors that affect the length of HIV clinical latency has a profound impact on the development of anti-HIV therapies. It has become a major anti-HIV strategy to protect patients from the progression of the infection, to maximize the length of the clinical latency, and to postpone the onset of AIDS (Deeks and Walker, 2007). In the 1990s, many studies showed the correlation between the virologic setpoint,



the viral load in plasma during the clinical latency, and the disease progression rate (Ho, 1996). In particular, Mellors et al. demonstrated that the virologic setpoint is directly related to the probability of developing AIDS in years (Mellors et al., 1996). These findings contributed to the initiation of the highly active antiretroviral therapy (HAART), which has remarkable efficacy to lower the setpoint and to lengthen the clinical latency, and has remained the major treatment for HIV till now.

The success of HAART seems to suggest that controlling the viral load is the right approach to combat HIV. Just a few years after the introduction of HAART, studies found a small group of HAART-naïve patients with impressive ability to control the viral load, and these patients also tended to have better prognosis than average people (Hubert et al., 2000; Goudsmit et al., 2002). This group of patients are termed “HIV controllers”, and can be divided into two subsets. One subset shows low plasma viremia ( $< 5000$  HIV-RNA copies/ml). They are termed long-term nonprogressors. People in the other subset are “elite controllers” or “natural controllers”, who have positive antibody tests, yet carry no measurable viral load ( $< 50$  HIV-RNA copies/ml) (Deeks and Walker, 2007; Saxena et al., 2007).

#### **1.4 Previous Works**

HIV heavily depends on the cellular machinery to replicate. Both of the clinical latency and the onset of AIDS are results of an extremely dynamic and complex interplay between HIV proteins and numerous host factors. To understand the variability among

individuals, the immediate attempt is to look for potential HIV host factors (Fauci, 1996; Haynes et al., 1996).

#### **1.4.1 GWA Studies**

The arrival of the genomic era advanced the HIV studies to a new stage. Recently there have been numerous GWAS conducted on HIV control (Fellay et al., 2007; Fellay et al., 2009; International HIV Controllers Study et al., 2010; Pelak et al., 2010; van Manen et al., 2011). These studies have not produced consistent results, but they agree unanimously on HLA genes (HLA-B and HLA-C in particular) being the strongest determinant for virologic set point.

HLA complex, or major histocompatibility complex (MHC), is a gene cluster located on Chromosome 6. It spans over 3.6 megabases, approximately from 29,000,000 to 33,000,000 bp from the pter (O'Brien et al., 2001). The HLA complex contains more than 200 genes, some of which, according to the HLA nomenclature website, have extremely large number of alleles (<http://hla.alleles.org/nomenclature/stats.html>). This highly complex structure of HLA cluster hinders our understanding of its roles in HIV control. Not all of these genes are related to immunity. The ones that are consist of two classes (class I and class II), both of which are expressed on the surface of cells and function in antigen presentation.

Class I molecules include HLA-A, HLA-B and HLA-C and are expressed on most somatic cells in the body. They bind to endogenous peptides, which are the product of proteasome degradation, on the luminal surface of endoplasmic reticulum (ER). These

include viral peptides if the cell is infected. Class I molecules present the antigen to CD8 receptors on the cytotoxic T cells (CD8<sup>+</sup> T cells). Numerous GWA studies confirmed them as one of the major determinants of HIV control. In Caucasian population, HLA-B\*5701 and HLA-B27 are closely associated with long survival time (Migueles et al., 2000; Gao et al., 2005; Fellay et al., 2007; Navis et al., 2007), whereas HLA-B35 is associated with accelerated progression to AIDS (Carrington and O'Brien, 2003; van Manen et al., 2011). In African American population, HLA-B\*5703 is associated with delayed onset of AIDS. Class II molecules include many more genes. In contrast to class I molecules, they are mainly expressed on B cells, activated T cells, macrophages, and dendritic cells. They bind to exogenous peptides degraded in the lysosome, and present them to CD4<sup>+</sup> T cells (Klein and Sato, 2000).

#### **1.4.2 RNAi Screens**

HIV relies heavily on the host factors. There have been several RNAi screens on the HIV host factors (Brass et al., 2008; Konig et al., 2008; Zhou et al., 2008; Yeung et al., 2009). In these studies, researchers firstly knock down cellular genes with small interfering RNAs (siRNAs). Then, the cells are infected by engineered reporter HIV virus, which signifies successful infection (Brass et al., 2008; Konig et al., 2008; Zhou et al., 2008). Another way to evaluate infectivity is to measure cell proliferation, assuming that HIV infection will lead to cell death (Yeung et al., 2009). However, these results have not been reproducible. Not a single gene is in common across all four studies. The result by Yeung et al. seems to agree the least with the rest, possibly due to the different

way to assess infectivity. MED6, MED7, and RELA are the common genes between Brass et al.'s, Zhou et al.'s and Konig et al.'s results. The following shows the pairwise intersection between these four studies.

Brass et al. with Zhou et al.: AKT1, CD4, CXCR4, DDX3X, JAK1, MED28, MED4, MED6, MED7, NUP155, RAB28, RELA, RNF26, TCEB3, WNK1.

Brass et al. with Konig et al.: CTDP1, DMXL1, IDH1, MAP4, MED6, MED7, MID1IP1, NUP153, RANBP2, RELA, TNPO3, TRIM55.

Brass et al. with Yeung et al.: EXOSC5, MR1, ZNF354A.

Zhou et al. with Konig et al.: ADRBK1, ANAPC2, CHST1, MED19, MED6, MED7, MRE11A, PRDM10, RELA.

Zhou et al. with Yeung et al.: CKLF, NFKB1.

Konig et al. with Yeung et al.: AES, DLGAP4, EPAS1.

### **1.4.3 Interactome Studies**

The monumental work in Dr. Nevan Krogan's lab presents a systematic characterization of the pairwise interactions between HIV proteins and host proteins (Jager et al., 2011). Jager et al. developed a computational method (MiST) to evaluate the interactions. Only interactions with sufficient reproducibility, specificity and abundance could pass the criteria. This study identified 435 host factors from HEK293 cells and Jurkat cells. Among the 435 proteins, 55 overlap with the proteins found in the four RNAi screens combined, with a hypergeometric p-value of  $2.7 \times 10^{-10}$ .

#### 1.4.4 CCR5 and CXCR4

CCR5 and CXCR4 are chemokine receptors mainly expressed on subsets of hematopoietic cell. Besides their functions in directing leukocyte migration in inflammation, CCR5 and CXCR4 are well-known facilitators of HIV infection by acting as a co-receptor (Springer, 1994; Mackay, 1996). So far, they are the only host factors identified with profuse experimental evidence.

In a mature HIV particle, the HIV genome is enclosed in a protein core, which is further protected by a lipid bilayer envelope. The envelope is acquired from the host cell membrane as the virion buds from the cell (Ganser-Pornillos et al., 2008). Embedded in the lipid envelope is the viral glycoprotein gp120/gp41. Gp120 needs to bind to both CD4, an HIV receptor expressed on the surface of leukocytes such as CD4<sup>+</sup> T cells, monocytes, macrophages and dendritic cells, and its co-receptor CCR5 (in some cases CXCR4). With the present of gp41, this interaction induces the fusion between the host cell membrane and the viral envelope. Hence, the intact CCR5 or CXCR4 is necessary for the viral entry (Choe et al., 1996; Bleul et al., 1997; Chan et al., 1997; Gallo et al., 2003).

The mid 1990s witnessed a breakthrough that a 32-bp deletion allele of CCR5 (CCR5 $\Delta$ 32) confers resistance to HIV infection. Dean et al. found that the CCR5 $\Delta$ 32 homozygotes were antibody-negative when exposed with HIV, and that the heterozygotes displayed slower progression to AIDS (Dean et al., 1996). Chemokine receptor antagonists such as (AOP)-RANTES and TAK-779 were immediately explored about their potential as anti-HIV therapy (Simmons et al., 1997; Baba et al., 1999). Various

other methods have also been tried, including siRNA and intracellular immunization against CCR5 (Steinberger et al., 2000; Qin et al., 2003).

### **1.5 Objectives**

In my study, I aim to apply the computational method developed in my lab to look for potential genes that might be associated with the control of HIV progression. The study will be particularly valuable if any gene outside of HLA complex can be implicated and supported with strong literature evidence. I will also compare my result with previous host factor screenings and assess if my method is more productive and accurate.

## **Chapter Two: Analysis and Results**

In order for the program to produce the most accurate result, I select a GWAS study and an eQTL study with best quality. Based on these two datasets, a list of genes, ranked by their likelihood of being a causal gene associated with the HIV controller phenotype, is generated by the program. I then compare my result with two datasets. One is the interactome data using mass spectroscopy (Jager et al., 2011), and the other is a combination of four RNAi results (Brass et al., 2008; Konig et al., 2008; Zhou et al., 2008; Yeung et al., 2009). I perform hypergeometric tests and rank-sum tests to assess how much the results agree. For the top ten genes on my list and the intersect of the three datasets, I conduct extensive literature search for its biological function.

### **2.1 Careful Choosing of the GWAS and eQTL Datasets**

Both GWAS and eQTL datasets are required to run the program. In order for the program to reach maximum accuracy, the complete result of the GWAS (i.e. the p-values of every single SNP tested in the entire genome) is required. Despite numerous GWAS studies on HIV control, I only find one study that publishes its complete GWAS result (Fellay et al., 2007). In addition, I am fortunate to obtain the complete result of the study by The International HIV Controllers Study et al. from my colleague Dr. Xin He. In my analysis, I choose to use the latter GWAS result because it has greater population size (974 controllers and 2648 progressors) and has data on ~ 1,300,000 SNPs. All of the human subjects in this study are treatment naive. The controllers are defined by three consecutive measurement of plasma viral load with < 2000 RNA copies/ml. The median

viral load of controllers is 241 RNA copies/ml, and that of the progressors is 61,698 RNA copies/ml.

An ideal eQTL dataset for this analysis would have the same population structure as the GWAS and is derived from a disease-related tissue. Among the six eQTL datasets that I have access to (Dixon et al., 2007; Myers et al., 2007; Duan et al., 2008; Schadt et al., 2008; Webster et al., 2009; Zeller et al., 2010), the study conducted by Duan et al., which contains 9830 genes, is the best for two reasons. Firstly, this study is one of the two studies (Dixon et al., 2007; Duan et al., 2008) performed on the lymphoblastoid cell line. It is the immortalized precursor of lymphocytes, which is mainly constituted of T lymphocytes and B lymphocytes. T lymphocyte is one of the major infection sites of HIV. Therefore, lymphoblastoid cell line is a tissue type that matches well with HIV infection. Secondly, Among the studies by Duan et al. and Dixon et al., the latter uses individuals from families associated with asthma. This might cause the genetic network to be rewired and hence a biased result.

## **2.2 Producing the List of Target Genes**

The parameters of the program are adjusted accordingly to the size of the GWAS and eQTL datasets. Table 1 (page 38) shows the specific parameters I used to produce the results.

The top 10 genes, starting from the highest likelihood, returned by the program are: CCHCR1, FAM20B, MOBKL2B (MOB3B), HLA-C, ATP5O, DPM3, HLA-B, MICA, MICB, SENP8. Among them, (Table 2 page 38) five are located within HLA



region. Amidst the other half, FAM20B, ATP5O, DPM3, and SENP8 all have the most significant GWAS SNP in HLA (Table 3 page 39). Though this result confirms the critical role of HLA genes in HIV infection, it is not particularly interesting because it does not reveal any gene that has both a specific biological function and a close connection to HIV infection that bodes well for a great potential target.

HLA related genes have been known to be ubiquitously involved in various immune diseases. The significant GWAS signals almost exclusively reside within HLA. When the HLA region, as well as CCR5, a known player in HIV control, is excised, the Q-Q plot of the p-values overlaps with null (International HIV Controllers Study et al., 2010). Given the dense signal, an immediate question is to distinguish real signals versus noises. Consider a single strong SNP in an intergenic region. There might be two genes in close proximity. Even more genes might be associated with it on the expression level. Therefore, the high density of GWAS signals in HLA makes it difficult to pinpoint the specific gene involved in HIV control. This could lead to the discovery of false causal genes simply because they are next to a strong signal. If this is the case, the false genes are likely to have no significant SNP support except for a few in the HLA region on Chromosome 6. FAM20B, ATP5O, DPM3, and SENP8 all satisfy this criterion.

### **2.3 Brief Summary of the Top Ten Genes Produced in 2.2**

HLA-C / HLA-B. Both of HLA-C and HLA-B are Class I HLA genes and have critical roles in antigen presentation. Many studies have confirmed their association with the control of HIV as discussed in Chapter One.

MICA / MICB. MICA and MICB are the human MHC (HLA) I chain-related (MIC) proteins. They do not belong to class I or II of HLA genes. These MIC proteins are normally expressed on the surface of gastrointestinal epithelium cells. They are ligands of an activating receptor NKG2D on the surface of natural killer (NK) cells. Different from HLA proteins, the MIC proteins do not present antigens. Their major function is to respond to the stress condition of cells by binding to NKG2D, which in turn activates the NK cells (Steinle et al., 2001; Stephens, 2001). In the GWAS done by the International HIV Controllers Study et al., MICA is suggested to be a potential target. A study in 2006 showed that the Nef protein of HIV down-regulates the cell-surface expression of MICA, hindering the activation of NK cells (Cerboni et al., 2007).

CCHCR1. CCHCR1 is closely associated with psoriasis, a skin disease due to an autoimmune disorder (Bowcock and Krueger, 2005). It is not uncommon for HIV infected patients to develop psoriasis (Morar et al., 2010). Psoriasis is thought of as a symptom of HIV by some patients. Interestingly enough, study has suggested that genetic variants which predispose patients to psoriasis might increase resistance against HIV (Chen et al., 2012). In the GWAS literature published by the International HIV Controllers Study et al., the authors point out that PSORS1C3, another gene implicated in psoriasis, can be a potential target associated with HIV control (International HIV Controllers Study et al., 2010). These information seem to suggest a loose connection between CCHCR1 and the control of HIV. However, no evidence could be found to rule out the possibility that CCHCR1 is inferred by the program simply because it is next to a strong SNP signal, given that the only strong signals are from within the HLA region.

In the rest of the top ten genes, FAM20B is a kinase that phosphorylates xylose at the 2-O position (Koike et al., 2009). MOBKL2B stands for MOB kinase activator 3B that has a role in MOB kinase regulation. ATP5O is a subunit of the ATP synthase. DPM3's function is to stabilize the dolichol-phosphate-mannose synthase complex. No literature evidence could be found to support the association between these genes and the control of HIV.

#### **2.4 Pruning the GWAS Dataset and Producing a Second List of Target Genes**

In order to shield the interference from the strong signals and to uncover any target gene outside of the HLA cluster, I acquire the chromosome coordinate for every SNP of the GWA study from the dbSNP database, and remove any SNP that falls within 28,000,000 and 34,000,000 on Chromosome 6.

I re-run the program using the same settings. The top genes, starting from the highest likelihood, are MOBKL2B, PRKCH, ANKDD1A, NAPRT1, TMPRSS3, CD151, LBX2, MED28, LAD1, and SEPN1 (Table 4, page 40). None of the top ten genes are from Chromosome 6. Out of the ten genes, six are entirely supported by *trans*-SNPs. The other four are ANKDD1A, NAPRT1, CD151, and LAD1. Interestingly, the signals of the *cis*-SNPs are weak. It would have been impossible to infer these genes based on their *cis*-SNPs alone due to their low GWAS significance (Table 5, page 41). For instance, rs5030780 and rs1108991 are the only two *cis*-SNPs for CD151. The chromosome coordinates of them are 838110 and 1537517 on Chromosome 11 (Genome Build: 37.3, Assembly: GRCh37.p5), with rs5030780 falling right within CD151. The GWAS p-

values are 0.167773 and 0.00633979 respectively, and are not the strongest signals for CD151 (rs518063 has a GWAS p-value of 0.00183253). It is a great manifestation of our philosophy that the supporting signals can be spread and divided into multiple SNPs, and that the significance adds up within the collection of SNPs. Literature search reveals that CD151 is involved in inhibiting HIV entry, which I will discuss in detail in the next section.

## **2.5 Brief Summary of the Top Ten Genes Produced in 2.4**

MED28. All SNPs for MED28 are *trans*-SNPs. The most strong GWAS signal is rs489105, with a GWAS p-value of 0.00106469. It is a perfect example of a gene that could easily have been missed by only looking at top GWAS signals and search for genes in *cis*. MED28 is a member of the multi-subunit mediator complex. The complex plays an active role in regulating the transcription activity by interacting with RNA pol II and the general transcription factors (Kornberg, 2005; Taatjes, 2010). Recently, the mediator complex has been implicated by many studies to be HIV dependency factors (Brass et al., 2008; Konig et al., 2008; Zhou et al., 2008). In Brass et al.'s study, the silencing of MED28 shows robust inhibitory effect on the early stage of HIV infection. Moreover, MED28 seems to be specific to HIV since the silencing of MED28 does not affect Murine Leukemia Virus (MLV) infection. Another two subunits are also worth mentioning. MED6 and MED7 have shown up in three studies on HIV host factors (Brass et al., 2008; Konig et al., 2008; Zhou et al., 2008). Brass et al. and Zhou et al. suggest that MED28, along with MED6 and MED7, directly mediate the interaction between

transcription factors and RNA pol II. The depletion of these mediator subunits inhibits HIV infection by affecting transcription activity. Konig et al.'s study seems to suggest MED7's role in reverse transcription. In conclusion, MED28 is a host factor with decent amount of literature evidence and has turned out to be one of the top candidate genes on my list. It is worth investigating with more follow-up studies.

CD151. CD151 is a member of the tetraspanins, a family of small transmembrane proteins that regulate cell migration, fusion and signaling events. Particularly, CD151 interacts with integrins and modulate cell fusion, integrin-dependent cell morphology and cell migration (Hemler, 2005). CD151's role in cell fusion could be extended to the inhibition of HIV entry. Several studies have implicated CD151's role in HIV infection, though the results have not been consistent (Pelchen-Matthews et al., 2003; Gordon-Alonso et al., 2006; Ho et al., 2006). Pelchen-Matthews et al. precipitated small amount of HIV particles with anti-CD151 antibody, suggesting that some virions incorporate this molecule into their envelopes. An siRNA study by Gordon-Alonso et al. showed that the interference of CD151 partially inhibited membrane fusion, but they did not detect an inhibitory effect on HIV entry. Ho et al. fused the extracellular domain of CD151 with glutathione S-transferase (GST) into a soluble chimeric protein. The study demonstrated a completely inhibited HIV infection of macrophages by the chimeric protein, suggesting that the infection of macrophage might involve some interaction between HIV and CD151.

SEPN1. SEPN1 is a selenoprotein that incorporates selenium. Although there is no direct evidence that SEPN1 is related to HIV control, selenium and selenoprotein have

been suggested to delay the progression of HIV infection. Some studies suggest that the oxidative stress could promote viral replication or activation from the proviral state. Thus, it is assumed that selenoprotein, an antioxidant, may inhibit the progression of HIV (Schwarz, 1996; Moghadaszadeh and Beggs, 2006).

PRKCH. PRKCH stands for protein kinase C eta (PKC $\eta$ ), an isoform of PKC. There has been no report on PRKCH being an HIV host factor or involved in HIV control, but some studies claim that a motif in the HIV protein gp41 demonstrates immunosuppressive capability by inhibiting the PKC-dependent T cell activation (Ruegg and Strand, 1991; Chen et al., 1995).

MOBKL2B has occurred in the top ten genes before the pruning. This MOB kinase activator 3B has not been shown to be related to HIV. The literature of the other half of the top ten genes cannot be found.

In addition to the top ten genes, A4GALT is also a notably interesting gene. A4GALT encodes for alpha 1,4-galactosyltransferase. It is the synthase of Gb<sub>3</sub>, a type of glycosphingolipid. Although it ranks the 11<sup>th</sup> on the list, with a bayesian factor of 164 (5.1039 on the natural log scale), it is observed that individuals with accumulated expression of Gb<sub>3</sub> due to a mutation demonstrate increased resistance to HIV infection (Lund et al., 2009; Branch, 2010). Moreover, the introduction of exogenous Gb<sub>3</sub> in Gb<sub>3</sub>-devoid Jurkat cells increases the resistance against HIV infection (Lund et al., 2009).

## 2.6 Overlap Studies

I examine the overlap between the interactome study (Jager et al., 2011), the RNAi screen results (Brass et al., 2008; Konig et al., 2008; Zhou et al., 2008; Yeung et al., 2009) and my result. I evaluate the significance of the overlap with hypergeometric test, and assess whether the overlaps concentrate more on the top of my result list. At the end, I discuss some overlap genes with interesting literature evidence

### 2.6.1 Significance of Overlap

Both of HLA-B and HLA-C have confirmed involvement in HIV control. After pruning, the rank of HLA-B on my result list is 872 and that of HLA-C is 210. I thus define a list of the top 872 genes and assess its enrichment of genes identified in interactome data and RNAi screen. 438 genes were identified in the interactome study and 1035 genes in the pooled RNAi studies. All genes in the three datasets are represented in HUGO gene names. Those that do not have a HUGO gene name have been discarded.

33 genes are in common between my result and the interactome data:

COX5A	FOXC1	SUMF2	ANAPC2	QARS	XP05	G3BP1	MRPS35	CSDE1	HLA-C	RER1
CUL4A	COPS8	TMEM43	COPS6	EXOC4	TMED4	CCDC47	WDR61	COPS4	DAGLB	RNF7
VAPA	ACBD3	EIF3M	GPS2	PRKCSH	ATL3	TBC1D15	MRPL11	DDX49	HARS2	HLA-B.

Assume there are 22,000 genes in the human genome, and that all the genes (438 in total) from the interactome study are true target genes, the significance of the hypergeometric test for the overlap is  $p = 0.000161$ . Similarly, 74 genes are in common between my result and the RNAi screen data:

PRKCH	MED28	A4GALT	SSU72	NRAS	ZNF436	CYBB	TOMM70A	PRCC	PSMA2	CFLAR
IFRD2	ANAPC2	CDC16	MAP3K14	RPP40	CDH22	SHCBP1	MRPL44	NUDT4	MAP3K7	PSMA5
BAHD1	GCLM	ETHE1	ETF1	USP39	PRMT3	DCBLD1	BTBD1	MRPL24	MRE11A	ATG5
UBE2L3	SPAST	GBAS	CEP68	TPR	SAMM50	ADAM10	FAM118A	SSR1	OCIAD1	NUMBL
KHDRBS1	SPEN	USP20	PIGY	RAD21	BTN3A3	ANKFY1	EIF2B5	BCR	INTS12	CLNS1A
PKN2	STARD3NL	IL1A	MED17	AMDHD2	PIP5K1C	TOM1	KLF5	TBC1D10A	PIK3CB	NEU1
NUP50	CSPP1	WNK1	DDX49	CHAF1A	LIN7C	PHF12	ACADSB.			

The significance of the overlap is  $3.33 \times 10^{-7}$ . A4GALT, mentioned above, is also within this intersection. Both p-values are sufficiently low to rule out the possibility of random overlap, indicating that genes in my result are more likely to also occur as result in other studies.

## 2.6.2 Rank-sum Test of the Overlap Genes

Another question is whether this computational method can effectively rank the target genes. If it does, the top genes should have a higher likelihood of being a target. Figure 2 (page 46) shows that for both datasets, the frequency of the rank does not show a decreasing trend, indicating that the top of the list is no more enriched.

I assess this aspect by Mann-Whitney rank-sum test. Two vectors are required to perform the test. One is the rank of the overlap genes in my result. I construct the other vector with the exact same range and the number of elements as the rank, but the elements are uniformly distributed. For instance, the rank of the overlap between my result and the interactome data is: 42, 74, 75, 114, 131, ..., 9603, 9641, 9680, 9729, 9821. This is tested against the constructed vector: 42, 68, 95, 121, 148, ..., 9714, 9741, 9767, 9794, 9821. Neither the overlap with the interactome (p-value = 0.5211) nor that of the RNAi screen (p-value = 0.7515) appears to be shifted upwards the top on the result list.



### **2.6.3 Overlap Between the Interactome, the RNAi, and My Result**

There are 55 overlapping genes between the interactome study and the RNAi screens. In these 55 genes, two also appear in my result: ANAPC2 and DDX49, the rank of which are 114 and 816 respectively.

ANAPC2 encodes for anaphase promoting complex subunit 2. This gene was firstly proposed by Konig et al. in their RNAi study due to the reduced viral DNA integration accompanying the knock down of this gene. Konig et al. noted that the cell's entering into mitosis might be associated with HIV viral DNA integration. Since ANAPC2 promotes the entrance in anaphase, it is reasonable to hypothesize a relationship between ANAPC2 and HIV infection (Konig et al., 2008). A later study following up on ANAPC2 confirms that knocking down ANAPC2 decreases viral integration, and that ANAPC2 is likely to be HIV specific (Ocwieja et al., 2011). So far there are few studies on the association between ANAPC2 and HIV infection. Yet this gene occurs in all three genome-wide gene screens, which I think promises a great potential to uncover a new aspect of this virus.

The other gene is DDX49, which encodes for one of the 42 identified DEAD-box RNA helicases. Proteins in the DDX family share 9 motifs, none of which is missing in DDX49 (Umate et al., 2011). There has been no specific studies on the relationship between DDX49 and HIV, but two other members of the DDX family, DDX1 and DDX3, have been implicated as co-factors of HIV Rev. Rev is an HIV accessory protein that targets any viral mRNA containing the Rev response element (RRE), and helps to export them from the nucleus to the cytoplasm for translation or packaging (Pollard and Malim,

1998). It is shown that DDX1 is associated with nuclear / cytoplasmic distribution of Rev in astrocytes, the most abundant cells in the brain (Fang et al., 2005). Coimmunoprecipitation studies suggest that in Jurkat cells and MT4 cells, DDX3 forms complex with REV and CRM1 and is required in the nuclear transportation performed by the viral protein REV (Yedavalli et al., 2004). This is later supported by an RNAi study in HEK 293T cells (Ishaq et al., 2008).

## **Chapter Three: Discussion and Self Assessment**

### **3.1 New Ways to Interpret GWAS May Lead to the Discovery of New Genes**

It is worth noting that our method is not limited to eQTL, but is able to leverage information in essentially any type of genomic data, such as various DNA methylation, acetylation and protein interaction data. In fact, the idea of combining different types of genomic data has existed for years, and many attempts have been made, including meta-analysis of multiple studies and conceiving new algorithms to infer genetic interaction from GWAS and protein complex data, the latter of which is particularly a forerunner of transforming the interpretation of GWAS (Bushman et al., 2009; Hannum et al., 2009). With our method, we hope to resolve the ambiguities in GWAS and to ultimately enable researchers to predict causal genes.

Comparing to RNAi screens, immunoprecipitation, or mass spectroscopy, GWAS has been less successful in identifying HIV related host genes. The old way to identify causal genes has failed because the signals outside HLA are generally weak, dwarfed by the HLA complex with extremely noisy and strong signals (International HIV Controllers Study et al., 2010). If one naively sets a threshold and picks some top SNPs to investigate, he or she is sure to only focus on the HLA and overlook the information that is not concentrated in certain regions but spread along the entire genome. Moreover, a highly signal-rich region does not make gene identification any easier. The noises decrease the resolution within HLA and make it difficult to pinpoint any causal gene. Essentially all GWA studies with concentrated strong signals are faced with this challenge.

Our new computational method partially solves these problems by combining GWAS with eQTL. The extra information in eQTL dataset lends the ability to integrate the information spread along the genome. As a result, genes outside of HLA complex (FAM20B, MOBKL2B, ATP5O, DPM3, SENP8) start to appear as top candidates. Pruning the GWAS data removes the strongest SNP signals and completely removes the HLA genes from the top of the result, enabling other interesting candidates to be revealed. In my result, both of MED28 and A4GALT are solely supported by weak *trans*-SNPs. They have never been identified in any GWA studies despite solid experimental evidences. In this study, the two genes are computational identified for the first time.

### **3.2 Multiple GWAS Datasets and eQTL Datasets Will Improve the Accuracy**

This study is solely based on one pair of GWAS-eQTL datasets, which limits the accuracy. Multiple GWAS and eQTL datasets are highly desired. There are several benefits. Firstly, multiple GWAS datasets enables one to perform meta-analysis, greatly improving the statistical power. Secondly, single GWAS is prone to false positives and false negatives. Using multiple GWAS may eliminate bias and produce more robust results. Thirdly, multiple GWA studies make it possible to do population stratification while maintaining a relatively large sample size. Unfortunately, most GWA studies do not publish their complete results, leading to a very limited selection range for my study. I believe that the value of GWAS is far beyond the top SNP signals. Sharing the complete results for each GWA study grant researchers with access to much richer information and fuller view of human genetics.

Access to various eQTL studies is also important. Ideally, the tissue type of the eQTL study should match the major phenotype. For example, an ideal eQTL employed in this study would have been conducted with CD4<sup>+</sup> T cells, or macrophages. In different tissues, similar phenotypes may be caused by different genes. The DEAD-box RNA helicases are perfect examples. Both of DDX1 and DDX3 are proposed to be HIV co-factors, but they function in brain tissue and T cells respectively. To push this study to even better accuracy, the GWAS and eQTL should be performed on individuals in the same developmental stage, as gene expression changes not only spatially but also temporally.

### **3.3 Critiques About This Work and Future Direction**

Careful self-assessment revealed several weaknesses, according to which I suggest a few possible future directions.

Firstly, although this study has brought up some genes worth attention, it fails to identify some of the well-recognized HIV co-factors. CD4 is required for HIV entry. It is not present in Duan et al.'s study, and is thus missed by my study. Similarly, I cannot find CCR5 in the eQTL data. CXCR4 is identified in my study, but only ranks the 1735<sup>th</sup> out of the 9830 genes. APOBEC3G, a non-HIV-specific antiviral factor that is normally suppressed by Vif in HIV-infected T cells, only ranks the 5896<sup>th</sup> (Stopak et al., 2003). Table 6 (page 44) shows the SNP support for the two genes. They both have relatively fewer SNP supports. Since the bayesian factor (BF) of a gene is the sum of the BFs of all its SNP supports, the BF of a gene is very sensitive to the number of its supporting SNPs.

It is likely that the gene's rank is influenced by artifacts in eQTL studies, leading to reduced number of SNP supports. It is also likely that the expression of the genes is not directly associated with SNPs, but is regulated by epigenetic factors. Therefore, one of the future direction could possibly be to incorporate epigenetic markers, such as DNA methylation and various histone code, into this algorithm.

Secondly, with a high mutation rate, there exist many strains of HIV. Using study results from different HIV strains could affect host factor detection. For instance, there are R5 and X4 HIVs, which utilizes CCR5 and CXCR4 for cell entry respectively. The HIV that uses CCR5 as co-receptors usually infects macrophages (Wu et al., 1997). There is also evidence that blocking CXCR4 reduces the infection of T cell lines (Murakami et al., 1997). This might explain the absence of CCR5 and the presence of CXCR4 in my result, as lymphoblastoid cell line are immortalized precursors of T cells. On the one hand, these strain information of HIV will not be available in GWA studies since the virus in the human body constitutes of multiple strains. On the other hand, some genes, like CCR5 and CXCR4, interact with specific HIV strains. This conflict is a big challenge to identifying genes associated with HIV control.

Thirdly, pruning the GWAS data seems to be effective in focusing on novel non-HLA complex and unveiling some interesting genes. Though my work shows a significant overlap between previous studies and my result after pruning, demonstrating non-significant overlap before pruning will be a solid proof that pruning is an effective method. Another future direction is to assess which genes have moved up or down the list after the pruning, and how much have the genes moved. This study will contribute to the

systematic understanding of this method, which may be used in similar cases in the future.

## Bibliography

- Altshuler, D., Daly, M.J., and Lander, E.S. (2008). Genetic mapping in human disease. *Science* 322, 881-888.
- Baba, M., Nishimura, O., Kanzaki, N., Okamoto, M., Sawada, H., Iizawa, Y., Shiraishi, M., Aramaki, Y., Okonogi, K., Ogawa, Y., Meguro, K., and Fujino, M. (1999). A small-molecule, nonpeptide CCR5 antagonist with highly potent and selective anti-HIV-1 activity. *Proc. Natl. Acad. Sci. U. S. A.* 96, 5698-5703.
- Barre-Sinoussi, F., Chermann, J.C., Rey, F., Nugeyre, M.T., Chamaret, S., Gruest, J., Dauguet, C., Axler-Blin, C., Vezinet-Brun, F., Rouzioux, C., Rozenbaum, W., and Montagnier, L. (1983). Isolation of a T-lymphotropic retrovirus from a patient at risk for acquired immune deficiency syndrome (AIDS). *Science* 220, 868-871.
- Bleul, C.C., Wu, L., Hoxie, J.A., Springer, T.A., and Mackay, C.R. (1997). The HIV coreceptors CXCR4 and CCR5 are differentially expressed and regulated on human T lymphocytes. *Proc. Natl. Acad. Sci. U. S. A.* 94, 1925-1930.
- Bowcock, A.M., and Krueger, J.G. (2005). Getting under the skin: the immunogenetics of psoriasis. *Nat. Rev. Immunol.* 5, 699-711.
- Branch, D.R. (2010). Blood groups and susceptibility to virus infection: new developments. *Curr. Opin. Hematol.* 17, 558-564.
- Brass, A.L., Dykxhoorn, D.M., Benita, Y., Yan, N., Engelman, A., Xavier, R.J., Lieberman, J., and Elledge, S.J. (2008). Identification of host proteins required for HIV infection through a functional genomic screen. *Science* 319, 921-926.
- Bushman, F.D., Malani, N., Fernandes, J., D'Orso, I., Cagney, G., Diamond, T.L., Zhou, H., Hazuda, D.J., Espeseth, A.S., Konig, R., *et al.* (2009). Host cell factors in HIV replication: meta-analysis of genome-wide studies. *PLoS Pathog.* 5, e1000437.
- Carrington, M., and O'Brien, S.J. (2003). The influence of HLA genotype on AIDS. *Annu. Rev. Med.* 54, 535-551.
- Cerboni, C., Neri, F., Casartelli, N., Zingoni, A., Cosman, D., Rossi, P., Santoni, A., and Doria, M. (2007). Human immunodeficiency virus 1 Nef protein downmodulates the ligands of the activating receptor NKG2D and inhibits natural killer cell-mediated cytotoxicity. *J. Gen. Virol.* 88, 242-250.
- Chan, D.C., Fass, D., Berger, J.M., and Kim, P.S. (1997). Core structure of gp41 from the HIV envelope glycoprotein. *Cell* 89, 263-273.



- Chen, H., Hayashi, G., Lai, O.Y., Dilthey, A., Kuebler, P.J., Wong, T.V., Martin, M.P., Fernandez Vina, M.A., McVean, G., Wabl, M., *et al.* (2012). Psoriasis patients are enriched for genetic variants that protect against HIV-1 disease. *PLoS Genet.* 8, e1002514.
- Chen, Y.H., Christiansen, A., and Dierich, M.P. (1995). HIV-1 gp41 selectively inhibits spontaneous cell proliferation of human cell lines and mitogen- and recall antigen-induced lymphocyte proliferation. *Immunol. Lett.* 48, 39-44.
- Choe, H., Farzan, M., Sun, Y., Sullivan, N., Rollins, B., Ponath, P.D., Wu, L., Mackay, C.R., LaRosa, G., Newman, W., *et al.* (1996). The beta-chemokine receptors CCR3 and CCR5 facilitate infection by primary HIV-1 isolates. *Cell* 85, 1135-1148.
- Chun, T.W., and Fauci, A.S. (1999). Latent reservoirs of HIV: obstacles to the eradication of virus. *Proc. Natl. Acad. Sci. U. S. A.* 96, 10958-10961.
- Cookson, W., Liang, L., Abecasis, G., Moffatt, M., and Lathrop, M. (2009). Mapping complex disease traits with global gene expression. *Nat. Rev. Genet.* 10, 184-194.
- Cooper, D.A., Gold, J., Maclean, P., Donovan, B., Finlayson, R., Barnes, T.G., Michelmore, H.M., Brooke, P., and Penny, R. (1985). Acute AIDS retrovirus infection. Definition of a clinical illness associated with seroconversion. *Lancet* 1, 537-540.
- Dean, M., Carrington, M., Winkler, C., Huttley, G.A., Smith, M.W., Allikmets, R., Goedert, J.J., Buchbinder, S.P., Vittinghoff, E., Gomperts, E., *et al.* (1996). Genetic restriction of HIV-1 infection and progression to AIDS by a deletion allele of the CKR5 structural gene. Hemophilia Growth and Development Study, Multicenter AIDS Cohort Study, Multicenter Hemophilia Cohort Study, San Francisco City Cohort, ALIVE Study. *Science* 273, 1856-1862.
- Deeks, S.G., and Walker, B.D. (2007). Human immunodeficiency virus controllers: mechanisms of durable virus control in the absence of antiretroviral therapy. *Immunity* 27, 406-416.
- Dixon, A.L., Liang, L., Moffatt, M.F., Chen, W., Heath, S., Wong, K.C., Taylor, J., Burnett, E., Gut, I., Farrall, M., *et al.* (2007). A genome-wide association study of global gene expression. *Nat. Genet.* 39, 1202-1207.
- Duan, S., Huang, R.S., Zhang, W., Bleibel, W.K., Roe, C.A., Clark, T.A., Chen, T.X., Schweitzer, A.C., Blume, J.E., Cox, N.J., and Dolan, M.E. (2008). Genetic architecture of transcript-level variation in humans. *Am. J. Hum. Genet.* 82, 1101-1113.
- Duerr, R.H., Taylor, K.D., Brant, S.R., Rioux, J.D., Silverberg, M.S., Daly, M.J., Steinhart, A.H., Abraham, C., Regueiro, M., Griffiths, A., *et al.* (2006). A genome-wide

association study identifies IL23R as an inflammatory bowel disease gene. *Science* 314, 1461-1463.

Embretson, J., Zupancic, M., Ribas, J.L., Burke, A., Racz, P., Tenner-Racz, K., and Haase, A.T. (1993). Massive covert infection of helper T lymphocytes and macrophages by HIV during the incubation period of AIDS. *Nature* 362, 359-362.

Emilsson, V., Thorleifsson, G., Zhang, B., Leonardson, A.S., Zink, F., Zhu, J., Carlson, S., Helgason, A., Walters, G.B., Gunnarsdottir, S., *et al.* (2008). Genetics of gene expression and its effect on disease. *Nature* 452, 423-428.

Fang, J., Acheampong, E., Dave, R., Wang, F., Mukhtar, M., and Pomerantz, R.J. (2005). The RNA helicase DDX1 is involved in restricted HIV-1 Rev function in human astrocytes. *Virology* 336, 299-307.

Fauci, A.S. (1996). Host factors and the pathogenesis of HIV-induced disease. *Nature* 384, 529-534.

Fellay, J., Ge, D., Shianna, K.V., Colombo, S., Ledergerber, B., Cirulli, E.T., Urban, T.J., Zhang, K., Gumbs, C.E., Smith, J.P., *et al.* (2009). Common genetic variation and the control of HIV-1 in humans. *PLoS Genet.* 5, e1000791.

Fellay, J., Shianna, K.V., Ge, D., Colombo, S., Ledergerber, B., Weale, M., Zhang, K., Gumbs, C., Castagna, A., Cossarizza, A., *et al.* (2007). A whole-genome association study of major determinants for host control of HIV-1. *Science* 317, 944-947.

Gallo, R.C., Sarin, P.S., Gelmann, E.P., Robert-Guroff, M., Richardson, E., Kalyanaraman, V.S., Mann, D., Sidhu, G.D., Stahl, R.E., Zolla-Pazner, S., Leibowitch, J., and Popovic, M. (1983). Isolation of human T-cell leukemia virus in acquired immune deficiency syndrome (AIDS). *Science* 220, 865-867.

Gallo, S.A., Finnegan, C.M., Viard, M., Raviv, Y., Dimitrov, A., Rawat, S.S., Puri, A., Durell, S., and Blumenthal, R. (2003). The HIV Env-mediated fusion reaction. *Biochim. Biophys. Acta* 1614, 36-50.

Ganser-Pornillos, B.K., Yeager, M., and Sundquist, W.I. (2008). The structural biology of HIV assembly. *Curr. Opin. Struct. Biol.* 18, 203-217.

Gao, X., Bashirova, A., Iversen, A.K., Phair, J., Goedert, J.J., Buchbinder, S., Hoots, K., Vlahov, D., Altfeld, M., O'Brien, S.J., and Carrington, M. (2005). AIDS restriction HLA allotypes target distinct intervals of HIV-1 pathogenesis. *Nat. Med.* 11, 1290-1292.

Gordon-Alonso, M., Yanez-Mo, M., Barreiro, O., Alvarez, S., Munoz-Fernandez, M.A., Valenzuela-Fernandez, A., and Sanchez-Madrid, F. (2006). Tetraspanins CD9 and CD81 modulate HIV-1-induced membrane fusion. *J. Immunol.* 177, 5129-5137.

Gottlieb, M.S., Schroff, R., Schanker, H.M., Weisman, J.D., Fan, P.T., Wolf, R.A., and Saxon, A. (1981). Pneumocystis carinii pneumonia and mucosal candidiasis in previously healthy homosexual men: evidence of a new acquired cellular immunodeficiency. *N. Engl. J. Med.* *305*, 1425-1431.

Goudsmit, J., Bogaards, J.A., Jurriaans, S., Schuitemaker, H., Lange, J.M., Coutinho, R.A., and Weverling, G.J. (2002). Naturally HIV-1 seroconverters with lowest viral load have best prognosis, but in time lose control of viraemia. *AIDS* *16*, 791-793.

Hannum, G., Srivas, R., Guenole, A., van Attikum, H., Krogan, N.J., Karp, R.M., and Ideker, T. (2009). Genome-wide association data reveal a global map of genetic interactions among protein complexes. *PLoS Genet.* *5*, e1000782.

Hardy, J., and Singleton, A. (2009). Genomewide association studies and human disease. *N. Engl. J. Med.* *360*, 1759-1768.

Haynes, B.F., Pantaleo, G., and Fauci, A.S. (1996). Toward an understanding of the correlates of protective immunity to HIV infection. *Science* *271*, 324-328.

Hemler, M.E. (2005). Tetraspanin functions and associated microdomains. *Nat. Rev. Mol. Cell Biol.* *6*, 801-811.

Hill, A.V., Allsopp, C.E., Kwiatkowski, D., Anstey, N.M., Twumasi, P., Rowe, P.A., Bennett, S., Brewster, D., McMichael, A.J., and Greenwood, B.M. (1991). Common west African HLA antigens are associated with protection from severe malaria. *Nature* *352*, 595-600.

Hill, A.V., Elvin, J., Willis, A.C., Aidoo, M., Allsopp, C.E., Gotch, F.M., Gao, X.M., Takiguchi, M., Greenwood, B.M., and Townsend, A.R. (1992). Molecular analysis of the association of HLA-B53 and resistance to severe malaria. *Nature* *360*, 434-439.

Ho, D.D. (1996). Viral counts count in HIV infection. *Science* *272*, 1124-1125.

Ho, S.H., Martin, F., Higginbottom, A., Partridge, L.J., Parthasarathy, V., Moseley, G.W., Lopez, P., Cheng-Mayer, C., and Monk, P.N. (2006). Recombinant extracellular domains of tetraspanin proteins are potent inhibitors of the infection of macrophages by human immunodeficiency virus type 1. *J. Virol.* *80*, 6487-6496.

Hsu, Y.H., Zillikens, M.C., Wilson, S.G., Farber, C.R., Demissie, S., Soranzo, N., Bianchi, E.N., Grundberg, E., Liang, L., Richards, J.B., *et al.* (2010). An integration of genome-wide association study and gene expression profiling to prioritize the discovery of novel susceptibility Loci for osteoporosis-related traits. *PLoS Genet.* *6*, e1000977.

Hubert, J.B., Burgard, M., Dussaix, E., Tamalet, C., Deveau, C., Le Chenadec, J., Chaix, M.L., Marchadier, E., Vilde, J.L., Delfraissy, J.F., Meyer, L., and Rouzioux. (2000).

Natural history of serum HIV-1 RNA levels in 330 patients with a known date of infection. The SEROCO Study Group. *AIDS* 14, 123-131.

International HIV Controllers Study, Pereyra, F., Jia, X., McLaren, P.J., Telenti, A., de Bakker, P.I., Walker, B.D., Ripke, S., Brumme, C.J., Pulit, S.L., *et al.* (2010). The major genetic determinants of HIV-1 control affect HLA class I peptide presentation. *Science* 330, 1551-1557.

Ishaq, M., Hu, J., Wu, X., Fu, Q., Yang, Y., Liu, Q., and Guo, D. (2008). Knockdown of cellular RNA helicase DDX3 by short hairpin RNAs suppresses HIV-1 viral replication without inducing apoptosis. *Mol. Biotechnol.* 39, 231-238.

Jager, S., Cimermancic, P., Gulbahce, N., Johnson, J.R., McGovern, K.E., Clarke, S.C., Shales, M., Mercenne, G., Pache, L., Li, K., *et al.* (2011). Global landscape of HIV-human protein complexes. *Nature* 481, 365-370.

Klein, J., and Sato, A. (2000). The HLA system. First of two parts. *N. Engl. J. Med.* 343, 702-709.

Klein, R.J., Zeiss, C., Chew, E.Y., Tsai, J.Y., Sackler, R.S., Haynes, C., Henning, A.K., SanGiovanni, J.P., Mane, S.M., Mayne, S.T., *et al.* (2005). Complement factor H polymorphism in age-related macular degeneration. *Science* 308, 385-389.

Koike, T., Izumikawa, T., Tamura, J., and Kitagawa, H. (2009). FAM20B is a kinase that phosphorylates xylose in the glycosaminoglycan-protein linkage region. *Biochem. J.* 421, 157-162.

Konig, R., Zhou, Y., Elleder, D., Diamond, T.L., Bonamy, G.M., Irelan, J.T., Chiang, C.Y., Tu, B.P., De Jesus, P.D., Lilley, C.E., *et al.* (2008). Global analysis of host-pathogen interactions that regulate early-stage HIV-1 replication. *Cell* 135, 49-60.

Kornberg, R.D. (2005). Mediator and the mechanism of transcriptional activation. *Trends Biochem. Sci.* 30, 235-239.

Levy, D., Ehret, G.B., Rice, K., Verwoert, G.C., Launer, L.J., Dehghan, A., Glazer, N.L., Morrison, A.C., Johnson, A.D., Aspelund, T., *et al.* (2009). Genome-wide association study of blood pressure and hypertension. *Nat. Genet.* 41, 677-687.

Lund, N., Olsson, M.L., Ramkumar, S., Sakac, D., Yahalom, V., Levene, C., Hellberg, A., Ma, X.Z., Binnington, B., Jung, D., Lingwood, C.A., and Branch, D.R. (2009). The human P(k) histo-blood group antigen provides protection against HIV-1 infection. *Blood* 113, 4980-4991.

Mackay, C.R. (1996). Chemokine receptors and T cell chemotaxis. *J. Exp. Med.* 184, 799-802.

- Mackay, T.F., Stone, E.A., and Ayroles, J.F. (2009). The genetics of quantitative traits: challenges and prospects. *Nat. Rev. Genet.* *10*, 565-577.
- Mellors, J.W., Rinaldo, C.R., Jr, Gupta, P., White, R.M., Todd, J.A., and Kingsley, L.A. (1996). Prognosis in HIV-1 infection predicted by the quantity of virus in plasma. *Science* *272*, 1167-1170.
- Migueles, S.A., Sabbaghian, M.S., Shupert, W.L., Bettinotti, M.P., Marincola, F.M., Martino, L., Hallahan, C.W., Selig, S.M., Schwartz, D., Sullivan, J., and Connors, M. (2000). HLA B\*5701 is highly associated with restriction of virus replication in a subgroup of HIV-infected long term nonprogressors. *Proc. Natl. Acad. Sci. U. S. A.* *97*, 2709-2714.
- Moghadaszadeh, B., and Beggs, A.H. (2006). Selenoproteins and their impact on human health through diverse physiological pathways. *Physiology (Bethesda)* *21*, 307-315.
- Morar, N., Willis-Owen, S.A., Maurer, T., and Bunker, C.B. (2010). HIV-associated psoriasis: pathogenesis, clinical features, and management. *Lancet Infect. Dis.* *10*, 470-478.
- Murakami, T., Nakajima, T., Koyanagi, Y., Tachibana, K., Fujii, N., Tamamura, H., Yoshida, N., Waki, M., Matsumoto, A., Yoshie, O., *et al.* (1997). A small molecule CXCR4 inhibitor that blocks T cell line-tropic HIV-1 infection. *J. Exp. Med.* *186*, 1389-1393.
- Myers, A.J., Gibbs, J.R., Webster, J.A., Rohrer, K., Zhao, A., Marlowe, L., Kaleem, M., Leung, D., Bryden, L., Nath, P., *et al.* (2007). A survey of genetic human cortical gene expression. *Nat. Genet.* *39*, 1494-1499.
- Navis, M., Schellens, I., van Baarle, D., Borghans, J., van Swieten, P., Miedema, F., Kootstra, N., and Schuitemaker, H. (2007). Viral replication capacity as a correlate of HLA B57/B5801-associated nonprogressive HIV-1 infection. *J. Immunol.* *179*, 3133-3143.
- Nica, A.C., Montgomery, S.B., Dimas, A.S., Stranger, B.E., Beazley, C., Barroso, I., and Dermitzakis, E.T. (2010). Candidate causal regulatory effects by integration of expression QTLs with complex trait genetic associations. *PLoS Genet.* *6*, e1000895.
- Nicolae, D.L., Gamazon, E., Zhang, W., Duan, S., Dolan, M.E., and Cox, N.J. (2010). Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. *PLoS Genet.* *6*, e1000888.
- O'Brien, S.J., Gao, X., and Carrington, M. (2001). HLA and AIDS: a cautionary tale. *Trends Mol. Med.* *7*, 379-381.

- Ocwieja, K.E., Brady, T.L., Ronen, K., Huegel, A., Roth, S.L., Schaller, T., James, L.C., Towers, G.J., Young, J.A., Chanda, S.K., *et al.* (2011). HIV integration targeting: a pathway involving Transportin-3 and the nuclear pore protein RanBP2. *PLoS Pathog.* 7, e1001313.
- Pantaleo, G., Graziosi, C., Demarest, J.F., Butini, L., Montroni, M., Fox, C.H., Orenstein, J.M., Kotler, D.P., and Fauci, A.S. (1993). HIV infection is active and progressive in lymphoid tissue during the clinically latent stage of disease. *Nature* 362, 355-358.
- Pelak, K., Goldstein, D.B., Walley, N.M., Fellay, J., Ge, D., Shianna, K.V., Gumbs, C., Gao, X., Maia, J.M., Cronin, K.D., *et al.* (2010). Host determinants of HIV-1 control in African Americans. *J. Infect. Dis.* 201, 1141-1149.
- Pelchen-Matthews, A., Kramer, B., and Marsh, M. (2003). Infectious HIV-1 assembles in late endosomes in primary macrophages. *J. Cell Biol.* 162, 443-455.
- Pollard, V.W., and Malim, M.H. (1998). The HIV-1 Rev protein. *Annu. Rev. Microbiol.* 52, 491-532.
- Qin, X.F., An, D.S., Chen, I.S., and Baltimore, D. (2003). Inhibiting HIV-1 infection in human T cells by lentiviral-mediated delivery of small interfering RNA against CCR5. *Proc. Natl. Acad. Sci. U. S. A.* 100, 183-188.
- Ruegg, C.L., and Strand, M. (1991). A synthetic peptide with sequence identity to the transmembrane protein GP41 of HIV-1 inhibits distinct lymphocyte activation pathways dependent on protein kinase C and intracellular calcium influx. *Cell. Immunol.* 137, 1-13.
- Saksena, N.K., Rodes, B., Wang, B., and Soriano, V. (2007). Elite HIV controllers: myth or reality? *AIDS. Rev.* 9, 195-207.
- Schadt, E.E., Molony, C., Chudin, E., Hao, K., Yang, X., Lum, P.Y., Kasarskis, A., Zhang, B., Wang, S., Suver, C., *et al.* (2008). Mapping the genetic architecture of gene expression in human liver. *PLoS Biol.* 6, e107.
- Schwarz, K.B. (1996). Oxidative stress during viral infection: a review. *Free Radic. Biol. Med.* 21, 641-649.
- Sille, F.C., Thomas, R., Smith, M.T., Conde, L., and Skibola, C.F. (2012). Post-GWAS functional characterization of susceptibility variants for chronic lymphocytic leukemia. *PLoS One* 7, e29632.
- Simmons, G., Clapham, P.R., Picard, L., Offord, R.E., Rosenkilde, M.M., Schwartz, T.W., Buser, R., Wells, T.N., and Proudfoot, A.E. (1997). Potent inhibition of HIV-1 infectivity in macrophages and lymphocytes by a novel CCR5 antagonist. *Science* 276, 276-279.

Speliotes, E.K., Willer, C.J., Berndt, S.I., Monda, K.L., Thorleifsson, G., Jackson, A.U., Allen, H.L., Lindgren, C.M., Luan, J., Magi, R., *et al.* (2010). Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nat. Genet.* *42*, 937-948.

Springer, T.A. (1994). Traffic signals for lymphocyte recirculation and leukocyte emigration: the multistep paradigm. *Cell* *76*, 301-314.

Steinberger, P., Andris-Widhopf, J., Buhler, B., Torbett, B.E., and Barbas, C.F.,3rd. (2000). Functional deletion of the CCR5 receptor by intracellular immunization produces cells that are refractory to CCR5-dependent HIV-1 infection and cell fusion. *Proc. Natl. Acad. Sci. U. S. A.* *97*, 805-810.

Steinle, A., Li, P., Morris, D.L., Groh, V., Lanier, L.L., Strong, R.K., and Spies, T. (2001). Interactions of human NKG2D with its ligands MICA, MICB, and homologs of the mouse RAE-1 protein family. *Immunogenetics* *53*, 279-287.

Stephens, H.A. (2001). MICA and MICB genes: can the enigma of their polymorphism be resolved? *Trends Immunol.* *22*, 378-385.

Stopak, K., de Noronha, C., Yonemoto, W., and Greene, W.C. (2003). HIV-1 Vif blocks the antiviral activity of APOBEC3G by impairing both its translation and intracellular stability. *Mol. Cell* *12*, 591-601.

Stranger, B.E., Stahl, E.A., and Raj, T. (2011). Progress and promise of genome-wide association studies for human complex trait genetics. *Genetics* *187*, 367-383.

Taatjes, D.J. (2010). The human Mediator complex: a versatile, genome-wide regulator of transcription. *Trends Biochem. Sci.* *35*, 315-322.

Tindall, B., and Cooper, D.A. (1991). Primary HIV infection: host responses and intervention strategies. *AIDS* *5*, 1-14.

Umate, P., Tuteja, N., and Tuteja, R. (2011). Genome-wide comprehensive analysis of human helicases. *Commun. Integr. Biol.* *4*, 118-137.

van Manen, D., Delaneau, O., Kootstra, N.A., Boeser-Nunnink, B.D., Limou, S., Bol, S.M., Burger, J.A., Zwinderman, A.H., Moerland, P.D., van 't Slot, R., *et al.* (2011). Genome-wide association scan in HIV-1-infected individuals identifying variants influencing disease course. *PLoS One* *6*, e22208.

Webster, J.A., Gibbs, J.R., Clarke, J., Ray, M., Zhang, W., Holmans, P., Rohrer, K., Zhao, A., Marlowe, L., Kaleem, M., *et al.* (2009). Genetic control of human brain transcript expression in Alzheimer disease. *Am. J. Hum. Genet.* *84*, 445-458.

Wu, L., Paxton, W.A., Kassam, N., Ruffing, N., Rottman, J.B., Sullivan, N., Choe, H., Sodroski, J., Newman, W., Koup, R.A., and Mackay, C.R. (1997). CCR5 levels and expression pattern correlate with infectability by macrophage-tropic HIV-1, in vitro. *J. Exp. Med.* *185*, 1681-1691.

Yedavalli, V.S., Neuveut, C., Chi, Y.H., Kleiman, L., and Jeang, K.T. (2004). Requirement of DDX3 DEAD box RNA helicase for HIV-1 Rev-RRE export function. *Cell* *119*, 381-392.

Yeung, M.L., Houzet, L., Yedavalli, V.S., and Jeang, K.T. (2009). A genome-wide short hairpin RNA screening of jurkat T-cells for human proteins contributing to productive HIV-1 replication. *J. Biol. Chem.* *284*, 19463-19473.

Zeller, T., Wild, P., Szymczak, S., Rotival, M., Schillert, A., Castagne, R., Maouche, S., Germain, M., Lackner, K., Rossmann, H., *et al.* (2010). Genetics and beyond--the transcriptome of human monocytes and disease susceptibility. *PLoS One* *5*, e10693.

Zhou, H., Xu, M., Huang, Q., Gates, A.T., Zhang, X.D., Castle, J.C., Stec, E., Ferrer, M., Strulovici, B., Hazuda, D.J., and Espeseth, A.S. (2008). Genome-scale RNAi screen for host factors required for HIV replication. *Cell. Host Microbe* *4*, 495-504.



## Appendix

**Table 1**

eQTL parameters		
N_expr	176	sample size of the eQTL data (need to change for any eQTL dataset)
cis_expr_prior	1.00E-03	the prior of a cis-eSNP
trans_expr_prior	5.00E-05	the prior of a trans-eSNP
sigma_a_expr	0.5	the effect size prior of an expression QTL
GWAS parameters		
is_pheno_binary	1	whether the phenotypic trait is binary or quantitative
N_pheno	3622	sample size of the GWAS data
phi	0.27	the proportion of cases in the GWAS
K	0.01	the disease prevalence
pheno_prior	1.00E-04	the prior of a phenotypic trait locus
sigma_a_pheno	0.2	the effect size prior of a phenotypic trait locus

**Table 2**

The chromosome coordinate information in this table is obtained based on genome build 37.3.

Gene	Description	log(BF)	Chromosome	Start (bp from pter)	End (bp from pter)
CCHCR1	coiled-coil alpha-helical rod protein 1	13.5958	6	31,110,216	31,126,015
FAM20B	family with sequence similarity 20, member B	9.40828	1	178,995,074	179,045,702
MOBKL2B	MOB kinase activator 3B	9.25255	9	27,325,207	27,529,850
HLA-C	major histocompatibility complex, class I, C	8.99442	6	31,236,526	31,239,913
ATP5O	ATP synthase, H <sup>+</sup> transporting, mitochondrial F1 complex, O	8.26566	21	35,275,757	35,288,158
DPM3	dolichyl-phosphate mannosyltransferase polypeptide 3	8.03168	1	155,112,367	155,112,996
HLA-B	major histocompatibility complex, class I, B	7.75934	6	31,321,649	31,324,989
MICA	MHC class I polypeptide-related sequence A	7.7352	6	31,367,561	31,383,090
MICB	MHC class I polypeptide-related sequence B	7.73257	6	31,465,855	31,478,901

SENP8	SUMO/sentrin specific peptidase family member 8	7.36874	15	72,406,599	72,433,311
-------	---	---------	----	------------	------------

**Table 3**

Due to space limitation, only SNPs with  $\log(\text{BF}) > 0.05$  are listed.

The chromosome coordinate information in this table is obtained based on genome build 36.3.

Gene	SNPs	Chromosome	Coordinate	cis / trans	$\log(\text{BF})$ of SNPs
<b>FAM20B <math>\log(\text{BF}) = 9.0104</math></b>					
FAM20B	rs300278	1	54807936	trans	0.059916
FAM20B	rs12075898	1	164097717	trans	0.0416094
FAM20B	rs6757018	2	199038919	trans	0.290194
FAM20B	rs163541	3	6815697	trans	0.0593867
FAM20B	rs6780762	3	24972516	trans	0.0467571
FAM20B	rs10804694	3	145620760	trans	0.0606078
FAM20B	rs1499807	3	178918446	trans	0.0863446
FAM20B	rs1011495	4	97159210	trans	0.307039
FAM20B	rs10032098	4	177379770	trans	0.394049
FAM20B	rs2169095	5	113811888	trans	0.0481513
FAM20B	rs1100580	6	22484054	trans	0.231732
FAM20B	rs3094208	6	31198651	trans	5.95951
FAM20B	rs9381530	6	47304549	trans	0.202839
FAM20B	rs6939322	6	114228177	trans	0.0468737
FAM20B	rs17061433	6	132922472	trans	0.0543157
FAM20B	rs7806365	7	2956535	trans	0.244419
FAM20B	rs6950340	7	12831783	trans	0.842979
FAM20B	rs11763159	7	57249189	trans	0.030444
FAM20B	rs7802743	7	81996518	trans	0.280295
FAM20B	rs1821892	9	6606648	trans	0.0941377
FAM20B	rs1940247	9	112562853	trans	0.0709472
FAM20B	rs129889	9	135532689	trans	0.670612
FAM20B	rs9299574	10	59089117	trans	0.580641
FAM20B	rs557309	11	60716712	trans	0.328257
FAM20B	rs3802893	11	84219051	trans	0.0947974
FAM20B	rs9572108	13	68671752	trans	0.0706718
FAM20B	rs1481420	14	84229253	trans	0.0544086
FAM20B	rs11071167	15	53192545	trans	3.05853
FAM20B	rs2671666	17	44891520	trans	0.110063

FAM20B	rs10502421	18	13454437	trans	0.0495132
FAM20B	rs6512158	19	16982533	trans	0.3009
FAM20B	rs10426205	19	21765001	trans	0.267686
FAM20B	rs3844453	19	59049576	trans	0.0540084
FAM20B	rs6048024	20	22256842	trans	0.0643483
FAM20B	rs735455	22	20743793	trans	0.066166
<b>ATP5O log(BF) = 7.95662</b>					
ATP5O	rs1264420	6	30683582	trans	7.17177
ATP5O	rs16897785	6	165719948	trans	0.754585
ATP5O	rs1616483	13	27291672	trans	0.524064
<b>DPM3 log(BF) = 7.44053</b>					
DPM3	rs3132488	6	31350674	trans	6.77716
DPM3	rs9938060	16	85022070	trans	0.63187
DPM3	rs1403528	17	44600805	trans	0.0310129
DPM3	rs200766	20	15561056	trans	0.110495
<b>SENP8 log(BF) = 7.08848</b>					
SENP8	rs13437088	6	31463098	trans	7.20186
SENP8	rs6939322	6	114228177	trans	0.0533025
SENP8	rs6470789	8	131084458	trans	0.0352814

**Table 4**

The chromosome coordinate information in this table is obtained based on genome build 37.3.

Gene	Description	log(BF)	Chromosome	Start(bp from pter)	End (bp from pter)
MOBKL2B	MOB kinase activator 3B	9.25255	9	27,325,207	27,529,850
PRKCH	Protein kinase C, eta	6.63506	14	61,788,515	62,017,698
ANKDD1A	ankyrin repeat and death domain containing 1A	6.62846	15	65,204,101	65,251,042
NAPRT1	Nicotinate phosphoribosyltransferase domain containing 1	5.9829	8	144,656,955	144,660,513
TMPRSS3	transmembrane protease, serine 3	5.53317	21	43,791,996	43,816,955
CD151	CD151 molecule (Raph blood group)	5.49685	11	832,952	838,835
LBX2	ladybird homeobox 2	5.39866	2	74,724,644	74,730,443
MED28	mediator complex subunit 28	5.38231	4	17,616,273	17,626,160
LAD1	ladinin 1	5.21392	1	201,349,966	201,368,669
SEPN1	selenoprotein N, 1	5.1874	1	26,126,667	26,144,713

**Table 5**

Due to space limitation, only SNPs with  $\log(\text{BF}) > 0.05$  are listed, except for rs5030780, which is specifically discussed in §2.4. All of the SNPs filtered out are *trans*-SNPs, except for rs6998917 (for NAPRT1 with a  $\log(\text{BF})$  of -0.32955).

The chromosome coordinate information in this table is obtained based on genome build 36.3.

Gene	SNPs	Chromosome	Coordinate	cis / trans	$\log(\text{BF})$ of SNPs
<b>MOBKL2B <math>\log(\text{BF}) = 9.25255</math></b>					
MOBKL2B	rs13024819	2	129615846	trans	7.01888
MOBKL2B	rs7079148	10	115320376	trans	0.0548581
MOBKL2B	rs713974	22	25540691	trans	2.03702
MOBKL2B	rs4821089	22	31448165	trans	0.057498
<b>PRKCH <math>\log(\text{BF}) = 6.63506</math></b>					
PRKCH	rs2801178	1	15080799	trans	0.108415
PRKCH	rs3806187	1	158017253	trans	0.362164
PRKCH	rs9821993	3	39264607	trans	0.163898
PRKCH	rs7650998	3	46582259	trans	0.580038
PRKCH	rs7666932	4	143478755	trans	0.0927929
PRKCH	rs11778620	8	3925428	trans	0.168607
PRKCH	rs1930144	10	55943558	trans	0.0945184
PRKCH	rs4943750	13	39736899	trans	1.59968
PRKCH	rs2296316	14	64589999	trans	1.64584
PRKCH	rs934537	15	53212148	trans	2.91533
<b>ANKDD1A <math>\log(\text{BF}) = 6.62846</math></b>					
ANKDD1A	rs6683133	1	44906148	trans	0.328271
ANKDD1A	rs10427335	2	14572045	trans	1.15259
ANKDD1A	rs2664095	3	9074653	trans	0.382531
ANKDD1A	rs7657630	4	90246443	trans	0.277017
ANKDD1A	rs17608937	6	12195974	trans	0.100787
ANKDD1A	rs1228412	10	125392451	trans	0.489565
ANKDD1A	rs11025102	11	19307344	trans	1.33793
ANKDD1A	rs1421566	11	99522626	trans	1.59316
ANKDD1A	rs1385951	15	63003010	cis	0.137471
ANKDD1A	rs8108252	19	55135873	trans	1.14917
ANKDD1A	rs2830437	21	27014160	trans	0.481143
ANKDD1A	rs13046217	21	46267004	trans	0.432625
<b>NAPRT1 <math>\log(\text{BF}) = 5.9829</math></b>					
NAPRT1	rs11131170	3	984495	trans	0.102164

NAPRT1	rs653316	3	184354656	trans	0.08354
NAPRT1	rs7631503	3	186512797	trans	0.130147
NAPRT1	rs2279525	4	23403350	trans	0.0888101
NAPRT1	rs91315	5	1908301	trans	0.417698
NAPRT1	rs17077288	5	173650577	trans	1.47139
NAPRT1	rs2429216	7	139605013	trans	0.0799407
NAPRT1	rs7932859	11	78872747	trans	0.783579
NAPRT1	rs7102251	11	120753111	trans	2.19781
NAPRT1	rs746690	12	14675065	trans	0.317195
NAPRT1	rs10860309	12	97136051	trans	0.0733323
NAPRT1	rs12825698	12	113470092	trans	0.237646
NAPRT1	rs17252387	15	66398543	trans	1.87507
NAPRT1	rs1468191	17	13076821	trans	0.229263
NAPRT1	rs6134639	20	12627627	trans	0.173953
NAPRT1	rs12053796	22	41944043	trans	0.232996
<b>TMPRSS3 log(BF) = 5.53317</b>					
TMPRSS3	rs7552599	1	30386259	trans	0.0958183
TMPRSS3	rs6334	1	155112857	trans	0.192863
TMPRSS3	rs6426551	1	224608672	trans	0.877136
TMPRSS3	rs40997	2	8016854	trans	0.0558463
TMPRSS3	rs10164749	2	20336328	trans	0.0867924
TMPRSS3	rs3900566	2	145445159	trans	0.0584469
TMPRSS3	rs3769931	2	165864479	trans	0.312595
TMPRSS3	rs10202550	2	173127236	trans	0.0886694
TMPRSS3	rs3138373	3	130631711	trans	0.158538
TMPRSS3	rs10007960	4	40114171	trans	0.158417
TMPRSS3	rs12513607	5	5036619	trans	1.14367
TMPRSS3	rs2287904	5	64617808	trans	0.0585229
TMPRSS3	rs7734266	5	163340895	trans	0.647947
TMPRSS3	rs6929735	6	1020055	trans	0.190189
TMPRSS3	rs2227234	6	103582559	trans	0.178188
TMPRSS3	rs17248900	7	7137460	trans	0.487562
TMPRSS3	rs2948929	7	152170108	trans	0.254564
TMPRSS3	rs434645	10	8161457	trans	0.112013
TMPRSS3	rs399593	10	30952036	trans	0.367894
TMPRSS3	rs2488647	10	86353619	trans	0.253093
TMPRSS3	rs17833422	14	58494680	trans	0.0726075
TMPRSS3	rs3736054	15	41109405	trans	0.533459
TMPRSS3	rs9928327	16	2190234	trans	3.08546
TMPRSS3	rs238142	18	3472786	trans	0.076077
TMPRSS3	rs9630890	19	2600218	trans	0.393992
TMPRSS3	rs1545117	20	53983182	trans	0.175123
TMPRSS3	rs4819596	22	16370606	trans	0.360458
TMPRSS3	rs5768034	22	46603224	trans	0.915409

<b>CD151 log(BF) = 5.49685</b>					
CD151	rs518063	1	47775708	trans	1.83833
CD151	rs6672824	1	232194894	trans	0.289559
CD151	rs10184722	2	143470773	trans	0.0517988
CD151	rs9289218	3	124547521	trans	0.122067
CD151	rs7806065	7	139879676	trans	0.343708
CD151	rs12543567	8	110096523	trans	0.974944
CD151	rs7069690	10	27298646	trans	0.0904704
CD151	rs5030780	11	828110	cis	0.042543
CD151	rs1108991	11	1494093	cis	1.96222
CD151	rs10521320	16	54133662	trans	0.630982
CD151	rs6025653	20	55613969	trans	0.292126
<b>LBX2 log(BF) = 5.39866</b>					
LBX2	rs891898	5	146600275	trans	3.92373
LBX2	rs9458808	6	163675242	trans	0.28768
LBX2	rs6942887	7	37154023	trans	0.122181
LBX2	rs13276508	8	72848378	trans	0.0859322
LBX2	rs2720972	10	129008874	trans	1.65381
<b>MED28 log(BF) = 5.38231</b>					
MED28	rs1373287	1	112319067	trans	0.141767
MED28	rs9871964	3	179856789	trans	0.331383
MED28	rs3800027	6	55486273	trans	0.224717
MED28	rs11196301	10	84624827	trans	1.30597
MED28	rs17119973	14	83982864	trans	0.163289
MED28	rs489105	15	53183925	trans	3.55254
<b>LAD1 log(BF) = 5.21392</b>					
LAD1	rs498795	1	4298856	trans	0.219071
LAD1	rs675508	1	199723677	cis	2.92835
LAD1	rs3011631	10	22402416	trans	1.4036
LAD1	rs670848	11	124846513	trans	0.707541
LAD1	rs1875051	13	62095838	trans	0.164085
LAD1	rs10152049	14	77865046	trans	0.118763
LAD1	rs7257503	19	55108776	trans	0.311779
<b>SEPN1 log(BF) = 5.1874</b>					
SEPN1	rs12023823	1	202312571	trans	1.2508
SEPN1	rs17045065	2	53731055	trans	0.0722753
SEPN1	rs1371238	4	44226286	trans	0.719195
SEPN1	rs2526977	7	105423382	trans	0.583362
SEPN1	rs2884594	12	4376087	trans	0.704122
SEPN1	rs9514497	13	105569397	trans	2.84168
SEPN1	rs1872159	14	22017743	trans	0.0579029
SEPN1	rs238142	18	3472786	trans	0.30284

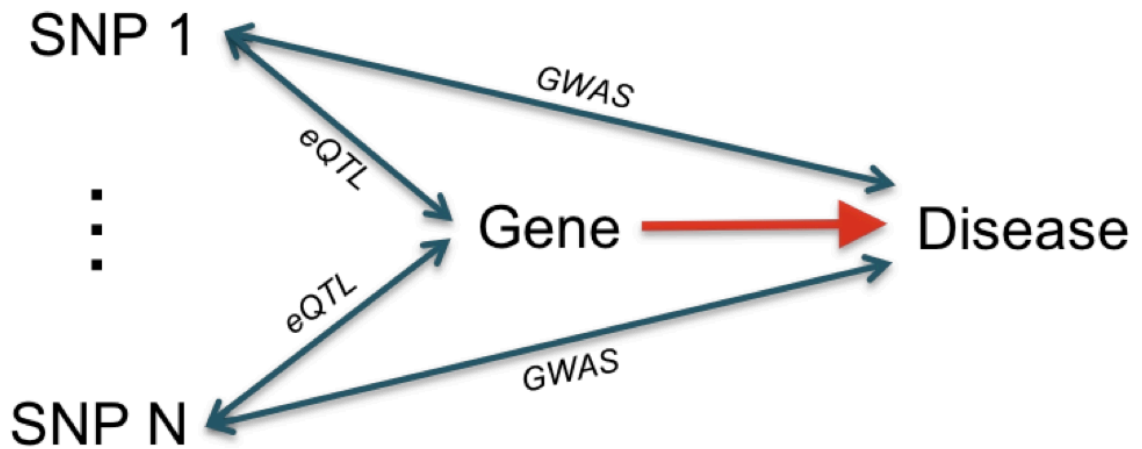
**Table 6**

The chromosome coordinate information in this table is obtained based on genome build 36.3.

Gene	SNPs	Chromosome	Coordinate	cis / trans	log(BF) of SNPs
<b>CXCR4 log(BF) = 0.834994</b>					
CXCR4	rs2372565	2	216066837	trans	-0.057943
CXCR4	rs12520069	5	168589997	trans	0.875323
CXCR4	rs10814443	9	36744763	trans	0.118574
CXCR4	rs1752156	9	125574333	trans	-0.0492948
CXCR4	rs7908645	10	112846415	trans	-0.00892216
CXCR4	rs2759303	13	36237281	trans	-0.0427429
<b>APOBEC3G log(BF) = -0.043849</b>					
APOBEC3G	rs11192130	10	106349130	trans	-0.00215559
APOBEC3G	rs4802561	19	54308942	trans	-0.0416935

**Figure 1**

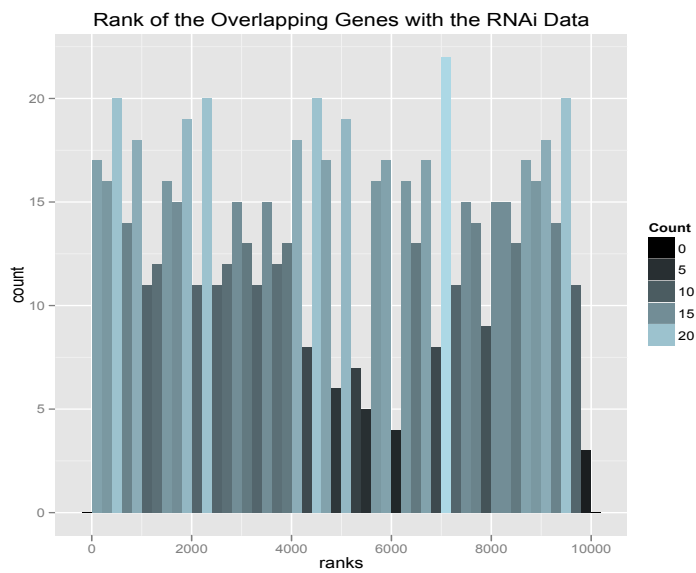
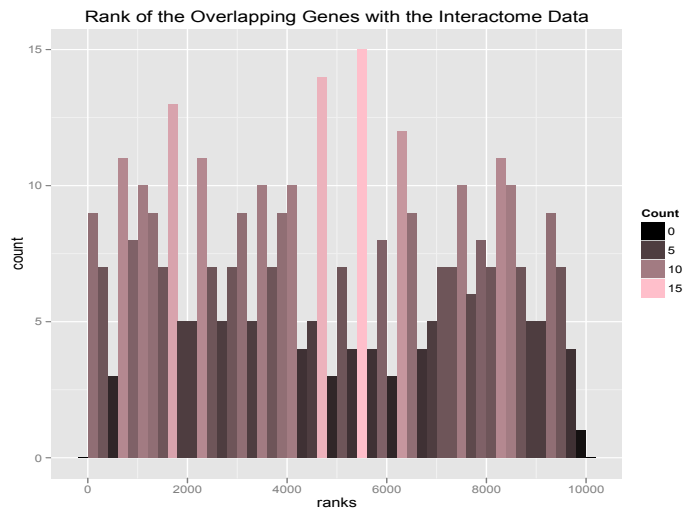
eQTL connects the variation of gene expression to SNPs, which is related to disease susceptibility by GWAS. The relationship we are trying to infer is represented by the red arrow. Bayesian statistics indicates that given the strengths of the eQTL relationship and the GWAS relationship (in teal blue), one can estimate the likelihood of the causal relationship between a gene and the disease (in red).





## Figure 2

The histograms show the overlap counts of my result with the interactome data and the RNAi data. The horizontal axis represents the rank of the overlapping genes. The plots demonstrate that the overlapping of the genes does not seem to be more concentrated in the area with higher scores.




**Publishing Agreement**

*It is the policy of the University to encourage the distribution of all theses, dissertations, and manuscripts. Copies of all UCSF theses, dissertations, and manuscripts will be routed to the library via the Graduate Division. The library will make all theses, dissertations, and manuscripts accessible to the public and will preserve these to the best of their abilities, in perpetuity.*

***Please sign the following statement:***

*I hereby grant permission to the Graduate Division of the University of California, San Francisco to release copies of my thesis, dissertation, or manuscript to the Campus Library to provide access and preservation, in whole or in part, in perpetuity.*

  
\_\_\_\_\_  
Author Signature

09/06/2012  
Date