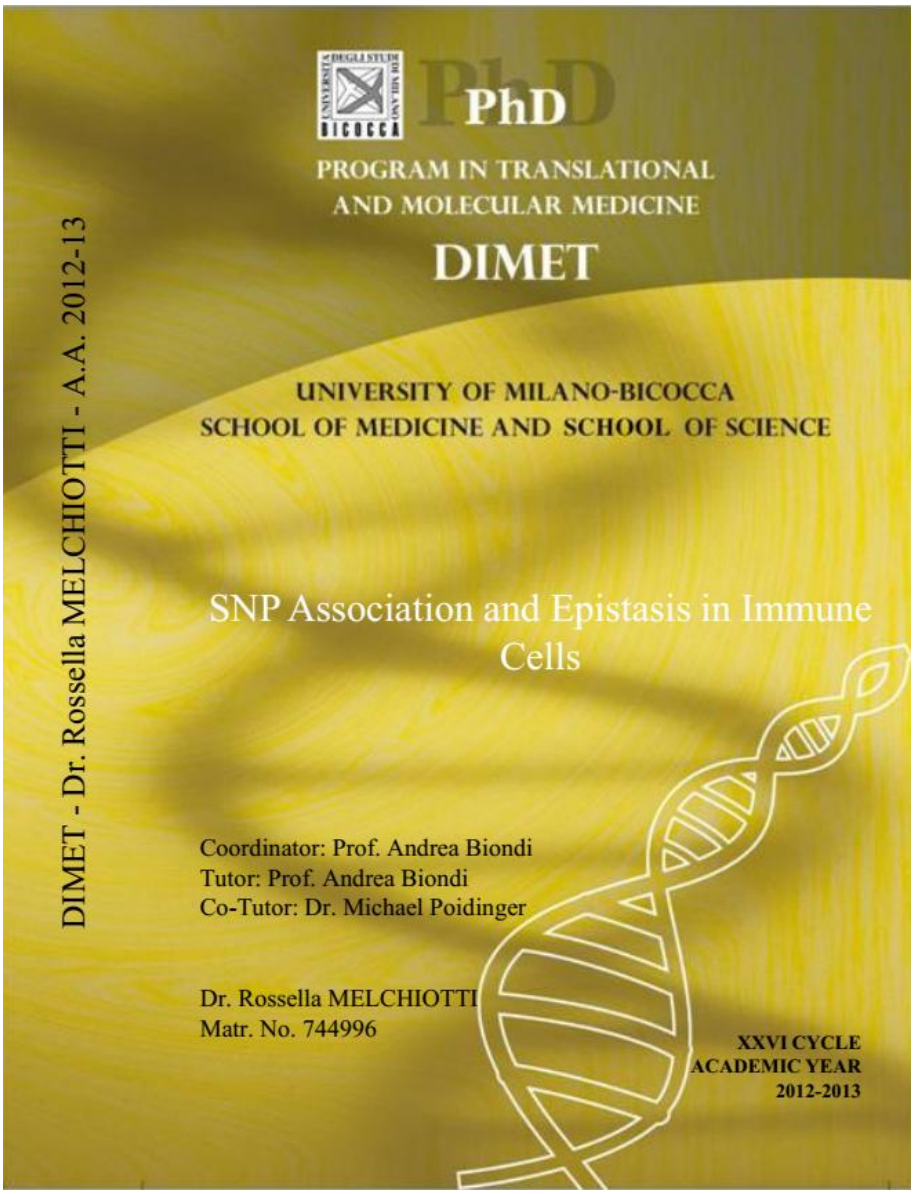

PhD
PROGRAM IN TRANSLATIONAL
AND MOLECULAR MEDICINE
DIMET
UNIVERSITY OF MILANO-BICOCCA
SCHOOL OF MEDICINE AND SCHOOL OF SCIENCE
SNP Association and Epistasis in Immune Cells
Coordinator: Prof. Andrea Biondi
Tutor: Prof. Andrea Biondi
Co-Tutor: Dr. Michael Poidinger
Dr. Rossella MELCHIOTTI
Matr. No. 744996
XXVI CYCLE
ACADEMIC YEAR
2012-2013
DIMET - Dr. Rossella MELCHIOTTI - A.A. 2012-13

To my family

Table of Contents

CHAPTER 1

GENERAL INTRODUCTION

The relationship between genotype and phenotype

The human genome is a complex structure containing approximately 3 billion base pairs (bp)[1,2]. Much of the genetic diversity between humans is due to the presence of single nucleotide polymorphisms (SNPs), mutations introduced into the human genome by replication errors[3]. As demonstrated by a recent analysis of the genome sequence of 1,092 individuals from 14 distinct populations, SNPs are not a rare phenomenon: more than 38 million SNPs have in fact been identified by this study[4]. Unfortunately elucidating the phenotypic impact of these polymorphisms is not straightforward due to the multiple ways mutations can affect phenotypes. Coding mutations (non-synonymous SNPs in particular) are easier to characterize because they directly impact protein structure or function. However protein-coding regions only account for a very small proportion of the entire sequence (~1.5%)[5]. The impact of SNPs on the remaining sequence is more difficult to assess but cannot be neglected since more than 80% of the genome was reported to have some biochemical function[6]. Mutations in non-coding regions can interfere with splicing events, impact transcription or translation and also affect post-translational modifications[7].

Due to the complexity of the relationship between genotype and phenotype, the identification of genetic variants involved in

susceptibility to complex diseases is extremely challenging. Genetic linkage studies involving family trios, where parents are used as controls for an affected offspring, were a traditional means to identify genomic regions harboring genes predisposing to disease[8]. This past decade has seen the development of another approach for the unbiased identification of susceptibility genes: genome-wide association studies (GWAS). The statistical concept behind association studies is simple: significant associations are identified by comparing allele or genotype frequencies between cases and controls (qualitative phenotypes) or by comparing mean phenotypic values between genotypes (quantitative phenotypes) (Figure 1)[9].

The popularity of GWAS was made possible by the development of affordable genotype microarray platforms that can now measure up to 5 million SNPs across the genome. As of April 2014, 13,156 SNPs were reported as being genetically associated to some trait/disease by the 1,902 published studies registered in the Catalog of Published Genome-Wide Association Studies[10]. Most of these studies were performed on individuals of Caucasian ethnicity, non European populations have so far not been well characterized[11].

In spite of the large number of SNPs reported by GWAS, understanding the origin of disease susceptibility remains in most cases problematic, many of the reported polymorphisms in fact exhibit small to moderate effects and the majority of them explain

only a small proportion of the estimated heritability for the studied trait[13] (heritability refers to the fraction of phenotypic variance in a population that can be additively explained by genetic factors[14]).

In addition most reported disease-associated polymorphisms do not

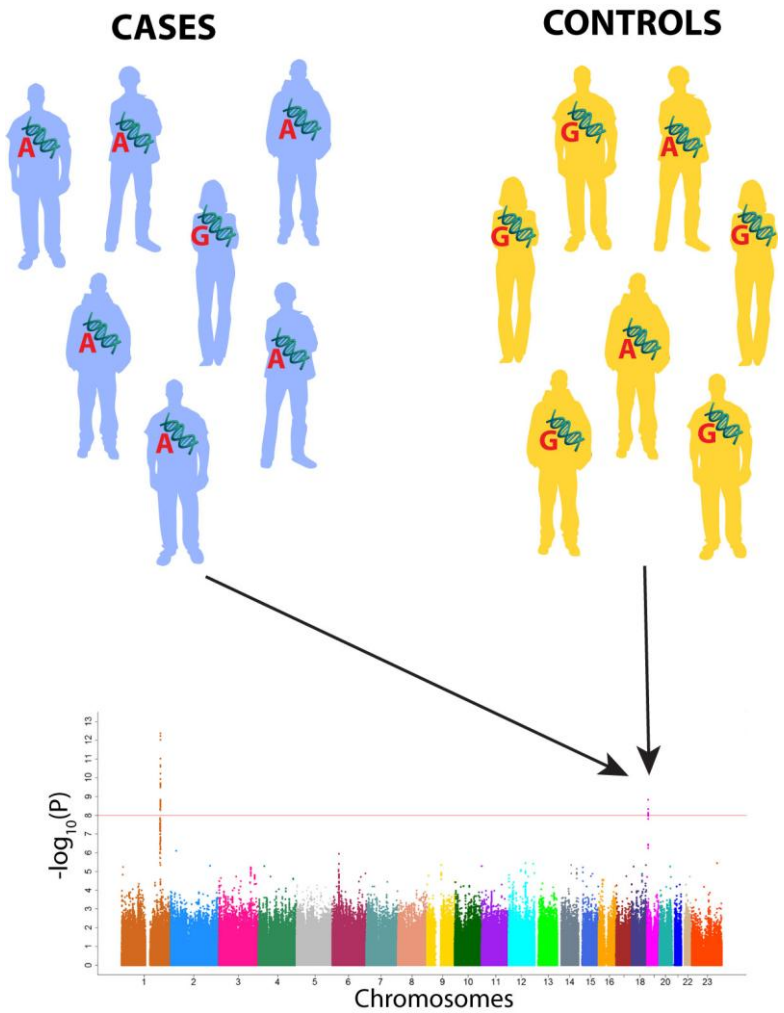


Figure 1. The concept behind genetic association studies for qualitative phenotypes. Allele frequencies are compared between cases and controls: the higher the difference for a particular polymorphism, the stronger the association of that SNP with disease susceptibility (modified from[12]).

replicate well across cohorts or populations. Only 16% to 30% of reported significant associations have been consistently replicated[15]. Multiple factors can explain this phenomenon such as population stratification, low statistical power, poor coverage, measurement errors and incorrect analysis assumptions that lead to false positives[16]. Another possible explanation for the low replication rate of GWAS concerns the presence of gene-gene and gene-environment interactions[17]. Complex systems are characterized by the interplay of multiple entities which often exert redundant functions to ensure the robustness of the system. It is therefore possible that the effect of a polymorphism on a particular phenotype might be missed if the genetic variant affecting a gene is considered in isolation, without taking into account the genetic state of its interaction partners[17]. This phenomenon where the effect of a SNP depends on its genetic background is referred to as epistasis[18]. The role of the environment should also not be ignored. The impact of a polymorphism on a certain trait might become evident only in the presence of certain environmental conditions or stimuli[19]. Genetic analyses across populations selected from different environments can in fact lead to false positives especially when cases and controls are not matched across the different conditions. It is therefore important to perform genetic studies in homogenous cohorts were the impact of environmental factors can be minimized.

GWAS are normally carried out using commercially available genotype platforms for the identification of associations between genetic variants and phenotypes. While this technology can now

measure up to 5 million polymorphisms it only assays a subset of all existing SNPs. Each SNP present on the arrays tags a certain number of other SNPs in a defined population. This is due to a phenomenon known as linkage disequilibrium (LD), the non-random association of alleles at two or more loci. SNPs closely located tend in fact to be inherited together so that their allelic states closely match. Each SNP on the array can therefore be used as a tag for all its closely linked SNPs. Once a tag SNP is identified as being associated with a particular phenotype, further analyses need to be performed to ascertain which polymorphism is functional among those tagged. Therefore significant associations identified by association studies in most cases do not reveal causative polymorphisms.

Moreover linkage disequilibrium patterns differ across populations. For this reason association signals identified in one population cannot be easily extended to other populations. Since association and epistasis studies have so far been strongly biased towards European populations there is a great need for genetic studies focusing on less characterized populations.

The immune system as a substrate for studying association and epistasis

The immune system is a complex network characterized by the interplay of multiple immune cell subsets with distinct functions. Multiple immune diseases have been shown to have a strong genetic component, evident from the higher concordance in monozygotic twins compared to their heterozygotic counterpart[20–25]. Considering

the high heritability of most immune diseases and the complexity of the interactions between cell subsets the immune system represent a good model for studying the impact of genetic variants on phenotype.

Because of its key role in promoting survival by protecting hosts against infections the immune system is characterized by a strong redundancy which ensures the robustness of the system[26–28]. It is therefore important to consider not only single mutations (association), which might have a small effect on complex immune phenotypes because of functional redundancy, but also combinations of mutations (epistasis).

Association

SNP association can be performed both on quantitative phenotypes such as gene or protein expression and qualitative phenotypes like disease status.

A high-throughput association analysis often performed in the context of quantitative phenotypes is the expression Quantitative Trait Locus (eQTL) analysis. This technique associates genetic variants with gene expression measurements assayed in a particular cell type or tissue (Figure 2). The term eQTL can refer to the association between a genetic variant (called eSNP) and the gene expression of a proximal (*cis* eQTL) or distal (*trans* eQTL) gene[29]. The exact definition of what constitutes a *cis* eQTL varies from study to study. In most cases *cis* eQTLs are defined as associations between gene expression

of one probe and allelic states of SNPs located in a ±1Mbp region spanning the probe midpoint[30–32].

The identification of eQTLs in different cell types is particularly valuable for understanding the relationship between polymorphisms and disease. Gene expression is in fact an important intermediate phenotype for multiple conditions. The advantage of associating mutations with variations in gene expression rather than susceptibility to immune diseases lies in the fact that the impact of a polymorphism on expression levels is direct. The impact on a complex disease is on the contrary more subtle since it is diluted by the complexity underlying biological networks which involve cross-talks between multiple genes and pathways.

EQTL analyses have been shown to be an important tool for identifying functional and causal SNPs from tag SNPs associated with disease susceptibility[30,33,34]. EQTLs are particularly useful in the context of SNPs located in intergenic or non-coding regions which cannot be directly connected with gene function[35]. Since eQTLs link genetic variants with their modulated genes they can be used to pinpoint proteins involved in disease pathogenesis[34].

Trait-associated SNPs have been shown to be more likely to be associated with gene expression confirming the important role played by transcriptional regulation in disease susceptibility or etiology[36]. EQTLs could therefore be used to limit the number of SNPs to include in genome-wide association studies for a particular disease in order to reduce the burden of multiple testing[37].

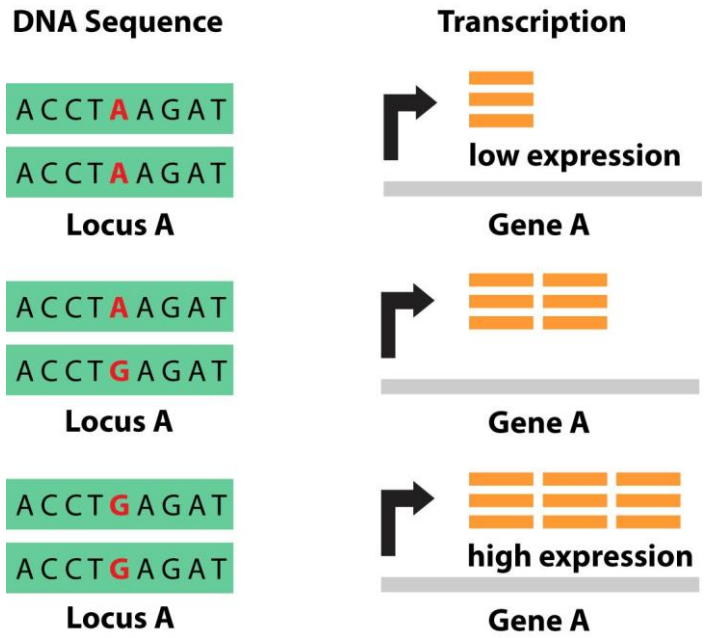


Figure 2. Expression Quantitative Trait Locus. Expression levels of Gene A are influenced by the genotype at Locus A. The A allele is associated with low transcriptional expression while the G allele is associated with higher transcriptional levels.

EQTL studies can also be used to propose new candidate genes for complex diseases. This is particularly relevant when eQTL studies are performed on well defined primary cell subsets already known to be involved in a particular disease. For example a study of autoimmune diseases, which are characterized by a strong T regulatory cells (Treg) involvement, could benefit from the identification of eQTLs in this CD4+ T cell population.

Detection of statistical association

From a statistical point of view eQTLs can be identified using a number of different tests. The most commonly used tests for

estimating significance of association are linear regression[38–40], Spearman correlation[31,39,41–43] and one-way analysis of variance (ANOVA or Kruskal-Wallis)[44]. While P-values estimate the significance of an association between a factor and an outcome they give no indication on the magnitude of the effect [45]. There are multiple metrics to estimate the strength of the association between genetic variants and quantitative phenotypes. The metric is usually related to the statistical test used to estimate the significance of the association. Effects sizes of an association estimated through one-way analysis of variance are usually reported as fold changes between the mean or median of the highest group and the mean or median of the lowest group. In the context of linear regression or correlation effect sizes are estimated by the regression coefficient β or by the correlation coefficient r^2.

Association can also be performed on qualitative phenotypes such as disease status. One of the most used tests for association in this context is the Pearson's chi-squared test, also known as the chi-square test for goodness-of-fit[46–48]. This test is applied to categorical variables and it estimates the association between row and column factors using a two-way table: one variable contains alleles or genotypes for a particular SNP; the other variable is the disease status (case/control). Each cell contains the count of the number of individuals with that particular factor combination. The null hypothesis H_0 assumes that there exist no association between the SNP under consideration and susceptibility to the disease under study.

Another commonly used test is logistic regression[48]. Logistic regression fits the following model to the data:

$$log\frac{p}{(1-p)} = \alpha + \beta x$$

(where p is the probability of being affected by the disease and x is a measure of the allele dosage for the SNP). The test for association evaluates whether the coefficient β is significantly different from zero. The strength of a statistical association in the context of qualitative phenotypes is usually measured using the metric odds ratio (OR). An odds ratio is a measure of association between an exposure (in statistical genetics this usually refers to an allele or a genotype) and an outcome (in statistical genetics this usually refers to the presence or absence of disease)[49].

Epistasis

As previously mentioned, single mutations are unlikely to account for the complexity of multifactorial diseases. Thus, the synergistic or antagonistic effect of multiple polymorphisms needs to be investigated in order to better understand the mechanisms underlying disease susceptibility.

Definition

The word epistasis was first introduced by Bateson to refer to the masking of the effects of one genetic locus by another locus[50]. A classical example of epistasis involves fur color in mice. Coat color is determined by two loci: locus A, which affects the early stages of the synthesis of an allele involved in pigment production, and locus B,

which determines whether the fur has bands or not. For locus A, the dominant allele results in normal pigment production while the recessive allele impairs synthesis. For locus B the dominant allele results in a brown fur with bands while the recessive allele results in a black fur without bands. Mice carrying the aa genotype for locus A are albinos regardless of their genotype at locus B. The aa genotype at locus A is therefore considered as epistatic to locus B (Table 1).

Bateson's definition of epistasis refers to what is nowadays considered as biological epistasis: the result of physical interactions between entities within a biological pathway or a gene regulatory network that make the effect of one gene on a phenotype dependent on the effect of one or more other genes[51].

Epistasis is an ubiquitous property of biological networks[52]. Biological networks are in fact extremely complex because of their numerous interacting entities. They are also characterized by multiple redundant pathways that confer a strong robustness to the system. This robustness is guaranteed by the scale-free nature of interaction networks where most molecules are involved in few interactions but few are involved in many different ones (hubs)[53].

Table 1 Example of biological epistasis. Mouse fur color is determined by two loci (A and B). A mouse carrying the aa genotype for locus A will be white regardless of its genotype at locus B. The genotype aa is therefore epistatic to locus B.

	AB	Ab	aB	ab
AB	AABB Brown	AABb Brown	AaBB Brown	AaBb Brown
Ab	AAbB Brown	AAbb Black	AabB Brown	Aabb Black
aB	aABB Brown	aABb Brown	aaBB White	aaBb White
ab	aAbB Brown	aAbb Black	aabB White	aabb White

It has been shown that a disruption of 80% of nodes in a scale-free network still allows it to function correctly since most pathways between any two nodes can still be drawn[52]. Due to the extreme robustness of biological networks it is thus very unlikely that a variant alone could have a huge impact on a complex phenotype. It is not clear why epistasis exists but it is likely that the redundant and robust nature of genetic networks, that buffer a phenotype against the effects of mutations, makes phenotypic changes visible only when multiple genes are disrupted in a pathway[54]. The ubiquity of interactions between molecular entities in gene regulation suggests that the relationship between genetic variants and phenotypes is not biunivocal but that complex traits are likely to involve synergy between multiple gene products[55].

The term epistasis is also often used to refer to statistical epistasis, a concept introduced by Fisher in 1918 with respect to deviations from additivity in a statistical model explaining the relationship between various loci and phenotypic variation in a particular population[56]. In the context of qualitative phenotypes Fisher's definition of epistasis would refer to the following model:

$$p_{ij} = \alpha_i + \beta_j$$

where p_{ij} is the penetrance (the probability of developing a disease given a certain genotype) associated to genotype i at locus A and genotype j at locus B, α_i is a parameter representing the effect of genotype i at locus A and β_j is a parameter representing the effect of genotype j at locus B.

Other authors assume that epistasis refers to a departure from a multiplicative model on the penetrance scale[57,58]. In this case penetrance would be defined as:

$$p_{ij} = \alpha_i * \beta_j$$

While biological epistasis happens at the cellular level in a particular individual, statistical epistasis is a phenomenon associated to an entire population[59].

The existence of different definitions of the term epistasis often creates confusion since it is not always clear from the context which type of epistasis is being analyzed in a particular study.

Statistical Epistasis vs. Biological Epistasis

One of the major challenges of multi-locus analyses consists in finding ways to conciliate the concept of statistical epistasis (that is

usually discovered through standard epistasis analyses) and biological epistasis (the phenomenon we are really interested in). Making hypotheses about biological function from statistical epistasis is extremely challenging due to the large number of confounding factors such as biases in the study design and analysis and the presence of linkage disequilibrium which complicates the identification of causative SNPs[15]. This challenge is even more pronounced in the case of humans due to the impossibility of conducting the double gene knockouts experiments that are possible in model organisms. Making the connection between biological and statistical epistasis therefore requires the integration of multiple data sources such as genomics, transcriptomics, proteomics and metabolomics. To complicate matters biological epistasis, which is a phenomenon occurring at the individual level, can exist in the absence of statistical epistasis, which is a phenomenon occurring at the population level, simply as a result of biases in the sample collection[51]. This is particularly true when the polymorphisms involved in the interaction are rare so that a random sample might not contain individuals with the deleterious variant combination. The reverse is also true, statistical epistasis does not always imply biological epistasis especially when there are biases in sample collection and cases and controls are not properly matched.

Models

There exist multiple models for epistasis. In the case of 2 biallelic loci 512 fully penetrant models can be identified[60], the most common

Table 2. Four common epistatic models: threshold model (T), dominant-dominant model (DD), recessive-dominant model (RD) and recessive-recessive model (RR)[60,61]

T Model			
SNP_1/SNP2	aa	aA	AA
bb	1	1	0
bB	1	0	0
BB	0	0	0

DD Model			
SNP_1/SNP2	aa	aA	AA
bb	1	1	0
bB	1	1	0
BB	0	0	0

RD Model			
SNP_1/SNP2	aa	aA	AA
bb	1	1	0
bB	0	0	0
BB	0	0	0

RR Model			
SNP_1/SNP2	aa	aA	AA
bb	1	0	0
bB	0	0	0
BB	0	0	0

ones including the threshold model (at least three risk alleles required for disease independently of which locus they come from), the jointly dominant-dominant model (at least one copy of the risk allele at both loci required for disease), the jointly dominant-recessive model (two copies of the risk allele from the first locus and at least one copy of the risk allele from the second locus required for disease) and the jointly recessive-recessive one (two copies of the risk allele at both loci required for disease) (Table 2)[60,61].

An important question to consider is whether there exist loci that jointly affect a phenotype without displaying marginal effects (also referred to as main effects)[62,63]. Marginal effects are present when a locus independently displays an association with a particular

phenotype. From a mathematical point of view it is possible to conceive epistatic models where both loci do not show any marginal association with the phenotype but it is still not clear if these models are biologically plausible and how common they are in humans[17]. An example of this would be the XOR (exclusive OR) model. This model assumes that individuals are at a high risk if they inherit the heterozygous genotype for one locus and the homozygous genotype for the other locus (AaBB, Aabb, AABb and aaBb are high risk combinations in this model)[64].

Role of epistasis

The role of epistasis has been recognized in multiple studies on model organisms such as yeast, *Escherichia coli* and mice[65–69] and in the pathogenesis of several human diseases such as cardiovascular diseases, cancer and other immune conditions[18,70–73].

Epistasis is often considered as a possible factor accounting for the so called missing heritability in complex diseases[74,75]. All risk loci identified so far by GWAS studies only explain a small proportion of the total estimated heritability[14]. It has been suggested that a large proportion of this missing heritability is due to an overestimation, from population data, of the total heritability. Most estimates of total heritability are in fact based on the assumption that there exist no interactions between loci. This assumption is unjustifiable since models involving interactions are also consistent with the observed data. Including interactions in the model would lead to smaller total heritability estimates. In a model allowing for interactions, variants

reported as significantly associated with a disease would therefore explain a larger proportion of the total heritability than initially thought[75].

Mechanisms underlying epistasis

Multiple molecular mechanisms can cause the emergence of epistasis (as reviewed by Ben Lehner[76]). Epistatic interaction can involve mutations in different genes (intermolecular epistasis) or multiple polymorphisms of the same gene (intramolecular epistasis). The simplest example of intermolecular epistasis involves molecules interacting directly. A mutation affecting a molecule can for example become phenotypically apparent only in the presence of a different mutation in the interacting partner protein. Similarly, deleterious mutations in one protein can be compensated by mutations in their interacting protein. This mechanism of epistasis is common among surface receptors and their respective ligands. In this context Baessler *et al.* described an epistatic interaction between the ghrelin ligand, an appetite inducing hormone secreted by the stomach, and its receptor affecting susceptibility to myocardial infarction and coronary artery disease[77].

Another biological mechanism that can cause epistasis is functional redundancy. In complex biological network such as those that characterize the immune system multiple proteins play a similar role or function. This redundancy is essential in assuring the robustness of the system. In this context phenotypic changes might become evident only in the presence of mutations affecting multiple

functionally similar proteins since the dysregulation of a single molecule can be easily compensated by other analogous molecules.

Functional redundancy can also involve mutations affecting genes in functionally similar pathways. For example if under certain conditions the same metabolite can be produced by two different pathways, mutations affecting a single pathway will have little effect on the synthesis of the metabolite[78].

Epistasis can also arise when mutations affect multiple genes belonging to the same molecular complex or pathway. In linear pathways or in molecular complexes a mutation disrupting a single gene affects the whole system. Subsequent mutations in the pathway or complex would have no further effect. Mutations in one gene would therefore mask all other mutations in the pathway. For example a loss of function mutation in an upstream molecule would have no effect on the pathway if another mutation constitutively activated a downstream target.

Another interesting molecular context for epistasis is linked to the concept of physical constraint. If a mutation forces a system to reach saturation further mutations might have smaller effects than expected by an additive model simply because the maximum or minimum effect has already been attained. Physical constraints can cause not only intermolecular epistasis but also intramolecular epistasis. Intermolecular epistasis can also occur in more complex non-linear systems such as feedback loops.

While most intermolecular epistatic interactions are synergistic the majority of intramolecular epistatic interactions are compensatory[79].

One of the mechanisms for intramolecular epistasis is threshold epistasis: each gene and protein has a particular stability and a single mutation might not be sufficient to cross the stability threshold[80]. Disruption of function is obtained only in the presence of multiple deleterious polymorphisms.

Epistasis can also be associated to conformational changes. The effect of a mutation in a receptor that increases affinity might become relevant only if another mutation changes the conformation of the protein so that the new residue can come in contact with the ligand.

Another interesting example of intramolecular epistasis comes from the interaction between coding mutations that affect function and non-coding mutations that affect transcription. A beneficial coding mutation could in fact have no noticeable phenotypic effect in the presence of a non-coding mutation switching off the transcription of the corresponding gene.

Detection of statistical epistasis

There exist multiple methods to detect statistical epistasis. The most common statistical test used for identifying epistatic interactions in the context of qualitative phenotypes is logistic regression. Logistic regression fits the following model to the data:

$$log\frac{p}{(1-p)} = \alpha + \beta x_A + \gamma x_B + \delta x_A x_B$$

(where p is the probability of being affected by the disease and x is a measure of the allele dosage for the SNP). The test for interaction evaluates whether the coefficient δ is significantly different from

zero. The regression coefficient δ refers to interaction as departure from multiplicativity. In the context of quantitative phenotypes logistic regression is replaced by linear regression. In this case the dependent variable is the measurement under study and the interaction term reflects departure from additivity[81]. The advantage of using logistic or linear models lies in the ease of incorporating covariates such as gender or age.

As for association, the strength of a statistical epistatic interaction in the context of qualitative phenotypes is usually measured using the metric odds ratio (OR). In this case the odds ratio is a measure of association between a genotype combination and the presence or absence of disease[49]. For quantitative phenotypes the value of the standardized regression coefficient for the interaction term in the regression model is usually a good estimation of the effect size.

The burden of multiple testing

Due to the large number of tests that needs to be performed to identify statistical association and epistasis at a genome-wide level, it is of uttermost importance to adopt strict multiple testing correction strategies during the discovery phase in order to reduce or control the proportion of false positives.

The simplest and most commonly adopted correction strategy in this context is Bonferroni correction[82,83]. This correction procedure has the advantage of being simple and computationally inexpensive but unfortunately it relies on the hypothesis of independency between tests which in most genetic analyses does not hold due to linkage

disequilibrium between SNPs. In the context of eQTL analysis probes are also not independent due to the correlation existing between co-regulated genes.

False-discovery-rate (FDR) is another method frequently used for multiple testing correction that controls the expected proportion of falsely rejected null hypothesis[84]. FDR is relatively fast, easy to compute and it is more powerful than other multiple testing correction strategies but it is less conservative than Bonferroni and leads to an increased proportion of type I errors (a type I error is the rejection of a true null hypothesis).

One of the most powerful strategies for multiple testing correction is permutation testing[85]. This strategy has been applied by many epistatic algorithms[86,87]. It is also extremely common in the context of eQTL studies[30,39,41]. Permutation-based corrections do not rely on any *a priori* hypothesis on the statistical distribution underlying the process under study. In a permutation test phenotype labels are randomly shuffled so that any real association between genotype and phenotype is broken and a statistics is computed using the same algorithm as for the original dataset. This strategy is used to build a null distribution that is then used to estimate significance.

The main drawback of this approach lies in the computational burden of obtaining permuted P-values. In the context of genome-wide epistasis studies this approach becomes impractical due to the extremely large numbers of pairwise tests to perform for each permutation.

Regardless of the strategy used for multiple testing correction, replication still remains one of the major validation methods to reduce the risk of false-positives[88]. An eQTL is defined as replicated if a significant association of the same SNP with the same probe/gene and the same direction of effect is detected in a similar population. A less stringent definition of replication can also be adopted. In this case a perfect match between SNPs is not required as long as the eSNPs detected in the two cohorts are in strong linkage disequilibrium. The concept of replication for epistatic results has not been clearly defined[37]. Some argue that replication should be at the SNP level and with the same direction of effect in the same population. Others adopt a less strict definition looking for replication of a particular gene or biological pathway.

The role played by linkage disequilibrium

Commercially available microarray platforms can now measure up to 5 million polymorphisms. Many of these genetic variants are not independent but jointly inherited due to a phenomenon known as linkage disequilibrium (LD). Linkage disequilibrium is the non-random association of allelic states at two or more loci. Knowledge of LD is extremely important in the context of SNP association/epistasis because variants identified as significantly associated with a particular phenotype in most cases are not causative but simply tag a causative variant. Commercially available microarrays in fact do not measure all SNPs but only a portion of them. Nonetheless these platforms were developed to tag a large number of variants. LD

therefore plays an important role in identifying causative variants from disease-associated variants.

The concept of LD is linked to the concept of haplotypes, combinations of alleles at adjacent loci that are inherited together.

Multiple metrics have been developed to estimate linkage disequilibrium from genotypic datasets. The most used metric in the context of genetic studies is r^2. Given two genetic polymorphic loci A (alleles A_1 and A_2) and B (alleles B_1 and B_2) four distinct haplotypes can be formed: A_1B_1 with frequency $f_{A_1B_1}$, A_1B_2 with frequency $f_{A_1B_2}$, A_2B_1 with frequency $f_{A_2B_1}$ and A_2B_2 with frequency $f_{A_2B_2}$. The metric D is defined as follows:

$$D = f_{A_1B_1} - f_{A_1}f_{B_1}$$

where f_{A_1} is the frequency of allele A_1 and f_{B_1} is the frequency of allele B_1. D can be normalized using allele frequencies to obtain the coefficient D':

$$D' = \frac{D}{D_{max}} \text{ where } \begin{cases} D_{max} = \min\left(f_{A_1}f_{B_1}, f_{A_2}f_{B_2}\right) if\ D < 0 \\ D_{max} = \min\left(f_{A_1}f_{B_2}, f_{A_2}f_{B_1}\right) if\ D > 0 \end{cases}$$

If D'=±1 only three out of the four possible haplotypes for SNP A and SNP B exist in the population under consideration.

Another common normalization for the coefficient D is the coefficient r^2 which is defined as followed:

$$r^2 = \frac{D^2}{f_{A_1}f_{B_1}f_{A_2}f_{B_2}}$$

where f_{A_2} is the frequency of the allele A_2 and f_{B_2} is the frequency of the allele B_2. This coefficient is a measure of the correlation between the allelic states at different loci. An r^2 of 1 between two SNPs

(perfect linkage disequilibrium) means that each SNP is a perfect tag for the other. It is therefore redundant to genotype both SNPs since the allelic state of one can be accurately predicted from the allelic state of the other. In this case only two out of the four possible haplotypes for SNP A and SNP B exist in the population under consideration. At the opposite side of the spectrum an r^2 of 0 would imply that the two SNPs are completely independent.

When a new mutation arises in proximity of another mutation it is initially limited to one haplotype (Figure 3). The genomic pool now contains three haplotypes, the two original ones and the one carrying the new mutation. A newly risen mutation is therefore linked to its adjacent mutation (D'=1). When a recombination event occurs between the two mutations a new haplotypes is created causing erosion in LD between the two polymorphisms: the larger the number of recombination events, the stronger the erosion[89]. LD thus decays with distance: SNPs closely located on the genome have a higher chance of being in strong LD because there is a lower chance of a recombination event affecting them[90].

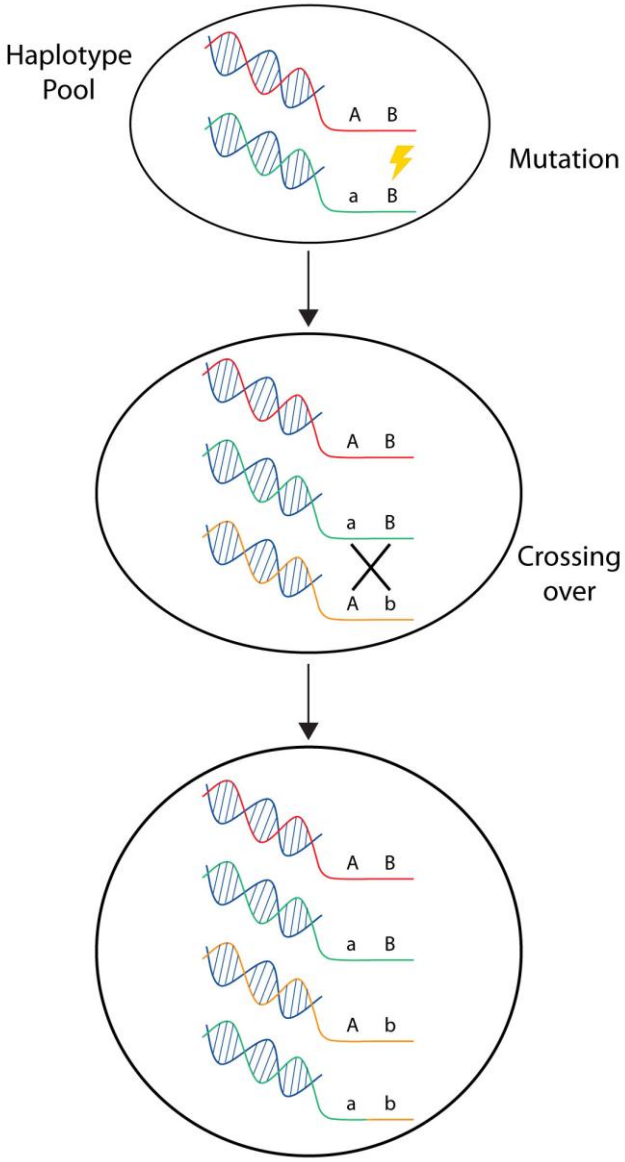


Figure 3. The origin of linkage disequilibrium. When a new mutation is introduced in a population (locus B) it is initially linked (D'=1) to all adjacent mutations (locus A). When a recombination event occurs between two adjacent mutations a new haplotype is introduced into the genetic pool and the linkage disequilibrium between the two loci is eroded.

One of the most widespread computational tools for the estimation and visualization of LD from genotypic datasets is Haploview[91]. This software utilizes triangular correlation plots to represent LD. SNPs are lined up on the horizontal axis and the color of the square connecting two SNPs is a measure of their pairwise LD (Figure 4).

As can be easily deduced from the plot this tool is not suitable for the analysis of large or highly polymorphic regions because the complexity of the visualization increases with the number of SNPs under consideration.

Another frequently used software for the analysis of LD is SNAP Proxy. The software supplies a user-friendly interface to retrieve lists

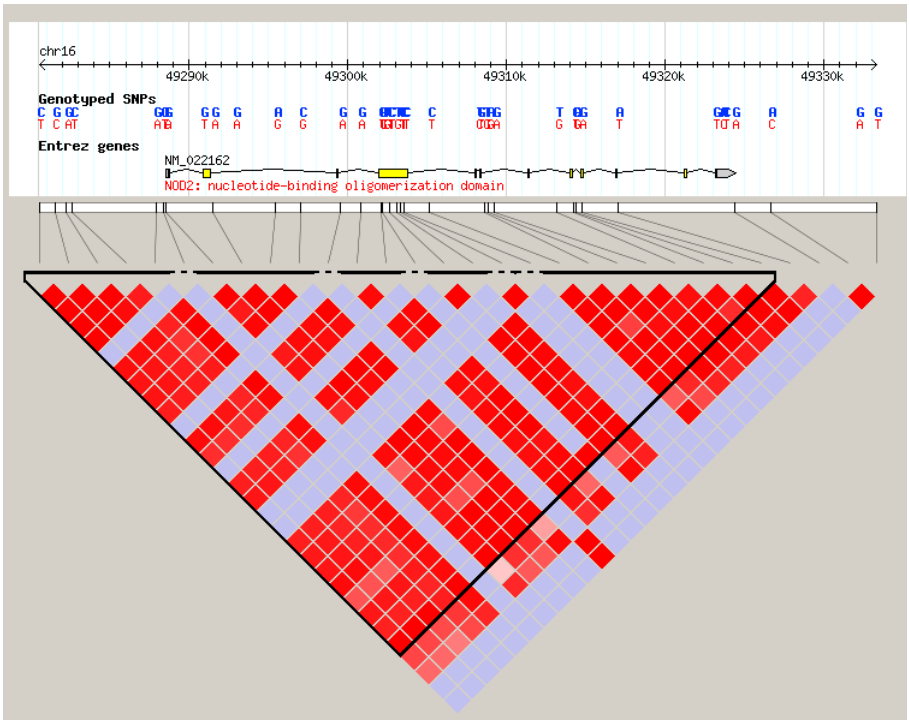


Figure 4. Linkage Disequilibrium plot for gene NOD2 as produced by the software Haploview[91]

of SNPs in LD with a particular tag SNP in a pre-defined population (HapMap[92–94] or 1000 Genomes Pilot project[95]). It also provides an option for building LD plots from a tag SNP. These plots are easy to interpret: SNPs are arranged according to their chromosomal position (x axis) and their pairwise r^2 with the SNP of interest (left y axis). Recombination rates can also be obtained from the plot (right y axis) (Figure 5).

A weakness of this visualization lies in the fact that SNP labels are not included in the plot but need to be retrieved using a different interface.

The main drawback of both these two tools is that they do not aid in the identification of causative or functional SNPs from tag SNPs, an

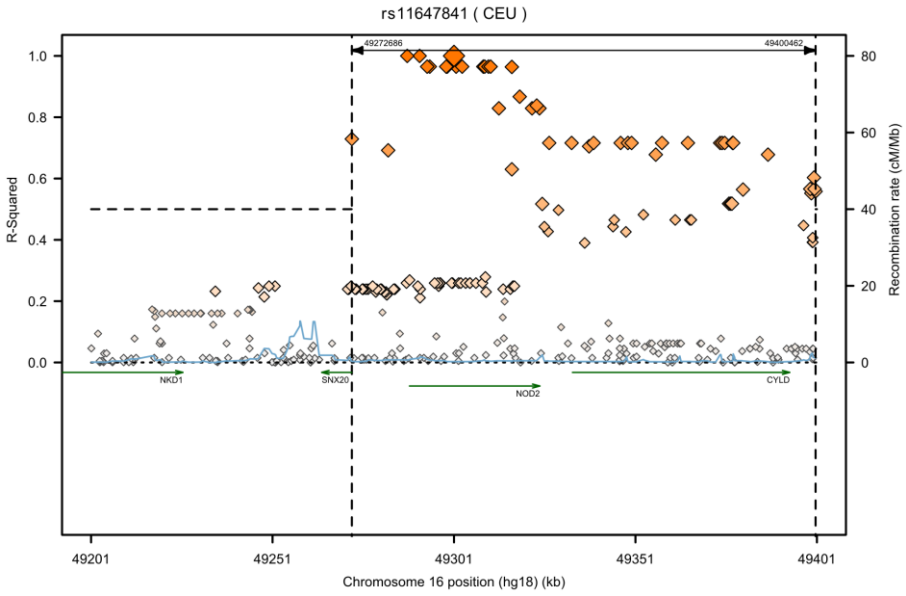


Figure 5. Linkage Disequilibrium plot for SNP rs11657841 as produced by SNAP Proxy[96]

essential step for understanding the biological mechanisms underlying a statistical association. Association and epistasis studies are in fact not performed on the totality of existent SNPs but are in most cases limited to those polymorphisms available on commercial genotyping platforms. As previously mentioned SNPs identified as being associated with a particular phenotype are rarely causative. The first step in pinpointing causative variants is to retrieve all SNPs in strong linkage with the polymorphism of interest. These variants then need to be ranked according to their functional potential before being carried forward for functional studies. This step requires the analysis of publicly available biological information. Relative position with respect to genes (intronic, exonic, promoter or intergenic regions) might for example hint at biologically interesting polymorphisms while publicly available experimental data like transcription-factor binding sites or open-chromatin regions might suggest a transcriptional involvement of the corresponding SNP.

None of the existing LD visualization tools provides a straightforward way to integrate linkage disequilibrium information (required for the identification of potential causative SNPs) with additional publicly available knowledge (essential for the identification of polymorphisms with functional potential). There is therefore a strong need for new tools offering this functionality to assist geneticists in the identification of causative SNPs from disease- or trait-associated polymorphisms in order to bridge the gap between statistical association and biological mechanisms.

EQTLs in Immune Cells

The first genome-wide eQTL study in humans was carried out by Stranger *et al.*[32] on Epstein-Barr virus–transformed lymphoblastoid cell lines developed for 60 unrelated individuals of Caucasian ancestry (CEU) collected by the HapMap project[92,93]. The analysis was limited to 630 genes. In a subsequent publication the investigation was extended to 13,643 autosomal genes for all 270 individuals genotyped by the HapMap project[93,94]. The consortium selected individuals from four geographically distinct populations: 30 mother-father-child trios of European ancestry living in Utah selected from the Centre d'Etude du Polymorphisme Humain (CEPH) collection (CEU), 45 unrelated Han Chinese individuals from Beijing, China (CHB), 45 unrelated Japanese individuals from Tokyo, Japan (JPT) and 30 trios from the Yoruba region in Ibadan, Nigeria (YRI). Using a 0.001 permutation threshold they detected eQTLs for 299 genes in CEU, 318 genes in CHB, 341 genes in JPT and 394 genes in YRI. About 37% of significant eQTLs indentified by the study were shared between two or more HapMap populations.

Since then multiple studies have been performed to elucidate the role of genetic variants on gene expression in different cell types. Dimas *et al.* performed a genome-wide scan for cis-eQTLs on three cell types (primary fibroblasts, Epstein-Barr-virus-immortalized B cells and T cells) in a cohort of 75 individuals of Caucasian ethnicity. Using a permutation threshold of 0.001 they reported 427, 442 and 430 genes with at least a significant eQTL in fibroblasts, Epstein-Barr virus immortalized B cells and T cells respectively. A large proportion of

eQTLs were cell-type specific even for genes which were not significantly differentially expressed across cell types[42]. Murphy *et al.* performed a *cis* eQTL analysis on isolated CD4+ lymphocytes from 200 individuals of self-reported non-Hispanic white ancestry. They identified significant eQTLs for 1585 distinct genes at a false discovery rate of 0.05. Their eQTLs were shown to be enriched for GWAS hits according to the GWAS catalog[97].

Zeller *et al.* used monocytes isolated from 1490 individuals of German origin to elucidate the role played by genetic variants in controlling gene expression in this immune cell subset[44]. Out of the 12,808 monocyte-expressed genes identified by the study, 2,745 harbored at least one SNP affecting their expression. Fairfax *et al.* performed a similar analysis on B cells and monocytes isolated from 283 European individuals[39]. 7,468 genes harbored a B cell specific eQTL, 6,831 harbored a monocytes specific eQTL and 1,323 harbored an eQTL shared by the two immune subsets. They showed that while many eQTLs with large effect sizes were shared across these two cell types, the majority of eQTLs identified for these two primary tissues were cell-type specific. Even in this case a strong overlap with disease-associated variants was detected.

Ferraro *et al.* profiled CD4+ conventional T cells and CD4+ Treg cells isolated from 168 donors, healthy or affected by type 1 diabetes and selected 65 individuals to perform a *cis* eQTL analysis[98]. They identified 105 genes harboring an eQTL for Treg cells and 110 genes for conventional CD4+ T cells (0.001 permutation threshold). Many of the eQTLs were shared across the two cell types.

Detection of eQTLs is not limited to isolated cells. Mehta *et al.* used a discovery cohort of 322 whole blood samples obtained from Caucasians and identified significant eQTLs for 363 genes after Bonferroni correction[38]. 98.6% of cis eQTLs were replicated in two independent cohorts. Westra *et al.* performed a similar analysis on 5,311 samples and identified significant *cis* eQTLs for 44% of all tested genes at a false discovery rate of 0.05[43].

A new field of analysis has recently emerged focusing on response-specific eQTLs (reQTLs)[99]. Barreiro *et al.* used dendritic cells isolated from 65 individuals to map eQTLs before and after *Mycobacterium Tuberculosis* (MTB) infection[99]. They were able to identify 198 response eQTL that were specific to either untreated or MTB-infected dendritic cells. A similar study was performed by Lee *et al.*[100]. They collected dendritic cells from 534 individuals from three different ethnicities (295 Caucasians, 122 African Americans, 117 East Asians) and stimulated them with three different stimuli: *Escherichia coli* lipopolysaccharide (LPS), influenza virus and interferon-β (IFN-β). They identified 121 genetic variants associated with the expression of their *cis* gene for at least one of the stimuli tested. A reQTL analysis was also performed in the context of monocytes[101]. Monocytes isolated from 432 healthy Europeans were stimulated with interferon gamma (IFN-γ) or LPS (2 or 24 hours). A large proportion of observed *cis* eQTLs modulated gene expression only upon stimulation. Conversely a large number of *cis* eQTLs present in naïve cells lost significance after stimulation.

Most of the studies published so far have been limited to individuals of Caucasian ethnicity. The few available studies involving Asian populations assayed Epstein-Barr virus–transformed lymphoblastoid cell lines developed from individuals collected by the HapMap project[30] and whole blood expression measurements from a small cohort (76 samples) of Japanese individuals[102]. In this small scale study significant eQTLs were detected only for 107 genes. Up to now no study has analyzed the impact of genetic variants on isolated immune cells in Asian populations. Considering the strong genetic differences between Caucasian and Asian populations it is essential to perform similar studies in Asian populations to estimate the percentage of population specific eQTLs.

Other studies have been performed trying to link genetic variants with protein expression (frequently referred to as pQTLs) demonstrating that the effect of genetic variants can be seen not only at the transcriptional level but also at the translational level. Melzer *et al.* measured the levels of 42 proteins in human serum and plasma from 1200 fasting individuals of European ancestry[103]. They identified 8 proteins with a *cis* eQTL. Lourdusamy *et al.* performed a similar study on 96 elderly Europeans. They assayed plasma levels of 778 proteins using a new aptamer-based proteomic technology[104]. They detected *cis*-associations for 60 proteins using a false discovery rate of 5%. 20 of these proteins were previously reported as being associated to diseases.

Epistasis in Immune Cells

The immune system carries out his functions through a complex network of diverse immune cell subsets. Effector function in response to stimuli is attained through the cross-talk between different cells and between distinct pathways inside a particular cell. Similar stimuli can, through the activation of diverse cascades of molecules, lead to different outcomes[18]. Because of its complexity and its functional redundancy the immune system is a good substrate for the study of epistasis in humans.

Due to the computational and statistical complexity associated to the detection of epistasis on a genome-wide scale most published works concerning epistatic interactions involving immune cells or playing a role in immune-related diseases have so far been limited to candidate genes or to variants that independently showed an effect on the trait or disease under consideration (main effect).

For example epistasis was detected between two genetic variants of the transcription factor ETS1 with strong main effects in the context of systemic lupus erythematosus (SLE)[105]. These two polymorphisms, previously shown as being associated with SLE in a Hong Kong Chinese population and three follow up cohorts (Anhui, Shanghai, Thailand)[106] and in an independent GWAS Chinese study[107], were shown by Zhang *et al.* to synergistically affect IL-17 production in SLE patients[105].

A complementary SLE study focused on interactions between genes involved in the mediation of lymphocyte B and T responses previously reported as possible GWAS loci for the disease[108]. The

study identified an interaction between TRAF1 and TNFAIP3 in both a Chinese and Caucasian cohort[109].

Martin *et al.* identified an epistatic interaction between KIR3DS1 and HLA-B delaying progression to depletion of CD4+ T cells in AIDS[110]. Both KIR3DS1 and HLA-B are good candidate genes for AIDS susceptibility.

Julià *et al.* performed a genome-wide epistatic association study using a Spanish cohort of rheumatoid arthritis cases and controls[111]. While none of the SNPs pairs reached genome-wide significance several of them were close to significance. Tao *et al.* performed a similar study for prostate cancer. They used a two-stage scan. In the first stage they tested all pairwise SNP combinations for association with the disease using a cohort of 1176 cases and 1101 controls. None of the SNPs reached genome-wide significance. The top interactions were evaluated in a different cohort of 1964 cases and 3172 controls. They identified 16 nominally significant SNPs (cutoff 0.01) but none of the interactions withstood multiple testing correction. Another genome-wide epistatic study was performed by Wei *et al.* using an Italian cohort[112]. The group correlated pairs of genetic variants with human serum uric acid levels. Uric acid is a key danger signal released by damaged cells[113]. Even in this case no pair reached significance after Bonferroni correction for multiple testing. Nevertheless the group identified some interesting candidate pairs. In particular, the solute carrier SLC2A9 was reported as interacting with multiple other genes across the genome. Two interactions

involving this gene reached the conventional genome-wide threshold ($5x10^{-8}$).

These three genome-wide studies exemplify the statistical issues associated to this kind of analysis. The large number of tests that need to be performed for genome-wide epistatic analyses requires much larger cohorts than those usually available.

For quantitative phenotypes one of the most interesting genome-wide epistatic analyses in the context of immune cells was performed by Becker *et al.*[114] on gene expression measured in 210 healthy individuals selected from HapMap[94]. The authors correlated pairs of genetic variants with gene expression in lymphoblastoid cell lines focusing on *cis-trans* epistatic eQTLs. For each expressed gene only pairs of variants where one SNP was located in proximity of the gene or on it (*cis*) and one SNP was distally located from the gene (*trans*) were considered. They reported *cis-trans* interactions for about 15% of the genes analyzed after Bonferroni correction and showed that *cis* variants with marginal effects were more likely to be part of an interaction. Even so the majority of *cis* variants involved in epistatic interactions showed no direct association with expression.

Neutrophils as a candidate cell subset for studying the role of SNP association in gene regulation

Neutrophils are the most abundant leukocyte subset in blood and constitute the first mechanism of defense against invading pathogens[115–117]. Neutrophils are essential for innate immunity. As opposed to adaptive immune cell subsets their response to infection

is immediate and does not depend on any previous exposure to microbes[118]. Neutrophils can act as phagocytic cells by engulfing bacteria and rapidly killing them with lytic enzymes stored in granules, oxygen reactive species and other antimicrobial proteins[119,120]. Bactericidal factors can also be released into the extracellular medium by degranulation[121]. Another mechanism often used by neutrophils to kill bacteria extracellularly are neutrophil extracellular traps, structures composed of granule proteins and chromatin. These extracellular matrices bind Gram-positive and Gram–negative bacteria preventing them from spreading and incapacitating them[119].

Pathological reduction in neutrophil numbers and defects in neutrophil signaling, intracellular killing and granule formation are responsible for multiple immunodeficiency syndromes, all characterized by an increased susceptibility to microbial and fungal infections[122].

In spite of the central role played by neutrophils in innate immunity no eQTL study was so far performed on this cell type. Neutrophils are particularly difficult to study because of their strong sensitivity to *ex-vivo* manipulation. These cells are in fact short-lived and easily activated[123,124]. Neutrophil mRNA has also been shown to be difficult to isolate because of the high lytic enzyme content of their lysosomal granules[125].

Because of their key role in innate immunity a better understanding of how genetic variants affect gene expression of important neutrophil mediators would be extremely valuable.

Allergic rhinitis and the immunosuppressive gene CD39 as candidates for studying the role of SNP association and epistasis in immune cells

As previously mentioned, the immune system is a complex structure that involves many different players. Due to its complexity the relationship between genotype and phenotype cannot be easily elucidated at the system level but needs to be tackled at a smaller scale. For this particular project we focused on an immune disease characterized by the interplay of multiple immune cell subsets, allergic rhinitis (AR).

Allergic rhinitis is an IgE-mediated allergic disease accompanied by the development of specific symptoms (rhinorrhea, nasal itching, sneezing and progressive blockage of the nasal passages) in response to allergen stimulation[126]. Epidemiological studies from several countries have indicated that the prevalence of AR is increasing worldwide and that the disease now affects between 10-30% of the total world population[127–129]. Prevalence is even higher in Singapore where this study took place. Andiappan *et al.* recently reported an AR rate higher than 40% for Singapore-born Chinese[130]. The city of Singapore therefore constitutes an optimal environment for the study of AR because of the high prevalence of the disease and because the allergic response is dominated by a single allergen class (dust mite)[130]. As a result the study population is strongly homogenous since the disease is always triggered by the same antigen.

While environmental factors clearly exert a strong influence on risk of allergic diseases, family-based and twin studies have established that there is also a genetic component for allergy prevalence[131–137]. However, linking genotype variants with phenotypic manifestations of allergic diseases has not been straightforward: as in most other complex diseases, gene-gene and gene-environment interactions complicate the identification of direct associations[138–140].

Allergic airway diseases are mediated by Th2 cytokines including IL-4, IL-5 and IL-13 which drive eosinophil infiltration to the nasal mucosa, promote IgE switching and stimulate mast cells to release important inflammatory mediators such as histamine, leukotrienes and prostaglandins[141–146]. Monocytes, important regulators of inflammation, can exhibit pro- as well as anti-inflammatory properties in the context of allergies and can strongly influence the course of the allergic reaction. Pro-inflammatory monocytes enhance the allergic reaction by producing cytokines like TNF-α and IL-6 but can be converted, under the influence of basophils, into anti-inflammatory monocytes to attenuate the reaction[147].

T regulatory cells, a subset of CD4+ lymphocytes with suppressive functions, also play an essential role in dampening the allergic inflammation by inhibiting CD4+ T effector proliferation and repressing the production of inflammatory Th2 cytokines (Figure 6)[148–151].

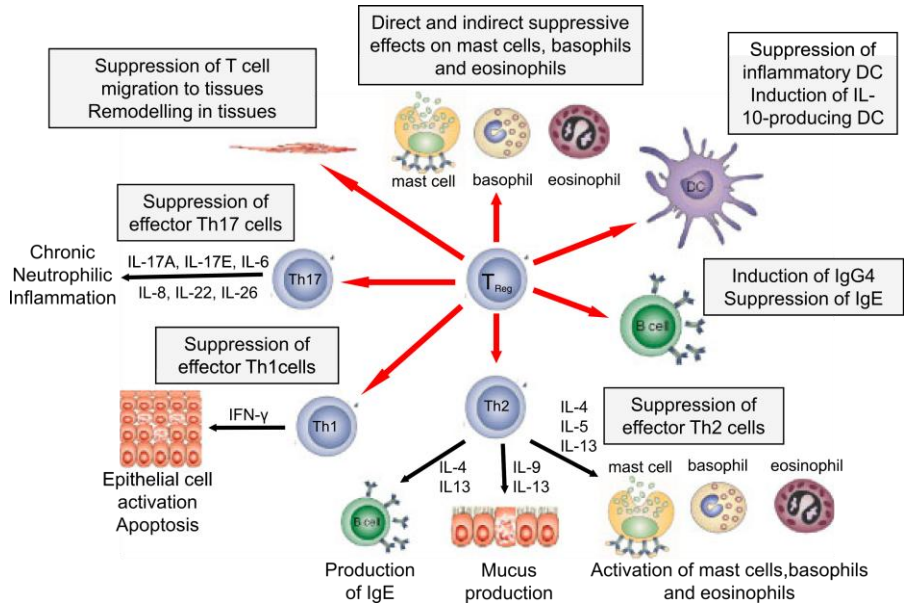


Figure 6. The role of Treg cells in allergic diseases[148]

Treg cells in fact maintain peripheral tolerance against self antigens and inhibit excessive inflammatory responses against dangerous pathogens[152–154]. It is not exactly clear how Treg cells control immune responses. Several mechanisms of suppression have been described and might reflect the heterogeneity of Treg subsets[155]. It is likely that immune tolerance and homeostasis is achieved through a combination of suppressive means including cell contact and secretion of inhibitory molecules (Figure 7). Molecules expressed on the surface of Treg cells such as cytotoxic T-lymphocyte-associated protein 4 (CTLA-4), lymphocyte-activation gene 3 (LAG3) and TGF-β expressed at cellular membrane are key players in the context of contact-dependent suppression[153,156,157]. CTLA-4 is constitutively expressed on Treg cells as opposed to naïve T cells, where it becomes expressed only upon stimulation. CTLA-4 suppresses T cells responses

both by down-regulating the expression of the co-stimulatory molecules CD80 and CD86 on antigen presenting cells, by promoting secretion of the immunosuppressive cytokine TGF-β by both naïve and CD4+ effector T cells and by enhancing activity of the immunosuppressive enzyme IDO on dendritic cells and monocytes[158–161]. LAG3, an adhesion molecule which binds to the major histocompatibility complex class II, is expressed on the surface of Treg cells upon activation and contributes to the suppressive function of these cells[156]. Surface-bound TGF-β is also highly expressed by Treg cells upon stimulation by antigen presenting cells and might mediate immune suppression of both T cell and B cell functions via interaction with the corresponding receptor on target cells[157].

Another mechanism through which Treg cells maintains immune homeostasis is the production of the immunosuppressive cytokines IL-10 and TGF-β[155]. Treg cells also perform cytotoxic functions by releasing perforin and granzyme A which promote monocytes, DCs and T cells death[162]. Another important but frequently ignored mechanism through which Treg cells exert their suppressive function is the hydrolysis of the pro-inflammatory signal adenosine triphosphate (ATP) by the ectonucleoside triphosphate diphosphohydrolase 1 (CD39)[163,164]. ATP is a key player in energy metabolism and it is present in all cells of the human body[165].

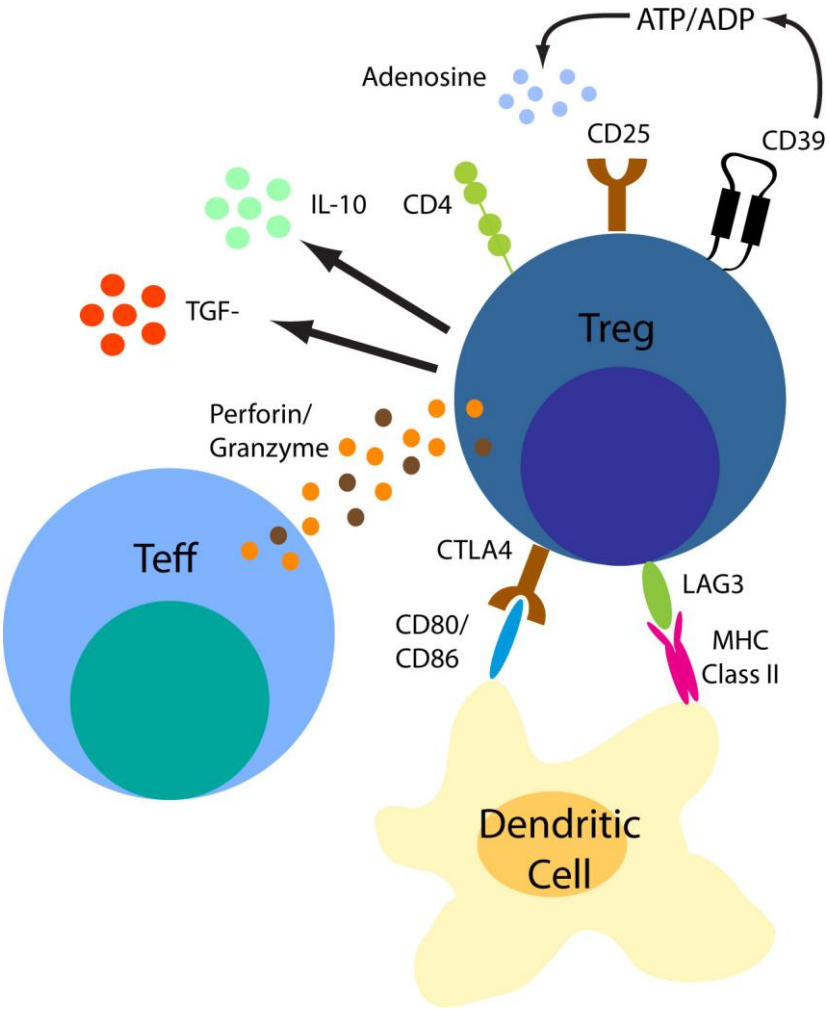


Figure 7. Mechanisms through which T regulatory cells maintain immune homeostasis.

Upon tissue damage or pathogen encounter ATP can leak in the extracellular compartment where it acts as a danger signal for the initiation of an immune response[166,167]. Extracellular ATP is therefore a potent damage-associated molecular pattern which induces a number of pro-inflammatory responses including leukocyte chemotaxis, dendritic cell maturation, and inflammasome-mediated

production of IL-1β[163,168,169]. ATP receptors such as P2X- and P2Y have been shown to be important mediators for the allergic airway inflammation. Blocking of these receptors in experimental asthma models was in fact shown to attenuate the allergic reaction[170,171].

CD39 is a member of the ectonucleoside triphosphate diphosphohydrolase (E-NTPDase) family that hydrolyzes extracellular ATP and adenosine diphosphate (ADP) into adenosine monophosphate (AMP). It also acts in concert with another ectonucleotidase (CD73) to produce adenosine[164], a nucleoside with suppressive and anti-proliferative properties[165,172]. Removal of inflammatory ATP by Treg-expressed CD39 contributes to the control of the inflammatory reaction[163]. However, the catalytic activity of CD39 is strongly influenced by the expression level of this protein on the cell surface of human Treg, which is highly variable between individuals[163].

The importance of CD39 expression on Treg cells for immune regulation is supported by multiple studies. Differences in CD39+ Treg cell percentages have in fact been reported for many complex immune and infectious diseases such as multiple sclerosis[163,173], hepatitis B[174], hepatitis C[175] and HIV-1[176]. Moreover CD39 genetic variants have already been associated with immune phenotypes such as susceptibility to inflammatory bowel disease[177] and HIV progression[176].

Because of its role in immunosuppression and its involvement in multiple immune diseases, CD39 is a strong gene candidate for the study of association and epistasis in the context of allergic rhinitis.

SCOPE OF THE THESIS

The scope of this project is to study the relationship between genotype and phenotype in immune cells both in the context of association and epistasis. Since the majority of GWAS and eQTL studies performed so far have been centered on individuals of European descent we decided to target a cohort of Chinese ethnicity to advance the genetic characterization of this population.

Due to the complexity of the immune system and the impossibility of tackling this question at the system level our study focused on two well-defined questions.

The first aim of our project was to characterize the eQTL landscape of resting neutrophils, a cell subset strongly involved in innate immunity. Neutrophils have so far not been studied in the context of eQTLs due to the technical difficulties associated with handling these cells which are considered fragile and have been shown to be easily activated[123,178]. We therefore carefully isolated neutrophils from 114 well-matched individuals of Chinese ethnicity and performed a genome-wide *cis* eQTL analysis to evaluate the impact of genetic variants on transcriptional regulation.

The second aim of our project was to study the role played by association and epistasis in a complex immune disease with a strong genetic component. Due to the availability of a Singapore Chinese cohort of allergic rhinitis patients and matched controls our analysis targeted this disease which is characterized by the interplay of multiple immune cell subsets. Allergic rhinitis has been shown to have a strong genetic component both by family-based and twin

studies and is therefore a good candidate for studying the impact of genetic variants on disease susceptibility[131–137].

Since the size of the cohort did not allow for a genome-wide analysis we centered our genetic study on a gene with a strong role in immune suppression, CD39. CD39 is a key inhibitory molecule used by T regulatory cells, a subset of CD4+ lymphocytes involved in immune regulation. T regulatory cells have been shown to be essential for controlling the allergic reaction upon allergen exposure[148,150,179]. Mutations disrupting the function of CD39 might therefore have an impact on the suppressive ability of these cells and consequently affect susceptibility to allergic rhinitis.

For tackling these specific questions it became essential to develop tools for identifying causative variants from disease-associated polymorphisms. These polymorphisms are in fact rarely causative: they simply tag causative variants through linkage disequilibrium (LD). The state-of-the-art tools for the visualization of linkage disequilibrium in human populations did not provide a way of integrating LD information with additional biological data that would help in identifying functional variants (for example open-chromatin regions, transcription factors binding sites, GWAS hits or previous published findings linked to a particular polymorphism). It therefore became necessary, for the interpretation of association and epistatic results, to develop a tool that would facilitate this integration. Thus, another important goal for this thesis became the development of a user-friendly tool for combining linkage disequilibrium plots with additional biological information with the aim of helping bridging the

gap between statistical association/epistasis and underlying biological mechanisms.

This thesis is organized as follows. Chapter 1 contains a broad introduction to the subject. Chapter 2 illustrates the genome-wide eQTL analysis performed on neutrophils isolated from 114 Singapore Chinese individuals. Chapter 3 focuses on the role played by CD39 polymorphisms in allergic rhinitis. Chapter 4 describes the software we developed for supporting the identification of causative variants from trait-associated tag variants. Chapter 5 summarizes the results and discusses future directions for the project.

REFERENCES

1. Venter, J. C. *et al.* The sequence of the human genome. *Science* **291,** 1304–51 (2001).
2. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921 (2001).
3. Malkova, A. & Haber, J. E. Mutations arising during repair of chromosome breaks. *Annu. Rev. Genet.* **46,** 455–73 (2012).
4. The 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491,** 56–65 (2012).
5. Giresi, P., Kim, J. & McDaniell, R. FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. *Genome Res.* **17,** 877–885 (2007).
6. Dunham, I. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489,** 57–74 (2012).
7. Conde, L., Bracci, P. M., Richardson, R., Montgomery, S. B. & Skibola, C. F. Integrating GWAS and Expression Data for Functional Characterization of Disease-Associated SNPs: An Application to Follicular Lymphoma. *Am. J. Hum. Genet.* **92,** 126–30 (2013).
8. Teare, M. D. & Barrett, J. Genetic linkage studies. *Lancet* **366,** 1036–1044 (2005).
9. Hirschhorn, J. N. & Daly, M. J. Genome-wide association studies for common diseases and complex traits. *Nat. Rev. Genet.* **6,** 95–108 (2005).

10. Hindorff, L. *et al.* A Catalog of Published Genome-Wide Association Studies. at <www.genome.gov/gwastudies>
11. Need, A. C. & Goldstein, D. B. Next generation disparities in human genomics: concerns and remedies. *Trends Genet.* **25,** 489–94 (2009).
12. Yang, X. *et al.* Genome-wide association study for serum complement C3 and C4 levels in healthy Chinese subjects. *PLoS Genet.* **8,** e1002916 (2012).
13. Manolio, T. A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461,** 747–753 (2009).
14. Visscher, P. M., Hill, W. G. & Wray, N. R. Heritability in the genomics era--concepts and misconceptions. *Nat. Rev. Genet.* **9,** 255–66 (2008).
15. Page, G. P., George, V., Go, R. C., Page, P. Z. & Allison, D. B. "Are we there yet?": Deciding when one has demonstrated specific genetic causation in complex diseases and quantitative traits. *Am. J. Hum. Genet.* **73,** 711–9 (2003).
16. Ioannidis, J. P. A. Non-replication and inconsistency in the genome-wide association setting. *Hum. Hered.* **64,** 203–213 (2007).
17. Cordell, H. J. Detecting gene-gene interactions that underlie human diseases. *Nat. Rev. Genet.* **10,** 392–404 (2009).
18. Rose, A. M. & Bell, L. C. K. Epistasis and immunity: the role of genetic interactions in autoimmune diseases. *Immunology* **137,** 131–8 (2012).

19. Hunter, D. J. Gene-environment interactions in human diseases. *Nat. Rev. Genet.* **6,** 287–98 (2005).
20. Zhernakova, A., van Diemen, C. C. & Wijmenga, C. Detecting shared pathogenesis from the shared genetics of immune-related diseases. *Nat. Rev. Genet.* **10,** 43–55 (2009).
21. Bonen, D. K. & Cho, J. H. The genetics of inflammatory bowel disease. *Gastroenterology* **124,** 521–36 (2003).
22. MacGregor, a J. *et al.* Characterizing the quantitative genetic contribution to rheumatoid arthritis using data from twins. *Arthritis Rheum.* **43,** 30–7 (2000).
23. De Craen, a J. M. *et al.* Heritability estimates of innate immunity: an extended twin study. *Genes Immun.* **6,** 167–70 (2005).
24. Cooper, G. S., Miller, F. W. & Pandey, J. P. The role of genetic factors in autoimmune disease: implications for environmental research. *Environ. Health Perspect.* **107 Suppl ,** 693–700 (1999).
25. Cooke, G. S. & Hill, a V. Genetics of susceptibility to human infectious disease. *Nat. Rev. Genet.* **2,** 967–77 (2001).
26. Casadevall, A. & Pirofski, L.-A. Exploiting the redundancy in the immune system: vaccines can mediate protection by eliciting “unnatural” immunity. *J. Exp. Med.* **197,** 1401–4 (2003).
27. Kitano, H. Biological robustness. *Nat. Rev. Genet.* **5,** 826–37 (2004).
28. Nembrini, C., Abel, B., Kopf, M. & Marsland, B. J. Strong TCR Signaling, TLR Ligands, and Cytokine Redundancies Ensure

Robust Development of Type 1 Effector T Cells. *J. Immunol.* **176,** 7180–7188 (2006).

29. Gilad, Y., Rifkin, S. a & Pritchard, J. K. Revealing the architecture of gene regulation: the promise of eQTL studies. *Trends Genet.* **24,** 408–15 (2008).
30. Stranger, B. E. *et al.* Population genomics of human gene expression. *Nat. Genet.* **39,** 1217–24 (2007).
31. Fu, J. *et al.* Unraveling the regulatory mechanisms underlying tissue-dependent genetic variation of gene expression. *PLoS Genet.* **8,** e1002431 (2012).
32. Stranger, B. E. *et al.* Genome-Wide Associations of Gene Expression Variation in Humans. *PLoS Genet.* **1,** 10 (2005).
33. Dixon, A. L. *et al.* A genome-wide association study of global gene expression. *Nat. Genet.* **39,** 1202–7 (2007).
34. Cookson, W., Liang, L., Abecasis, G., Moffatt, M. & Lathrop, M. Mapping complex disease traits with global gene expression. *Nat. Rev. Genet.* **10,** 184–194 (2009).
35. Hsu, Y.-H. *et al.* 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,** 16 (2010).
36. Nicolae, D. L. *et al.* Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. *PLoS Genet.* **6,** e1000888 (2010).

37. Ritchie, M. D. Using biological knowledge to uncover the mystery in the search for epistasis in genome-wide association studies. *Ann. Hum. Genet.* **75,** 172–182 (2011).

38. Mehta, D. *et al.* Impact of common regulatory single-nucleotide variants on gene expression profiles in whole blood. *Eur. J. Hum. Genet.* **21,** 48–54 (2013).

39. Fairfax, B. P. *et al.* Genetics of gene expression in primary immune cells identifies cell type-specific master regulators and roles of HLA alleles. *Nat. Genet.* **44,** 502–510 (2012).

40. Veyrieras, J.-B. *et al.* High-Resolution Mapping of Expression-QTLs Yields Insight into Human Gene Regulation. *PLoS Genet.* **4,** 15 (2008).

41. Nica, A. C. *et al.* The architecture of gene regulatory variation across multiple human tissues: the MuTHER study. *PLoS Genet.* **7,** e1002003 (2011).

42. Dimas, A. S. *et al.* Common regulatory variation impacts gene expression in a cell type-dependent manner. *Science* **325,** 1246–50 (2009).

43. Westra, H.-J. *et al.* Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nat. Genet.* **45,** 1238–1243 (2013).

44. Zeller, T. *et al.* Genetics and beyond-the transcriptome of human monocytes and disease susceptibility. *PLoS One* **5,** e10693 (2010).

45. Nakagawa, S. & Cuthill, I. C. Effect size, confidence interval and statistical significance: a practical guide for biologists. *Biol. Rev. Camb. Philos. Soc.* **82,** 591–605 (2007).
46. Pearson, K. On the criterion that a given system of deviations from the probable in the case of correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philos. Mag.* **50,** 157–175 (1900).
47. Wang, K. Statistical tests of genetic association for case-control study designs. *Biostatistics* **13,** 724–33 (2012).
48. Clarke, G. M. *et al.* Basic statistical analysis in genetic case-control studies. *Nat. Protoc.* **6,** 121–33 (2011).
49. Szumilas, M. Explaining odds ratios. *J. Can. Acad. Child Adolesc. Psychiatry* **19,** 227–9 (2010).
50. Bateson, W. Facts limiting the theory of heredity. *Science* **26,** 649–660 (1907).
51. Moore, J. H. & Williams, S. M. Traversing the conceptual divide between biological and statistical epistasis: systems biology and a more modern synthesis. *BioEssays news Rev. Mol. Cell. Dev. Biol.* **27,** 637–646 (2005).
52. Tyler, A. L., Asselbergs, F. W., Williams, S. M. & Moore, J. H. Shadows of complexity: what biological networks reveal about epistasis and pleiotropy. *BioEssays news Rev. Mol. Cell. Dev. Biol.* **31,** 220–7 (2009).
53. Lehner, B., Crombie, C., Tischler, J., Fortunato, A. & Fraser, A. G. Systematic mapping of genetic interactions in

Caenorhabditis elegans identifies common modifiers of diverse signaling pathways. *Nat. Genet.* **38,** 896–903 (2006).

54. Moore, J. H. A global view of epistasis. *Nat. Genet.* **37,** 13–14 (2005).

55. Moore, J. H. The ubiquitous nature of epistasis in determining susceptibility to common human diseases. *Hum. Hered.* **56,** 73–82 (2003).

56. Fisher, R. XV.—The Correlation between Relatives on the Supposition of Mendelian Inheritance. *Trans. R. Soc. Edinburgh* **52,** 399–433 (1919).

57. Wilson, S. Epistasis and its possible effects on transmission disequilibrium tests. *Ann. Hum. Genet.* **65,** 565–575 (2001).

58. Morris, a & Whittaker, J. Generalization of the extended transmission disequilibrium test to two unlinked disease loci. *Genet. Epidemiol.* **17 Suppl 1,** S661–6 (1999).

59. Moore, J. H. & Williams, S. M. Epistasis and its implications for personal genetics. *Am. J. Hum. Genet.* **85,** 309–320 (2009).

60. Li, W. & Reich, J. A complete enumeration and classification of two-locus disease models. *Hum. Hered.* **50,** 334–349 (1999).

61. Dong, C. *et al.* Exploration of gene-gene interaction effects using entropy-based methods. *Eur. J. Hum. Genet. EJHG* **16,** 229–35 (2008).

62. Culverhouse, R., Suarez, B. K., Lin, J. & Reich, T. A perspective on epistasis: limits of models displaying no main effect. *Am. J. Hum. Genet.* **70,** 461–471 (2002).

63. Upstill-Goddard, R., Eccles, D., Fliege, J. & Collins, A. Machine learning approaches for the discovery of gene-gene interactions in disease data. *Brief. Bioinform.* **14,** (2012).

64. Gayán, J. *et al.* A method for detecting epistasis in genome-wide studies using case-control multi-locus association analysis. *BMC Genomics* **9,** 360 (2008).

65. Segrè, D., Deluna, A., Church, G. M. & Kishony, R. Modular epistasis in yeast metabolism. *Nat. Genet.* **37,** 77–83 (2005).

66. Khan, A. I., Dinh, D. M., Schneider, D., Lenski, R. E. & Cooper, T. F. Negative epistasis between beneficial mutations in an evolving bacterial population. *Science* **332,** 1193–6 (2011).

67. Remold, S. K. & Lenski, R. E. Pervasive joint influence of epistasis and plasticity on mutational effects in Escherichia coli. *Nat. Genet.* **36,** 423–6 (2004).

68. Boone, C., Bussey, H. & Andrews, B. J. Exploring genetic interactions and networks with yeast. *Nat. Rev. Genet.* **8,** 437–49 (2007).

69. Shimomura, K. *et al.* Genome-wide epistatic interaction analysis reveals complex genetic determinants of circadian behavior in mice. *Genome Res.* **11,** 959–980 (2001).

70. Tao, S. *et al.* Genome-wide two-locus epistasis scans in prostate cancer using two European populations. *Hum. Genet.* **131,** 1225–34 (2012).

71. Ritchie, M. D. *et al.* Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism

genes in sporadic breast cancer. *Am. J. Hum. Genet.* **69,** 138–147 (2001).

72. Pessi, T. *et al.* CRP and FCGR2A genes have an epistatic effect on carotid artery intima-media thickness: the Cardiovascular Risk in Young Finns Study. *Int. J. Immunogenet.* **36,** 39–45 (2009).

73. Howson, J. M. M. *et al.* Evidence of Gene-Gene Interaction and Age-at-Diagnosis Effects in Type 1 Diabetes. *Diabetes* **61,** 3012–7 (2012).

74. Eichler, E. E. *et al.* Missing heritability and strategies for finding the underlying causes of complex disease. *Nat. Rev. Genet.* **11,** 446–450 (2010).

75. Zuk, O., Hechter, E., Sunyaev, S. R. & Lander, E. S. The mystery of missing heritability: Genetic interactions create phantom heritability. *Proc. Natl. Acad. Sci.* **109,** 1193–1198 (2012).

76. Ben Lehner. Molecular mechanisms of epistasis within and between genes. *Trends Genet.* **27,** 323–331 (2011).

77. Baessler, A. *et al.* Epistatic interaction between haplotypes of the ghrelin ligand and receptor genes influence susceptibility to myocardial infarction and coronary artery disease. *Hum. Mol. Genet.* **16,** 887–99 (2007).

78. Papp, B., Pál, C. & Hurst, L. D. Metabolic network analysis of the causes and evolution of enzyme dispensability in yeast. *Nature* **429,** 661–664 (2004).

79. Costanzo, M. *et al.* The genetic landscape of a cell. *Science* **327,** 425–31 (2010).

80. Bershtein, S., Segal, M., Bekerman, R., Tokuriki, N. & Tawfik, D. S. Robustness-epistasis link shapes the fitness landscape of a randomly drifting protein. *Nature* **444,** 929–32 (2006).

81. Knol, M. J., van der Tweel, I., Grobbee, D. E., Numans, M. E. & Geerlings, M. I. Estimating interaction on an additive scale between continuous determinants in a logistic regression model. *Int. J. Epidemiol.* **36,** 1111–8 (2007).

82. Tang, W., Wu, X., Jiang, R. & Li, Y. Epistatic Module Detection for Case-Control Studies: A Bayesian Model with a Gibbs Sampling Strategy. *PLoS Genet.* **5,** 18 (2009).

83. Emily, M., Mailund, T., Hein, J., Schauser, L. & Schierup, M. H. Using biological networks to search for interacting loci in genome-wide association studies. *Eur. J. Hum. Genet.* **17,** 1231–1240 (2009).

84. Benjamini, Y. & Yekutieli, D. The control of the false discovery rate in multiple testing under dependency. *Ann. Stat.* **29,** 1165–1188 (2001).

85. Pesarin, F. *Multivariate Permutation Tests: With Applications to Biostatistics*. *Book* (Wiley, 2001).

86. Li, M., Romero, R., Fu, W. J. & Cui, Y. Mapping Haplotype-haplotype Interactions with Adaptive LASSO. *BMC Genet.* **11,** 79 (2010).

87. Zhang, X., Huang, S., Zou, F. & Wang, W. Tools for efficient epistasis detection in genome-wide association study. *Source Code Biol. Med.* **6,** 1 (2011).

88. Milne, R. L., Fagerholm, R., Nevanlinna, H. & Benítez, J. The importance of replication in gene-gene interaction studies: multifactor dimensionality reduction applied to a two-stage breast cancer case-control study. *Carcinogenesis* **29,** 1215–8 (2008).

89. Ardlie, K. G., Kruglyak, L. & Seielstad, M. Patterns of linkage disequilibrium in the human genome. *Nat. Rev. Genet.* **3,** 299–309 (2002).

90. Stephens, M. & Scheet, P. Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. *Am. J. Hum. Genet.* **76,** 449–62 (2005).

91. Barrett, J. C., Fry, B., Maller, J. & Daly, M. J. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* **21,** 263–265 (2005).

92. The International HapMap Consortium. The International HapMap Project. *Nature* **426,** 789–796 (2003).

93. The International HapMap Consortium. A haplotype map of the human genome. *Nature* **437,** 1299–1320 (2005).

94. The International HapMap Consortium. A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449,** 851–861 (2007).

95. The 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature* **467,** 1061–1073 (2010).

96. Johnson, A. D. *et al.* SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap. *Bioinformatics* **24,** 2938–9 (2008).

97. Murphy, A. *et al.* Mapping of numerous disease-associated expression polymorphisms in primary peripheral blood CD4+ lymphocytes. *Hum. Mol. Genet.* **19,** 4745–57 (2010).

98. Ferraro, a. *et al.* Interindividual variation in human T regulatory cells. *Proc. Natl. Acad. Sci.* **111,** E1111–E1120 (2014).

99. Barreiro, L. & Tailleux, L. Deciphering the genetic architecture of variation in the immune response to Mycobacterium tuberculosis infection. *Proc. Natl. Acad. Sci.* **109,** 1204–1209 (2012).

100. Lee, M. N. *et al.* Common genetic variants modulate pathogen-sensing responses in human dendritic cells. *Science* **343,** 1246980 (2014).

101. Fairfax, B. P. *et al.* Innate immune activity conditions the effect of regulatory variants upon monocyte gene expression. *Science* **343,** 1246949 (2014).

102. Sasayama, D. *et al.* Identification of single nucleotide polymorphisms regulating peripheral blood mRNA expression with genome-wide significance: an eQTL study in the Japanese population. *PLoS One* **8,** e54967 (2013).

103. Melzer, D. *et al.* A genome-wide association study identifies protein quantitative trait loci (pQTLs). *PLoS Genet.* **4,** e1000072 (2008).

104. Lourdusamy, A. *et al.* Identification of cis-regulatory variation influencing protein abundance levels in human plasma. *Hum. Mol. Genet.* **21,** 3719–26 (2012).

105. Zhang, J. *et al.* Epistatic Interaction between Genetic Variants in Susceptibility Gene ETS1 Correlates with IL-17 Levels in SLE Patients. *Ann. Hum. Genet.* **1,** 1–7 (2013).

106. Yang, W. *et al.* Genome-wide association study in Asian populations identifies variants in ETS1 and WDFY4 associated with systemic lupus erythematosus. *PLoS Genet.* **6,** e1000841 (2010).

107. Han, J.-W. *et al.* Genome-wide association study in a Chinese Han population identifies nine new susceptibility loci for systemic lupus erythematosus. *Nat. Genet.* **41,** 1234–7 (2009).

108. Gregersen, P. & Olsson, L. Recent advances in the genetics of autoimmune disease. *Annu. Rev. Immunol.* **27,** 363–391 (2009).

109. Zhou, X. *et al.* Gene-gene interaction of BLK, TNFSF4, TRAF1, TNFAIP3, and REL in systemic lupus erythematosus. *Arthritis Rheum.* **64,** 222–31 (2012).

110. Martin, M. P. *et al.* Epistatic interaction between KIR3DS1 and HLA-B delays the progression to AIDS. *Nat. Genet.* **31,** 429–34 (2002).

111. Julià, A. *et al.* Genome-wide association study of rheumatoid arthritis in the Spanish population: KLF12 as a risk locus for rheumatoid arthritis susceptibility. *Arthritis Rheum.* **58,** 2275–86 (2008).

112. Wei, W. *et al.* Characterisation of genome-wide association epistasis signals for serum uric acid in human population isolates. *PLoS One* **6,** e23836 (2011).

113. Shi, Y., Evans, J. E. & Rock, K. L. Molecular identification of a danger signal that alerts the immune system to dying cells. *Nature* **425,** 516–521 (2003).

114. Becker, J., Wendland, J. R., Haenisch, B., Nöthen, M. M. & Schumacher, J. A systematic eQTL study of cis-trans epistasis in 210 HapMap individuals. *Eur. J. Hum. Genet.* **20,** 97–101 (2011).

115. Mollinedo, F., Borregaard, N. & Boxer, L. Novel trends in neutrophil structure, function and development. *Immunol. Today* 535–537 (1999).

116. Von Vietinghoff, S. & Ley, K. Homeostatic Regulation of Blood Neutrophil Counts. *J. Immunol.* **181,** 5183–5188 (2008).

117. Nathan, C. Neutrophils and immunity: challenges and opportunities. *Nat. Rev. Immunol.* **6,** 173–82 (2006).

118. Kobayashi, S. D. & Deleo, F. R. Towards a comprehensive understanding of the role of neutrophils in innate immunity: a systems biology-level approach. *Wiley Interdiscip Rev Syst Biol Med* **1,** 309–333 (2012).

119. Brinkmann, V. *et al.* Neutrophil extracellular traps kill bacteria. *Science* **303,** 1532–5 (2004).

120. Mantovani, A., Cassatella, M. a, Costantini, C. & Jaillon, S. Neutrophils in the activation and regulation of innate and adaptive immunity. *Nat. Rev. Immunol.* **11,** 519–31 (2011).

121. Gallin, J. I., Goldstein, I. M. & Snyderman, R. *Inflammation: Basic Principles and Clinical Correlate*. 511–539 (Raven Press, 1992).

122. Mackay, I. Immunodeficiency diseases caused by defects in phagocytes. *N. Engl. J. Med.* **343,** 1703–1714 (2000).

123. Pillay, J. *et al.* A subset of neutrophils in human systemic inflammation inhibits T cell responses through Mac-1. *J. Clin. Invest.* **122,** 327–336 (2012).

124. Beutler, B. Innate immunity: an overview. *Mol. Immunol.* **40,** 845–859 (2004).

125. Fulvio, M. Di & Gomez-cambronero, J. Phospholipase D (PLD) gene expression in human neutrophils and HL-60 differentiation. *J. Leukoc. Biol.* **77,** 999–1007 (2005).

126. Bousquet, J. *et al.* Allergic Rhinitis and its Impact on Asthma (ARIA) 2008 update (in collaboration with the World Health Organization, GA(2)LEN and AllerGen). *Allergy* **63,** 8–160 (2008).

127. Pawankar, R., Canonica, G. W., Holgate, S. T. & Lockey, R. F. *WAO White Book on Allergy*. (2011).

128. Eriksson, J. *et al.* Update of prevalence of self-reported allergic rhinitis and chronic nasal symptoms among adults in Sweden. *Clin Respir J* **6,** 159–168 (2012).

129. De Marco, R. *et al.* Trends in the prevalence of asthma and allergic rhinitis in Italy between 1991 and 2010. *Eur Respir J* **39,** 883–892 (2012).

130. Andiappan, A. K. *et al.* Allergic airway diseases in a tropical urban environment are driven by dominant mono-specific sensitization against house dust mites. *Allergy* (2014). doi:10.1111/all.12364

131. Dold, S., Wjst, M., von Mutius, E., Reitmeir, P. & Stiepel, E. Genetic risk for asthma, allergic rhinitis, and atopic dermatitis. *Arch Dis Child* **67,** 1018–1022 (1992).

132. Duffy, D. L., Martin, N. G., Battistutta, D., Hopper, J. L. & Mathews, J. D. Genetics of asthma and hay fever in Australian twins. *Am Rev Respir Dis* **142,** 1351–1358 (1990).

133. Räsäsnen, M., Laitinen, T., Kaprio, J., Koskenvuo, M. & Laitinen, L. A. Hay fever-a Finnish nationwide study of adolescent twins and their parents. *Allergy* **53,** 885–890 (1998).

134. Lichtenstein, P. & Svartengren, M. Genes, environments, and sex: factors of importance in atopic diseases in 7-9-year-old Swedish twins. *Allergy* **52,** 1079–1086 (1997).

135. Fagnani, C. *et al.* Heritability and shared genetic effects of asthma and hay fever: an Italian study of young twins. *Twin Res Hum Genet* **11,** 121–131 (2008).

136. Van Beijsterveldt, C. E. M. & Boomsma, D. I. Genetics of parentally reported asthma, eczema and rhinitis in 5-yr-old twins. *Eur Respir J* **29,** 516–21 (2007).

137. Thomsen, S. F. *et al.* Findings on the atopic triad from a Danish twin registry. *Int J Tuberc Lung Dis* **10,** 1268–72 (2006).

138. Maksimovic, N. *et al.* Risk factors of allergic rhinitis: a case-control study. *HealthMed* **4,** 63–70 (2010).

139. Ng, T. P. & Tan, W. C. Epidemiology of allergic rhinitis and its associated risk factors in Singapore. *Int J Epidemiol* **23,** 553–558 (1994).

140. Vercelli, D. Discovering susceptibility genes for asthma and allergy. *Nat. Rev. Immunol.* **8,** 169–182 (2008).

141. Li, L. *et al.* Effects of Th2 cytokines on chemokine expression in the lung: IL-13 potently induces eotaxin expression by airway epithelial cells. *J. Immunol.* **162,** 2477–2487 (1999).

142. Durham, S. R. *et al.* Cytokine messenger RNA expression for IL-3, IL-4, IL-5, and granulocyte/macrophage-colony-stimulating factor in the nasal mucosa after local allergen provocation: relationship to tissue eosinophilia. *J Immunol* **148,** 2390–2394 (1992).

143. Bischoff, S. C. *et al.* IL-4 enhances proliferation and mediator release in mature human mast cells. *Proc. Natl. Acad. Sci.* **96,** 8080–8085 (1999).

144. Murray, J. J. *et al.* Release of prostaglandin D2 into human airways during acute antigen challenge. *N. Engl. J. Med.* **315,** 800–804 (1986).

145. Schleimer, R. P. *et al.* Characterization of inflammatory mediator release from purified human lung mast cells. *Am Rev Respir Dis* **133,** 614–617 (1986).

146. MacGlashan, D. W. *et al.* Generation of leukotrienes by purified human lung mast cells. *J. Clin. Invest.* **70,** 747–751 (1982).

147. Egawa, M. *et al.* Inflammatory monocytes recruited to allergic skin acquire an anti-inflammatory M2 phenotype via basophil-derived interleukin-4. *Immunity* **38,** 570–580 (2013).

148. Palomares, O. *et al.* Role of Treg in immune regulation of allergic diseases. *Eur. J. Immunol.* **40,** 1232–1240 (2010).

149. Ling, E. M. *et al.* Relation of CD4+ CD25+ regulatory T-cell suppression of allergen-driven T-cell activation to atopic status and expression of allergic disease. *Lancet* **363,** 608–615 (2004).

150. Robinson, D. S., Larché, M. & Durham, S. R. Tregs and allergic disease. *J. Clin. Invest.* **114,** 1389–1397 (2004).

151. Bellinghausen, I., Klostermann, B., Knop, J. & Saloga, J. Human CD4+CD25+ T cells derived from the majority of atopic donors are able to suppress TH1 and TH2 cytokine production. *J Allergy Clin Immunol* **111,** 862–868 (2003).

152. Sakaguchi, S. *et al.* Foxp3+ CD25+ CD4+ natural regulatory T cells in dominant self-tolerance and autoimmune disease. *Immunol. Rev.* **212,** 8–27 (2006).

153. Sakaguchi, S. Naturally arising Foxp3-expressing CD25+CD4+ regulatory T cells in immunological tolerance to self and non-self. *Nat. Immunol.* **6,** 345–52 (2005).

154. Bacchetta, R., Gambineri, E. & Roncarolo, M.-G. Role of regulatory T cells and FOXP3 in human diseases. *J. Allergy Clin. Immunol.* **120,** 227–35; quiz 236–7 (2007).

155. Miyara, M. & Sakaguchi, S. Natural regulatory T cells: mechanisms of suppression. *Trends Mol. Med.* **13,** 108–16 (2007).

156. Huang, C.-T. *et al.* Role of LAG-3 in regulatory T cells. *Immunity* **21,** 503–513 (2004).

157. Nakamura, K., Kitani, A. & Strober, W. Cell contact-dependent immunosuppression by CD4(+)CD25(+) regulatory T cells is mediated by cell surface-bound transforming growth factor beta. *J. Exp. Med.* **194,** 629–644 (2001).

158. Chambers, C. a & Allison, J. P. Costimulatory regulation of T cell function. *Curr. Opin. Cell Biol.* **11,** 203–10 (1999).

159. Wing, K. *et al.* CTLA-4 control over Foxp3+ regulatory T cell function. *Science* **322,** 271–5 (2008).

160. Chen, W., Jin, W. & Wahl, S. Engagement of cytotoxic T lymphocyte–associated antigen 4 (CTLA-4) induces transforming growth factor β (TGF-β) production by murine CD4+ T cells. *J. Exp. Med.* **188,** (1998).

161. Miwa, N. *et al.* IDO expression on decidual and peripheral blood dendritic cells and monocytes/macrophages after treatment with CTLA-4 or interferon-gamma increase in normal pregnancy but decrease in spontaneous abortion. *Mol. Hum. Reprod.* **11,** 865–70 (2005).

162. Grossman, W. J. *et al.* Human T regulatory cells can use the perforin pathway to cause autologous target cell death. *Immunity* **21,** 589–601 (2004).

163. Borsellino, G. *et al.* Expression of ectonucleotidase CD39 by Foxp3+ Treg cells: hydrolysis of extracellular ATP and immune suppression. *Blood* **110,** 1225–1232 (2007).

164. Deaglio, S. *et al.* Adenosine generation catalyzed by CD39 and CD73 expressed on regulatory T cells mediates immune suppression. *J. Exp. Med.* **204,** 1257–1265 (2007).

165. Bours, M. J. L., Swennen, E. L. R., Di Virgilio, F., Cronstein, B. N. & Dagnelie, P. C. Adenosine 5'-triphosphate and adenosine as endogenous signaling molecules in immunity and inflammation. *Pharmacol. Ther.* **112,** 358–404 (2006).

166. Virgilio, F. Di *et al.* Extracellular ATP, P2 receptors, and inflammation. *Drug Dev. Res.* **59,** 171–174 (2003).

167. Sala, A. la & Ferrari, D. Alerting and tuning the immune response by extracellular nucleotides. *J. Leukoc. Biol.* **73,** 339–343 (2003).

168. Jacob, F., Novo, C. P., Bachert, C. & Van Crombruggen, K. Purinergic signaling in inflammatory cells: P2 receptor expression, functional effects, and modulation of inflammatory responses. *Purinergic Signal.* **9,** 285–306 (2013).

169. Latz, E., Xiao, T. S. & Stutz, A. Activation and regulation of the inflammasomes. *Nat. Rev. Immunol.* **13,** 397–411 (2013).

170. Kouzaki, H., Iijima, K., Kobayashi, T., O'Grady, S. M. & Kita, H. The danger signal, extracellular ATP, is a sensor for an airborne allergen and triggers IL-33 release and innate Th2-type responses. *J. Immunol.* **186,** 4375–4387 (2011).

171. Idzko, M. *et al.* Extracellular ATP triggers and maintains asthmatic airway inflammation by activating dendritic cells. *Nat. Med.* **13,** 913–919 (2007).

172. Thiel, M., Caldwell, C. C. & Sitkovsky, M. V. The critical role of adenosine A2A receptors in downregulation of inflammation and immunity in the pathogenesis of infectious diseases. *Microbes Infect.* **5,** 515–526 (2003).

173. Fletcher, J. M. *et al.* CD39+Foxp3+ regulatory T cells suppress pathogenic Th17 cells and are impaired in multiple sclerosis. *J. Immunol.* **183,** 7602–7610 (2009).

174. Tang, Y., Jiang, L., Zheng, Y., Ni, B. & Wu, Y. Expression of CD39 on FoxP3+ T regulatory cells correlates with progression of HBV infection. *BMC Immunol.* **13,** 17 (2012).

175. Kared, H., Fabre, T., Bédard, N., Bruneau, J. & Shoukry, N. H. Galectin-9 and IL-21 mediate cross-regulation between Th17 and Treg cells during acute hepatitis C. *PLoS Pathog.* **9,** e1003422 (2013).

176. Nikolova, M. *et al.* CD39/Adenosine Pathway Is Involved in AIDS Progression. *PLoS Pathog.* **7,** e1002110 (2011).

177. Friedman, D. J. *et al.* CD39 deletion exacerbates experimental murine colitis and human polymorphisms increase susceptibility to inflammatory bowel disease. *Proc. Natl. Acad. Sci.* **106,** 16788–16793 (2009).

178. Glasser, L. & Fiederlein, R. L. The effect of various cell separation procedures on assays of neutrophil function. A critical appraisal. *Am. J. Clin. Pathol.* **93,** 662–669 (1990).

179. Nouri-Aria, K. T. & Durham, S. R. Regulatory T cells and allergic disease. *Inflamm Allergy Drug Targets* **7,** 237–252 (2008).

CHAPTER 2

GENOME-WIDE EQTL STUDY OF NEUTROPHILS IN A CHINESE COHORT SUGGESTS A ROLE FOR NEUTROPHIL EQTLS IN DERMATOLOGICAL DISEASES

Rossella Melchiotti[a,b,2], Anand Kumar Andiappan[a,2], Nah Xin Hui[c], Tuang Yeow Poh[a], Kia Joo Puan[a,2], Boris San Luis[a], Bernett Lee[a], Elena Vigano[a], Josephine Lum[a], Anusha Vasudev[a], Alessandra Mortellaro[a], Francesca Zolezzi[a], Anis Larbi[a], Michael Poidinger[a,1], De Yun Wang[d,1] , Olaf Rötzschke[a,1]

[a]SIgN (Singapore Immunology Network), A*STAR (Agency for Science, Technology and Research), Singapore 138648, Singapore; [b]Doctoral School in Translational and Molecular Medicine (DIMET), University of Milano-Bicocca, Milan 20126, Italy; [c]Department of Biological Sciences, National University of Singapore, Singapore 117543, Singapore; [d]Department of Otolaryngology, National University of Singapore, Singapore 119228, Singapore.

[1]To whom correspondence may be addressed. E-mail: olaf_rotzschke@immunol.a-star.edu.sg, entwdy@nus.edu.sg, or michael_poidinger@immunol.a-star.edu.sg.

[2]Authors contributed equally to the study

Manuscript in preparation

Abstract

Background

Cis Expression Quantitative Trait Loci (eQTL) analysis links genotypes at polymorphic loci with expression levels of a nearby gene. While numerous studies have been performed on isolated immune cell subsets, so far none focused on neutrophils, key players of the innate immune system and first line of defense against invading pathogens. In addition, the majority of eQTL studies performed so far have employed cohorts of Caucasian descent. Due to differences in linkage disequilibrium (LD) across populations, findings detected in one ethnicity cannot be easily transferred to others.

Methods

We therefore isolated neutrophils from 114 individuals of Chinese descent and performed a genome-wide *cis* eQTL analysis to identify significant associations between single nucleotide polymorphisms (SNPs) and gene expression levels. Significant eQTLs were then compared with Genome-Wide Association Studies (GWAS) signals and analyzed for enrichment of particular diseases to estimate their role in disease susceptibility.

Results

Using a permutation threshold of 0.001 we identified 971 probes with at least a significant eQTL spanning 832 distinct HUGO genes. The majority of our eQTLs were previously reported in whole blood where neutrophils are abundant giving support to the validity of our eQTLs. While the majority of eQTLs detected were cell type specific we nonetheless identified a strong overlap with monocyte and B cell

eQTLs from a published Caucasian cohort. Moreover our eQTLs overlapped with GWAS signals suggesting a role for these variants in disease susceptibility. Neutrophils eQTLs were also found to be enriched for genes involved in dermatological diseases by QIAGEN's Ingenuity® Pathway Analysis (IPA). A large proportion of eQTL genes reported by IPA as being associated with psoriasis were discovered to be differentially expressed between lesional skin from cases and normal skin from controls in two independent published cohorts suggesting a role for neutrophil eQTLs in this disease.

Conclusions

Through a genome-wide *cis* eQTL analysis we identified numerous genes whose expression was modulated by the presence of polymorphisms. This dataset constituted a valuable resource both for linking disease-associated genetic variants with function, thanks to the strong overlap between GWAS signals and SNPs participating in eQTLs, or for identifying diseases where a dysregulation of key neutrophil genes might play a role. Enrichment and differential gene expression analysis in fact suggested a involvement for neutrophil eQTLs in psoriasis, which might be worth to further investigate.

Introduction

Gene expression, the primary mechanism by which information encoded in the genome is translated into function, serves as a good proxy for the phenotypic state of cells[1–3]. Basal transcript levels are highly variable across individuals and have been shown to be modulated not only by the environment, through different signals, but also by genetic variants in regulatory regions which affect transcription[4].

Variations in gene expression controlled by genetic polymorphisms are a key factor for disease susceptibility, the primary focus of Genome-Wide Association Studies (GWAS) which statistically associate genetic variants with complex phenotypes[5,6]. Unfortunately most of the variants reported by these studies fall in non coding regions and cannot therefore be easily linked with function. The identification of genetic variants controlling transcription might help bridging the gap between statistical association and biological function by directly linking disease-associated polymorphisms with variation in transcript abundance[5,7].

Expression quantitative trait loci (eQTL) studies associate genetic variation with gene expression variability, thus identifying polymorphic loci regulating transcription. Most eQTLs studies performed so far have focused on the impact of single nucleotide polymorphisms (SNPs) on messenger RNA (mRNA) levels. Numerous eQTLs have been described both for specific tissues such as normal brain samples[8], liver samples[9], skin biopsies[10,11] and adipose tissues[10,12] and for isolated cell subsets including primary

fibroblasts[13], Epstein-Barr virus-immortalized B cells (also referred to as lymphoblastoid cell lines)[11,14], T-cells[13,15,16], monocytes[17,18] and B-cells[18]. In addition numerous eQTL studies have been carried out on whole blood specimens[19,20]. However, most of the studies performed so far have been limited to individuals of European descent. Unfortunately findings reported for Caucasians are not easily extendable to other populations due to differences in linkage disequilibrium patterns across ethnicities. There is therefore a strong need to characterize the eQTL landscape of other populations.

Neutrophils, key players of the innate arm of the immune system, have so far not been analyzed in the context of eQTL studies[21]. These potent inflammatory cells are the first to migrate to the site of infection where they recruit other immune cells and interact with both the innate and adaptive immune system to overcome the infectious threat[21]. Through phagocytosis, antimicrobial activity and neutrophil extracellular trap (NET), these cells are able to contain and clear both bacterial and fungal infections[21,22].

In spite of their key role in immune defense neutrophils have so far been poorly studied due to the challenges associated with handling them in vitro. These cells are in fact considered fragile and have been shown to be easily activated[23,24].

Here we isolated neutrophils from 114 samples of Chinese descent and performed a genome-wide eQTL analysis to characterize the impact of *cis* polymorphisms on neutrophil transcriptional regulation. The aim is to acquire a better understanding of how genetically

determined changes in gene expression relate to disease susceptibility and immune dysfunctions.

Materials and methods

Ethics Statement

This study has been performed with the approval of the Institutional Review Board (IRB, Reference - NUS07-023 and NUS10-343) of the National University of Singapore and is in compliance with the Helsinki declaration. DNA samples used in this study were collected from ethnic Chinese participants following standard protocols of informed consent. The consent obtained was a “written consent” collected using the Participant Information Sheet containing information about the study.

Study subjects

The cohort studied consists of 114 individuals of Chinese ethnicity, collected from random recruitment drives from January to August 2011 from the National University of Singapore (NUS). Blood was drawn from each individual at steady state, in absence of any known illness. Additional information on prior medical history was recorded.

Neutrophil isolation

Neutrophils were isolated from blood collected from the cohort samples using Ficoll. The procedure can be summarized as follows: separation of blood, processing of peripheral blood mononuclear cells (PBMCs), processing of polymorphonuclear leukocytes (PMNs) and washing. 5 ml of blood was pipetted from BD Vacutainer and

diluted with 5 ml of sterile phosphate buffer saline (PBS) and reverse layer with 5 ml of Ficoll Paque PLUS in a 15 ml Falcon tube using a 23mm UV sterilized glass Pasteur pipette. Tubes were centrifuged at 1800 g, room temperature (R.T) for 20 mins. PBMCs were removed via a sterile dropper and transferred into a 15 ml Falcon tube. PMNs were isolated with as little erythrocytes as possible and transferred to a 15 ml Falcon tube with 6 ml of 1x BD Pharm Lyse added into each tube. The contents were mixed by vortexing gently and left to incubate for 15 mins at R.T. Afterward content was topped up to 15 ml with RPMI+10% FBS and centrifuged at 1200 rpm, 4°C, 5 mins. The supernatant was aspirated. Cell pellet was resuspended with 10 ml of RPMI+10% FBS and centrifuged at 1200 rpm, 4°C, 5mins. The supernatant was aspirated as before and 500 µl of RPMI+10% FBS was added and mixed. The cell suspension, of 10 µl was pipetted into 90 µl of PBS inside a 96-well format. Cell count was performed on MACSQuant with an uptake volume of 25 µl, speed:fast, mode: fast. The samples were mixed manually by pipetting before uptake of cell suspension by MACSQuant. The cell concentration was made up to $5x10^6$ cells/ml with RPMI+10% FBS. The samples were aliquoted into sterile 96-well U-bottom plate and the remaining cell suspensions were centrifuged at 1200 rpm, 4°C, 5 mins. Th Supernatant was aspirated up to approximately the 100 µl mark and 300 µl of Trizol LS reagent was added and lastly, transferred to a 1.5 ml microcentrifuge tube and stored at -80°C.

RNA extraction

Total RNA was extracted from neutrophils using Trizol separation followed by RNeasy micro kit extraction. Homogenised samples (1 ml) were incubated in TRIZOL reagent for 5 mins at R.T, 0.2 ml of chloroform was added per 1 ml of TRIZOL, incubated for 2 to 3 mins, and centrifuged at 12000 g for 15 mins at 4°C. Aqueous phase was separated and the following procedures used RNeasy micro kit extraction. Equal amounts of 70% ethanol (EtOH) was added slowly and mixed gently. Samples were loaded onto Qiagen columns and centrifuged at 9000 g for 15 sec. Buffer RW1 (350 µl) was added to the RNeasy MiniElute spin column, and centrifuged at 9000 g for 15 sec, flow-through was discarded and collection tube was changed. Buffer RPE (500 µl) was added to the spin column, and centrifuged at 9000 g for 15 sec to wash, flow-through was discarded and collection tube was reused. The step was repeated with 80% Ethanol, and collection tube was replaced. The tube was centrifuged at 12000 g for 5 mins and the RNeasy column was transferred into a new 1.5 ml collection tube. RNase-free water (10-15 µl) was pipetted onto the RNeasy membrane, incubated for a minute and centrifuged at 12000 g for 1 minute to elute the RNA. Elution was repeated with the eluate into the same collection tube.

Genome-wide gene expression

Illumina® Ambion TotalPrep™ 96 RNA Amplification kit (Part Number 4393543) was used for RNA preparation and amplification for hybridisation. Illumina® HumanHT-12-v4 Expression BeadChip Kit

(Catalog number: BD-103-0204) was used to quantify RNA from the 114 individuals. Illumina® Beadchip uses direct hybridisation assay. The starting amount of RNA used was 40 ng. The protocol can be summarized as follows: synthesis of first strand, synthesis of second strand, cDNA purification, in vitro transcription (IVT), cDNA purification, and hybridization using 750 ng of RNA at 58°C for 16 hours.

Probe sequences provided by Illumina were mapped to genome build hg18 using the pipeline RUM (http://cbil.upenn.edu/RUM/). Only 41,469 uniquely mapped to the genome. 39,596 probes were located on autosomes and retained for the analysis. Expression values were $\log_2$-transformed. Probes containing common SNPs (MAF≥1%) according to the 1000 Genome Pilot Project[25] for CHBJPT were excluded from the analysis (2,467 probes).

DNA extraction

DNA was extracted from the blood samples. The kit used for the extraction of DNA from the blood was Qiagen DNeasy® Blood & Tissue Kit (Catalogue number: 69504).

Genome-wide SNP genotyping and quality control

Illumina® Human Omni5Quad was used for SNP genotyping. It is able to detect up to 4.3 million SNP markers. Genome coordinates provided by Illumina were converted to genome build hg18 (for consistency with the annotation used for the Illumina probes) using the UCSC's tool *liftOver*[26]. 1,067 SNPs could not be converted and were removed from the analysis. SNPs located on sex chromosomes,

random chromosomes and mitochondrial chromosomes were also excluded from the computations. Monomorphic and low call rate SNPs (<95%) were excluded as well. 2,031,824 SNPs were retained for genome-wide *cis* eQTL analysis.

High Resolution Melt (HRM) curve

HRM analysis is a technique to genotype SNPs using primers, running real time-PCR and finally, analyzing the melt curve to identify the SNP variant. Samples lacking genotype information for a particular polymorphism were genotyped using the HRM technique.

Reverse transcription of RNA to cDNA

RNA was reverse transcribed to complementary DNA (cDNA) for detection through quantitative real time polymerase chain reaction (qRT-PCR). RNA was isolated and reverse transcription was performed using Qiagen® QuantiTect® Reverse Transcription. The amount of RNA used was 100 ng for reverse transcription into cDNA for the whole experiment. The protocol is described Qiagen® QuantiTect® Reverse Transcription handbook. The main steps are summarized as follows: elimination of gDNA and reverse transcription. The cycler used was Eppendorf Mastercycler and the final volume of cDNA was 20 µl. cDNA were stored in a -20°C fridge.

TaqMan® Gene Expression Assay

TaqMan® Gene Expression Assays were used to detect single gene expression. Probes for qRT-PCR from TaqMan® have a FAM fluorochrome at the 5’ end, and a quencher TAMRA at the 3’ end.

Each reaction mix was 20 µl, using 2.5 ng of cDNA per reaction, with 1 µl of probe, 10 µl of TaqMan® Master Mix (P/N: 4369016), and water to make the total volume of 20 µl. The cycle conditions for the qRT-PCR reaction was 50°C for 2 minutes, followed by 90°C for 10 sec, 40 cycles of 90°C for 10 seconds and 60°C for 2 minutes. The cycler used was CFX96™ Real-Time system, C1000™ Thermal cycler.
The TaqMan® probes used are described in Table S1.

Validation of gene expression using real-time PCR

Real-time PCR was performed using Bio Rad CFX96™ Real-Time system, C1000™ Thermal cycler. The reaction mix contained: 10 µl of the SsoFast™ EvaGreen® Supermix, 0.5 µl of the forward and reverse primer respectively, and 40 ng of genomic DNA (gDNA), and topped up to 20 µl with water. The reaction mix was heated to 98°C for 2 mins, followed by 40 cycles of amplification comprising of denaturation step at 98°C for 5 sec and annealing and elongation step at temperatures dependent on primer optimization for 5 sec. Standard DNA melt was performed by increasing the temperatures from 75°C to 95°C, holding at each temperature for 5 sec. Lastly, a high resolution melt step of increasing temperatures between 75°C to 95°C for 2 sec. The results were analyzed by Bio Rad Precision Melt Analysis.

Statistical Analysis

Genome-wide statistical association between SNPs and gene expression in the discovery cohort was evaluated using linear regression as implemented by the Apache Commons Mathematics

Library. Genotypes were coded based on allele counts (0, 2 for homozygous genotypes and 1 for heterozygous genotypes). P-values reported as zero were recomputed using the python function linregress from the library scipy to achieve a better precision. For each probe only SNPs located ± 250kbp from the probe midpoint were tested for association. Significance was determined by 10,000 phenotype permutations (label swapping) as described in[14]. A <SNP,probe> pair was considered significant if its nominal P-value was lower than the 0.001 tail of the distribution of minimal permuted P-values across all SNPs tested for a particular probe. Each probe was analyzed independently. Samples with missing genotypes for a particular SNP were excluded from the analysis of the corresponding SNP-probe pair. Illumina SNP IDs coded as kgp were mapped to the corresponding rs IDs by chromosomal position.

Statistical association between genotype and gene expression in the validation cohort was computed using a Kruskal-Wallis one-way analysis of variance (3 genotypes) or Mann-Whitney U test (2 genotypes). Multiple comparisons were performed using Dunn's multiple comparisons test (Graphpad Prism 6).

Estimation of the overlap between eQTL datasets and between neutrophil eQTLs and GWAS signals

The overlap between neutrophil eQTLs and monocyte, B cell and whole blood eQTLs was computed using a "double expansion" strategy. For every probe, SNPs associated to its expression were selected for each dataset independently. The set of SNPs in our

cohort was expanded to also include all the SNPs in linkage disequilibrium with the initial set according to a certain r^2 threshold using LD computed from the 60 CHBJPT samples sequenced by 1000 Genomes Pilot Project[25]. A similar expansion was done for monocyte, B cell and whole blood eQTLs using the 60 CEU samples sequenced by the 1000 Genomes Pilot Project[25].

A similar analysis was performed to estimate the overlap between neutrophil eQTLs and GWAS signals. Even in this case the expansion of GWAS signals was performed using the 60 CEU samples sequenced by the 1000 Genomes Pilot Project[25] on the assumption that most GWAS studies in the catalog were performed on individuals of Caucasian ethnicity.

Identification of differentially expressed genes in psoriasis and atopic dermatitis

GEO dataset Series Matrix files for two psoriasis studies (GSE13355[27] and GSE14905[28]) and one atopic dermatitis one (GSE5667[29]) were obtained from NCBI GEO. Data were then processed in the R statistical language 2.15.2[30] using the R package GEOquery. Comparisons between the lesional skin from cases and normal skin from controls were made using the Limma package. Multiple testing correction was performed using the method of Benjamini and Hochberg[31] and tests were considered significant if the multiple testing corrected P-value was lower than 0.05.

Results

eQTL discovery and validation

A cohort of 114 individuals of Chinese ethnicity, whose genotype and gene expression profiles were measured by Illumina® Omni5Quad and HumanHT-12-v4 expression beadchip platforms respectively, was used for discovering *cis* eQTLs in neutrophils (±250kbp from the probe midpoint). Association between allelic states at a particular locus and gene expression measurements was evaluated using linear regression for 39,596 probes and 2,031,824 SNPs. A total of 15,537,788 tests were performed. At a 0.001 permutation significance threshold our analysis identified 21,210 eQTLs involving 971 distinct probes and spanning 832 HUGO genes. Table 1 reports the top 20 most significant eQTL genes. Figure 1 depicts the Manhattan and the Q-Q plots associated with the analysis and provide information on the distribution of P-values across the various chromosomes and the deviation of the distribution of observed P-values from the expected distribution respectively.

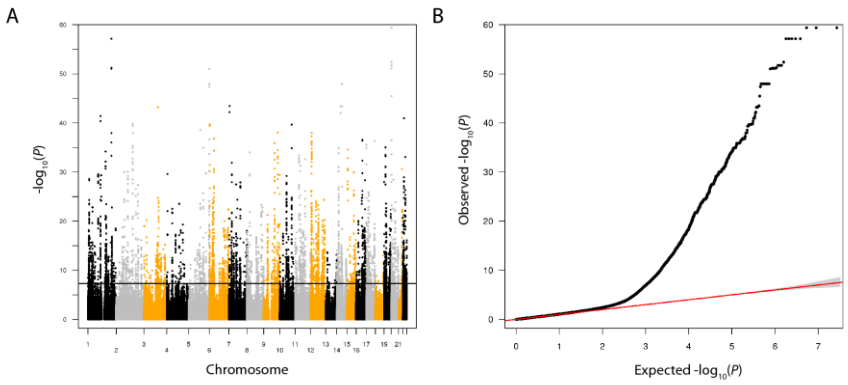


Figure 1. Manhattan plot and QQ plot.

Since high-throughput microarray platforms are known to produce experimental artifacts which might generate false positive eQTLs [19,32] we selected a subset of <SNP,probe> pairs for validation. Gene expression was in this case quantified using quantitative real-time polymerase chain reaction (qRT-PCR) with TaqMan® Gene Expression probes. This technique measures expression of single genes thus allowing for a more sensitive and accurate measurement compared to genome-wide assays. Figure S1 summarizes the results of the genes tested for validation: all the probes tested showed a similar distribution across genotypes as in the discovery cohort.

Table 1. Top 20 genes harboring a neutrophil eQTL.

Symbol	Probe	SNP ID	*P*	FC	R^2
LOC654055	ILMN_1742442	rs2209313	4.28×10^{-60}	3.77	0.91
C4BPA	ILMN_1810752	kgp8310661	7.21×10^{-58}	18.11	0.90
BTNL3	ILMN_1783795	rs4700772	9.72×10^{-52}	5.20	0.87
CHURC1	ILMN_1798177	rs7143432	1.13×10^{-48}	5.06	0.85
C7orf28B	ILMN_2246083	kgp11715520	3.42×10^{-44}	4.44	0.82
ERO1L	ILMN_1744963	rs12590590	4.09×10^{-44}	2.52	0.82
GOLGB1	ILMN_1747935	kgp5002661	6.43×10^{-44}	2.86	0.82
KIAA1324	ILMN_1771482	rs649539	3.93×10^{-42}	9.03	0.81
TFIP11	ILMN_2408102	kgp3038838	1.13×10^{-41}	2.63	0.81
LOC401233	ILMN_1674285	kgp12462332	2.05×10^{-40}	3.70	0.80
USMG5	ILMN_1773313	rs12220267	2.22×10^{-40}	2.27	0.80
PAM	ILMN_2313901	kgp9149121	$2.79 \times$	4.55	0.79

			10^{-39}		
CLEC4C	ILMN_1682259	rs10845821	5.25×10^{-38}	2.09	0.77
TREML4	ILMN_2205322	rs6458200	1.53×10^{-37}	3.27	0.77
LPCAT2	ILMN_1796335	rs1502000	2.83×10^{-37}	4.23	0.77
LOC100130520	ILMN_3272741	kgp10871931	5.07×10^{-37}	5.15	0.77
HIATL1	ILMN_1737964	kgp2340295	1.35×10^{-36}	2.69	0.76
C17orf97	ILMN_1707137	kgp9481735	2.91×10^{-36}	2.17	0.76
DNASE2	ILMN_1796245	rs7249143	9.67×10^{-36}	2.53	0.75
PPP2R3C	ILMN_1662617	kgp1904514	1.44×10^{-35}	2.19	0.75

Overlap between neutrophil eQTLs, whole blood and other immune cell subsets

Neutrophils are the most abundant leukocyte cell subset in the blood[33–35]. We therefore decided to evaluate the percentage of neutrophil eQTLs that could also be captured in whole blood. We compared our eQTLs with eQTLs reported by a large study performed on a cohort of 5,311 individuals of European descent[36]. A large proportion of significant eQTLs detected in neutrophils were also reported in whole blood: 525 out of 971 probes (54.07%) had at least a shared eQTL in the two datasets based on a perfect match between SNP IDs (Figure 2A). The number of probes with at least a shared eQTL increased to 574 (59.11%) after lowering the linkage disequilibrium threshold used for computing the overlap (r^2=0.5). To estimate the impact of effect size on the overlap we also compared our results with eQTLs discovered in a smaller whole blood cohort

composed of 322 individuals[20]. Only 95 out of 971 probes (9.78%) shared an eQTL in the two datasets.

To estimate the overlap between neutrophil eQTLs and other isolated immune subsets we selected the dataset published by Fairfax *et al.* reporting *cis* eQTLs identified in B cells and monocytes isolated from 288 Caucasian individuals[18]. Of the 971 probes with at least a significant eQTL, 248 (25.54%) shared an eQTL with both monocytes and B cells, 159 (16.37%) shared an eQTL with monocytes only and 28 (2.88%) shared an eQTL with B cells only based on a linkage disequilibrium threshold of r^2=0.8 for both populations (Figure 2B). While a strong overlap with monocytes and B cells was detected, the majority of neutrophil eQTLs (55.20%) was found to be unique to this cell subset.

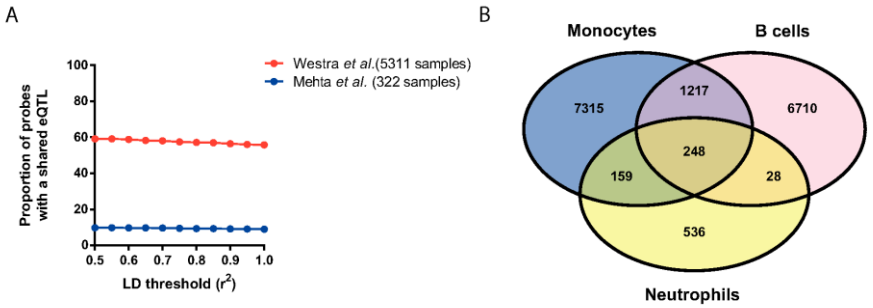


Figure 2. Overlap of neutrophil *cis* eQTLs with published eQTL datasets. (A) Overlap of neutrophil eQTLs with two whole blood datasets of different sample sizes (5311 samples[36] versus 322 samples[20]). The overlap was minimal for the smaller cohort while the majority of neutrophil eQTLs was detected in the larger cohort. Using different LD thresholds for computing the overlap did not have a big impact on the proportion of probes with a shared eQTL (B) Overlap of neutrophil eQTLs with monocytes and B cells eQTLs detected in a cohort of 288 individuals of European descent[18]. The overlap was computed using a LD threshold of r^2=0.8. A large proportion of eQTLs was detected both in monocytes and B cells. A stronger overlap was detected for neutrophils and monocytes compared to neutrophils and B cells, a finding in line with their shared myeloid lineage.

Overlap between neutrophil eQTLs and GWAS signals

The majority of GWAS signals falls in non-coding regions and cannot therefore be easily linked with function. These mutations do not in fact alter the amino acid sequence and consequently do not directly affect protein structure or function. It is nonetheless possible for these GWAS SNPs to modulate expression. This might be a potential mechanism through which these associated variants lead to disease susceptibility. We estimated the overlap between all the SNPs involved in at least a significant neutrophil eQTL (eSNPs) and SNPs

reported as being associated with diseases or traits by the Catalog of Published Genome-Wide Association Studies[37]. The overlap was estimated using a linkage disequilibrium threshold of r^2=0.8. GWAS variants associated to 89 diseases/traits were reported to be in linkage disequilibrium with at least one neutrophil eSNP (Table S2). For some of the diseases more than one eQTL-associated SNP was reported (Figure 3A). In addition multiple genes with a significant eQTL were found to be associated to GWAS variants implicated in more than one disease/trait (Figure 3B).

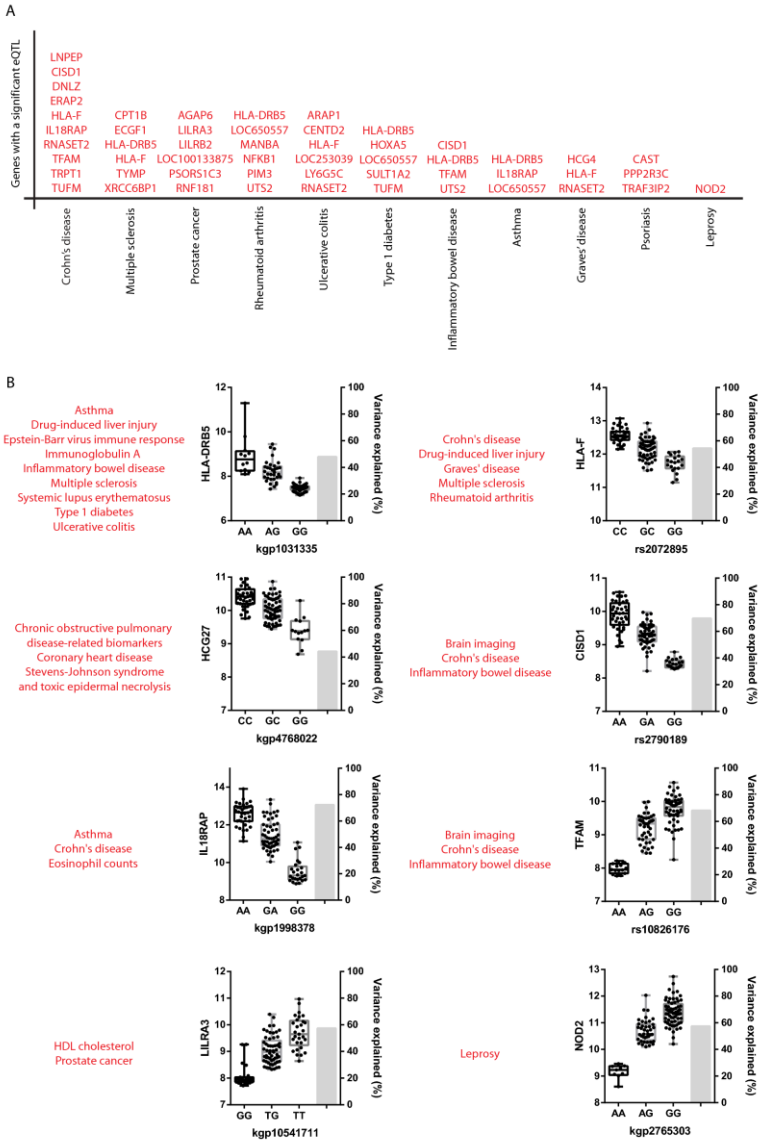


Figure 3. Overlap between neutrophil eQTLs and GWAS signals. Overlap between GWAS signals and neutrophil eQTLs was computed using a LD threshold of r^2=0.8. Only genome-wide significant findings were used to compute the overlap. (A) Representative list of diseases associated by GWAS to multiple neutrophil eQTLs. (B) Expression plots of representative neutrophil eQTLs reported as being significantly associated to diseases/traits by GWAS studies. Numerous neutrophil eQTLs were found to be associated to multiple diseases.

Enrichment for diseases and disorders

To investigate the functional impact of neutrophil eQTLs on disease we performed an enrichment analysis for the 971 Illumina probes with at least a significant eQTL using QIAGEN's Ingenuity® Pathway Analysis (IPA®, QIAGEN Redwood City, www.qiagen.com/ingenuity). Top hits for the category Diseases and Disorders are reported in Table 2.

As expected considering the role played by neutrophils in immune defense, genetically regulated probes were enriched for genes participating in the inflammatory response and associated with the resolution of organismal injury. Surprisingly eQTL probes were also enriched for dermatological diseases and conditions. The software reported 43 genes associated with psoriasis, 35 genes associated with dermatitis, 26 genes associated with atopic dermatitis and 2 genes associated with dermatitis of the ear (Table S3). To evaluate if

Table 2. IPA enrichment for diseases and disorders

Disease or Disorder	p-value	# molecules
Inflammatory Response	$2.39 \times 10^{-6} – 3.50 \times 10^{-2}$	130
Dermatological Diseases and Conditions	$1.46 \times 10^{-4} – 2.48 \times 10^{-2}$	69
Inflammatory Disease	$1.46 \times 10^{-4} – 3.23 \times 10^{-2}$	94
Cardiovascular Disease	$1.87 \times 10^{-4} – 3.23 \times 10^{-2}$	57
Organismal Injury and Abnormalities	$4.77 \times 10^{-4} – 3.34 \times 10^{-2}$	35

these reported genes showed differential expression between cases and controls we used two publicly available GEO datasets (GSE5667 for dermatitis[29] and GSE13355 for psoriasis[27]). None of the 35 genes tested for dermatitis showed differential expression between involved skin from 6 affected individuals and normal skin from 5 control samples after multiple testing correction (FDR). Interestingly 108 Affymetrix probes tagging 41 distinct genes out of the 43 reported for psoriasis were differentially expressed between involved skin from 58 psoriatic samples and normal skin from 64 control samples after multiple testing correction (Figure 4). 23 probes tagging 16 psoriasis-associated genes showed instead similar expression patterns across conditions ($P_{Fisher}= 5.98 \times 10^{-6}$).

Similar results were obtained using a second cohort composed of 21

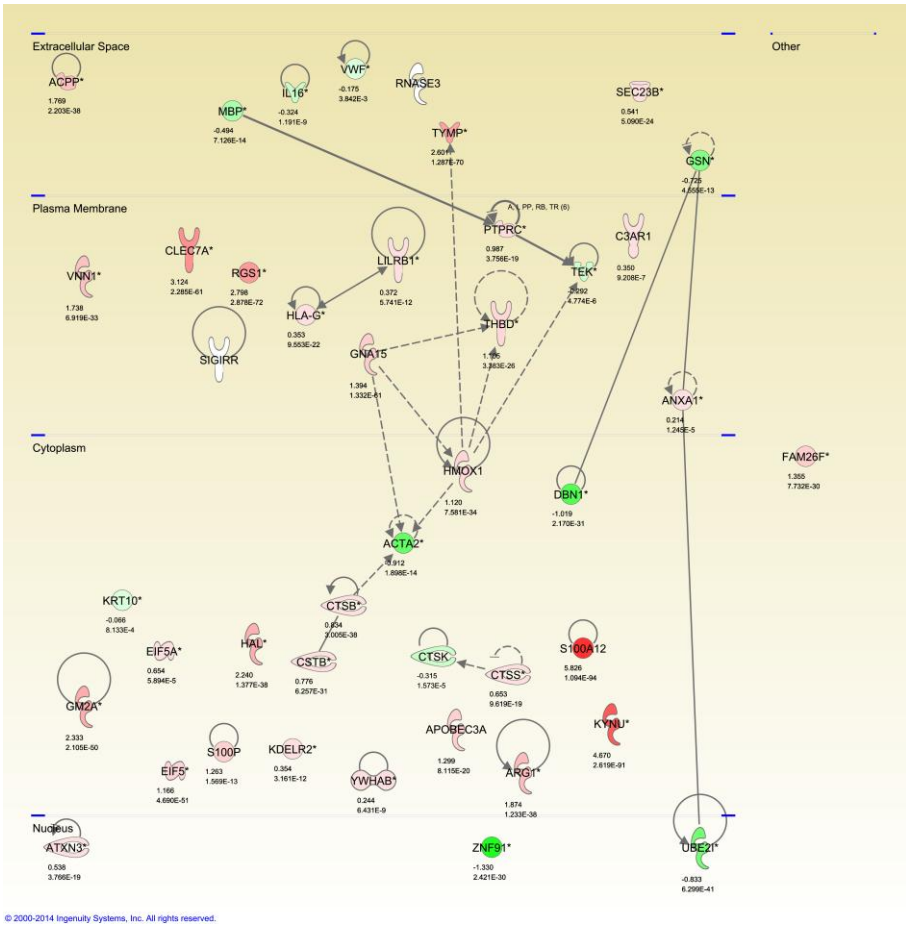


Figure 4. Comparison between affected skin from psoriatic patients and healthy skin from controls for the 43 neutrophil eQTL genes associated with psoriasis. The GEO dataset GSE13355 Series Matrix files were obtained from NCBI GEO. Comparison between the lesional skin from cases and normal skin from controls was made using the Limma R package. Multiple testing correction was performed using the method of Benjamini and Hochberg and a comparison was considered to be significant if the multiple testing corrected P value was less than 0.05.41 out of 43 genes were differentially expressed.

healthy donors and 28 psoriatic patients (GEO dataset GSE14905[28]) where 76 probes tagging 34 distinct genes were differentially expressed after FDR correction (Figure S2). 55 probes tagging 28 distinct genes showed instead similar expression patterns across conditions (P_{Fisher}= 3.18 x 10^{-8}). In both cohorts differentially expressed probes were therefore enriched for IPA psoriasis-associated genes.

We then extended our analysis to include all neutrophil eQTLs. Of the 1790 probes tagging our 832 genes harboring an eQTL 1309 were differentially expressed in the GEO dataset GSE13355 while 481 were not (P_{Fisher}<2.2 x 10^{-16}). Similarly in the GEO dataset GSE14905 876 probes were differentially expressed while 923 were not (P_{Fisher}<2.2 x 10^{-16}). Differentially expressed probes between psoriatic skin samples and control skin samples were therefore strongly enriched for neutrophil eQTLs in both GEO datasets.

Discussion

In this study we investigated the impact of *cis* polymorphisms on gene expression levels in neutrophils using a cohort of 114 Singapore Chinese individuals. To our knowledge this is the first eQTL study performed on this immune subset and one of the few studies employing an Asian cohort[14,38,39]. Neutrophils are notoriously difficult to isolate because fragile and easily activated[23,24]. This dataset therefore constitutes a valuable resource for the study of this cell subset.

Our genome-wide analysis identified 971 Illumina probes with at least a significant eQTL, spanning 832 distinct HUGO genes. The majority of these eQTLs was also detectable in whole blood, an easier to access biological resource. However a comparison between two whole blood studies based on cohorts of different sizes (5,311[36] vs 322 individuals[20]) showed that a 50% overlap was only obtained for the biggest cohort while the overlap with the smaller cohort was minimal.

This suggests that much larger sample sizes are needed to detect neutrophil eQTLs from blood compared to isolated cells in spite of the high frequency of neutrophils in this medium.

Comparison with monocytes and B cells isolated from 288 individuals of Caucasian descent revealed that while there exists some overlap between neutrophils and monocytes/B cells the majority of neutrophil eQTLs are still cell type specific confirming previous reports on other immune cell subsets[13,18]. As expected the overlap was stronger for monocytes than B cells because neutrophils and monocytes are more closely related (common myeloid lineage). According to our study the proportion of eQTLs shared between neutrophils and monocytes and neutrophils and B cells is larger than the proportion of eQTLs reported as being shared between monocytes and B cells (see Figure 2). This is easily explained considering that our cohort is smaller than the cohort used to detect eQTLs in monocytes and B cells. Since the power of the analysis increases with sample size, larger cohorts allow for the detection of smaller effect size eQTLs. In addition, it has been reported that cell

type-specific eQTLs tend to have smaller effect sizes than shared ones[13,18].

Our study also identified numerous neutrophil eSNP tagging GWAS signals. As described in Figure 3 numerous variants implicated in complex diseases were found to be significant eQTLs in neutrophils. For example GWAS variants associated with Crohn's disease were shown to modulate the expression of 10 different genes. On the other hand, multiple genetically regulated genes were associated to multiple diseases or traits.

In addition, genes harboring neutrophil eQTLs were shown to be enriched for inflammatory responses and diseases by an unbiased pathway analysis. This enrichment was expected considering the key role played by neutrophils in innate immunity. Enrichment was also detected for organismal injury, a finding consistent with the role of neutrophils as first line of defense against pathogens upon tissue damage[40]. Interestingly our eQTL genes were also enriched for dermatological diseases (dermatitis and psoriasis in particular). Analysis of two publicly available datasets revealed that the majority of eQTL genes reported for psoriasis are differentially expressed between affected skin collected from cases and normal skin collected from controls. This finding is in line with other reports supporting a role for neutrophils in the etiology of psoriasis, a immune-mediated disorder characterized by epidermal hyperproliferation[41]. Neutrophils have in fact been shown to accumulate under the stratum corneum of highly-inflamed psoriatic lesions where they influence the activation state of T cells and promote the growth of epidermal

keratinocytes[42,43]. Interestingly a link between neutrophils and psoriasis was also supported by the overlap between GWAS signals for this disease and our eQTLs. Three of the SNPs reported for this disease in the GWAS catalog are in fact eSNPs in our cohort.

A similar analysis on atopic dermatitis (AD) did not yield any significant result. None of the genes reported by IPA as being associated with the disease were differentially expressed between affected skin from cases and normal skin from controls. This is not surprising considering that while both atopic dermatitis and psoriasis affect the skin, the biological mechanisms underlying skin inflammation are quite different[42]. Atopic dermatitis skin lesions are in fact predominantly accompanied by infiltration of macrophages, dendritic cells, eosinophils and Th2 CD4+ lymphocytes[44,45] while psoriatic plaques are polarized towards a Th1 response and are characterized by an accumulation of T cells, monocytes and neutrophils[46].

In summary using a genome-wide *cis* eQTL analysis we identified numerous genes whose expression was affected by the presence of polymorphic loci. This dataset constitutes a valuable resource both for linking disease-associated genetic variants with function, as demonstrated by the strong overlap between eQTLs and GWAS signals, or for identifying diseases where a dysregulation of key neutrophil genes might play a role, as in psoriasis.

Supplementary files

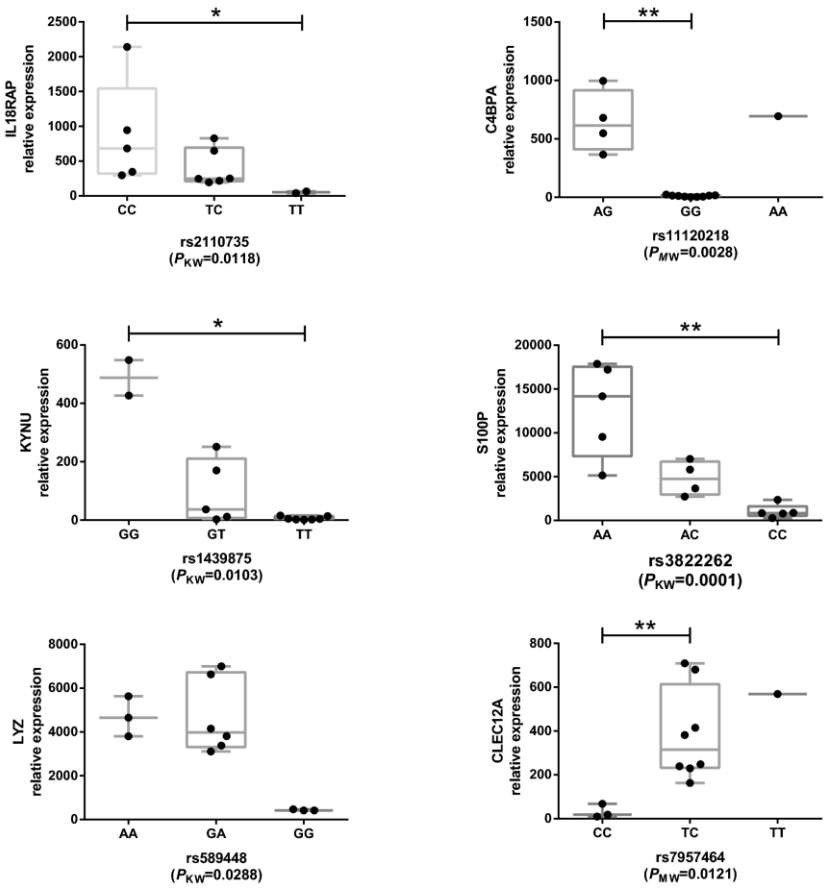


Figure S1. Technical validation of a subset of neutrophil eQTLs using TaqMan® Probe-Based Gene Expression Analysis. Validation was performed on a subset of 14 samples selected from the original cohort. P-values were computed using Kruskal-Wallis one-way analysis of variance in the case of loci with three available genotypes followed by Dunn's multiple comparisons test and Mann-Whitney U test in the case of loci with two available genotypes. P-values lower than 0.05 were considered significant.

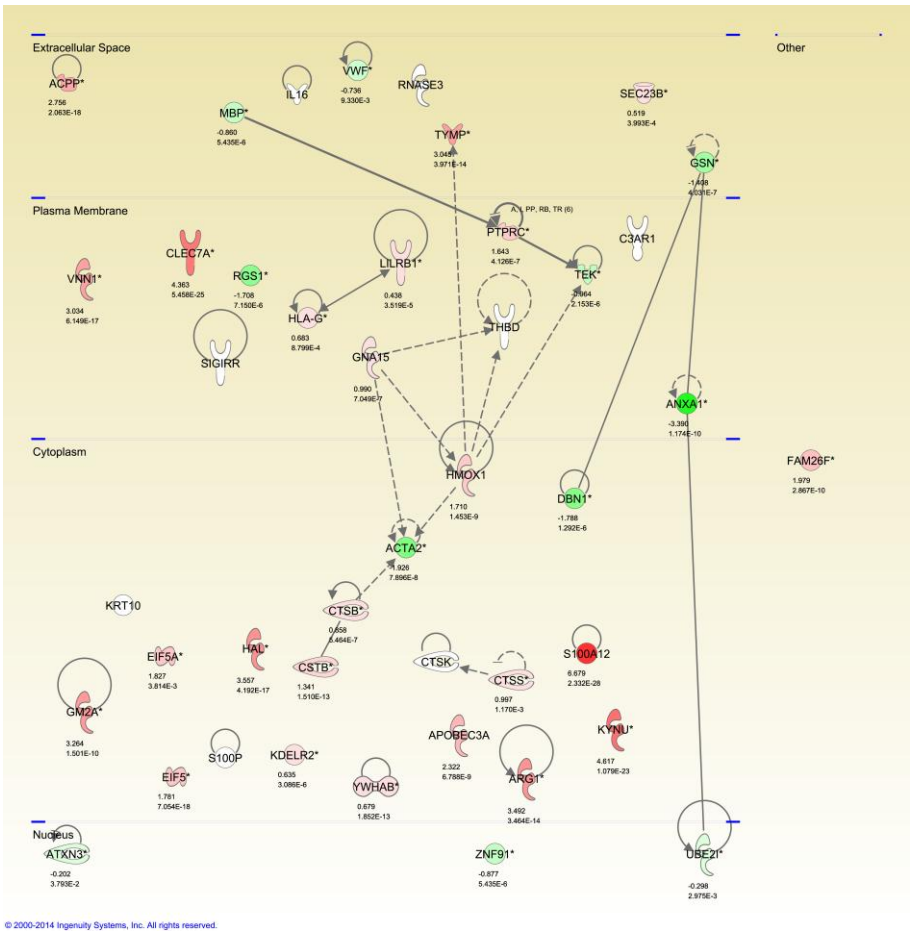


Figure S2. Differentially expressed genes between lesional skin from psoriatic patients and normal skin from healthy controls. The GEO dataset GSE14905 Series Matrix files were obtained from NCBI GEO. Comparison between the lesional skin from cases and normal skin from controls was made using the Limma R package. Multiple testing correction was performed using the method of Benjamini and Hochberg and a comparison was considered to be significant if the multiple testing corrected P value was less than 0.05.

Table S1. Catalogue no. of TaqMan® Gene Expression Assays and Amplicon size after qRT-PCR

Gene	Assay ID	Catalogue No.	Amplicon Size
PRKAR1A	Hs00267597_m1	#4453320	73
CLEC12A	Hs01074333_m1	#4448892	94
C4BPA	Hs00426339_m1	#4448892	105
IL18RAP	Hs00977695_m1	#4448892	151
KYNU	Hs01114099_m1	#4448892	102
LYZ	Hs00426232_m1	#4448892	67
S100P	Hs00195584_m1	#4453320	73
GAPDH	Hs03929097_g1	#4331182	93

Table S2. Overlap between neutrophil eQTLs and GWAS signals group by disease.

disease	# eQTL genes	eQTL genes	Study	GWAS SNP
Alzheimer's disease	1	MS4A6A	[47]	rs610932
Alzheimer's disease (late onset)	1	MS4A6A	[48]	rs4938933
Ankylosing spondylitis	1	CAST	[49,50]	rs27434, rs30187
Asthma	3	HLA-DRB5, IL18RAP, LOC650557	[51,52]	rs3771166, rs9272346, rs9273349
Behcet's disease	1	CCR3	[53]	rs7616215
Beta-2 microglubulin plasma levels	1	BAT5	[54]	rs2596466
Blond vs. brown hair color	1	TPCN2	[55]	rs35264875
Blood pressure	3	C15orf17, CSK, MTHFR	[56]	rs1378942, rs17367504

Body mass index	3	MYBPC3, SPI1, TUFM	[57–59]	rs10838738, rs3817334, rs7498665
Brain imaging	2	CISD1, TFAM	[60]	rs16912145
Brain structure	2	HRK, TMEM118	[61]	rs7294919
Breast cancer	1	ECHDC1	[62]	rs2180341
C4b binding protein levels	1	C4BPA	[63]	rs3813948
Celiac disease	1	RGS1	[64,65]	rs2816316
Cholesterol, total	2	N.A., CPNE1	[66]	rs2072183, rs2277862
Chronic obstructive pulmonary disease-related biomarkers	2	HCG27, PSORS1C3	[67]	rs1265093, rs2074488
Cleft lip	1	TRAF3IP3	[68]	rs10863790
Coffee consumption	1	CSK	[69]	rs6495122
Coronary heart disease	1	HCG27	[70]	rs3869109
Crohn's disease	10	N.A., CISD1, DNLZ, ERAP2, HLA-F, IL18RAP, RNASET2, TFAM, TRPT1, TUFM	[71–73]	rs151181, rs1819658, rs2058660, rs2301436, rs2549794, rs4077515, rs415890, rs694739, rs9258260
Dengue shock syndrome	1	N.A.	[74]	rs3132468
Diastolic blood pressure	3	C15orf17, CSK, LY6G5C	[75–77]	rs1378942, rs6495122, rs805303
Drug-induced liver injury (amoxicillin-clavulanate)	2	HLA-DRB5, HLA-F	[78]	rs2523822, rs9274407
Endometriosis	2	N.A., CDC42	[79]	rs10917151, rs4654783, rs2235529
Eosinophil counts	1	IL18RAP	[80]	rs1420101
Epstein-Barr	1	HLA-DRB5	[81]	rs477515

virus immune response (EBNA-1)				
Eye color traits	1	C17orf90	[82]	rs9894429
Glaucoma (primary open-angle)	1	PIK3C2A	[83]	rs11024102
Graves' disease	3	HCG4, HLA-F, RNASET2	[84,85]	rs3893464, rs9355610
HDL cholesterol	8	CDK2AP1, KCTD10, LILRA3, LOC100133875, MYBPC3, PPP1R1B, SPI1, UBE2L3	[66,86–88]	rs11869286, rs181362, rs2338104, rs386000, rs4759375, rs7120118
Height	9	CCBL2, CLIC4, FIG4, N4BP2L2, OCEL1, PFAAP5, PIK3C2A, RFP, SMPD2	[89]	rs1046943, rs1330, rs2279008, rs3129109, rs4601530, rs6699417, rs7332115
Hippocampal volume	2	HRK, TMEM118	[61]	rs7294919
Homocysteine levels	1	MFN2	[90]	rs1801133
Hypertension	1	LY6G5C	[75]	rs805303
Hypothyroidism	1	VAV3	[91]	rs4915077
IgA nephropathy	2	CD68, HCG4	[92]	rs2523946, rs4227
Immunoglobulin A	1	HLA-DRB5	[93]	rs9271366
Inflammatory bowel disease	4	CISD1, HLA-DRB5, TFAM, UTS2	[94–96]	rs2790216, rs35675666, rs477515, rs9271366
Iris color	1	HERC2	[97]	rs916977
LDL cholesterol	1		[66]	rs2072183
Leprosy	1	NOD2	[98]	rs9302752
Lipoprotein-associated phospholipase A2 activity and mass	1	MS4A6A	[99]	rs600550
Liver enzyme	1	NRBF2	[100]	rs10761779,

levels				rs12355784
Liver enzyme levels (gamma-glutamyl transferase)	3	CCBL2, DDT, THBS3	[101]	rs10908458, rs12145922, rs2739330
Lymphoma	1	TAP2	[102]	rs2621416
Magnesium levels	1	THBS3	[103]	rs4072037
Mean corpuscular hemoglobin	1	DNASE2	[104]	rs11085824
Mean corpuscular volume	1	DNASE2	[104]	rs7255045
Mean platelet volume	2	C12orf47, NRBF2	[105,106]	rs2393967, rs6490294
Melanoma	2	CTSK, DEF8	[107,108]	rs4785763, rs7412746
Menarche (age at onset)	2	ARNTL, RBM6	[109]	rs6762477, rs900145
Menopause (age at onset)	1	LY6G5C	[110,111]	rs1046089
Metabolic traits	1	TRAF3IP2	[112]	rs7760535
Metabolite levels	2	CLTCL1, TRAF3IP2	[113]	rs6900341, rs807669
Multiple sclerosis	6	CPT1B, ECGF1, HLA-DRB5, HLA-F, TYMP, XRCC6BP1	[114–117]	rs12368653, rs140522, rs2523393, rs703842, rs9271366
Multiple sclerosis (OCB status)	1	HLA-DRB5	[118]	rs9271640
Nasopharyngeal carcinoma	1	HCG4	[119,120]	rs2517713, rs2860580
Natriuretic peptide levels	1	MTHFR	[121]	rs1023252
Orofacial clefts	1	TRAF3IP3	[122]	rs861020
Other erythrocyte phenotypes	1	EPHB4	[104]	rs2075671
Pain	1	ZNF493	[123]	rs2562456
Parkinson's disease	1	BST1	[124,125]	rs11724635, rs4538475

Plasminogen activator inhibitor type 1 levels (PAI-1)	1	EPHB4	[126]	rs6976053
Platelet counts	6	AKAP10, KIAA2013, MFN2, NRBF2, PIK3C2A, PLOD1	[127]	rs10761731, rs13300663, rs2336384, rs397969
Primary biliary cirrhosis	1	MANBA	[128]	rs7665090
Proinsulin levels	2	MYBPC3, SPI1	[129]	rs10838687
Prostate cancer	6	AGAP6, LILRA3, LILRB2, LOC100133875, PSORS1C3, RNF181	[130–135]	rs10187424, rs103294, rs10993994, rs130067, rs3123078
Prostate-specific antigen levels	1	AGAP6	[136],	rs10993994
Psoriasis	3	CAST, PPP2R3C, TRAF3IP2	[137,138]	rs12586317, rs240993, rs27524
QT interval	1	CNOT1	[139]	rs7188697
Red blood cell traits	3	CPT1B, ECGF1, TYMP	[140]	rs140522
Renal function-related traits (BUN)	1	THBS3	[141]	rs2049805
Rheumatoid arthritis	6	ARAP1, CENTD2, HLA-F, LOC253039, LY6G5C, RNASET2	[142–147]	rs1610677, rs3093023, rs3761847, rs3781913, rs805297, rs881375
Sarcoidosis	1	TRPT1	[148]	rs479777
Schizophrenia	1	LSM1	[149]	rs16887244
Sphingolipid levels	1	LOC255167	[150]	rs1566039
Stevens-Johnson syndrome and toxic epidermal	1	HCG27	[151]	rs3130501

necrolysis (SJS-TEN)				
Systemic lupus erythematosus	2	HLA-DRB5, UBE2L3	[152,153]	rs131654, rs9270984, rs9271100
Systolic blood pressure	5	C15orf17, CSK, LY6G5C, MTHFR, USMG5	[75–77]	rs1004467, rs1378942, rs17367504, rs805303
Triglycerides	1	NRBF2	[66]	rs10761731
Tumor biomarkers	1	FAM3B	[154]	rs441810
Type 1 diabetes	5	HLA-DRB5, HOXA5, LOC650557, SULT1A2, TUFM	[155–157]	rs4788084, rs7804356, rs9272346
Type 1 diabetes autoantibodies	2	SULT1A2, TUFM	[158]	rs4788084
Type 2 diabetes	2	MAEA, PSTPIP1	[159–161]	rs6815464, rs7178572
Ulcerative colitis	6	HLA-DRB5, LOC650557, MANBA, NFKB1, PIM3, UTS2	[95,162–164]	rs35675666, rs3774959, rs5771069, rs6927022, rs9271366
Urate levels	2	SF1, THBS3	[165,166]	rs11264341, rs606458
Ventricular conduction	1	UBE2L3	[167]	rs13165478
Vitiligo	2	HCG4, RNASET2	[168,169]	rs2236313, rs6904029
Weight	1	TUFM	[58]	rs7498665

Table S3. Genes reported by IPA for the category Dermatological Diseases and Conditions

Diseases or Functions Annotation	P-Value	Molecules	# Molecules
Psoriasis	8.05×10^{-4}	ACPP,ACTA2,ANXA1,APOBEC3A,ARG1,ATXN3,C3AR1,CLEC7A,CSTB,CTSB,CTSK,CTSS,DBN1,EIF5,EIF5A,FAM26F,GM2A,GNA15,GSN,HAL,HLA-G,HMOX1,IL16,KDELR2,KRT10,KYNU,LILRB1,MBP,PTPRC,RGS1,RNASE3,S100A12,S100P,SEC23B,SIGIRR,TEK,THBD,TYMP,UBE2I,VNN1,VWF,YWHAB,ZNF91	43
Dermatitis	1.46×10^{-4}	ACADVL,ANXA1,ANXA5,C3AR1,CAPZB,CCR3,CPNE1,CSK,CYTIP,F2RL1,FAS,FCER1G,FLOT1,GNB1L,GSN,GSR,HLA-DRB5,HNRNPR,IL16,KRT10,NFKB1,PGLYRP1,POLE4,RABGEF1,RNASE3,RNF138,RPS6KA4,SELL,SIGIRR,SPAG1,SPI1,STAT6,SYK,THBD,UBE2I	35
Atopic Dermatitis	1.78×10^{-2}	ACADVL,ANXA1,ANXA5,CAPZB,CCR3,CPNE1,F2RL1,FCER1G,FLOT1,GNB1L,GSN,HLA-DRB5,HNRNPR,KRT10,RNF138,SPAG1,SPI1,SYK,UBE2I	19
Dermatitis of ear	2.48×10^{-2}	CYTIP,STAT6	2

REFERENCES

1. Stranger, B. E. *et al.* Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science* **315,** 848–53 (2007).
2. Dermitzakis, E. T. & Stranger, B. E. Genetic variation in human gene expression. *Mamm. Genome* **17,** 503–8 (2006).
3. Storey, J. D. *et al.* Gene-expression variation within and among human populations. *Am. J. Hum. Genet.* **80,** 502–9 (2007).
4. Hulse, A. M. & Cai, J. J. Genetic variants contribute to gene expression variability in humans. *Genetics* **193,** 95–108 (2013).
5. Cookson, W., Liang, L., Abecasis, G., Moffatt, M. & Lathrop, M. Mapping complex disease traits with global gene expression. *Nat. Rev. Genet.* **10,** 184–194 (2009).
6. Knight, J. C. Regulatory polymorphisms underlying complex disease traits. *J. Mol. Med. (Berl).* **83,** 97–109 (2005).
7. Dixon, A. L. *et al.* A genome-wide association study of global gene expression. *Nat. Genet.* **39,** 1202–7 (2007).
8. Myers, A. J. *et al.* A survey of genetic human cortical gene expression. *Nat. Genet.* **39,** 1494–9 (2007).
9. Schadt, E. E. *et al.* Mapping the genetic architecture of gene expression in human liver. *PLoS Biol.* **6,** e107 (2008).
10. Grundberg, E. *et al.* Mapping cis- and trans-regulatory effects across multiple tissues in twins. *Nat. Genet.* **44,** 1084–1089 (2012).

11. Ding, J. *et al.* Gene expression in skin and lymphoblastoid cells: Refined statistical method reveals extensive overlap in cis-eQTL signals. *Am. J. Hum. Genet.* **87,** 779–89 (2010).
12. Chen, Y. *et al.* Variations in DNA elucidate molecular networks that cause disease. *Nature* **452,** 429–35 (2008).
13. Dimas, A. S. *et al.* Common regulatory variation impacts gene expression in a cell type-dependent manner. *Science* **325,** 1246–50 (2009).
14. Stranger, B. E. *et al.* Population genomics of human gene expression. *Nat. Genet.* **39,** 1217–24 (2007).
15. Murphy, A. *et al.* Mapping of numerous disease-associated expression polymorphisms in primary peripheral blood CD4+ lymphocytes. *Hum. Mol. Genet.* **19,** 4745–57 (2010).
16. Ferraro, a. *et al.* Interindividual variation in human T regulatory cells. *Proc. Natl. Acad. Sci.* **111,** E1111–E1120 (2014).
17. Zeller, T. *et al.* Genetics and beyond-the transcriptome of human monocytes and disease susceptibility. *PLoS One* **5,** e10693 (2010).
18. Fairfax, B. P. *et al.* Genetics of gene expression in primary immune cells identifies cell type-specific master regulators and roles of HLA alleles. *Nat. Genet.* **44,** 502–510 (2012).
19. Fehrmann, R. S. N. *et al.* Trans-eQTLs reveal that independent genetic variants associated with a complex phenotype converge on intermediate genes, with a major role for the HLA. *PLoS Genet.* **7,** e1002197 (2011).

20. Mehta, D. *et al.* Impact of common regulatory single-nucleotide variants on gene expression profiles in whole blood. *Eur. J. Hum. Genet.* **21,** 48–54 (2013).
21. Kumar, V. & Sharma, a. Neutrophils: Cinderella of innate immune system. *Int. Immunopharmacol.* **10,** 1325–34 (2010).
22. Lehrer, R. I., Ganz, T., Selsted, M. E., Babior, B. M. & Curnutte, J. T. Neutrophils and host defense. *Ann. Intern. Med.* **109,** 127–142 (1988).
23. Pillay, J. *et al.* A subset of neutrophils in human systemic inflammation inhibits T cell responses through Mac-1. *J. Clin. Invest.* **122,** 327–336 (2012).
24. Beutler, B. Innate immunity: an overview. *Mol. Immunol.* **40,** 845–859 (2004).
25. The 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature* **467,** 1061–1073 (2010).
26. Hinrichs, A. S. *et al.* The UCSC Genome Browser Database: update 2006. *Nucleic Acids Res.* **34,** D590–D598 (2006).
27. Nair, R. P. *et al.* Genome-wide scan reveals association of psoriasis with IL-23 and NF-kappaB pathways. *Nat. Genet.* **41,** 199–204 (2009).
28. Yao, Y. *et al.* Type I interferon: potential therapeutic target for psoriasis? *PLoS One* **3,** e2737 (2008).
29. Plager, D. a *et al.* Early cutaneous gene transcription changes in adult atopic dermatitis and potential clinical implications. *Exp. Dermatol.* **16,** 28–36 (2007).

30. R Development Core Team. *R: A Language and Environment for Statistical Computing.* (2012).
31. Benjamini, Y. & Yekutieli, D. The control of the false discovery rate in multiple testing under dependency. *Ann. Stat.* **29,** 1165–1188 (2001).
32. Shi, L. *et al.* The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nat. Biotechnol.* **24,** 1151–61 (2006).
33. Mollinedo, F., Borregaard, N. & Boxer, L. Novel trends in neutrophil structure, function and development. *Immunol. Today* 535–537 (1999).
34. Von Vietinghoff, S. & Ley, K. Homeostatic Regulation of Blood Neutrophil Counts. *J. Immunol.* **181,** 5183–5188 (2008).
35. Nathan, C. Neutrophils and immunity: challenges and opportunities. *Nat. Rev. Immunol.* **6,** 173–82 (2006).
36. Westra, H.-J. *et al.* Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nat. Genet.* **45,** 1238–1243 (2013).
37. Welter, D. *et al.* The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.* **42,** D1001–6 (2014).
38. Wang, X. *et al.* Mapping of hepatic expression quantitative trait loci (eQTLs) in a Han Chinese population. *J. Med. Genet.* **51,** 319–26 (2014).

39. Sasayama, D. *et al.* Identification of single nucleotide polymorphisms regulating peripheral blood mRNA expression with genome-wide significance: an eQTL study in the Japanese population. *PLoS One* **8,** e54967 (2013).

40. Martin, P. & Leibovich, S. J. Inflammatory cells during wound repair: the good, the bad and the ugly. *Trends Cell Biol.* **15,** 599–607 (2005).

41. Gaspari, A. a. Innate and adaptive immunity and the pathophysiology of psoriasis. *J. Am. Acad. Dermatol.* **54,** S67–80 (2006).

42. Nomura, I. *et al.* Cytokine Milieu of Atopic Dermatitis, as Compared to Psoriasis, Skin Prevents Induction of Innate Immune Response Genes. *J. Immunol.* **171,** 3262–3269 (2003).

43. Terui, T., Ozawa, M. & Tagami, H. Role of neutrophils in induction of acute inflammation in T-cell-mediated immune dermatosis, psoriasis: a neutrophil-associated inflammation-boosting loop. *Exp. Dermatol.* **9,** 1–10 (2000).

44. Leung, D. Y. Atopic dermatitis: new insights and opportunities for therapeutic intervention. *J Allergy Clin Immunol* **105,** 860–876 (2000).

45. Leung, D. Y. M. & Bieber, T. Atopic dermatitis. *Lancet* **361,** 151–60 (2003).

46. Christophers, E. & Henseler, T. Contrasting disease patterns in psoriasis and atopic dermatitis. *Arch. Dermatol. Res.* **279 Suppl,** S48–S51 (1987).

47. Hollingworth, P. *et al.* Common variants at ABCA7, MS4A6A/MS4A4E, EPHA1, CD33 and CD2AP are associated with Alzheimer's disease. *Nat. Genet.* **43,** 429–35 (2011).

48. Naj, A. C. *et al.* Common variants at MS4A4/MS4A6E, CD2AP, CD33 and EPHA1 are associated with late-onset Alzheimer's disease. *Nat. Genet.* **43,** 436–41 (2011).

49. Reveille, J. D. *et al.* Genome-wide association study of ankylosing spondylitis identifies non-MHC susceptibility loci. *Nat. Genet.* **42,** 123–7 (2010).

50. Evans, D. M. *et al.* Interaction between ERAP1 and HLA-B27 in ankylosing spondylitis implicates peptide handling in the mechanism for HLA-B27 in disease susceptibility. *Nat. Genet.* **43,** 761–7 (2011).

51. Moffatt, M. F. *et al.* A large-scale, consortium-based genomewide association study of asthma. *N. Engl. J. Med.* **363,** 1211–21 (2010).

52. Lasky-Su, J. *et al.* HLA-DQ strikes again: genome-wide association study further confirms HLA-DQ in the diagnosis of asthma among adults. *Clin. Exp. Allergy* **42,** 1724–33 (2012).

53. Kirino, Y. *et al.* Genome-wide association analysis identifies new susceptibility loci for Behçet's disease and epistasis between HLA-B*51 and ERAP1. *Nat. Genet.* **45,** 202–207 (2013).

54. Tin, A. *et al.* Genome-wide association study identified the human leukocyte antigen region as a novel locus for plasma beta-2 microglobulin. *Hum. Genet.* **132,** 619–27 (2013).

55. Sulem, P. *et al.* Two newly identified genetic determinants of pigmentation in Europeans. *Nat. Genet.* **40,** 835–7 (2008).
56. Wain, L. V *et al.* Genome-wide association study identifies six new loci influencing pulse pressure and mean arterial pressure. *Nat. Genet.* **43,** 1005–11 (2011).
57. Speliotes, E. K. *et al.* Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nat. Genet.* **42,** 937–48 (2010).
58. Thorleifsson, G. *et al.* Genome-wide association yields new sequence variants at seven loci that associate with measures of obesity. *Nat. Genet.* **41,** 18–24 (2009).
59. Willer, C. J. *et al.* Six new loci associated with body mass index highlight a neuronal influence on body weight regulation. *Nat. Genet.* **41,** 25–34 (2009).
60. Shen, L. *et al.* Whole genome association study of brain-wide imaging phenotypes for identifying quantitative trait loci in MCI and AD: A study of the ADNI cohort. *Neuroimage* **53,** 1051–63 (2010).
61. Stein, J. L. *et al.* Identification of common variants associated with human hippocampal and intracranial volumes. *Nat. Genet.* **44,** 552–561 (2012).
62. Gold, B. *et al.* Genome-wide association study provides evidence for a breast cancer risk locus at 6q22.33. *Proc. Natl. Acad. Sci. U. S. A.* **105,** 4340–4345 (2008).
63. Buil, A. *et al.* C4BPB/C4BPA is a new susceptibility locus for venous thrombosis with unknown protein S-independent

mechanism: results from genome-wide association and gene expression analyses followed by case-control studies. *Blood* **115,** 4644–4650 (2010).

64. Hunt, K. A. *et al.* Newly identified genetic risk variants for celiac disease related to the immune response. *Nat. Genet.* **40,** 395–402 (2008).

65. Dubois, P. C. a *et al.* Multiple common variants for celiac disease influencing immune gene expression. *Nat. Genet.* **42,** 295–302 (2010).

66. Teslovich, T. M. *et al.* Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* **466,** 707–13 (2010).

67. Kim, D. K. *et al.* Genome-wide association analysis of blood biomarkers in chronic obstructive pulmonary disease. *Am. J. Respir. Crit. Care Med.* **186,** 1238–47 (2012).

68. Beaty, T. H. *et al.* A genome-wide association study of cleft lip with and without cleft palate identifies risk variants near MAFB and ABCA4. *Nat. Genet.* **42,** 525–529 (2010).

69. Amin, N. *et al.* Genome-wide association analysis of coffee drinking suggests association with CYP1A1/CYP1A2 and NRCAM. *Mol. Psychiatry* **17,** 1116–29 (2012).

70. Davies, R. W. *et al.* A genome-wide association study for coronary artery disease identifies a novel susceptibility locus in the major histocompatibility complex. *Circ. Cardiovasc. Genet.* **5,** 217–225 (2012).

71. Kenny, E. E. *et al.* A genome-wide scan of ashkenazi jewish crohn's disease suggests novel susceptibility loci. *PLoS Genet.* **8,** (2012).

72. Barrett, J. C. *et al.* Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. *Nat. Genet.* **40,** 955–62 (2008).

73. Franke, A. *et al.* Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nat. Genet.* **42,** 1118–25 (2010).

74. Khor, C. C. *et al.* Genome-wide association study identifies susceptibility loci for dengue shock syndrome at MICB and PLCE1. *Nat. Genet.* **43,** 1139–1141 (2011).

75. Ehret, G. B. *et al.* Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk. *Nature* **478,** 103–109 (2011).

76. Newton-Cheh, C. *et al.* Genome-wide association study identifies eight loci associated with blood pressure. *Nat. Genet.* **41,** 666–676 (2009).

77. Levy, D. *et al.* Genome-wide association study of blood pressure and hypertension. *Nat. Genet.* **41,** 677–687 (2009).

78. Lucena, M. I. *et al.* Susceptibility to amoxicillin-clavulanate-induced liver injury is influenced by multiple HLA class i and II alleles. *Gastroenterology* **141,** 338–347 (2011).

79. Albertsen, H. M., Chettier, R., Farrington, P. & Ward, K. Genome-Wide Association Study Link Novel Loci to Endometriosis. *PLoS One* **8,** (2013).

80. Gudbjartsson, D. F. *et al.* Sequence variants affecting eosinophil numbers associate with asthma and myocardial infarction. *Nat. Genet.* **41,** 342–7 (2009).

81. Rubicz, R. *et al.* A genome-wide integrative genomic study localizes genetic factors influencing antibodies against Epstein-Barr virus nuclear antigen 1 (EBNA-1). *PLoS Genet.* **9,** e1003147 (2013).

82. Liu, F. *et al.* Digital quantification of human eye color highlights genetic association of three new loci. *PLoS Genet.* **6,** 34 (2010).

83. Vithana, E. N. *et al.* Genome-wide association analyses identify three new susceptibility loci for primary angle closure glaucoma. *Nat. Genet.* **44,** 1142–1146 (2012).

84. Chu, X. *et al.* A genome-wide association study identifies two new risk loci for Graves' disease. *Nat. Genet.* **43,** 897–901 (2011).

85. Nakabayashi, K. *et al.* Identification of independent risk loci for Graves' disease within the MHC in the Japanese population. *J. Hum. Genet.* **56,** 772–778 (2011).

86. Kathiresan, S. *et al.* Common variants at 30 loci contribute to polygenic dyslipidemia. *Nat. Genet.* **41,** 56–65 (2009).

87. Sabatti, C. *et al.* Genome-wide association analysis of metabolic traits in a birth cohort from a founder population. *Nat. Genet.* **41,** 35–46 (2009).

88. Willer, C. J. *et al.* Newly identified loci that influence lipid concentrations and risk of coronary artery disease. *Nat. Genet.* **40,** 161–169 (2008).

89. Lango Allen, H. *et al.* Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature* **467,** 832–838 (2010).

90. Paré, G. *et al.* Novel associations of CPS1, MUT, NOX4, and DPEP1 with plasma Homocysteine in a healthy population a genome-wide evaluation of 13 974 participants in the women's genome health study. *Circ. Cardiovasc. Genet.* **2,** 142–150 (2009).

91. Eriksson, N. *et al.* Novel associations for hypothyroidism include known autoimmune risk loci. *PLoS One* **7,** (2012).

92. Yu, X.-Q. *et al.* A genome-wide association study in Han Chinese identifies multiple susceptibility loci for IgA nephropathy. *Nat. Genet.* **44,** 178–182 (2011).

93. Ferreira, R. C. *et al.* Association of IFIH1 and other autoimmunity risk alleles with selective IgA deficiency. *Nat. Genet.* **42,** 777–80 (2010).

94. Okada, Y. *et al.* HLA-Cw*1202-B*5201-DRB1*1502 haplotype increases risk for ulcerative colitis but reduces risk for Crohn's disease. *Gastroenterology* **141,** 864–871.e1–5 (2011).

95. Jostins, L. *et al.* Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature* **491,** 119–24 (2012).

96. Kugathasan, S. *et al.* Loci on 20q13 and 21q22 are associated with pediatric-onset inflammatory bowel disease. *Nat. Genet.* **40,** 1211–1215 (2008).

97. Kayser, M., Liu, F. & Janssens, A. Three Genome-wide Association Studies and a Linkage Analysis Identify HERC2 as a Human Iris Color Gene. *Am. J. Hum. Genet.* **82,** 411–423 (2008).

98. Zhang, F. R. *et al.* Genomewide association study of leprosy. *N Engl J Med* **361,** 2609–2618 (2009).

99. Chu, A. Y. *et al.* Genome-wide association study evaluating lipoprotein-associated phospholipase A2 mass and activity at baseline and after rosuvastatin therapy. *Circ. Cardiovasc. Genet.* **5,** 676–85 (2012).

100. Yuan, X. *et al.* Population-Based Genome-wide Association Studies Reveal Six Loci Influencing Plasma Levels of Liver Enzymes. *Am. J. Hum. Genet.* **83,** 520–528 (2008).

101. Chambers, J. C. *et al.* Genome-wide association study identifies loci influencing concentrations of liver enzymes in plasma. *Nat. Genet.* **43,** 1131–8 (2011).

102. Vijai, J. *et al.* Susceptibility Loci Associated with Specific and Shared Subtypes of Lymphoid Malignancies. *PLoS Genet.* **9,** (2013).

103. Meyer, T. E. *et al.* Genome-wide association studies of serum magnesium, potassium, and sodium concentrations identify six Loci influencing serum magnesium levels. *PLoS Genet.* **6,** (2010).

104. Ganesh, S. K. *et al.* Multiple loci influence erythrocyte phenotypes in the CHARGE Consortium. *Nat. Genet.* **41,** 1191–1198 (2009).

105. Soranzo, N. *et al.* A genome-wide meta-analysis identifies 22 loci associated with eight hematological parameters in the HaemGen consortium. *Nat. Genet.* **41,** 1182–1190 (2009).

106. Qayyum, R. *et al.* A meta-analysis and genome-wide association study of platelet count and mean platelet volume in African Americans. *PLoS Genet.* **8,** (2012).

107. Bishop, D., Demenais, F. & Iles, M. Genome-wide association study identifies three loci associated with melanoma risk. *Nat. Genet.* **41,** 920–925 (2009).

108. MacGregor, S. *et al.* Genome-wide association study identifies a new melanoma susceptibility locus at 1q21.3. *Nat. Genet.* **43,** 1114–1118 (2011).

109. Elks, C. E. *et al.* Thirty new loci for age at menarche identified by a meta-analysis of genome-wide association studies. *Nat. Genet.* **42,** 1077–1085 (2010).

110. Stolk, L. *et al.* Meta-analyses identify 13 loci associated with age at menopause and highlight DNA repair and immune pathways. *Nat. Genet.* **44,** 260–268 (2012).

111. Perry, J. R. B. *et al.* A genome-wide association study of early menopause and the combined impact of identified variants. *Hum. Mol. Genet.* **22,** 1465–1472 (2013).

112. Suhre, K. *et al.* Human metabolic individuality in biomedical and pharmaceutical research. *Nature* **477,** 54–60 (2011).

113. Kettunen, J. *et al.* Genome-wide association study identifies multiple loci influencing human serum metabolite levels. *Nat. Genet.* **44,** 269–276 (2012).

114. Sawcer, S., Hellenthal, G. & Pirinen, M. Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. *Nature* **476,** 214–219 (2011).

115. M, B. *et al.* Genome-wide association study identifies new multiple sclerosis susceptibility loci on chromosomes 12 and 20. *Nat. Genet.* **41,** 824–8 (2009).

116. De Jager, P. L. *et al.* Meta-analysis of genome scans and replication identify CD6, IRF8 and TNFRSF1A as new multiple sclerosis susceptibility loci. *Nat. Genet.* **41,** 776–782 (2009).

117. Nischwitz, S. *et al.* Evidence for VAV2 and ZNF433 as susceptibility genes for multiple sclerosis. *J. Neuroimmunol.* **227,** 162–166 (2010).

118. Mero, I. L. *et al.* Oligoclonal Band Status in Scandinavian Multiple Sclerosis Patients Is Associated with Specific Genetic Risk Alleles. *PLoS One* **8,** (2013).

119. Tse, K. P. *et al.* Genome-wide Association Study Reveals Multiple Nasopharyngeal Carcinoma-Associated Loci within the HLA Region at Chromosome 6p21.3. *Am. J. Hum. Genet.* **85,** 194–203 (2009).

120. Bei, J.-X. *et al.* A genome-wide association study of nasopharyngeal carcinoma identifies three new susceptibility loci. *Nat. Genet.* **42,** 599–603 (2010).

121. Del Greco, M. F. *et al.* Genome-wide association analysis and fine mapping of NT-proBNP level provide novel insight into the role of the MTHFR-CLCN6-NPPA-NPPB gene cluster. *Hum. Mol. Genet.* **20,** 1660–1671 (2011).

122. Ludwig, K. U. *et al.* Genome-wide meta-analyses of nonsyndromic cleft lip with or without cleft palate identify six new risk loci. *Nat. Genet.* **44,** 968–971 (2012).

123. Kim, H., Ramsay, E. & Lee, H. Genome-wide association study of acute post-surgical pain in humans. *Pharmacogenomics* **10,** 171–179 (2009).

124. Satake, W. *et al.* Genome-wide association study identifies common variants at four loci as genetic risk factors for Parkinson's disease. *Nat. Genet.* **41,** 1303–1307 (2009).

125. Nalls, M. A. *et al.* Imputation of sequence variants for identification of genetic risks for Parkinson's disease: a meta-analysis of genome-wide association studies. *Lancet* **377,** 641–649 (2011).

126. Huang, J. *et al.* Genome-wide association study for circulating levels of PAI-1 provides novel insights into its regulation. *Blood* **120,** 11–13 (2012).

127. Gieger, C. *et al.* New gene functions in megakaryopoiesis and platelet formation. *Nature* **480,** 201–208 (2011).

128. Mells, G. F. *et al.* Genome-wide association study identifies 12 new susceptibility loci for primary biliary cirrhosis. *Nat. Genet.* **43,** 329–332 (2011).

129. Strawbridge, R. J. *et al.* Genome-wide association identifies nine common variants associated with fasting proinsulin levels and provides new insights into the pathophysiology of type 2 diabetes. *Diabetes* **60,** 2624–34 (2011).

130. Takata, R. *et al.* Genome-wide association study identifies five new susceptibility loci for prostate cancer in the Japanese population. *Nat. Genet.* **42,** 751–754 (2010).

131. Xu, J. *et al.* Genome-wide association study in Chinese men identifies two new prostate cancer risk loci at 9q31.2 and 19q13.4. *Nat. Genet.* **44,** 1231–1235 (2012).

132. Eeles, R. A. *et al.* Identification of seven new prostate cancer susceptibility loci through a genome-wide association study. *Nat. Genet.* **41,** 1116–1121 (2009).

133. Thomas, G. *et al.* Multiple loci identified in a genome-wide association study of prostate cancer. *Nat. Genet.* **40,** 310–315 (2008).

134. Eeles, R. A. *et al.* Multiple newly identified loci associated with prostate cancer susceptibility. *Nat. Genet.* **40,** 316–321 (2008).

135. Kote-Jarai, Z. *et al.* Seven prostate cancer susceptibility loci identified by a multi-stage genome-wide association study. *Nat. Genet.* **43,** 785–791 (2011).

136. Gudmundsson, J. *et al.* Genetic correction of PSA values using sequence variants associated with PSA levels. *Sci. Transl. Med.* **2,** 62ra92 (2010).

137. Stuart, P., Nair, R., Ellinghaus, E. & Ding, J. Genome-wide association analysis identifies three psoriasis susceptibility loci. *Nat. Genet.* **42,** 1000–1004 (2010).

138. Strange, A. *et al.* A genome-wide association study identifies new psoriasis susceptibility loci and an interaction between HLA-C and ERAP1. *Nat. Genet.* **42,** 985–90 (2010).

139. Pfeufer, A. *et al.* Common variants at ten loci modulate the QT interval duration in the QTSCD Study. *Nat. Genet.* **41,** 407–414 (2009).

140. Van der Harst, P. *et al.* Seventy-five genetic loci influencing the human red blood cell. *Nature* **492,** 369–75 (2012).

141. Okada, Y. *et al.* Meta-analysis identifies multiple loci associated with kidney function-related traits in east Asian populations. *Nat. Genet.* **44,** 904–9 (2012).

142. Hu, H. J. *et al.* Common variants at the promoter region of the APOM confer a risk of rheumatoid arthritis. *Exp. Mol. Med.* **43,** 613–21 (2011).

143. Stahl, E. A. *et al.* Genome-wide association study meta-analysis identifies seven new rheumatoid arthritis risk loci. *Nat. Genet.* **42,** 508–514 (2010).

144. Okada, Y. *et al.* Meta-analysis identifies nine new loci associated with rheumatoid arthritis in the Japanese population. *Nat. Genet.* **44,** 511–516 (2012).

145. Eleftherohorinou, H., Hoggart, C. J., Wright, V. J., Levin, M. & Coin, L. J. M. Pathway-driven gene stability selection of two rheumatoid arthritis GWAS identifies and validates new susceptibility genes in receptor mediated signalling pathways. *Hum. Mol. Genet.* **20,** 3494–3506 (2011).

146. Gregersen, P. K. *et al.* REL, encoding a member of the NF-kappaB family of transcription factors, is a newly defined risk locus for rheumatoid arthritis. *Nat. Genet.* **41,** 820–823 (2009).

147. Plenge, R. M. *et al.* TRAF1-C5 as a risk locus for rheumatoid arthritis--a genomewide study. *N. Engl. J. Med.* **357,** 1199–1209 (2007).

148. Fischer, A. *et al.* A Novel Sarcoidosis Risk Locus for Europeans on Chromosome 11q13.1. *Am. J. Respir. Crit. Care Med.* **186,** 877–885 (2012).

149. Shi, Y. *et al.* Common variants on 8p12 and 1q24.2 confer risk of schizophrenia. *Nat. Genet.* **43,** 1224–1227 (2011).

150. Demirkan, A. *et al.* Genome-wide association study identifies novel loci associated with circulating phospho- and sphingolipid concentrations. *PLoS Genet.* **8,** (2012).

151. Génin, E. *et al.* Genome-wide association study of Stevens-Johnson Syndrome and Toxic Epidermal Necrolysis in Europe. *Orphanet J Rare Dis* **6,** 52 (2011).

152. Han, J.-W. *et al.* Genome-wide association study in a Chinese Han population identifies nine new susceptibility loci for systemic lupus erythematosus. *Nat. Genet.* **41,** 1234–7 (2009).

153. Yang, W. *et al.* Meta-analysis followed by replication identifies loci in or near CDKN1B, TET3, CD80, DRAM1, and ARID5B as associated with systemic lupus erythematosus in Asians. *Am. J. Hum. Genet.* **92,** 41–51 (2013).

154. He, M. *et al.* A genome wide association study of genetic loci that influence tumour biomarkers cancer antigen 19-9, carcinoembryonic antigen and α fetoprotein and their associations with cancer risk. *Gut* 143–151 (2013).

155. Barrett, J. C. *et al.* Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. *Nat. Genet.* **41,** 703–7 (2009).

156. Wellcome, T., Case, T. & Consortium, C. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447,** 661–78 (2007).

157. Cooper, J. D. *et al.* Meta-analysis of genome-wide association study data identifies additional type 1 diabetes risk loci. *Nat. Genet.* **40,** 1399–401 (2008).

158. Plagnol, V. *et al.* Genome-wide association analysis of autoantibody positivity in type 1 diabetes cases. *PLoS Genet.* **7,** e1002216 (2011).

159. Kooner, J. S. *et al.* Genome-wide association study in individuals of South Asian ancestry identifies six new type 2 diabetes susceptibility loci. *Nat. Genet.* **43,** 984–9 (2011).

160. Cho, Y. S. *et al.* Meta-analysis of genome-wide association studies identifies eight new loci for type 2 diabetes in east Asians. *Nat. Genet.* **44,** 67–72 (2012).

161. Perry, J. R. B. *et al.* Stratifying type 2 diabetes cases by BMI identifies genetic risk variants in LAMA1 and enrichment for risk variants in lean compared to obese cases. *PLoS Genet.* **8,** (2012).

162. Franke, A. *et al.* Genome-wide association study for ulcerative colitis identifies risk loci at 7q22 and 22q13 (IL17REL). *Nat. Genet.* **42,** 292–294 (2010).

163. Yang, S.-K. *et al.* Genome-wide association study of ulcerative colitis in Koreans suggests extensive overlapping of genetic susceptibility with Caucasians. *Inflamm. Bowel Dis.* **19,** 954–66 (2013).

164. Anderson, C. A. *et al.* Meta-analysis identifies 29 additional ulcerative colitis risk loci, increasing the number of confirmed associations to 47. *Nat. Genet.* **43,** 246–252 (2011).

165. Tin, A. *et al.* Genome-wide association study for serum urate concentrations and gout among African Americans identifies genomic risk loci and a novel URAT1 loss-of-function allele. *Hum. Mol. Genet.* **20,** 4056–4068 (2011).

166. Köttgen, A. *et al.* Genome-wide association analyses identify 18 new loci associated with serum urate concentrations. *Nat. Genet.* **45,** 145–54 (2013).

167. Sotoodehnia, N. *et al.* Common variants in 22 loci are associated with QRS duration and cardiac ventricular conduction. *Nat. Genet.* **42,** 1068–1076 (2010).

168. Quan, C. *et al.* Genome-wide association study for vitiligo identifies susceptibility loci at 6q27 and the MHC. *Nat. Genet.* **42,** 614–618 (2010).

169. Jin, Y. *et al.* Variant of TYR and autoimmunity susceptibility loci in generalized vitiligo. *N. Engl. J. Med.* **362,** 1686–1697 (2010).

CHAPTER 3

GENETIC ANALYSIS OF AN ALLERGIC RHINITIS COHORT REVEALS AN INTERCELLULAR EPISTASIS BETWEEN FAM134B AND CD39

Rossella Melchiotti[a,b,2], Kia Joo Puan[a,2], Anand Kumar Andiappan[a,2], Tuang Yeow Poh[a], Mireille Starke[a], Lydia Li Zhuang[a], Kerstin Petsch[a], Adrian Tuck Siong Lai[a], Fook Tim Chew[c], Anis Larbi[a], De Yun Wang[d,1], Michael Poidinger[a,1], Olaf Rötzschke[a,1]

[a]SIgN (Singapore Immunology Network), A*STAR (Agency for Science, Technology and Research), Singapore 138648, Singapore; [b]Doctoral School in Translational and Molecular Medicine (DIMET), University of Milano-Bicocca, Milan 20126, Italy; [c]Department of Biological Sciences, National University of Singapore, Singapore 117543, Singapore; [d]Department of Otolaryngology, National University of Singapore, Singapore 119228, Singapore.

[1]To whom correspondence may be addressed. E-mail: olaf_rotzschke@immunol.a-star.edu.sg, entwdy@nus.edu.sg, or michael_poidinger@immunol.a-star.edu.sg.

[2]Authors contributed equally to the study

Accepted by BMC Medical Genetics

Abstract

Background

Extracellular ATP is a pro-inflammatory molecule released by damaged cells. Regulatory T cells (Treg) can suppress inflammation by hydrolysing this molecule via ectonucleoside triphosphate diphosphohydrolase 1 (ENTPD1), also termed as CD39. Multiple studies have reported differences in $CD39^+$ Treg percentages in diseases such as multiple sclerosis, Hepatitis B and HIV-1. In addition, CD39 polymorphisms have been implicated in immune-phenotypes such as susceptibility to inflammatory bowel disease and AIDS progression. However none of the studies published so far has linked disease-associated variants with differences in CD39 Treg surface expression. This study aims at identifying variants affecting CD39 expression on Treg and at evaluating their association with allergic rhinitis, a disease characterized by a strong Treg involvement.

Methods

Cohorts consisting of individuals of different ethnicities were employed to identify any association of CD39 variants to surface expression. Significant variant(s) were tested for disease association in a published GWAS cohort by one-locus and two-locus genetic analyses based on logistic models. Further functional characterization was performed using existing microarray data and quantitative RT-PCR on sorted cells.

Results

Our study shows that rs7071836, a promoter SNP in the CD39 gene region, affects the cell surface expression on Treg cells but not on other CD39+ leukocyte subsets. Epistasis analysis revealed that, in conjunction with a SNP upstream of the FAM134B gene (rs257174), it increased the risk of allergic rhinitis (P = 1.98 x 10^{-6}). As a promoter SNP, rs257174 controlled the expression of the gene in monocytes but, notably, not in Treg cells. Whole blood transcriptome data of three large cohorts indicated an inverse relation in the expression of the two proteins. While this observation was in line with the epistasis data, it also implied that a functional link must exist. Exposure of monocytes to extracellular ATP resulted in an up-regulation of FAM134B gene expression, suggesting that extracellular ATP released from damaged cells represents the connection for the biological interaction of CD39 on Treg cells with FAM134B on monocytes.

Conclusions

The interplay between promoter SNPs of CD39 and FAM134B results in an intercellular epistasis which influences the risk of a complex inflammatory disease.

Background

Allergic Rhinitis (AR) is a common airway disease where allergen exposure triggers an IgE-mediated immune response. The typical symptoms include nasal itchiness, rhinorrhea, sneezing and progressive blockage of the inflamed nasal passages[1]. The disease is driven by a complex interplay of various leukocytes, including mast cells, eosinophils and basophils but also CD4+ T cells, IgE-producing B cells and dendritic cells. Th2 cytokines such as IL-4, IL-5 and IL-13 drive IgE production, promote eosinophil infiltration to the nasal mucosa, and stimulate mast cell release of key vasoactive mediators such as histamine[2–4]. In this context also monocytes are important effectors and regulators of inflammation[5]. While pro-inflammatory monocytes can fuel the allergic reaction by releasing cytokines such as TNF-α and IL-6, they can be converted into anti-inflammatory monocytes to dampen the reaction[6].

Central to the prevention or attenuation of pro-inflammatory immune responses are CD4+ Foxp3+ T regulatory cells (Treg). They can inhibit the proliferation of CD4+ effector T cells and impair the production of various Th2 cytokines[7–10]. A key mechanism by which Treg exert their regulatory function is the expression of the ectoenzyme CD39[11,12]. CD39 is involved in the hydrolysis of extracellular ATP, which is typically released from cells following tissue damage. ATP-sensors such as P2X- and P2Y-receptors are important mediators of allergic airway inflammation and their blockade has been shown to strongly suppress allergic reactions in experimental asthma models[13,14]. Treg-expressed CD39 thus

contributes to the control of inflammation via the removal of ATP[11,12]. Genetic and phenotypic correlations of CD39 variants have already revealed strong associations with inflammatory bowel disease[15] and multiple sclerosis[11,16], as well as with viral infections including HIV[17] and Hepatitis B[18]. However, a link between CD39 and allergic diseases has not previously been identified.

In the current report we demonstrate that variation in cell surface expression in Treg cells is associated with a genetic polymorphism (rs7071836) located in the promoter region of CD39. On its own this polymorphism had no direct impact on risk of AR but associated with disease risk through an epistatic interaction with rs257174, a promoter SNP of the *cis*-Golgi protein FAM134B. As rs257174 alters the gene expression in monocytes but not in Treg cells this represents the first example of an intercellular epistasis.

Methods

Ethics statement

This study has been approved by the Institutional Review Board of the National University of Singapore (IRB ref. NUS 07-023, NUS 10-343 and NUS 09-256) and complies with the Helsinki declaration. Written informed consent was obtained from all donors prior to sample collection.

Study populations

Case-control cohorts

We used two age-matched cohorts of Singapore Chinese individuals[19], including 456 AR cases and 486 non-atopic controls for

investigative discovery and a separate cohort of 676 AR cases, 511 non-atopic controls and 1647 atopic individuals without AR symptoms for validating our statistical interaction and estimating the role played by atopy in the epistasis. These two cohorts were used to evaluate the role played by the variant rs7071836 in AR risk both in the context of disease association and epistatic phenomena. Details of sample collection for each cohort and the genotyping and quality control filters applied for the discovery cohort, are described in[19]. Cases were defined as individuals displaying a positive skin prick test for at least one of the two house dust mite allergens tested (*Dermatophagoides pteronyssinus, Blomia tropicalis*) and exhibiting two or more symptoms of nasal blockage, sneezing, nasal itching, and rhinorrhea. Controls were defined as skin-prick test negative individuals with no history of allergic disease or AR symptoms. All individuals were genotyped using the Illumina HumanHap 550k BeadChip version 3 (Illumina, San Diego, California). Due to constraints in the number of available Sequenom slots 20 SNPs were selected for replication. Genotyping of the 20 SNPs chosen for replication in the validation cohort was performed using Sequenom's MassARRAY system and iPLEX technology (Sequenom Inc, San Diego) on 2834 samples. The experiment was carried out according to the manufacturer's guidelines. SNPs were called using the Sequenom TYPER software and were checked for deviation from Hardy-Weinberg equilibrium (variants with an adjusted P-value <0.05 in controls only or across all samples were excluded from the analysis; rs2900474, $P_{unaffected}$= 6.01 x 10^{-10}; rs4862396, $P_{unaffected}$= 8.53 x 10^{-4}).

Functional cohorts

An independent cohort of 165 ethnic Chinese volunteers was recruited to validate the genotype-phenotype association between rs7071836 and CD39 protein expression. Samples were collected in a similar manner and were age- and gender-matched to the case-control cohorts. Genotyping was performed on a genome-wide scale using the Illumina HumanOmni5-Quad chip (Illumina, San Diego, California) on DNA extracted from blood following standard protocols. SNP calling was carried out using the Genome Studio genotyping module (Illumina, San Diego, California). The same quality control filters were applied as those described for the case-control cohorts.

A small cohort of 41 self-reported Chinese individuals and 22 self-reported Caucasian subjects was recruited internally and then phenotyped for CD39 protein expression by T regulatory cells to evaluate the relative frequency of the 'CD39lo' phenotype in each ethnic group.

Published whole blood cohorts

Three published cohorts (Kora F4[20], DILGOM[21], SHIP-TREND[22]) for which whole blood gene expression measurements were available were used to correlate CD39/FAM134B expression levels. The three cohorts are composed respectively of 993, 518 and 991 healthy individuals of Caucasian ethnicity. A detailed description of how each cohort was collected and samples were processed can be found in the respective publications. Gene expression processed values were

downloaded from the corresponding public online repositories Array Express (E-MTAB-1708, E-TABM-1036) and Gene Expression Omnibus (GEO) (GSE36382).

FACS analysis

In both functional cohorts, CD39 expression was determined by FACS staining of PBMCs pre-incubated with LIVE/DEAD Fixable Blue Dead Cell Stain kit (Invitrogen) to identify viable cells. The cells were then incubated with anti-CD39 APC (clone TÜ66), anti-CTLA-4 PE (clone BNI3), anti-CCR6 PerCP-Cy5.5 (clone 11A9), anti-CD4 APC-Cy7 (clone RPA-T4), anti-CD25 PE-Cy7 (clone M-A251), anti-CD45RA eFluor605 (clone H100) mAbs. Intracellular staining of Treg was conducted using the anti-FoxP3 eFluor450 (clone PCH101) Staining kit (eBioscience). Adult peripheral blood T cells either express CD45RA or CD45RO and few cells are double positive or double negative[23]. In our cohort, CD45RA- FoxP3+ CD25+ CD4+ T cells are termed as CD45RA- T regulatory (Treg) cells (CD45RO+) whereas CD45RA- FoxP3- CD25- CD4+ are CD4+ effector T (Teff) cells (CD45RO+). The level of CD39 expression was measured using a BD LSR II flow cytometer (BD Biosciences). The gating strategy for Treg, T effector cells, B cells and monocytes is outlined in Additional file 1: Figure S1. The variability of CD39 expression among human Treg from different donors was established by calculating the ratio of CD39 geometric mean fluorescence intensity relative to the CD39 geometric mean fluorescence of donor-matched B cells (which constitutively express high levels of CD39). Samples were classified as CD39 high expressing

Treg ('CD39hi') or CD39 low expressing Treg ('CD39lo') using the unsupervised clustering method *k*-means on the log_2-transformed ratios. The analysis was performed using the function *k*-means in R 2.15.1[24], with the number of clusters set to two. Each population was considered and classified separately. CD39 expression values for the two groups as clustered by the *k*-means method are depicted in Additional file 2: Figure S2.

Cell sorting

Human blood was collected into BD K_2 EDTA vacutainers (Becton, Dickinson and Company) and the PBMCs were isolated by centrifugation over Ficoll-Paque density gradients (GE Healthcare) for 30 min at 400 x g. PBMCs were then re-suspended in FACS buffer (0.5% bovine serum albumin, 2mM EDTA in PBS) and incubated at 4˚C for 15 min with anti-CD49d FITC (clone MZ18-24A9, Miltenyi Biotec), CD127 PE (clone HIL-7R-M21), CCR6 PerCP-Cy5.5 (clone 11A9), CD4 APC-Cy7 (clone RPA-T4) CD25 PE-CY7 (clone M-A251), and CD19 Alexa700 (clone HIB19). All mAbs were purchased from BD Biosciences unless otherwise stated. After incubation, the cells were washed and re-suspended in FACS buffer at 1- 1.5 x 10^7 cells/ml for cell sorting of Treg, T effector cells, B cells and monocytes using a BD FACS Aria II cell sorter (BD Biosciences). See Additional file 3: Figure S3 for cell sorting strategy.

Expression analysis

FACS-sorted Treg, T effector cells, B cells and monocytes were obtained from 15 healthy donor blood samples selected from the

discovery cohort. Target ssDNA was prepared starting with 50 ng total RNA (RIN ≥ 7.1) using the Ambion WT Expression Kit and the Affymetrix WT Terminal Labelling kit. Fragmented ssDNA was hybridized to the Affymetrix Human Exon 1.0ST Arrays. The GeneChip arrays were washed and stained using the GeneChip Fluidics Station 450. After staining, the GeneChip arrays were scanned using a GeneChip Scanner 3000 at the BSF Microarray Facility. Array QC was conducted using the Affymetrix Expression Console Software. Raw data were normalized using the Robust Multi-Array Average (RMA) at the gene level[25].

Monocyte ATP treatment

Monocytes from individuals heterozygous for rs257174 were purified from PBMCs by positive selection using MACS human CD14 MicroBeads (Miltenyi Biotec) according to manufacturer's instructions. Monocytes were plated at 0.5 x 10^6 cells/well in 24-well tissue culture plates and then incubated with or without cell culture-grade ATP disodium salt hydrate (Sigma-Aldrich) in RPMI-1640 supplemented with 10% fetal bovine serum, 2 mM L-glutamine, 1 mM sodium pyruvate, 100 units/ml of penicillin and 100 µg/ml of streptomycin. For the inhibition of ATP the incubation was carried out in the presence of 10 µM of the purinergic receptor inhibitor A-438079 (Sigma-Aldrich). Monocytes were incubated for 2h before harvesting using cell scrapers. The treated monocytes were stored in TRIzol (Invitrogen Life Technologies) at -80˚C until further analysis.

Genotyping

Genotyping was performed by PCR using SsoFast EvaGreen Supermix (Bio-Rad) on a CFX96 Real-Time System (Bio-Rad). DNA was isolated using DNeasy Blood & Tissue Kits (Qiagen) according to the manufacturer's instructions. The following primers were used for the analysis: FAM134B rs257174 forward primer: GCACGCTTTTGCCTTTGTAAT; FAM134B rs257174 reverse primer: CACCCACTGGGAGAAAAGAC. Amplification was carried out using the following protocol: 3 min at 95°C, 40 cycles of 5s at 95°C, 5s at 58°C, and final extension for 10s at 95°C. A melt curve was generated from 65 to 95°C (in 0.2°C increments) with 10s/step. Genotype analysis was performed using Bio-Rad Precision Melt Analysis software.

Quantitative RT-PCR

Total RNA was isolated using TRIzol and RNeasy RNA isolation Kits (Qiagen) according to the manufacturer's instructions. Reverse transcription was performed using QuantiTect Reverse Transcription Kits (Qiagen). Expression analysis was performed by real-time PCR on a CFX96 Real-Time System (Bio-Rad). The analysis was carried out using the following protocol: 30s at 95°C, 40 cycles of 5s at 95°C, and ending with 5s at 62°C. HPRT was used as the housekeeping reference gene for normalization. qRT-PCR was run using the following primers: FAM134B forward primer for the long isoform (Exon1,2): CTGCTGTTCTGGTTCCTTGC; FAM134B reverse primer for the long isoform (Exon1,2): CGCCCAAGTATCATGACGGA; FAM134B forward primer for the short isoform (Exon4,5):

GCAGCCTTTGCCACTGTTATTAT; FAM134B reverse primer for the short isoform (Exon4,5): ATAACTTCCCAGCTTTTGCCTG; HPRT forward primer: CTCAACTTTAACTGGAAAGAATGTC; HPRT reverse primer: TCCTTTTCACCAGCAAGCT.

Statistical analysis

Association analysis was performed using the command --logistic in PLINK v.1.07[26] software including sex as a covariate (--sex). Due to the homogeneity of the cohort (which consisted of Singapore Chinese university students only), age and population were not included as factors in the model. Gene-gene interactions were evaluated in PLINK v.1.07 [26] conditioning on rs7071836 using the command --condition, fitting a logistic model (--logistic --interaction) and including sex as a covariate (--sex).

We fitted the following model:

$$Y \sim \beta_0+\beta_1A+\beta_2B+\beta_3AB+\beta_4S+\varepsilon$$

where A represents the allele dosage for the first SNP, B represents the allele dosage for the second SNP, AB is the interaction term, and S is the sex covariate.

Association between rs7071836 and CD39 cell surface expression was evaluated using Kruskal-Wallis testing on geometric mean intensity values. Association between genotypes and gene expression was evaluated using one-way ANOVA on $\log_2$-transformed expression values considering each genotype class as a distinct group (Graphpad Prism 6).

Differences in gene expression between control samples and samples treated with ATP were evaluated using repeated-measures ANOVA on Δc(t) values with the assumption of sphericity. The mean of each group was compared to the mean of the control samples using Tukey's multiple comparisons test. Concentration specific effects were evaluated using the post test for linear trends in Graphpad Prism 6. Normality was assessed using the Shapiro-Wilk normality test as implemented by the function shapiro.test in R 2.15.1[24]. Homoscedasticity was tested using the Bartlett test of homogeneity of variances as implemented by the function bartlett.test in R 2.15.1[24]. Clustering of whole blood CD39 gene expression into three groups (CD39lo, CD39int, CD39hi) was performed using the unsupervised clustering method *k*-means as implemented by the function kmeans in R 2.15.1[24] with the number of clusters set to three. Differences in FAM134B expression across the three groups (CD39lo, CD39int, CD39hi) were estimated using Kruskal-Wallis. Pairwise differences between CD39lo and CD39hi groups were tested using Dunn's multiple comparison test (Graphpad Prism 6). Correlation coefficients and significance were computed using Spearman correlation and the fitting line was evaluated using linear regression (Graphpad Prism 6). Each cohort was analyzed independently.

Linkage disequilibrium plots were built using the software ArchiLD[27] on LD values estimated from the 1000 Genomes Pilot Project for CHB+JPT[28].

Power estimation

Power estimation for epistatic interactions was performed using QUANTO software[29] with a significance threshold of 5×10^{-8} for the discovery study and 0.0025 for the validation study (Bonferroni correction for 20 tests). For both genes, the mode of inheritance was hypothesized to be log-addictive. Prevalence was set to 13% in both the discovery and validation cohorts. For each epistatic pair the allele frequency for the first SNP participating in the interaction was set to 0.25 (i.e. the allele frequency of rs7071836 in the discovery cohort) and the allele frequency of the second SNP was set to the allele frequency, in the discovery cohort, of the epistatic partner under consideration. R_{GH}, the interaction effect, was set to the OR of the interaction in the discovery cohort.

Results

CD39 expression on Treg is influenced by promoter SNP rs7071836

In humans, CD39 is expressed by effector/memory-like Treg cells[11]. For this study we assessed CD39 cell surface expression in a small cohort of 41 Chinese and 22 Caucasian blood donors. As previously reported in Caucasians[11] we detected substantial inter-individual differences in CD39 surface expression in blood samples from volunteers of Chinese ethnicity (Figure 1A). However, while only ~20% of Caucasians were 'CD39lo', more than half of Chinese donors exhibited this phenotype (Figure 1B, upper panel). An analysis of CD39 SNP frequencies as reported for both ethnicities by the HapMap project[30–33] suggested that rs7071836 could be associated

with the CD39 Treg phenotype (Figure 1B, lower panel). The frequency of the 'CD39lo' Treg phenotype in the Caucasian and Chinese cohorts closely resembled the frequency of the 'AA' genotype for this SNP.

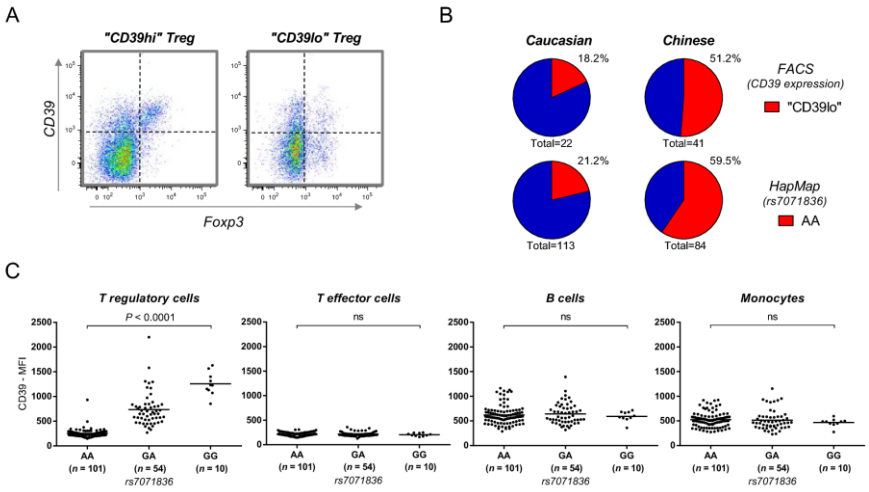


Figure 1. Promoter SNP rs7071836 influences Treg expression of CD39. (A) Illustrative example of the 'CD39hi' and 'CD39lo' Treg phenotypes. Intracellular staining of the Treg-associated transcription factor FoxP3 vs. cell surface expression of CD39 is shown for the CD45RO+CD4+ T cell compartment in two separate individuals of Chinese ethnicity. (B) Frequency of the 'CD39lo' Treg phenotype correlated with the frequency of the rs7071836-AA genotype (calculated using samples from HapMap3 release #27). (C) Genotype/phenotype association in a cohort of Chinese ethnicity (*n* = 165) showing CD39 expression in subsets of peripheral blood Treg, CD4+ effector T cells, B cells and monocytes. Significance was evaluated by Kruskal-Wallis test. Gating of the different cell populations is shown in Additional file 1: Figure S1.

Accordingly, FACS analysis of a larger cohort of 165 ethnic Chinese blood donors confirmed that rs7071836 SNP variants are indeed strongly associated with the phenotype of Treg cells (Figure 1C).

Notably, the correlation was restricted to Treg, since CD39 expression on monocytes, B cells and T effector cells appeared unaffected by the allelic state of the SNP. In line with our finding a variant in strong linkage disequilibrium (LD) with rs7071836 was recently reported as being associated with the percentage of CD39+ activated CD4+ Treg in a Caucasian population[34].

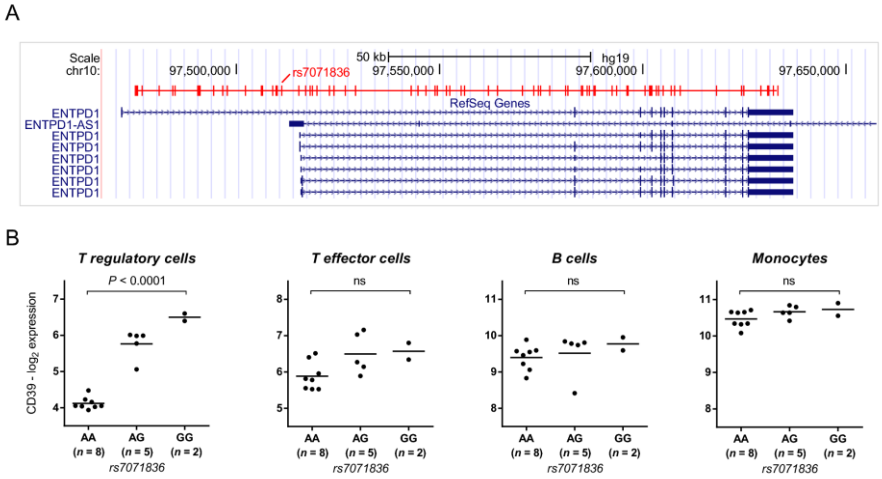


Figure 2. Association of rs7071836 with CD39 mRNA levels in Treg but not other CD39+ human leukocytes. (A) The position of the SNP-cluster which is in perfect linkage disequilibrium (r^2 = 1) with rs7071836 in Asian individuals (CHB/JPT) is shown here with respect to the transcripts of the gene CD39 (ENTPD1). Linkage disequilibrium was estimated using the 60 samples sequenced by the 1000 Genomes Pilot Project. This SNP-cluster contains 82 SNPs spanning the entire gene. The position of the tag-SNP rs7071836 is indicated. (B) Impact of rs7071836 on CD39 gene expression in Treg, CD4+ T effector cells, B cells and monocytes isolated by FACS sorting from 15 genotyped individuals. The CD39 mRNA content is plotted with reference to rs7071836 genotype. Statistical analysis revealed that the polymorphism is associated with CD39 cell surface expression in T regulatory cells only ($P < 0.0001$). Gene expression was measured using the Affymetrix Human Exon 1.0ST Array and data were normalized using the Robust Multichip Average (RMA) at the gene level. P-values for the association were evaluated using one-way ANOVA on $\log_2$-transformed data (n = 15).

SNP rs7071836 tags a cluster of perfectly-linked SNPs

In Asian individuals rs7071836 tags a group of 82 perfectly-linked SNPs (SNP-cluster[27]) that cover the entire gene locus (Figure 2A).

Some of these SNPs have been associated with risk of inflammatory disorders and viral infections. Specifically, the variants rs10748643 and rs11188513 have been respectively linked in Caucasians with risk of inflammatory bowel disease[15] and HIV[17]. In this population, both rs10748643 and rs7071836 are part of the same cluster, with linkage of r^2 = 0.68 to rs11188513. mRNA analysis in FACS-sorted cells confirmed that the pattern of CD39 cell surface expression was directly reflected by the amount of mRNA detected in the respective cell types (Figure 2B). While CD39 mRNA levels in Treg cells varied according to rs7071836 genotype, expression levels in B cells, monocytes and CD4+ T effector cells were unaffected by the polymorphism.

SNP rs7071836 affects AR risk via epistatic interaction with rs257174

To evaluate the role of the rs7071836 cluster in determining AR risk we assessed a Singapore Chinese cohort that comprised 456 atopic individuals affected by AR and 486 non-atopic asymptomatic controls (described in a previous publication)[19]. Using standard logistic regression models, we were unable to find any evidence of a direct association of rs7071836 with the risk of AR (*P* = 0.82, Table 1). We therefore explored possible epistatic interactions of rs7071836 with other polymorphic sites. Of the 550,000 SNPs detected by the Illumina HumanHap 550k array, a total of 447,081 tag SNPs passed quality control assessment as described in[19]. Using a cut-off of 10^{-4} we identified 58 candidate interactions with rs7071836 that

Table 1. Association results for rs7071836 and AR in a Singapore Chinese population

Gene	SNP	Test	Freq cases	Freq controls	Alleles[a]	MAF[b]	OR[c] [CI[d] 95%]	*P*
CD39	rs7071836	LOGISTIC	0.25	0.25	G/A	0.25	1.03 [0.83-1.27]	0.82

[a]Minor allele listed first. [b]Minor Allele Frequency. [C]Odds Ratio. [d]Confidence Interval.

exhibited odds ratios (ORs) of 0.17 - 2.94 with P-values ranging from 10^{-4} to 10^{-6} (Additional file 4: Table S1).

To identify the 'true' epistatic partners of rs7071836 an independent Singapore Chinese replication cohort consisting of 676 atopic AR cases and 511 non-atopic asymptomatic controls was employed[19]. Twenty SNPs were selected for replication and genotyped by Sequenom (Sequenom Inc, San Diego). After Bonferroni correction, the power to detect a significant interaction was estimated to be 0.88 - 0.99 (Additional file 5: Table S2). This finally allowed us to confirm an interaction of rs7071836 with a second SNP located ~12 kbp upstream of gene FAM134B (rs257174).

The combined P-value of discovery and validation cohorts was determined to be 1.98×10^{-6} with an odds ratio of 0.53 (Table 2). In order to exclude the possibility that the statistical interaction was driven by a direct influence of rs257174 on the risk of AR, we also tested this SNP for primary association with disease presentation. While a weak association of rs257174 with AR risk was detected in

Table 2. Summary of the statistical interaction between rs7071836 and rs257174 across all cohorts

SNP1	SNP2	Discovery		Validation		Combined	
		OR^{a}_{int} [CI[b] 95%]	P_{int}	OR^{a}_{int} [CI[b] 95%]	P_{int}	OR^{a}_{int} [CI[b] 95%]	P_{int}
rs7071836	rs257174	0.45 [0.31- 0.66]	**3.64 x 10^{-5}**	0.59 [0.41-0.84]	**4.02 x 10^{-3}**	0.53 [0.41- 0.69]	**1.98 x 10^{-6}**

[a]Odds Ratio. [b]Confidence Interval.

the discovery cohort (*P* = 0.01, OR = 1.32), neither the validation cohort (*P* = 0.35, OR = 0.91) nor the combined cohort produced any statistical significance (*P* = 0.38, OR = 1.07), indicating that the effect of rs257174 on incidence of AR is evident only when the variant is considered in combination with rs7071836 (marginal P-values and ORs are provided in Additional file 6: Table S3).

Epistatic interaction between rs7071836 and rs257174 is associated with AR risk but not atopy

In the tropical urban environment of Singapore AR is strongly associated with the sensitization against house dust mite (HDM) allergens[35]. We therefore sought to determine whether the epistatic effect of rs257174 and rs7071836 was associated with the production of HDM-specific IgE (atopy), the key mediator of AR pathology, or rather with a downstream event influencing the manifestation of the clinical symptoms.

Table 3. Atopy contribution to the statistical interaction between rs7071836 and rs257174 in the validation cohort

SNP1	SNP2	AR cases	Samples with atopy but no AR	OR[a] [CI[b] 95%]	P-value
rs7071836	rs257174	676	1647	0.69 [0.52-0.91]	**9.60×10^{-3}**

SNP1	SNP2	Samples with atopy but no AR	Non atopic samples	OR[a] [CI[b] 95%]	P-value
rs7071836	rs257174	1647	511	0.87 [0.66-1.15]	0.34

[a]Odds Ratio. [b]Confidence Interval.

All AR cases were atopic (HDM-IgE positive as defined by skin prick test), whereas all non-symptomatic controls were non-atopic. We therefore used the replication cohort described in[19] to assess the putative epistatic interaction by comparing the atopic AR group (AR+) comprising 676 individuals with 1647 individuals that were also atopic but did not show any AR symptoms (AR-). Epistasis of rs257174 and rs7071836 was still evident when comparing atopic AR+ cases with atopic AR- cases ($P = 9.6 \times 10^{-3}$), but this effect was lost when comparing atopic AR- individuals with a healthy non-atopic control group comprising 511 individuals ($P = 0.34$) (Table 3). These data indicated that epistasis of rs7071836 and rs257174 affects the manifestation of clinical symptoms but does not associate with the production of HDM specific IgE.

SNP rs257174 influences FAM134B expression levels in monocytes

While statistical analysis of the GWAS data revealed an epistatic interaction of rs7071836 with rs257174, it still had to be shown that the SNP was actually associated with function. Polymorphism rs257174 is located ~12 kbp upstream of the FAM134B gene. Like its epistatic partner rs7071836, it belongs to a cluster of perfectly-linked SNPs (Figure 3A). To establish a quantitative effect on FAM134B gene expression we used mRNA samples from the same cell sources as in our rs7071836 analyses (Figure 2B). The analysis revealed that the rs257174-cluster indeed represents an 'expression quantitative trait locus' (eQTL). As in the case of the rs7071836-cluster the allelic effect was cell-type specific. Surprisingly, however, it was most evident in monocytes while the expression of FAM134B in Treg cells was unaffected (Figure 3B).

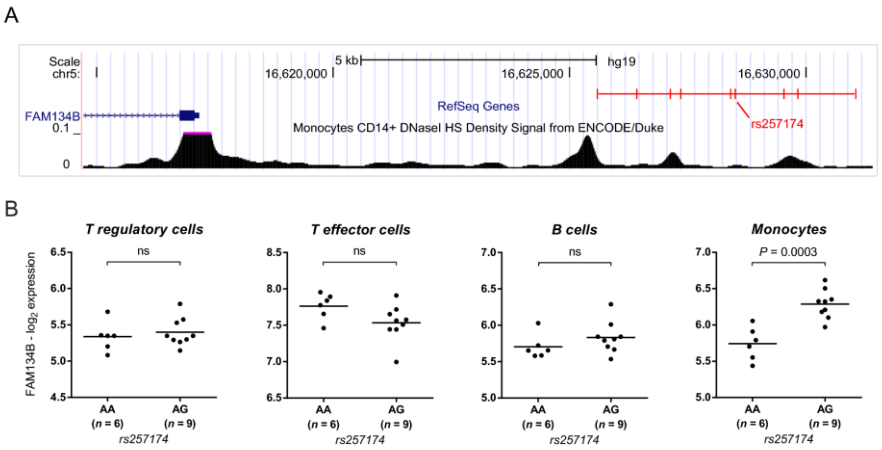


Figure 3. Variations in monocyte expression of FAM134B are associated with SNP rs257174. (A) The position of the SNP-cluster which is in perfect linkage disequilibrium (r^2 = 1) with rs257174 in Asian individuals (CHB/JPT) is shown here with respect to the transcripts of the gene FAM134B. This SNP-cluster contains 9 SNPs located ~9 kbp upstream of the transcriptional start site. The position of the tag SNP rs257174 is indicated. Several SNPs of this cluster are located in a region of open chromatin (identified by DNase I Hypersensitivity site analysis of CD14+ monocytes in the Encode Project), suggesting a regulatory role on gene expression. (B) Impact of rs257174 on FAM134B gene expression in Treg, CD4+ T effector cells, B cells and monocytes (same samples as in Figure 2B). FAM134B mRNA content is plotted with reference to rs257174 genotype. Statistical analysis revealed that the polymorphism is associated with FAM134B gene expression in monocytes only (P = 0.0003). Gene expression was measured using the Affymetrix Human Exon 1.0ST Array and data were normalized using the Robust Multichip Average (RMA) at the gene level. P-values for the association were evaluated using one-way ANOVA on $\log_2$-transformed data (n = 15).

In-line with our findings, the influence of rs257174 on FAM134B expression has also been detected by other investigators conducting transcriptome studies of human monocytes[36,37]. For this cell type, 3 of the SNPs in the rs257174-cluster are located in close proximity to an open chromatin region (DNase I Hypersensitivity peak), which, by ENCODE criteria, suggests a functional role in the regulation of gene expression[38] (Figure 3A).

CD39 expression is negatively correlated with FAM134B expression in whole blood

The epistatic interaction between rs7071836 and rs257174 suggested an inverse relation between the expression of CD39 and FAM134B. In order to confirm the relevance of this finding the expression levels of CD39 and FAM134B were compared in three Caucasian cohorts of healthy individuals (Kora F4[20], DILGOM[21], SHIP-TREND[22]) for which whole blood gene expression measurements were published. In all three cohorts CD39 expression values negatively correlated with

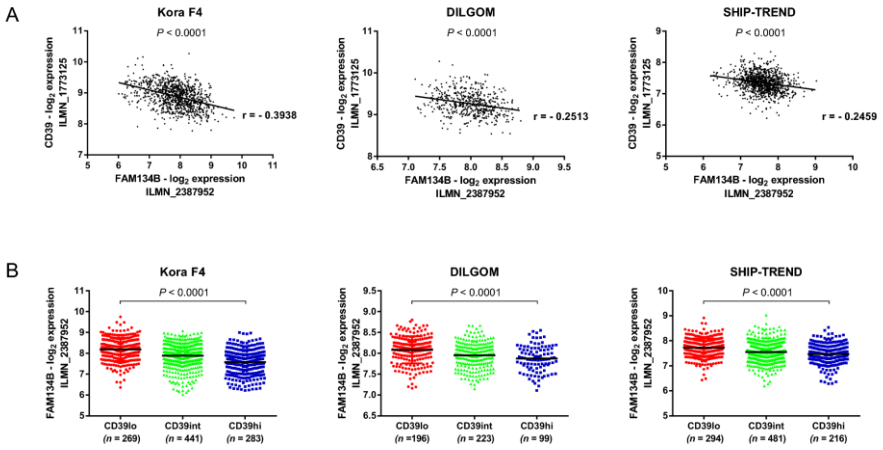


Figure 4. CD39 expression negatively correlates with FAM134B expression in whole blood. (A) Correlation between CD39 and FAM134B expression in whole blood in three Caucasian cohorts. The most significant FAM134B Illumina probe is plotted here with respect to the unique CD39 Illumina probe. A consistent negative correlation between the two genes was apparent in all cohorts ($P < 0.001$). (B) FAM134B expression is affected by CD39 levels in whole blood. Samples were classified according to their CD39 expression levels (CD39lo, CD39int, CD39hi) using unsupervised clustering (see Additional file 7: Figure S4) and FAM134B expression across the three groups was compared using Kruskal-Wallis test followed by Dunn's multiple comparison tests. CD39hi individuals were characterized by a significantly lower FAM134B expression than CD39lo individuals ($P_{krusal\text{-}wallis} < 0.0001$, $P_{CD39lo\ vs\ CD39hi} < 0.0001$).

FAM134B expression values ($P_{spearman} < 0.001$, $r_{Kora\ F4} = -0.3938$, $r_{DILGOM} = -0.2513$, $r_{SHIP\text{-}TREND} = -0.2459$) (Figure 4A). Similarly, FAM134B was differentially expressed across samples clustered according to their CD39 expression (CD39lo, CD39int, CD39hi, Additional file 7: Figure S4). Thus, a high expression of CD39 was inversely correlated to the expression of FAM134B (Figure 4B).

Extracellular ATP enhances FAM134B expression

Since variations in CD39 Treg expression affect the concentration of ATP in the extracellular space[11], we hypothesized that this molecule could represent the functional link for the epistasis of CD39 and FAM134B. This would imply however, that FAM134B expression would be controlled by the amount of extracellular ATP. In order to test this hypothesis, we isolated CD14+ blood monocytes and incubated these cells for 2h in the presence or absence of ATP before assessing mRNA levels of FAM134B (Figure 5A). The experiment confirmed that extracellular ATP indeed enhances the expression of FAM134B (Figure 5B). In monocytes exposed to either 0.2 mM or 1 mM ATP we observed a significant increase in FAM134B mRNA (P = 0.0089). Specificity of the ATP-mediated effect was confirmed by a partial block of the FAM134B expression by the addition of the ATP-antagonist A-438079 (Additional file 8: Figure S5). The stimulatory effect of ATP on FAM134B expression was dose-dependent (P = 0.0035, post test for assessment of linear trends), and observed only for the long isoform of FAM134B, which is encoded ~9 kbp downstream of the rs257174 LD block. Expression levels of a less abundant shorter isoform, which is encoded more than 100 kbp distal from the rs257174 LD cluster, were not affected by ATP exposure. Thus, in human monocytes the expression of FAM134B is modulated both by a genetic *cis* polymorphism and by extracellular ATP levels.

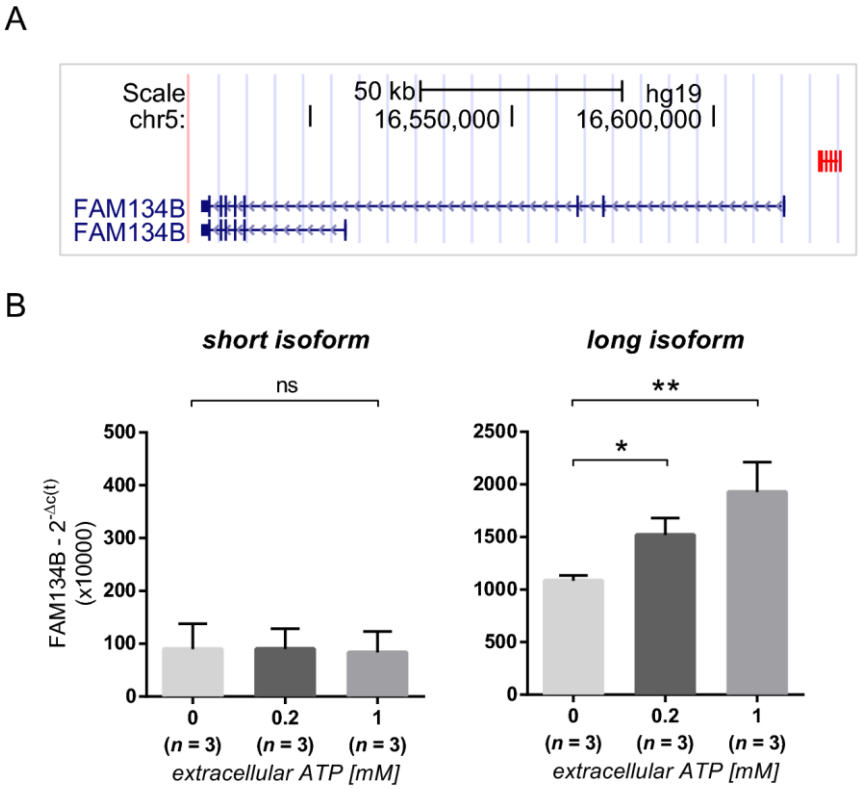


Figure 5. Extracellular ATP enhances monocyte expression of FAM134B. (A) Location of the rs257174 SNP-cluster with reference to the two common splice variants of FAM134B. The FAM134B transcript exists in two isoforms, a long isoform (~9 kbp down-stream of the rs257174 SNP-cluster), and a less abundant short isoform (more than 100 kbp distal from the rs257174 SNP-cluster). (B) CD14+ monocytes were isolated from 3 separate individuals and exposed to variable doses of ATP (0, 0.2 and 1 mM). After 2h incubation, expression of FAM134B was measured by RT-PCR and normalized by the $2^{-\Delta c(t)}$ method using HPRT as the housekeeping reference gene. FAM134B expression in monocytes was significantly and dose-dependently up-regulated in response to extracellular ATP treatment. Differences between the 3 treatments were estimated using repeated measures ANOVA (P = 0.0089) with Tukey's post-hoc comparison test on Δc(t) values. Error bars represent standard deviation across biological replicates.

Discussion

In this study we provide the first evidence that intercellular epistatic interactions can influence risk of complex human diseases. In this case, rs7071836, which modulated CD39 expression by human Treg, interacted with rs257174, which altered FAM134B gene expression in blood monocytes to affect the risk of AR. Both are functionally connected via extracellular ATP, a damage-associated molecule involved in eliciting a range of host inflammatory responses.

In a cohort of Singapore Chinese volunteers this novel epistatic interaction is associated to risk of allergic rhinitis. With an estimated disease prevalence of 13%[39] and an OR of 0.45, our study exhibited only 5% power to detect our CD39/FAM134B interaction in the discovery cohort at a significance threshold $\alpha = 5 \times 10^{-8}$. We therefore elected for a replication approach to improve the power of the study. By this approach, our power for detection was increased to 89% (α = 0.0025 after Bonferroni correction for 20 tests) with a combined P-value of 1.98×10^{-6}. While the epistasis was anchored on rs7071836, an eQTL known to regulate the surface expression of CD39, the epistasis-partner revealed by the statistical analysis also turned out to be functional. Polymorphism rs257174 was part of a SNP-cluster associated with the expression of the *cis*-Golgi protein FAM134B. Moreover, while the allelic combination of the epistasis suggested that the expression of CD39 and of FAM134B were inversely linked, this correlation could actually be validated from the expression data of a full blood transcriptome analysis.

CD39 is an ecto-ATPase that depletes extracellular ATP from the local microenvironment. Since extracellular ATP is a key indicator of host tissue damage, the efficacy of ATP removal by CD39 has been associated with various inflammatory conditions including autoimmune diseases, viral infections and cancer progression. The CD39 promoter SNP rs7071836 responsible for the effect is part of a large cluster of perfectly linked SNPs containing polymorphisms already implicated in Crohn's disease[15] and in progression of AIDS[17]. Our study indicates that in conjunction with a cluster controlling FAM134B it is also an important component of AR risk. While the biological function of CD39 is well-established, the exact role played by the epistatic partner remains enigmatic. As a newly identified *cis*-Golgi protein, it has only been shown to be expressed on few cell types mostly associated with the neural system[40–42]. While the protein was detected in autonomic and sensory ganglia, deleterious mutations in this gene have been shown to cause hereditary sensory and autonomic neuropathy type II (HSAN II), a severe genetic disease characterized by a dysfunction of the autonomic system and impaired nociception[40]. FAM134B knock-down in a mouse N2a neuroblastoma cell line resulted in a smaller *cis*-Golgi compartment and impaired cell proliferation, and FAM134B knock-down in cultured dorsal root ganglion mouse neurons resulted in apoptosis of nociceptive neurons[40]. FAM134B may therefore be involved in mediating multiple cellular pathways that affect the maturation and export of protein precursors and cell surface receptors.

Although FAM134B has been primarily associated with the nervous system, the expression of this gene is in fact far more widespread. Overexpression of FAM134B has been reported before in human esophageal squamous cell carcinoma[43] and this study provides evidence for a constitutive expression in multiple different populations of human leukocytes. The FAM134B molecule thus seems likely to play a significant role in host immune protection and inflammatory responses. Further work will now be required to fully dissect the role played by FAM134B in the numerous different leukocyte subsets that comprise the human immune system. Considering that FAM134B is located in the *cis*-Golgi compartment, this protein could potentially be involved in vesicle trafficking and may influence cytokine secretion by monocytes in response to external stimuli including ATP.

Conclusions

Epistasis has been recognized as a natural phenomenon that commonly occurs between SNPs that affect components of the same biological pathway[44–46]. Here we propose a novel mechanism of epistasis based on the interaction of two 'unrelated' molecules that are regulated by polymorphisms in different cell types. Hence, epistasis can also arise from functional links that facilitate cross-talk between disparate biological pathways. In the current report, the putative mediator of this inter-lineage epistasis is ATP. While it modulates monocyte expression of FAM134B, it is also depleted from the environment by Treg via the ectonucleotidase CD39, whose

expression is influenced by rs7071836 (Figure 6). The amount of FAM134B in monocytes is thus modulated both by a monocyte-specific *cis* polymorphism that determines basal expression levels and by a *trans*-polymorphism that affects CD39 expression in Treg. To our knowledge, this is the first report that intercellular genetic epistasis can play a role in susceptibility to a complex human disease.

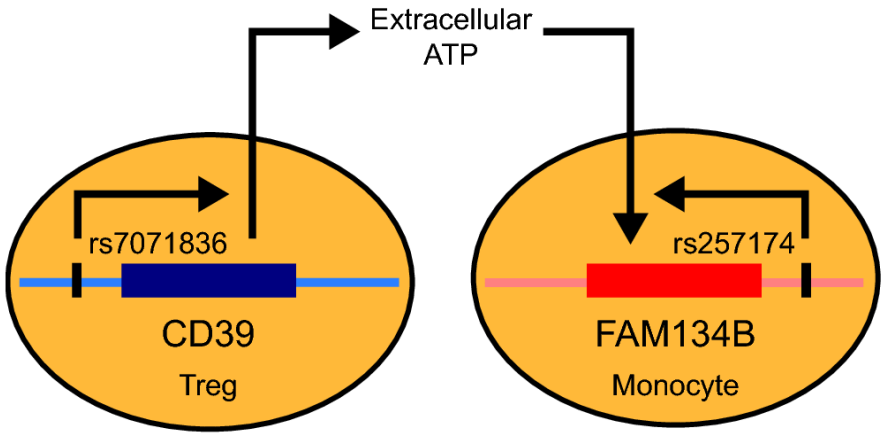


Figure 6. Schematic representation of the epistatic interaction between SNPs rs7071836 and rs257174. While promoter SNP rs7071836 regulates CD39 expression in Treg, the allelic state of rs257174 determines basal expression of FAM134B in monocytes. CD39 depletes extracellular ATP by hydrolyzing this damage-associated molecular pattern but the extent of ATP depletion depends directly on the allelic state of rs7071836. In turn, monocyte basal expression of FAM134B is regulated by rs257174 and is modulated by the concentration of ATP in the extracellular space. Host cell damage-associated ATP is thus the putative functional link that supports epistatic interaction of rs7071836 and rs257174 in determining AR risk.

List of abbreviations

Treg: regulatory T cells; AR: Allergic Rhinitis; SNP: Single Nucleotide Polymorphism; LD: Linkage Disequilibrium; ATP: Adenosine Triphosphate; HDM: House Dust Mite; HPRT: Hypoxanthine Guanine Phosphoribosyltransferase; IgE: Immunoglobulin E; PBMC: Peripheral Blood Mononuclear Cell;

Competing interests

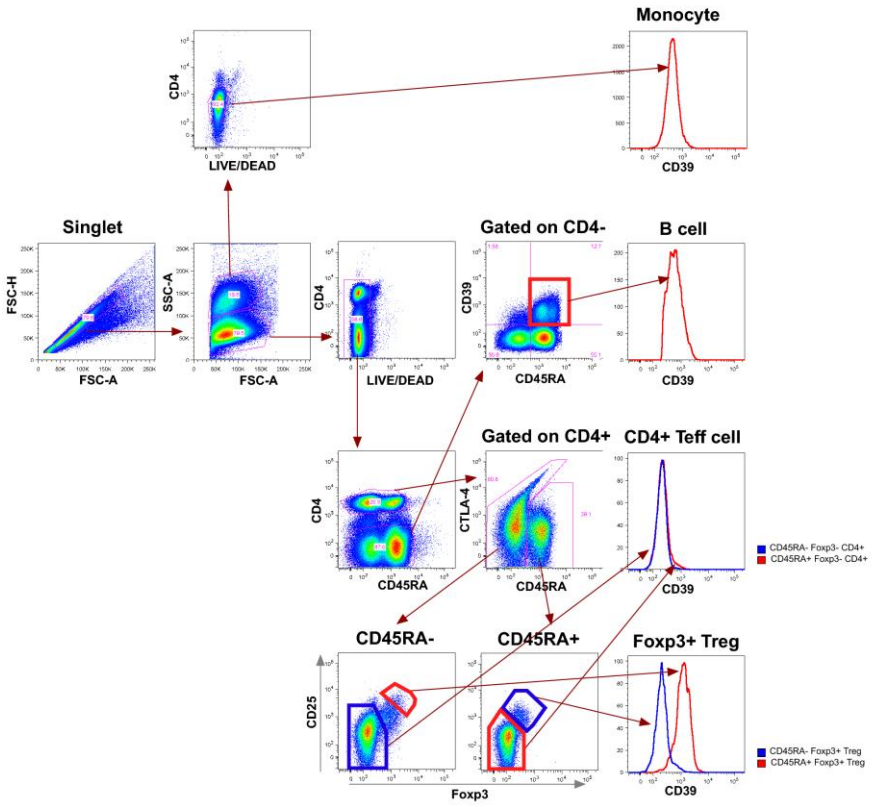
The authors declare that they have no competing interests.

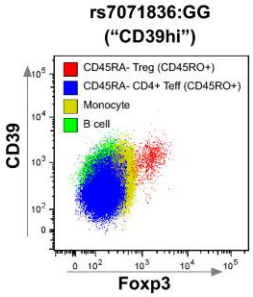
Authors' contributions

Author contributions: RM, KJP, AKA, MS, MP and OR designed the research; KJP, TYP, MS, LLZ, KP, ATSL and AL performed the biological experiments; FTC, DYW and AKA collected the samples and the epidemiological data and supervised the genotyping; RM analyzed the data; RM, KJP, AKA, TYP, MP and OR wrote the paper; all the authors read and critically reviewed the manuscript.

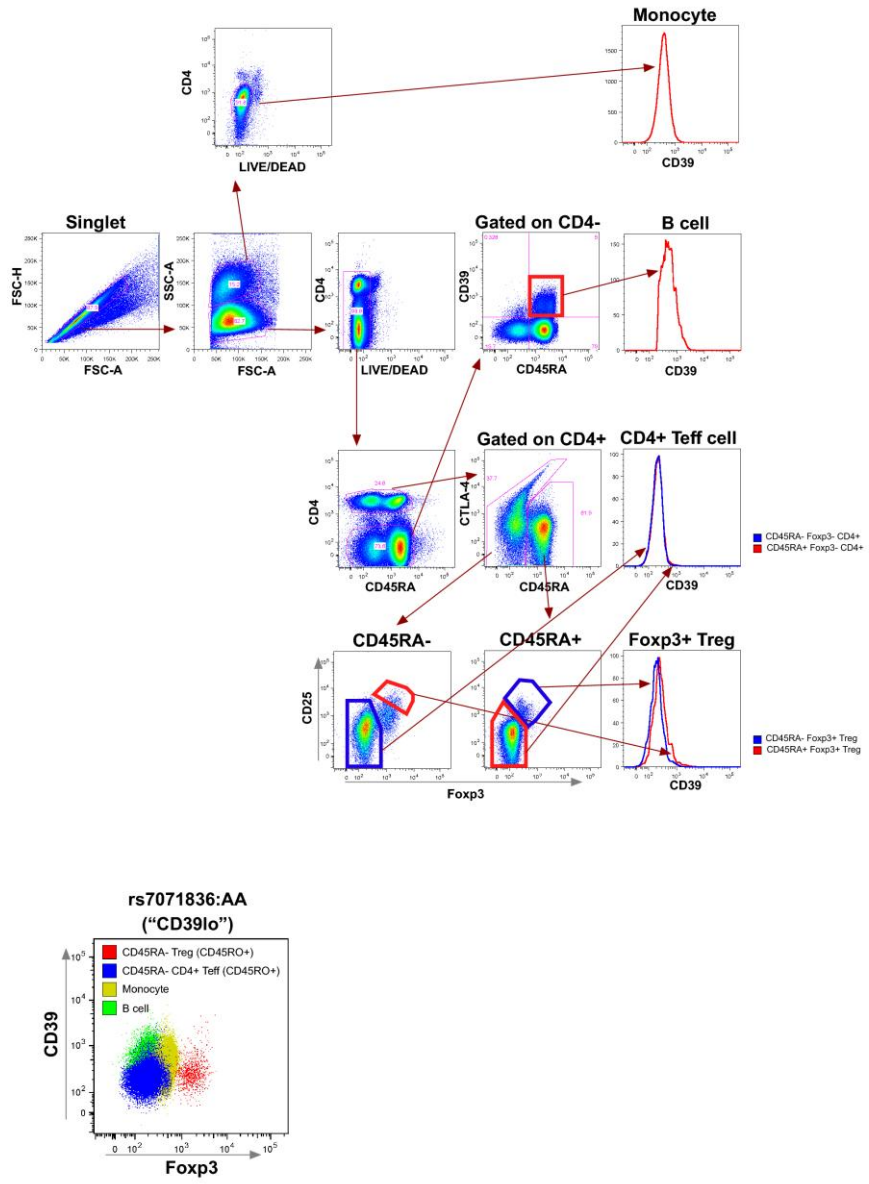
Additional data files

A



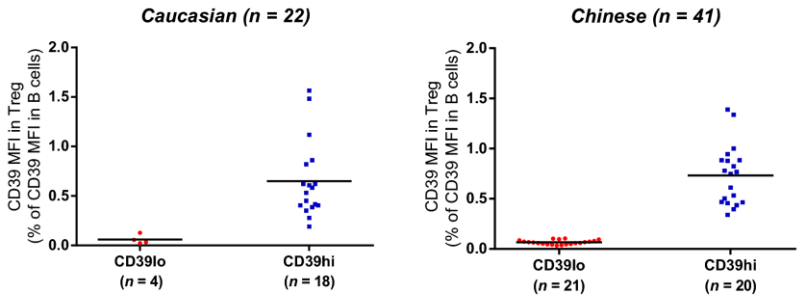


B



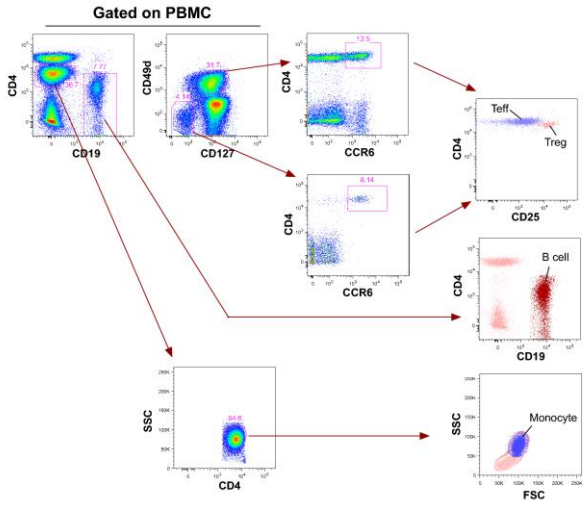
Supplementary Figure 1. Gating strategy for Treg, CD4+ T effector cells, B cells and monocytes. 1-2 x 10^6 PBMCs from each donor were stained using the LIVE/DEAD Fixable Blue Dead kit followed by surface staining with anti-CD4, CD25, CD39, CD45RA and CCR6 mAb, FoxP3 and CTLA-4 were stained intracellularly as

described in Materials and methods. Monocytes and B cells constitutively expressed CD39 at steady-state. B cells were gated based on CD4-, CD39+, and CD45RA+ cells whereas effector CD4+ T cells (Teff) were CD4+ and FoxP3-. CD39 was not expressed on CD45RA+ or CD45RA- Teff cells. (A) rs7071836 GG genotype resulted in a higher CD39 expression on a specific subset of Treg cells, CD45RA-FoxP3+ Treg (CD45RO+). (B) In contrast, rs7071836 AA genotype was associated with a CD39 low phenotype on the same population of Treg cells.

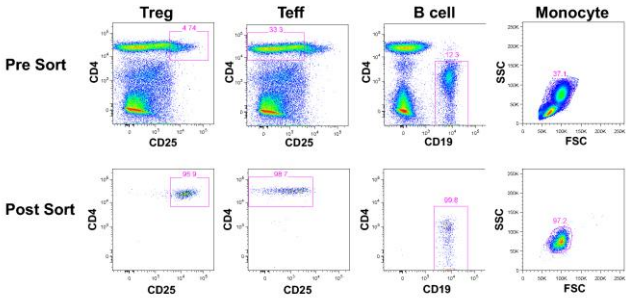


Supplementary Figure 2. Variation in CD39 cell surface expression in Caucasian and Chinese populations. CD39 expression is stable in human B cells (see Figure 1) but variable in Treg. Surface expression of CD39 in Treg was therefore normalized to the expression levels observed in donor-matched B cells in order to reduce possible batch effects. CD39 cell surface expression on the two subsets was determined by FACS analysis and individuals were clustered into two groups: CD39lo and CD39hi. Donor classification was confirmed using the unsupervised clustering k-means method on log2-transformed ratios, setting the number of clusters to two. While no major differences in clustering were observed between ethnicities, the relative frequency of CD39lo individuals was higher in the Chinese cohort.

A



B



Supplementary Figure 3. FACS Sorting strategy for Teff, Treg, B cells, and monocytes. (A) Gating strategy for the pre-sort sample. PBMCs were incubated with CD49d, CD127, CD4, CD25, CD19, and CCR6 mAbs in MACS buffer for 15 min. After staining, the samples were washed with MACS buffer and re-suspended at 40-50 x106 cells per ml in MACS buffer and applied to 70 uM Pre-separation filters (Miltenyi Biotec). The filtered samples were applied to a FACSAriaII cell sorter. Teff, Treg, B cells were collected in 5 ml Falcon polystyrene tube containing 1 ml of FACS sorting media (RPMI1640 supplemented with 10% fetal bovine serum and 10µg/ml gentamicin). For Treg cells, these cells were collected in 1.5 ml Eppendorf tube containing 1 ml of FACS sorting media. (B) Post-sort analysis of Treg, Teff, B cell, and monocyte. Post sort analysis was performed to determine the purity of Treg, Teff, B cells, and monocytes.

Supplementary Table 1. CD39 epistatic interactions identified in the discovery cohort (*P* < 0.0001)

SNP	CHR	BP	Gene	Location	OR	L95	U95	P
rs12745984	1	107905374	VAV3	downstream (9930bp)	0.3748	0.2464	0.57	4.50E-06
rs2311740	10	89818869	PTEN	downstream (100356bp)	0.4222	0.2938	0.6067	3.14E-06
rs12247999	10	89820726	PTEN	downstream (102213bp)	0.4033	0.2616	0.6219	3.96E-05
rs6502938	17	6086382	WSCD1, KIAA0523	downstream (117910bp)	0.5297	0.3904	0.7188	4.50E-05
rs518569	13	79663798	SPRY2	downstream (144314bp)	1.972	1.424	2.731	4.36E-05
rs12408799	1	107900866	VAV3	downstream (14438bp)	0.3937	0.2477	0.6256	7.98E-05
rs10881471	1	107897140	VAV3	downstream (18164bp)	0.3727	0.2451	0.5667	3.89E-06
rs9322777	6	96788552	FUT9	downstream (18342bp)	2.46	1.579	3.832	6.91E-05
rs1975732	12	48544213	FAIM2, KIAA0950	downstream (2733bp)	2.719	1.658	4.459	7.38E-05
rs2324999	3	86241575	BC040985	downstream (36870bp)	0.3784	0.2453	0.5838	1.12E-05
rs1394695	13	75829626	AX747676	downstream (473677bp)	2.178	1.521	3.121	2.18E-05
rs4862396	4	185780666	CASP3	downstream (5177bp)	0.3077	0.1717	0.5513	7.45E-05
rs9377629	6	96836588	FUT9	downstream (66378bp)	2.32	1.522	3.535	9.07E-05
rs6068671	20	51852394	SUMO1P1	downstream (72052bp)	0.4793	0.336	0.6836	4.91E-05
rs12674330	7	5070801	RBAK	exon, intron	0.469	0.3333	0.66	1.39E-05
rs3749430	3	72231809	AK097190	exon	2.048	1.456	2.881	3.82E-05
rs10460527	2	37358609	PRKD3	exon	0.4939	0.3517	0.6937	4.69E-05
rs4149117	12	20902747	SLCO1B3, LST3, LST-3TM12	exon	0.4763	0.3315	0.6845	6.10E-05
rs1328199	1	107929418	VAV3	intron	0.3928	0.2658	0.5803	2.71E-06
rs7619493	3	86133485	CADM2	intron	0.358	0.2319	0.5527	3.54E-06
rs4914950	1	107977810	VAV3	intron	2.132	1.539	2.955	5.40E-06
rs9385283	6	123385160	RLBP1L2	intron	0.2075	0.1008	0.4271	1.96E-05
rs2783495	1	93416265	TMED5, UNQ397	intron	2.19	1.528	3.14	1.98E-05

rs10206380	2	37307790	CEBPZ	intron	0.48	0.3425	0.6727	2.02E-05
rs2576203	1	215139683	ESRRG	intron	2.943	1.792	4.834	2.02E-05
rs444594	20	14414618	MACROD2	intron	0.4329	0.2943	0.6369	2.14E-05
rs2783490	1	93400117	TMED5, UNQ397	intron	2.174	1.517	3.115	2.33E-05
rs6549436	3	72229210	AK097190	intron	2.089	1.484	2.939	2.40E-05
rs11154131	6	123361081	RLBP1L2	intron	0.2114	0.1027	0.435	2.44E-05
rs3821144	2	37362985	PRKD3	intron	0.4786	0.3397	0.6744	2.54E-05
rs4777717	15	90839700	C15orf32	intron	0.4884	0.3495	0.6827	2.73E-05
rs2815429	1	93363256	MTF2	intron	2.169	1.511	3.115	2.73E-05
rs8030836	15	90834290	C15orf32	intron	0.4825	0.343	0.6787	2.84E-05
rs10122718	9	6989846	JMJD2C	intron	0.4716	0.3312	0.6716	3.09E-05
rs408086	20	14427272	MACROD2	intron	0.4465	0.3051	0.6534	3.33E-05
rs7021928	9	118583985	ASTN2	intron	0.51	0.3707	0.7017	3.53E-05
rs4384764	2	37443788	QPCT	intron	0.4808	0.3393	0.6814	3.86E-05
rs9866277	3	86188030	CADM2	intron	0.4566	0.3141	0.6638	4.01E-05
rs11031122	11	30504014	MPPED2	intron	0.1685	0.07174	0.3958	4.36E-05
rs7537311	1	107938539	VAV3	intron	0.4524	0.3092	0.662	4.42E-05
rs11031125	11	30509905	MPPED2	intron	0.1692	0.07175	0.3989	4.91E-05
rs2900474	12	20901315	SLCO1B3, LST3, LST-3TM12	intron	0.477	0.3328	0.6838	5.60E-05
rs9812103	3	86069410	CADM2	intron	0.3424	0.2021	0.5804	6.85E-05
rs1410406	1	107957883	VAV3	intron	2.049	1.439	2.916	6.86E-05
rs12545504	8	4337921	CSMD1	intron	0.5008	0.3558	0.7051	7.41E-05
rs2300888	2	37347840	PRKD3	intron	0.5039	0.3587	0.7079	7.76E-05
rs4629138	2	37371231	PRKD3	intron	0.5047	0.3593	0.7088	7.93E-05
rs4670685	2	37374689	PRKD3	intron	0.508	0.3626	0.7117	8.26E-05
rs12563120	1	93563233	CR609342, AL832786	intron	2.017	1.42	2.865	9.00E-05

rs7015436	8	76363384	BC062758	upstream (10132bp)	0.5307	0.3904	0.7213	5.21E-05
rs257174	5	16681511	FAM134B	upstream (11392bp)	0.4496	0.3076	0.657	3.64E-05

Supplementary Table 2. CD39 epistatic interactions tested in the replication cohort

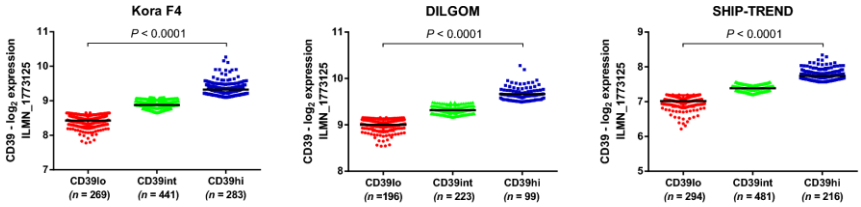
SNP	OR_discovery	L95_discovery	U95_discovery	P_discovery	MAF_discovery	Power to detect in the validation cohort	OR_validation	L95_validation	U95_validation	P_validation	OR_combined	L95_combined	U95_combined	P_combined
rs6502938	0.53	0.39	0.72	4.50E-05	0.47	0.88	1.04	0.79	1.36	7.96E-01	0.77	0.63	0.94	9.40E-03
rs257174	0.45	0.31	0.66	3.64E-05	0.23	0.89	0.59	0.41	0.84	4.02E-03	0.53	0.41	0.69	1.98E-06
rs12247999	0.40	0.26	0.62	3.96E-05	0.18	0.90	0.90	0.65	1.26	5.49E-01	0.67	0.51	0.87	2.70E-03
rs2900474	0.48	0.33	0.68	5.60E-05	0.28	0.90	1.05	0.76	1.45	7.85E-01	0.73	0.57	0.93	1.02E-02
rs4862396	0.31	0.17	0.55	7.45E-05	0.11	0.91	0.99	0.64	1.52	9.62E-01	0.64	0.45	0.90	1.14E-02
rs4777717	0.49	0.35	0.68	2.73E-05	0.33	0.92	0.90	0.66	1.23	5.15E-01	0.67	0.54	0.84	5.98E-04
rs6782778	1.95	1.39	2.73	9.81E-05	0.32	0.93	0.78	0.58	1.04	8.93E-02	1.16	0.94	1.44	1.68E-01
rs3821144	0.48	0.34	0.67	2.54E-05	0.35	0.94	1.08	0.82	1.42	5.93E-01	0.78	0.63	0.96	2.15E-02
rs10206380	0.48	0.34	0.67	2.02E-05	0.37	0.94	1.06	0.81	1.39	6.74E-01	0.78	0.63	0.95	1.62E-02
rs10881471	0.37	0.25	0.57	3.89E-06	0.20	0.96	1.00	0.70	1.43	9.98E-01	0.65	0.50	0.85	1.53E-03
rs9377629	2.32	1.52	3.54	9.07E-05	0.19	0.96	0.94	0.66	1.32	7.02E-01	1.35	1.04	1.75	2.65E-02
rs12674330	0.47	0.33	0.66	1.39E-05	0.42	0.97	1.11	0.84	1.46	4.57E-01	0.79	0.64	0.98	2.98E-02
rs1394695	2.18	1.52	3.12	2.18E-05	0.25	0.97	0.91	0.66	1.25	5.47E-01	1.37	1.08	1.72	9.00E-03
rs1328199	0.39	0.27	0.58	2.71E-06	0.23	0.97	1.04	0.75	1.44	8.11E-01	0.69	0.54	0.88	2.65E-03
rs7619493	0.36	0.23	0.55	3.54E-06	0.19	0.97	1.15	0.82	1.62	4.13E-01	0.72	0.56	0.93	1.25E-02
rs6549436	2.09	1.48	2.94	2.40E-05	0.32	0.97	0.82	0.61	1.10	1.91E-01	1.24	1.00	1.54	5.28E-02

rs9285176	2.04	1.47	2.84	2.38E-05	0.42	0.97	0.94	0.71	1.24	6.53E-01	1.29	1.05	1.59	1.71E-02
rs2311740	0.42	0.29	0.61	3.14E-06	0.31	0.98	0.99	0.75	1.31	9.52E-01	0.74	0.59	0.92	6.15E-03
rs2783495	2.19	1.53	3.14	1.98E-05	0.29	0.98	1.06	0.78	1.43	7.02E-01	1.44	1.15	1.81	1.82E-03
rs4914950	2.13	1.54	2.96	5.40E-06	0.49	0.99	1.00	0.76	1.30	9.78E-01	1.38	1.12	1.69	2.13E-03

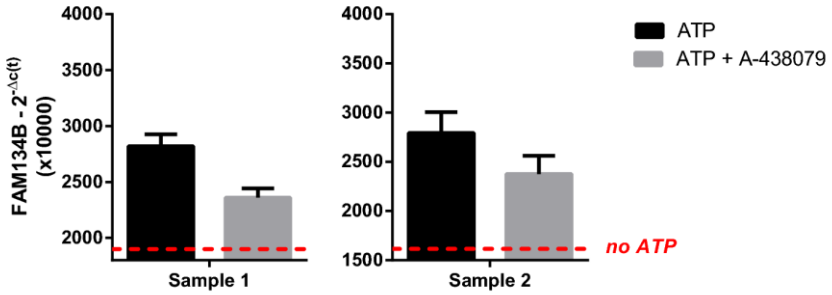
Supplementary Table 3. Marginal and interaction P-values and ORs for the statistical interaction between rs7071836 and rs257174 across all cohorts

	rs7071836		rs257174		Interaction	
	OR[a] [CI[b] 95%]	*P*	OR[a] [CI[b] 95%]	*P*	OR[a] [CI[b] 95%]	*P*
Discovery	1.47 [1.11-1.93]	6.42×10^{-3}	2.01 [1.48-2.71]	6.18×10^{-6}	0.45 [0.31-0.66]	3.64×10^{-5}
Validation	1.06 [0.84-1.34]	0.62	1.15 [0.88-1.51]	0.30	0.59 [0.41-0.84]	4.02×10^{-3}
Combined	1.21 [1.01-1.44]	3.92×10^{-2}	1.46 [1.20-1.78]	2.05×10^{-4}	0.53 [0.41-0.69]	1.98×10^{-6}

[a]Odds Ratio. [b]Confidence Interval.



Supplementary Figure 4. CD39 expression clustering in whole blood to simulate the effect of polymorphism rs7071836. Gene expression processed data were downloaded from the respective repositories and used as input for the clustering algorithm *k*-means. The number of clusters was set to three. Significance was assessed using Kruskal-Wallis test followed by Dunn's multiple comparison tests. Each cohort was analyzed independently. In all cohorts the three CD39 expression clusters (CD39lo, CD39int, CD39hi) were strongly separated.



Supplementary Figure 5. Inhibition of the ATP-mediated induction of FAM134B by a purinergic receptor antagonist. Monocytes of two donors were incubated for 2h with 1mM ATP in the absence or presence of the ATP antagonist 10 µM A-438079. After the incubation the mRNA levels of FAM134B were determined by RT-PCR. The analysis revealed a 50% and 36% reduction in reference to the basal expression in the absence of ATP (red dashed line).

Acknowledgments

The authors would like to thank Bernett Lee and Pavandip Singh Wasan for their statistical advice, Josephine Lum and Francesca Zolezzi for performing the microarray experiments, Raffaele Calogero for normalizing the expression arrays, Neil McCarthy of Insight Editing London for critical review of the manuscript, and all the volunteers and their family members who participated in this study. All the Singapore Immunology Network authors are supported by the A*STAR/Singapore Immunology Network core grant. This study was supported by grants from the Singapore Immunology Network (SIgN-06-006, SIgN-08-020 and SIgN-10-029), Singapore; the National Medical Research Council (NMRC/1150/2008), Singapore; the Biomedical Research Council, Singapore; Agency for Science,

Technology and Research (A*STAR), Singapore, the National University of Singapore (NUS) for the Graduate Research Scholarship and the A*STAR Research Attachment Program (ARAP) for the students involved in the study. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

REFERENCES

1. Bousquet, J. *et al.* Allergic Rhinitis and its Impact on Asthma (ARIA) 2008 update (in collaboration with the World Health Organization, GA(2)LEN and AllerGen). *Allergy* **63,** 8–160 (2008).
2. Li, L. *et al.* Effects of Th2 cytokines on chemokine expression in the lung: IL-13 potently induces eotaxin expression by airway epithelial cells. *J. Immunol.* **162,** 2477–2487 (1999).
3. Durham, S. R. *et al.* Cytokine messenger RNA expression for IL-3, IL-4, IL-5, and granulocyte/macrophage-colony-stimulating factor in the nasal mucosa after local allergen provocation: relationship to tissue eosinophilia. *J Immunol* **148,** 2390–2394 (1992).
4. Bischoff, S. C. *et al.* IL-4 enhances proliferation and mediator release in mature human mast cells. *Proc. Natl. Acad. Sci.* **96,** 8080–8085 (1999).
5. Shi, C. & Pamer, E. G. Monocyte recruitment during infection and inflammation. *Nat. Rev. Immunol.* **11,** 762–774 (2011).
6. Egawa, M. *et al.* Inflammatory monocytes recruited to allergic skin acquire an anti-inflammatory M2 phenotype via basophil-derived interleukin-4. *Immunity* **38,** 570–580 (2013).
7. Palomares, O. *et al.* Role of Treg in immune regulation of allergic diseases. *Eur. J. Immunol.* **40,** 1232–1240 (2010).

8. Ling, E. M. *et al.* Relation of CD4+ CD25+ regulatory T-cell suppression of allergen-driven T-cell activation to atopic status and expression of allergic disease. *Lancet* **363,** 608–615 (2004).
9. Robinson, D. S., Larché, M. & Durham, S. R. Tregs and allergic disease. *J. Clin. Invest.* **114,** 1389–1397 (2004).
10. Bellinghausen, I., Klostermann, B., Knop, J. & Saloga, J. Human CD4+CD25+ T cells derived from the majority of atopic donors are able to suppress TH1 and TH2 cytokine production. *J Allergy Clin Immunol* **111,** 862–868 (2003).
11. Borsellino, G. *et al.* Expression of ectonucleotidase CD39 by Foxp3+ Treg cells: hydrolysis of extracellular ATP and immune suppression. *Blood* **110,** 1225–1232 (2007).
12. Deaglio, S. *et al.* Adenosine generation catalyzed by CD39 and CD73 expressed on regulatory T cells mediates immune suppression. *J. Exp. Med.* **204,** 1257–1265 (2007).
13. Kouzaki, H., Iijima, K., Kobayashi, T., O'Grady, S. M. & Kita, H. The danger signal, extracellular ATP, is a sensor for an airborne allergen and triggers IL-33 release and innate Th2-type responses. *J. Immunol.* **186,** 4375–4387 (2011).
14. Idzko, M. *et al.* Extracellular ATP triggers and maintains asthmatic airway inflammation by activating dendritic cells. *Nat. Med.* **13,** 913–919 (2007).
15. Friedman, D. J. *et al.* CD39 deletion exacerbates experimental murine colitis and human polymorphisms increase susceptibility to inflammatory bowel disease. *Proc. Natl. Acad. Sci.* **106,** 16788–16793 (2009).

16. Fletcher, J. M. *et al.* CD39+Foxp3+ regulatory T cells suppress pathogenic Th17 cells and are impaired in multiple sclerosis. *J. Immunol.* **183,** 7602–7610 (2009).

17. Nikolova, M. *et al.* CD39/Adenosine Pathway Is Involved in AIDS Progression. *PLoS Pathog.* **7,** e1002110 (2011).

18. Tang, Y., Jiang, L., Zheng, Y., Ni, B. & Wu, Y. Expression of CD39 on FoxP3+ T regulatory cells correlates with progression of HBV infection. *BMC Immunol.* **13,** 17 (2012).

19. Andiappan, A. K. *et al.* Genome-wide association study for atopy and allergic rhinitis in a Singapore Chinese population. *PLoS One* **6,** e19719 (2011).

20. Westra, H.-J. *et al.* Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nat. Genet.* **45,** 1238–1243 (2013).

21. Inouye, M. *et al.* An immune response network associated with blood lipid levels. *PLoS Genet.* **6,** e1001113 (2010).

22. Mayerle, J. *et al.* Identification of genetic loci associated with Helicobacter pylori serologic status. *JAMA* **309,** 1912–1920 (2013).

23. De Rosa, S. C. Multicolor immunophenotyping: human mature immune system. *Methods Cell Biol.* **75,** 577–594 (2004).

24. R Development Core Team. *R: A Language and Environment for Statistical Computing.* (2012).

25. Irizarry, R. A. *et al.* Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **4,** 249–264 (2003).

26. Purcell S *et al.* PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *Am. J. Hum. Genet.* **81,** 559–575 (2007).
27. Melchiotti, R., Rötzschke, O. & Poidinger, M. ArchiLD: Hierarchical Visualization of Linkage Disequilibrium in Human Populations. *PLoS One* **9,** e86761 (2014).
28. The 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature* **467,** 1061–1073 (2010).
29. Gauderman, W. J. Sample size requirements for association studies of gene-gene interaction. *Am. J. Epidemiol.* **155,** 478–484 (2002).
30. The International HapMap Consortium. The International HapMap Project. *Nature* **426,** 789–796 (2003).
31. The International HapMap Consortium. A haplotype map of the human genome. *Nature* **437,** 1299–1320 (2005).
32. The International HapMap Consortium. A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449,** 851–861 (2007).
33. The International HapMap Consortium. Integrating common and rare genetic variation in diverse human populations. *Nature* **467,** 52–58 (2010).
34. Orrù, V. *et al.* Genetic Variants Regulating Immune Cell Levels in Health and Disease. *Cell* **155,** 242–256 (2013).
35. Andiappan, A. K. *et al.* Allergic airway diseases in a tropical urban environment are driven by dominant mono-specific

sensitization against house dust mites. *Allergy* (2014). doi:10.1111/all.12364

36. Fairfax, B. P. *et al.* Genetics of gene expression in primary immune cells identifies cell type-specific master regulators and roles of HLA alleles. *Nat. Genet.* **44,** 502–510 (2012).
37. Zeller, T. *et al.* Genetics and beyond-the transcriptome of human monocytes and disease susceptibility. *PLoS One* **5,** e10693 (2010).
38. Thurman, R. E. *et al.* The accessible chromatin landscape of the human genome. *Nature* **489,** 75–82 (2012).
39. Wang, D. Y. & Gordon, B. R. Management of persistent allergic rhinitis in the tropics: Singapore experiences. *Clin. Exp. Allergy* **8,** 37–44 (2008).
40. Kurth, I. *et al.* Mutations in FAM134B, encoding a newly identified Golgi protein, cause severe sensory and autonomic neuropathy. *Nat. Genet.* **41,** 1179–1181 (2009).
41. Verpoorten, N., De Jonghe, P. & Timmerman, V. Disease mechanisms in hereditary sensory and autonomic neuropathies. *Neurobiol. Dis.* **21,** 247–255 (2006).
42. Kong, M., Kim, Y. & Lee, C. A strong synergistic epistasis between FAM134B and TNFRSF19 on the susceptibility to vascular dementia. *Psychiatr. Genet.* **21,** 37–41 (2011).
43. Tang, W. K. *et al.* Oncogenic properties of a novel gene JK-1 located in chromosome 5p and its overexpression in human esophageal squamous cell carcinoma. *Int. J. Mol. Med.* **19,** 915–923 (2007).

44. Ritchie, M. D. Using biological knowledge to uncover the mystery in the search for epistasis in genome-wide association studies. *Ann. Hum. Genet.* **75,** 172–182 (2011).

45. Pattin, K. A. & Moore, J. H. Exploiting the proteome to improve the genome-wide genetic analysis of epistasis in common human diseases. *Hum. Genet.* **124,** 19–29 (2008).

46. Emily, M., Mailund, T., Hein, J., Schauser, L. & Schierup, M. H. Using biological networks to search for interacting loci in genome-wide association studies. *Eur. J. Hum. Genet.* **17,** 1231–1240 (2009).

CHAPTER 4

ARCHILD: HIERARCHICAL VISUALIZATION OF LINKAGE DISEQUILIBRIUM IN HUMAN POPULATIONS

Rossella Melchiotti[1, 2], Olaf Rötzschke[1] and Michael Poidinger[1,*]

Address: [1]Singapore Immunology Network (SIgN), Agency for Science, Technology and Research (A*STAR), Singapore 138648, Singapore, [2]Doctoral School in Translational and Molecular Medicine (DIMET), University of Milano-Bicocca, Milan 20126, Italy

[*]Corresponding author

Melchiotti R, Rötzschke O, Poidinger M (2014) ArchiLD: Hierarchical Visualization of Linkage Disequilibrium in Human Populations. PLoS ONE 9(1): e86761.

Abstract

Linkage disequilibrium (LD) is an essential metric for selecting single-nucleotide polymorphisms (SNPs) to use in genetic studies and identifying causal variants from significant tag SNPs. The explosion in the number of polymorphisms that can now be genotyped by commercial arrays makes the interpretation of triangular correlation plots, commonly used for visualizing LD, extremely difficult in particular when large genomics regions need to be considered or when SNPs in perfect LD are not adjacent but scattered across a genomic region. We developed ArchiLD, a user-friendly graphical application for the hierarchical visualization of LD in human populations. The software provides a powerful framework for analyzing LD patterns with a particular focus on blocks of SNPs in perfect linkage as defined by r^2. Thanks to its integration with the UCSC Genome Browser, LD plots can be easily overlapped with additional data on regulation, conservation and expression. ArchiLD is an intuitive solution for the visualization of LD across large or highly polymorphic genomic regions. Its ease of use and its integration with the UCSC Genome Browser annotation potential facilitates the interpretation of association results and enables a more informed selection of tag SNPs for genetic studies.

Introduction

The development and diffusion of high-throughput technologies for the analysis of genetic variants, such as single-nucleotide polymorphism (SNP) microarrays and next-generation sequencing, has lead to a substantial increase in the number of variants that can be included in population-based genetic studies[1]. Commercially available microarray platforms can now measure up to 5 million SNPs on a single chip.

A large number of these variants are not independent but correlated through linkage disequilibrium (LD), the non-random association of alleles at two genomic locations. Knowledge of LD plays an important role in the selection of SNPs to be tested for association with a particular phenotype. This metric can in fact be used to alleviate the burden of multiple testing by pruning out redundant information. It is also useful for identifying possible causative variants from relevant tag polymorphisms.

The completion of the International HapMap project[2–5] and the 1000 Genomes Pilot Project[6] has provided scientists with a high-density map of genetic variation across the major human populations making it possible to evaluate patterns of LD on a genome-wide scale.

Triangular correlation plots for the visualization of pairwise linkage disequilibrium across genome regions, as implemented in Haploview[7], are the most used method to report LD information in population-based genetic studies[8]. However, the extensive number of variants now available on commercial arrays makes the

interpretation of triangular correlation plot difficult especially when SNPs in high linkage are not adjacent but interspersed between other SNPs. Furthermore the complexity of triangular correlation plots increases with the size of the region of interest which complicates the analysis of LD patterns across large genes. It is therefore necessary to develop alternative ways to plot LD for genome regions characterized by a high density of SNPs or fragmented LD blocks.

We conceived a new way of visualizing LD patterns across the genome as blocks of SNPs in perfect LD (r^2=1) that can be hierarchically clustered based on their pairwise linkage. ArchiLD is the implementation of this new concept in a user-friendly Java application which integrates the UCSC Genome Browser[9] visualization potential with a simple and intuitive tool for building LD blocks.

Implementation

ArchiLD comes in two versions: a client-server application (ArchiLD1k) for the analysis of LD across the four populations sequenced by the 1000 Genomes Pilot Project (CEU, CHB, JPT and YRI)[6] and a standalone application (ArchiLDCustom) for the analysis of LD across custom genotypic datasets.

ArchiLD1k computes blocks of SNPs in perfect LD (r^2=1), called clusters, from pre-calculated pairwise LD measures obtained using the software Haploview[7] on the four populations provided by the 1000 Genomes Pilot Project[6]. Due to the low number of CHB and JPT samples sequenced by the consortium (30 samples each) when

compared with the number of YRI and CEU samples (respectively 59 and 60), the two Asian populations have been merged together as previously done by others[6,10]. This allows for an easier comparison of LD patterns across populations due to their similar sample size. The software will soon be updated to include all 1092 individuals (14 distinct populations) sequenced by the 1000 Genome Project phase I[11]. More individuals/populations will be added as they become available.

ArchiLDCustom computes LD clusters from custom genotyping datasets imported by the user. The software accepts pedigree data and marker information in the standard linkage format used by Haploview[7]. Each chromosome needs to be loaded independently and r^2 values are estimated using the software Haploview[7].

In both versions of the software clusters are visualized as custom tracks in an integrated instance of the UCSC Genome Browser[9].

ArchiLD1k is implemented as a client-server application developed in Java v1.6 (client-side) and Java EE v1.7 (server-side). All computations are carried out by a Java servlet deployed on Apache Tomcat (v7.0.30). Hierarchical clustering of LD blocks is performed using R v2.15.1 by the function hclust from the package {stats} utilizing an agglomeration method based on average[12]. The distance matrix used for the analysis is defined from LD pairwise measures as $1-r^2$. Analysis parameters can be selected by means of a query interface on the client side and are then dispatched to the servlet which computes LD blocks and generates the custom Browser Extensible Data (BED) tracks used for plotting clusters. BED is a file

format used by the UCSC Genome Browser[9] to define genomic regions: a description of the required fields can be found on the UCSC Genome Bioinformatics website[13]. Custom tracks are automatically imported in the browser when the user selects which architecture to visualize. Gene annotations were downloaded from the UCSC Genome Browser database[14] and SNP annotations were provided by the 1000 Genome Pilot Project (genome build hg18)[6]. Hg19 positions were obtained using the liftOver tool provided by the UCSC Genome Browser[15]. Minor allele frequencies were computed using the software PLINK[16] on all samples sequenced by the 1000 Genomes Pilot Project (60 CEU, 60 CHB+JPT and 59 YRI)[6]. Each population was analyzed independently. Tables containing LD information and gene/SNP annotations are managed on the server side using MySQL v 5.5.27.

ArchiLDCustom is completely implemented in Java v1.6. Hierarchical clustering of LD blocks is performed by a local instance of R using the same algorithm described for ArchiLD1k. Annotation tables and custom datasets are managed through a MySQL database. Connection parameters to the MySQL server and the complete path to the local installation of Haploview and R need to be set before any analysis can be performed. All the computations are run locally. LD plots can be visualized using an integrated instance of the UCSC Genome Browser[9] but contrarily to the client-server version tracks containing LD plots need to be uploaded manually in the browser. All plots are stored as BED files.

All graphical interfaces were implemented using the Qt libraries for Java[17].

Results

LD architectures

LD architectures, sets of clusters of SNPs in perfect LD (r^2=1), can be built starting from four distinct genomic elements: genes, SNPs, chromosomes (ArchiLD1k only) and genomic regions (ArchiLDCustom only).

Gene-centered architectures are composed of LD blocks with at least one SNP located inside a gene or in proximity of it. The maximum distance between the gene transcription start/end and the SNPs to be included in the analysis can be adjusted by the user. The tool accepts three different identifiers for gene names: Entrez IDs, RefSeq IDs and HUGO gene symbols.

SNP-centered architectures contain all clusters in LD with a selected SNP (reference SNP). To reduce the size of the tables used for the computations and the processing time required to generate the clusters only variants with an r^2>=0.5 with the reference SNP are considered.

Chromosome-centered architectures contain all LD blocks located on a specific chromosome. Region-centered architectures focus on the genomic region described in the genotypic dataset imported by the user. Only SNPs included in the file are used to build clusters.

Regardless of the type of analysis selected, users can set a minor allele frequency threshold to exclude rare SNPs from the analysis and

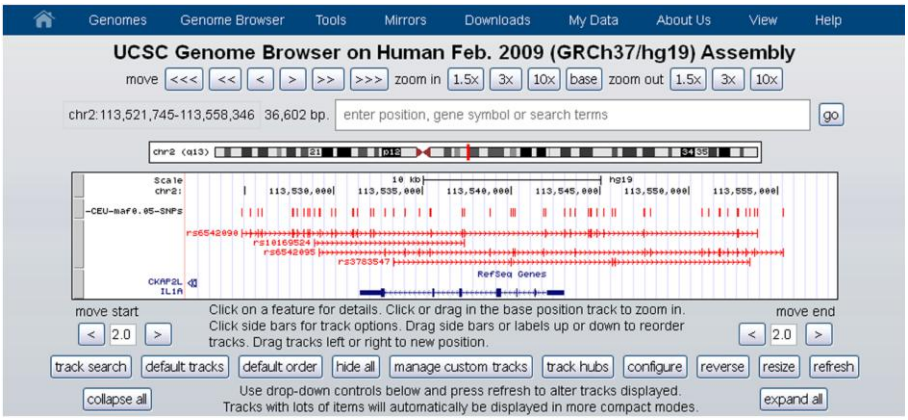


Figure 1. An example of gene-centered architecture. The first track contains the names and positions of all SNPs considered in the analysis. The second track contains all clusters associated to the gene.

choose which genome build (hg18/hg19) to use for the visualization (ArchiLD1k only).

Visualization

LD architectures are visualized as custom BED files in the integrated instance of the UCSC Genome Browser[9]. SNPs in perfect LD (r^2=1) are joined by a horizontal line which constitutes a cluster.

Gene-centered architectures are represented by two distinct tracks, one containing the name and position of individual SNPs and one containing all blocks in the region. Clusters are identified by the name of their first SNP (Figure 1). Users can decide to include only SNPs belonging to a cluster or all the SNPs in the region. If SNPs not belonging to any LD block (singletons) are included, a new track is added to the visualization.

It is also possible to build a hierarchical tree of clusters spanning the gene. Blocks of SNPs in perfect LD are clustered according to their pairwise r^2. Hierarchical plots are displayed next to the Genome Browser window to facilitate the interpretation of LD patterns. When this option is selected each cluster is represented by a distinct track with the order of the tracks following the order of the clusters in the hierarchical tree (Figure 2).

Region-centered architectures are similarly visualized but only SNPs included in a custom genotypic dataset are used for the analysis.

SNP-centered architectures are represented by multiple tracks. The first track contains SNP names. The second track displays the reference SNP and all its perfectly linked SNPs. All other tracks are ordered by decreasing r^2 values (the corresponding r^2 is included in

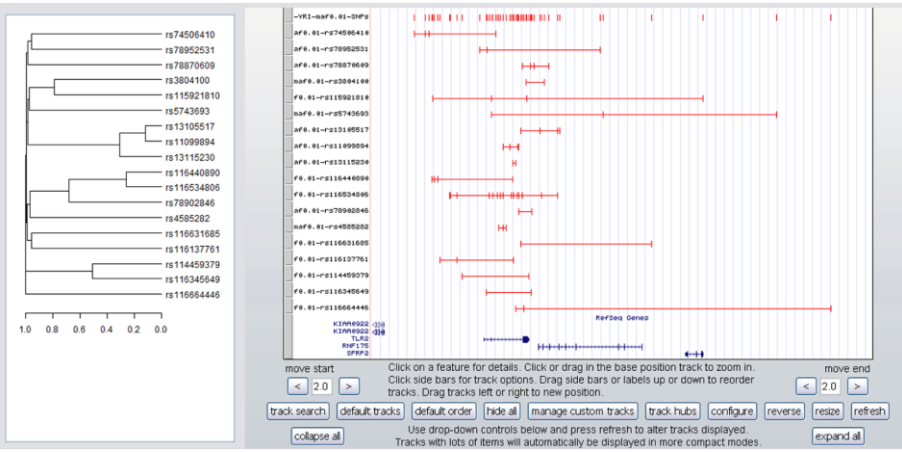


Figure 2. An example of hierarchical clustering. On the left the hierarchical clustering of all SNP blocks associated to a particular gene. On the right a graphical representation of a gene-centered architecture. The first track contains the names and positions of all SNPs considered in the analysis. The following tracks are ordered as they appear in the hierarchical clustering plot.

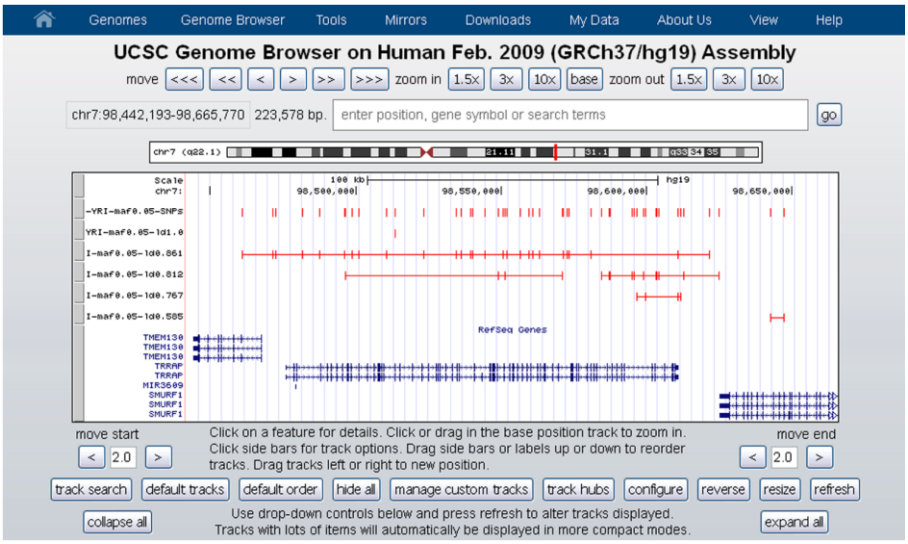


Figure 3. An example of SNP-centered architecture. The first track contains the names and positions of all SNPs considered in the analysis. The second track contains the reference SNP and its linked variants. The following tracks are ordered by descending r^2 with respect to the reference SNP.

the track name, Figure 3). Singletons can be included and added to the track with the appropriate r^2 value.

For gene-centered, SNP-centered and region-centered architectures the user has the option to color SNPs and LD blocks by minor allele frequency. This simplifies the identification of clusters of SNPs in linkage disequilibrium with similar allele distributions.

For chromosome-centered architectures all clusters are added to the same track (Figure 4). When singletons are included an additional track containing only these SNPs is added to the visualization.

BED tracks and filenames containing the IDs of SNPs in a particular cluster can be easily exported for external use using the navigation tree. Multiple plots can be uploaded simultaneously as UCSC tracks

to facilitate the comparison of LD patterns across different populations.

In ArchiLD1k, gene-centered and SNP-centered architectures can be automatically loaded into the browser by clicking on the corresponding item in the navigation tree. Due to their large size chromosome-centered architectures cannot be automatically loaded but need to be exported first and then manually imported. ArchiLDCustom requires users to export BED tracks and manually import them in the browser regardless of the type of architecture generated.

Discussion

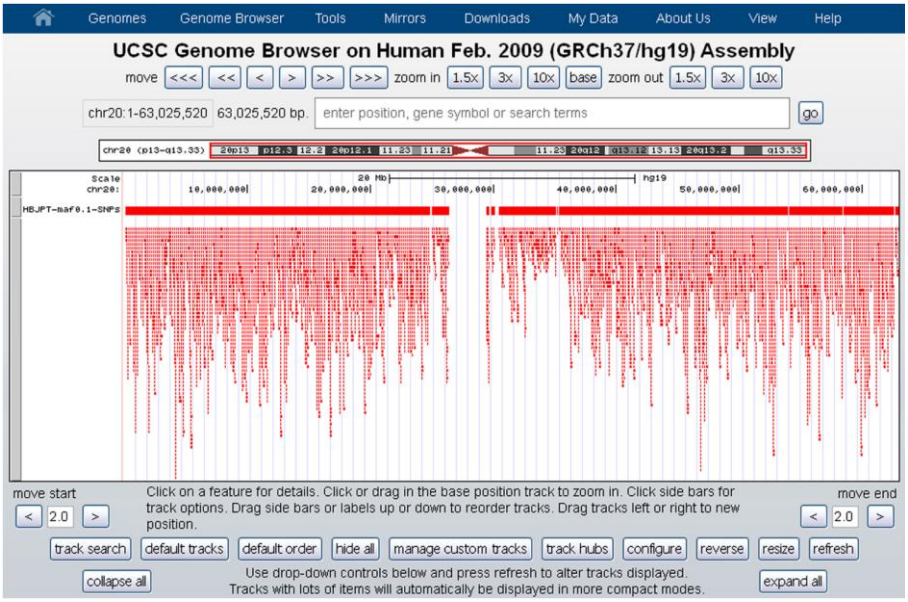


Figure 4. An example of chromosome-centered architecture. The first track contains the names and positions of all SNPs considered in the analysis. The second track contains all clusters located on the chromosome.

ArchiLD is a powerful software for producing highly interpretable plots that can be used to select tag SNPs in the context of association studies or to prioritize SNPs for functional studies. Its integration with the UCSC Genome Browser[9] makes it easy to overlap additional information about regulation, conservation across species, phenotype and disease associations aiding the users in the interpretation of their results.

Query performance

Generation of SNP-centered architectures requires few seconds. For gene-centered architectures the process can take from few seconds to several minutes according to the length of the gene and the size of the upstream/downstream region selected by the user. Chromosome-centered architectures are pre-computed but due to their large size require minutes to be uploaded into the UCSC Genome Browser[9]. The import of custom datasets can be time-consuming depending on the number of SNPs included in the file: the limiting factor is the computational time required by Haploview[7] to generate pairwise LD measures.

Comparison with similar software

Multiple tools have so far been developed to tackle the complexity of linkage disequilibrium visualization in human populations. Haploview[7] is one of the most used software for the computation and visualization of LD. It is extremely powerful for visualizing small genomic regions where highly linked SNPs are organized in compact blocks and it is therefore strongly used for tag selection in candidate

gene studies. Triangular correlation plots become difficult to interpret when the region of interest contains a large number of SNPs or when highly linked SNPs are not adjacent (Figure 5). ArchiLD facilitates the analysis of LD across these regions by joining perfectly linked SNPs in visual clusters and by ordering clusters with respect to their relative LD: SNP-centered architectures are ordered by decreasing r^2 with respect to a reference SNP (index SNP) while gene-centered and region-centered architectures are ordered according to their hierarchical tree with strongly linked blocks clustered together. In Haploview LD computations are done on the fly, a time and memory consuming process. ArchiLD1k on the other hand uses pre-computed r^2 values for building and hierarchically organizing LD blocks and can thus be used on a very large scale.

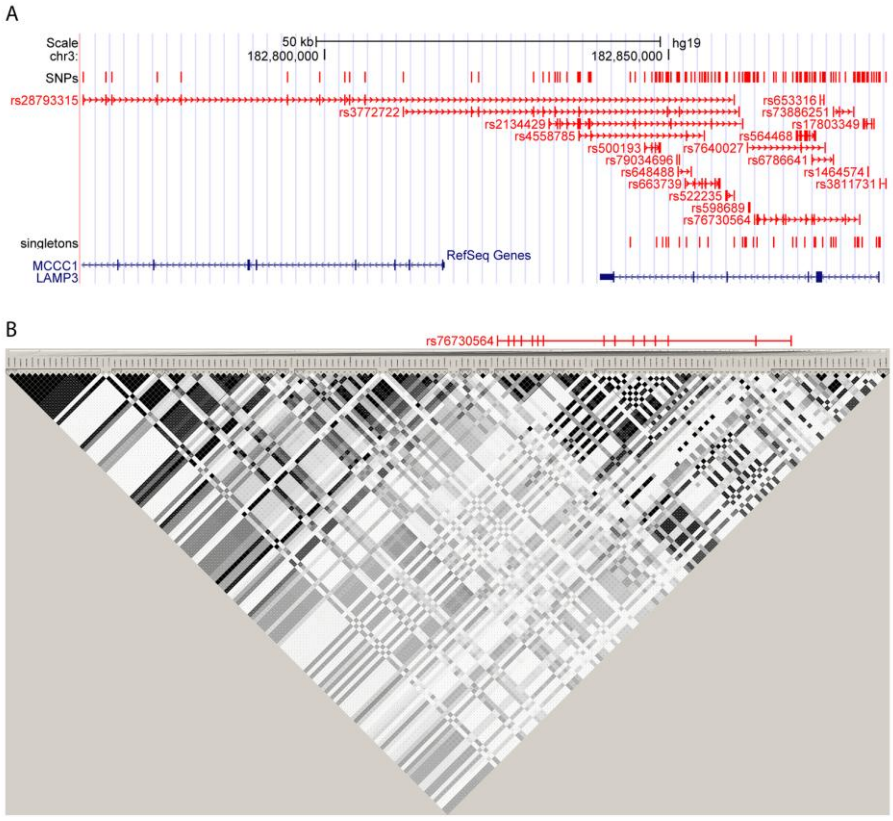


Figure 5. Comparison between ArchiLD and Haploview. (A) Gene-centered visualization for the gene LAMP3 in CEU as provided by ArchiLD1k. (B) Analogous visualization in Haploview. The large number of variants analyzed makes it difficult, for example, to identify the position of the SNPs in perfect LD with rs76730564.

Another interesting way of computing LD clusters is offered by LD-based clumping, a technique implemented, for example, in PLINK[16]. The clumping procedure is straightforward: it requires a SNP dataset and an input file containing variant names and association p-values. SNPs with an assigned p-value lower than a certain threshold (user-based) are taken as index SNPs. All the other SNPs in the region are assigned to their most closely linked index SNP. An r^2 threshold can

be selected to ignore variants that are weakly linked to index SNPs. The output of the procedure is a text file containing the list of index SNPs (with their respective association p-values) and the list of their linked SNPs. The information provided by the clumping procedure is extremely valuable for the interpretation of association results and the selection of genetic variants to use in downstream analysis but the visualization of the results is not straightforward. ArchiLD does not offer any clumping functionality since our visualization of linkage disequilibrium is independent of the availability of association results. Nonetheless clumping could be performed in two steps: by manually analyzing an association file to identify index SNPs and by creating a SNP-centered architecture for each index SNP.

Another well used software for the analysis of LD patterns across the genome is SNAP[10]. The plots produced by SNAP are very similar to the SNP-centered plots produced by ArchiLD but our software has the advantage of including individual SNP labels and visualizing the gene structure (number of transcripts, positions of the exons) (Figure 6). SNAP can also be used to integrate association results with LD information. The LD plot is by default centered on the SNP with the highest association signal (this can be modified by the user). The tool does not offer any way of focusing on a particular gene/region. This limitation is overcome by other tools such as LocusZoom[18] where the user can choose a gene or a genomic region to analyze. As for SNAP the plot is centered on the SNP with the lowest p-value but this can be easily modified by the user. Both HapMap[2–5] and the 1000 Genome Project datasets[6,11] can be used for computing LD. The

advantage of this tool with respect to ArchiLD lies in the availability of pre-loaded GWAS datasets. Custom association datasets can also be loaded for the visualization. The integration of LD plots with pre-loaded GWAS datasets is also offered by Ricopili, a tool developed by the Broad Institute[19]. As for LocusZoom the association/LD plot can be centered on a particular gene or a particular genomic region. The advantage of Ricopili with respect to LocusZoom is that more than one index SNP can be used for the visualization (clumping). When more than one reference variant is selected the most associated SNPs are chosen as index SNPs and all the other SNPs are assigned to their most closely linked reference SNP. SNPs are colored according to their pairwise LD with the reference SNP they are assigned to. The main disadvantage of this kind of visualization is that it does not provide any information about the pairwise LD of SNPs with the same color. It is impossible to say from the plot if these SNPs are totally independent or are strongly linked. ArchiLD tackles this limitation by joining SNPs which are perfectly linked with a horizontal line. While ArchiLD does not provide any pre-loaded GWAS dataset, custom association p-values can be imported as bedGraph custom tracks[13] and easily overlapped with the LD plot in the integrated browser. Except for Haploview all of the tools mentioned above generate SNP-centered plots: the plot can be centered on a particular gene but one or more reference SNPs need to be selected before a visualization can be created. ArchiLD not only provides options to generate gene-centered LD patterns but also offers the possibility of clustering the resulting LD blocks using a hierarchical tree.

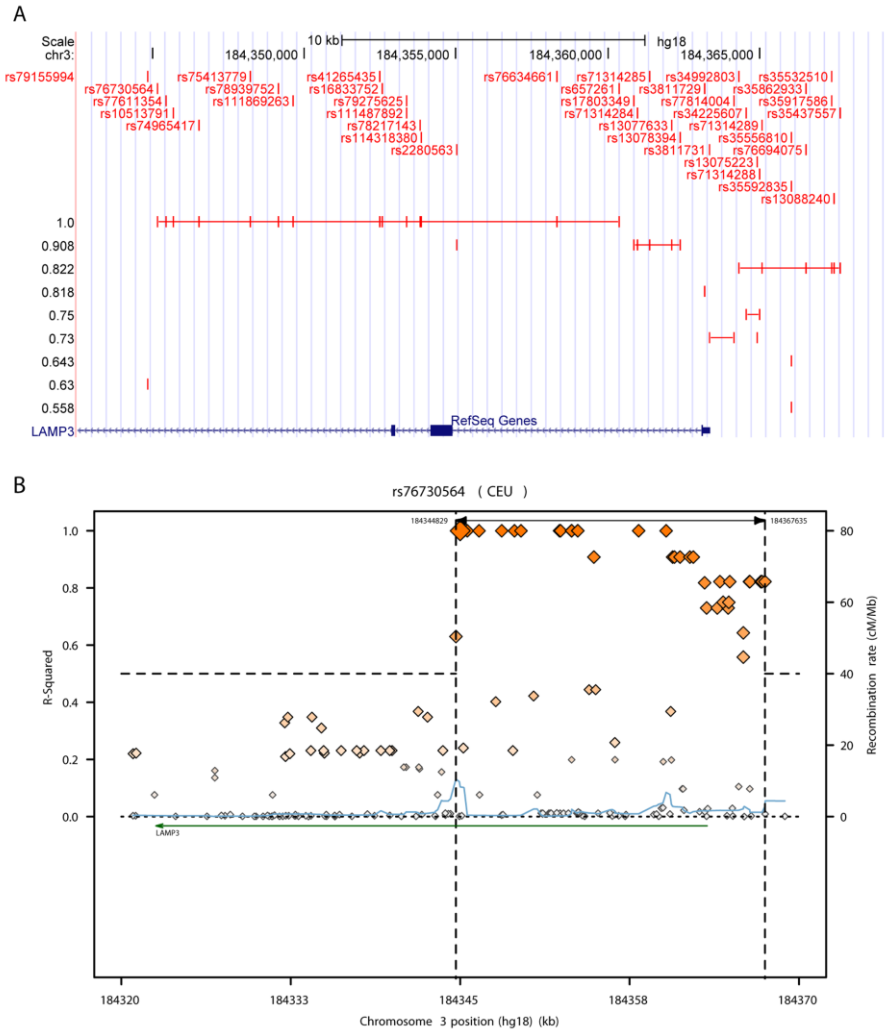


Figure 6. Comparison between ArchiLD and SNAP. (A) SNP-centered visualization for the variant rs76730564 in CEU as provided by ArchiLD1k. Clusters are ordered by descending values of r^2 with the reference SNP. (B) Analogous visualization in SNAP. The plot does not contain any SNP label nor provide any information on the location of exons.

A strong advantage of ArchiLD over the aforementioned tools is that it does not produce static pdf plots but interactive plots thanks to its

integration with the UCSC Genome Browser[9]. Plots can be modified in real-time by zooming in/out or shifting the visualization upstream/downstream. Once the user is satisfied with the visualization a pdf file can be created.

The major drawback of all the tools here described when compared with ArchiLD consists in the difficulty of adding functional annotations to LD plots (besides association p-values or in the case of Ricopili NHGRI GWAS catalog variants) : ArchiLD integration with the UCSC Genome Browser[9] provides a solution to this limitation thanks to the large number of free annotation tracks available.

Availability

Precompiled binaries for both ArchiLD1k and ArchiLDCustom (Windows, Linux and Mac OS) can be downloaded from the project website (http://archild.sign.a-star.edu.sg)[20]. ArchiLD1k binaries are distributed as zip files containing a runnable jar and all the libraries required by the application. ArchiLDCustom binaries are distributed as zip files containing a runnable jar, all the libraries and annotation tables required by the application and two sample datasets that can be used to test the software. The two sample datasets have been generated using genotype data from the 1000 Genomes Pilot Project[6] for the first 500kbp of chromosome 7 (CEU) and chromosome 12 (CHB+JPT) respectively. ArchiLDCustom requires R v 2.15.1 and access to a MySQL database (MySQL v 5.5.27 or higher).

Instructions on how to launch the software under different operating systems and how to use the different functionalities of ArchiLD are

described in the manual, available for download on the project website. Source files can also be obtained from the website and easily imported as java projects in Eclipse IDE for Java Developers. The software is released under the GNU General Public License (GPL) version 3.

Conclusions

ArchiLD is a user-friendly application for the visualization of linkage disequilibrium in human populations. The software was developed to aid geneticists in selecting SNPs to include in genetic studies and in identifying putative causative SNPs from relevant tag variants. Its ease of use and high interpretability make ArchiLD a powerful addition to every geneticist's toolbox.

Acknowledgements

We would like to thank Solomonraj Wilson (SIgN), Yoong Hin Tay and Wai Kok Kenny Hoi (A*STAR) for technical advice and the set up of the virtual machine server.

REFERENCES

1. Kofler, R. & Schlötterer, C. Gowinda: unbiased analysis of gene set enrichment for genome-wide association studies. *Bioinformatics* **28,** 2084–2085 (2012).
2. The International HapMap Consortium. The International HapMap Project. *Nature* **426,** 789–796 (2003).
3. The International HapMap Consortium. A haplotype map of the human genome. *Nature* **437,** 1299–1320 (2005).
4. The International HapMap Consortium. A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449,** 851–861 (2007).
5. The International HapMap Consortium. Integrating common and rare genetic variation in diverse human populations. *Nature* **467,** 52–58 (2010).
6. The 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature* **467,** 1061–1073 (2010).
7. Barrett, J. C., Fry, B., Maller, J. & Daly, M. J. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* **21,** 263–265 (2005).
8. Bush, W. S., Dudek, S. M. & Ritchie, M. D. Visualizing SNP statistics in the context of linkage disequilibrium using LD-Plus. *Bioinformatics* **26,** 578–579 (2010).
9. Kent, W. J. *et al.* The Human Genome Browser at UCSC. *Genome Res.* **12,** 996–1006 (2002).

10. Johnson, A. D. *et al.* SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap. *Bioinformatics* **24,** 2938–9 (2008).
11. The 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491,** 56–65 (2012).
12. R Development Core Team. *R: A Language and Environment for Statistical Computing.* (2012).
13. UCSC Genome Bioinformatics. at <http://genome.ucsc.edu/FAQ/FAQformat.html>
14. Meyer, L. R. *et al.* The UCSC Genome Browser database: extensions and updates 2013. *Nucleic Acids Res.* **41,** D64–D69 (2013).
15. Hinrichs, A. S. *et al.* The UCSC Genome Browser Database: update 2006. *Nucleic Acids Res.* **34,** D590–D598 (2006).
16. Purcell S *et al.* PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *Am. J. Hum. Genet.* **81,** 559–575 (2007).
17. Qt Jambi. at <http://qt-jambi.org/>
18. Pruim, R. J. *et al.* LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics* **26,** 2336–7 (2010).
19. Ricopili. (2010). at <http://www.broadinstitute.org/mpg/ricopili/>
20. ArchiLD. (2013). at <http://archild.sign.a-star.edu.sg/>

CHAPTER 5

SUMMARY, CONCLUSIONS AND FUTURE PERSPECTIVES

The connection between genotype and phenotype is not straightforward due to the multiple mechanisms through which DNA mutations can affect phenotypic manifestations and the numerous ways polymorphisms can interact to produce a final outcome. This is particularly true in the context of the immune system which is characterized by a complex interplay between multiple cell subsets. Because of the redundancy of the system and the number of players involved it is rare for a single polymorphism to have a measurable direct impact on a complex phenotype such as disease susceptibility. In most cases in fact polymorphisms do not act alone but synergistically contribute to the emergence of certain phenotypes. The identification of these interacting variants is not easy due to the large number of entities in the system which poses a statistical challenge because many tests needs to be performed with relatively small cohorts. It is therefore useful to also focus on intermediate phenotypes such as gene expression in addition to complex phenotypes like disease susceptibility to try and elucidate the impact of variants on complex diseases.

For this project we adopted a knowledge driven approach for the identification of association and epistasis events in the context of

immune cells and immune diseases. The aim of our work was two-fold: to characterize the influence of genetic variants on transcriptional regulation in an immune cell subset; and to study the role played by association and epistasis in a complex immune disease characterized by the interaction of multiple cell types. Considering that both expression quantitative trait loci and genome wide association/epistasis studies have so far been mostly performed on populations of European descent we decided to center our characterization on a less studied population: Singapore Chinese.

To evaluate the impact of polymorphisms on expression in an immune cell subset we focused on neutrophils, key players of the innate immune system. Neutrophils have so far not been characterized for eQTLs most likely because of the technical difficulties inherent in handling this cell subset. These cells are in fact considered fragile and have been shown to be easily activated[1,2]. For this study we carefully isolated neutrophils from 114 well-matched samples of Chinese ethnicity and performed a genome-wide eQTL study to identify *cis* polymorphisms associated with gene expression levels. Using a permutation significance threshold of 0.001 we identified 21,210 significant eQTL probe/SNP pairs involving 971 distinct probes (832 distinct HUGO genes). Out of the 971 probes with a significant eQTL 525 were also reported by a large scale whole blood eQTL study backing our claim of genetic regulation for those probes[3]. In addition numerous SNPs involved in neutrophil eQTLs were previously described as being associated to diseases/traits as listed by the GWAS catalog supporting the hypothesis that

transcriptional dysregulation can in some cases lead to increased disease susceptibility[4]. An enrichment analysis for the 971 probes with a significant eQTLs suggested an involvement for neutrophil eQTLs in dermatological diseases (psoriasis and dermatitis in particular). An analysis of two GEO psoriasis datasets revealed that differentially expressed genes between lesional skin from cases and normal skin from controls were enriched for neutrophil eQTLs. No enrichment was found instead for atopic dermatitis, a disease characterized by different infiltrates than psoriasis. While atopic dermatitis skin lesions are predominantly accompanied by infiltration of macrophages, dendritic cells, eosinophils and Th2 CD4+ lymphocytes[5,6] psoriatic plaques are polarized towards a Th1 response and are characterized by an accumulation of T cells, monocytes and neutrophils[7].

Future work will focus on functionally characterizing neutrophil eQTLs involving differentially expressed genes in psoriasis and other interesting eQTL signals associated with disease susceptibility by GWAS. A better grasp of the impact of genetic variants on transcriptional regulation might help in acquiring a deeper understanding of diseases characterized by a dysregulation in neutrophil functions or numbers. It would also be interesting to apply a similar approach to study how the eQTL landscape of neutrophils changes upon treatment of these cells with different stimuli to better characterize their behavior under inflammatory conditions.

As regards the translational contribution of this project neutrophils are the most abundant leukocyte cell subset in blood and the first

line of defense against invading pathogens[8–10]. A better characterization of the impact of mutations on transcriptional regulation of key neutrophil genes might lead to a deeper understanding of some of the mechanisms underlying disease susceptibility, especially for those immune conditions characterized by a strong neutrophil involvement.

To study the role played by association and epistasis on a complex immune disease we concentrated on a published allergic rhinitis (AR) cohort of individuals of Chinese ethnicity[11]. Due to the relatively small size of our discovery and validation cohorts we limited our search to an important gene known to play a role in immune suppression by T regulatory cells (Treg), CD39. Treg cells have in fact been shown to play a very important role in allergic diseases[12,13]. They are essential in dampening the allergic reaction by suppressing CD4+ T effector cells and inhibiting the production of Th2 cytokines driving the allergic reaction[12,14–16]. The removal of extracellular ATP, a pro-inflammatory danger signal, by CD39 is one of the known mechanisms through which Treg cells promote immune suppression[17]. Differences in CD39 expression on Treg cells have been described for multiple immune diseases such as multiple sclerosis and cancer as well as for infectious diseases such as Hepatitis B, Hepatitis C and HIV[17–22]. CD39 polymorphisms have already been described in association with inflammatory bowel disease (IBD) and HIV progression[18,23]. A closely linked polymorphism to those described in the context of IBD and HIV was found to be

associated with the frequency of CD39+ activated Treg with respect to their parental population[24].

In our study we showed that another highly linked CD39 polymorphism (rs7071836) was associated with the surface expression of this molecule on Treg cells but not on other CD39-expressing leukocyte subsets. Together with another polymorphism in the promoter region of FAM134B (rs257174), SNP rs7071836 was found to impact susceptibility to AR. A comparison between symptomatic AR cases and asymptomatic atopic controls established that the interaction was associated with disease but not with predisposition to atopy. Polymorphism rs257174 was shown to affect expression of its *cis* gene in monocytes but notably not in Treg cells. Using three distinct large cohorts we were also able to show that the expression of CD39 and FAM134B was inversely correlated in whole blood implying that the genetic interaction had an effect *in vivo*. In addition we were able to suggest a mechanism for the interaction of CD39 in Treg cells and FAM134B in monocytes by showing that extracellular ATP levels, which inversely correlate with CD39 expression, regulated FAM134B expression in monocytes. We also demonstrated that in the presence of an ATP-receptor antagonist this ATP-induced upregulation of FAM134B was reduced.

The role of FAM134B in allergic rhinitis remains to be elucidated since not much is known about this gene. FAM134B is a *cis*-Golgi protein so far mainly described in the context of the nervous system. Loss-of-function mutations in this gene have been associated to severe sensory and autonomic neuropathy and this gene has been

implicated in the long term survival of nociceptive and autonomic neurons[25]. The role of this gene in the context of the immune system still needs to be elucidated. We showed that this gene is expressed in multiple immune cell subsets and genetically regulated in monocytes. It would be interesting to investigate the function of this gene in this cell subset and see how this could relate to AR susceptibility. Future work will therefore focus on studying how this gene may affect protein trafficking and secretion in response to different stimuli and on revealing how this might relate to AR symptoms. Since the interaction between CD39 and FAM134B did not seem to be associated with allergy predisposition in our population it is unlikely that this molecule disrupts the IgE pathway. The protein is most probably involved in the development of a symptomatic reaction in response to allergen challenge.

As regards the translational contribution of this project we were able to propose a new candidate gene for allergic rhinitis, FAM134B. A deep characterization of this not well-described protein might lead to novel insights into AR susceptibility. In addition a similar approach based on selecting a gene known to be involved in a disease and looking for genome-wide epistatic partners, might be easily applied to other immune conditions in order to detect new candidate genes for disease susceptibility.

Throughout this thesis we did not limit our analysis to the identification of statistical association or epistasis but, whenever possible, we tried to characterize the biological impact of statistical findings. In order to do this we needed a tool for the identification of

potential functional SNPs from disease/trait-associated polymorphisms. Due to a phenomenon known as linkage disequilibrium (the non random association of allelic states at two or more loci) statistically significant SNPs are in fact usually not causative but simply tag functional SNPs. To identify biologically important genetic variants from statistically relevant ones it is therefore necessary to integrate linkage disequilibrium information with public biological knowledge. None of the available tools for the visualization of linkage disequilibrium provided an option for incorporating available biological knowledge. We therefore conceived a novel tool with the capability of overlapping biological knowledge with linkage disequilibrium plots to try and bridge the gap between statistical association and biological mechanisms. We developed a software called ArchiLD[26] that integrates a user-friendly graphical representation of linkage disequilibrium with the annotation potential of the UCSC Genome Browser[27]. Linkage disequilibrium plots, which can be generated using both custom datasets and data from the four populations sequenced by the 1000 Genomes Pilot Project[28], can now be easily overlapped with transcription factor binding sites, open chromatin regions, conservation plots and published findings. This is a great tool for the discovery of causative variants from tag polymorphisms reported as being associated with a particular immune phenotype or disease. The identification and characterization of causative variants is the first step in understanding how a particular mutation affects a phenotype, a prerequisite for developing pharmaceutical solutions to complex

diseases. Future work will focus on adding new populations to the software like the fourteen populations sequenced by the 1000 Genomes Pilot Project[29] and the three provided by the Singapore Genome Variation Project[30]. Moreover it would be useful to integrate into the software eQTL information derived both by published studies and by in-house cohorts. With this additional data tag SNPs or their linked SNPs could be directly linked to changes in expression across different immune cells.

In this thesis we propose a biologically-oriented statistical approach for the identification of association and epistasis in the context of the immune cells and apply it both in the context of gene regulation (by characterizing the eQTL landscape of neutrophils) and in the context of complex diseases (by studying the impact of CD39 variants on AR susceptibility). Due to the large number of tests performed for both analyses and the relatively small size of our cohorts, whenever possible, we tried to validate our statistical results using replication cohorts, published studies or additional biological experiments to reduce the possibility of false positives. In addition we often tried to link statistical findings with biological mechanisms in order to increase the translational potential of our findings.

REFERENCES

1. Pillay, J. *et al.* A subset of neutrophils in human systemic inflammation inhibits T cell responses through Mac-1. *J. Clin. Invest.* **122,** 327–336 (2012).
2. Glasser, L. & Fiederlein, R. L. The effect of various cell separation procedures on assays of neutrophil function. A critical appraisal. *Am. J. Clin. Pathol.* **93,** 662–669 (1990).
3. Westra, H.-J. *et al.* Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nat. Genet.* **45,** 1238–1243 (2013).
4. Welter, D. *et al.* The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.* **42,** D1001–6 (2014).
5. Leung, D. Y. Atopic dermatitis: new insights and opportunities for therapeutic intervention. *J Allergy Clin Immunol* **105,** 860–876 (2000).
6. Leung, D. Y. M. & Bieber, T. Atopic dermatitis. *Lancet* **361,** 151–60 (2003).
7. Christophers, E. & Henseler, T. Contrasting disease patterns in psoriasis and atopic dermatitis. *Arch. Dermatol. Res.* **279 Suppl,** S48–S51 (1987).
8. Mollinedo, F., Borregaard, N. & Boxer, L. Novel trends in neutrophil structure, function and development. *Immunol. Today* 535–537 (1999).

9. Von Vietinghoff, S. & Ley, K. Homeostatic Regulation of Blood Neutrophil Counts. *J. Immunol.* **181,** 5183–5188 (2008).
10. Nathan, C. Neutrophils and immunity: challenges and opportunities. *Nat. Rev. Immunol.* **6,** 173–82 (2006).
11. Andiappan, A. K. *et al.* Genome-wide association study for atopy and allergic rhinitis in a Singapore Chinese population. *PLoS One* **6,** e19719 (2011).
12. Palomares, O. *et al.* Role of Treg in immune regulation of allergic diseases. *Eur. J. Immunol.* **40,** 1232–1240 (2010).
13. Dimeloe, S., Nanzer, A., Ryanna, K. & Hawrylowicz, C. Regulatory T cells, inflammation and the allergic response-The role of glucocorticoids and Vitamin D. *J. Steroid Biochem. Mol. Biol.* **120,** 86–95 (2010).
14. Ling, E. M. *et al.* Relation of CD4+ CD25+ regulatory T-cell suppression of allergen-driven T-cell activation to atopic status and expression of allergic disease. *Lancet* **363,** 608–615 (2004).
15. Robinson, D. S., Larché, M. & Durham, S. R. Tregs and allergic disease. *J. Clin. Invest.* **114,** 1389–1397 (2004).
16. Bellinghausen, I., Klostermann, B., Knop, J. & Saloga, J. Human CD4+CD25+ T cells derived from the majority of atopic donors are able to suppress TH1 and TH2 cytokine production. *J Allergy Clin Immunol* **111,** 862–868 (2003).
17. Borsellino, G. *et al.* Expression of ectonucleotidase CD39 by Foxp3+ Treg cells: hydrolysis of extracellular ATP and immune suppression. *Blood* **110,** 1225–1232 (2007).

18. Nikolova, M. *et al.* CD39/Adenosine Pathway Is Involved in AIDS Progression. *PLoS Pathog.* **7,** e1002110 (2011).
19. Tang, Y., Jiang, L., Zheng, Y., Ni, B. & Wu, Y. Expression of CD39 on FoxP3+ T regulatory cells correlates with progression of HBV infection. *BMC Immunol.* **13,** 17 (2012).
20. Fletcher, J. M. *et al.* CD39+Foxp3+ regulatory T cells suppress pathogenic Th17 cells and are impaired in multiple sclerosis. *J. Immunol.* **183,** 7602–7610 (2009).
21. Mandapathil, M. *et al.* Increased ectonucleotidase expression and activity in regulatory T cells of patients with head and neck cancer. *Clin. Cancer Res.* **15,** 6348–57 (2009).
22. Kared, H., Fabre, T., Bédard, N., Bruneau, J. & Shoukry, N. H. Galectin-9 and IL-21 mediate cross-regulation between Th17 and Treg cells during acute hepatitis C. *PLoS Pathog.* **9,** e1003422 (2013).
23. Friedman, D. J. *et al.* CD39 deletion exacerbates experimental murine colitis and human polymorphisms increase susceptibility to inflammatory bowel disease. *Proc. Natl. Acad. Sci.* **106,** 16788–16793 (2009).
24. Orrù, V. *et al.* Genetic Variants Regulating Immune Cell Levels in Health and Disease. *Cell* **155,** 242–256 (2013).
25. Kurth, I. *et al.* Mutations in FAM134B, encoding a newly identified Golgi protein, cause severe sensory and autonomic neuropathy. *Nat. Genet.* **41,** 1179–1181 (2009).

26. Melchiotti, R., Rötzschke, O. & Poidinger, M. ArchiLD: Hierarchical Visualization of Linkage Disequilibrium in Human Populations. *PLoS One* **9,** e86761 (2014).
27. Kent, W. J. *et al.* The Human Genome Browser at UCSC. *Genome Res.* **12,** 996–1006 (2002).
28. The 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature* **467,** 1061–1073 (2010).
29. The 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491,** 56–65 (2012).
30. Teo, Y., Sim, X., Ong, R. & Tan, A. Singapore Genome Variation Project: a haplotype map of three Southeast Asian populations. *Genome Res.* **19,** 2154–2162 (2009).

PUBLICATIONS

Melchiotti R, Rötzschke O, Poidinger M (2014) ArchiLD: Hierarchical Visualization of Linkage Disequilibrium in Human Populations. PLoS ONE 9(1): e86761.