Review

Bench-to-bedside review: Genetics and proteomics: deciphering gene association studies in critical illness

Sachin Yende[1], Candace M Kammerer[2] and Derek C Angus[1]

[1]The CRISMA (Clinical Research, Investigation, and Systems Modeling of Acute Illness) Laboratory, Department of Critical Care Medicine, University of Pittsburgh, Pittsburgh, PA, USA
[2]Department of Human Genetics, University of Pittsburgh, PA, USA

Corresponding author: Sachin Yende, yendes@upmc.edu

Abstract

There is considerable interest in understanding genetic determinants of critical illness to improve current risk stratification models, provide individualized therapies, and improve our current understanding of disease mechanisms. This review provides a broad overview of genetic nomenclature, different study designs, and problems unique to each of these study designs in critical illnesses. Well designed genetic studies with careful attention to these issues during the planning phase, use of rigorous statistical methods during analysis, replication of these results in different cohorts will lead to more robust results and improved understanding of genetics of critical care.

The completion of the Human Genome draft in 2000 has been accompanied by an explosion of studies examining genetic determinants of disease [1,2]. In critical care, current prediction models based on socio-demographic and clinical risk factors fail to explain fully why a particular patient either develops or succumbs to disease. Consequently, physicians have tried to understand if genetic variation affects susceptibility and outcome of critical illnesses. Genetics may also provide insights into biological mechanisms and allow more precise use of interventions. Using targeted therapy based on an individual's genetic makeup, rather than using it on all patients, is an appealing strategy. But conflicting results from early studies in genetics of critical illness have led the scientific community to view these results with skepticism [3]. For example, there has been little consensus regarding genetic markers associated with a tumor necrosis factor (TNF) hypersecretor response. In particular, contradictory reports have been published for the association between the -308 guanine to adenine transition within the promoter region of the TNF gene and its expression and severe sepsis susceptibility [4,5]. This article will provide a broad outline of study designs to ascertain the role of genetic

variation in critical care and focus on gene association studies, the most common study design in critical care. The article also addresses both problems generic to genetic studies and those unique to genetics of critical illness.

Mendelian and complex traits

Mendelian traits or diseases, such as sickle cell disease or cystic fibrosis, are influenced by a single gene. In contrast, most critical illnesses are multifactorial diseases, and called 'complex traits' in genetic parlance. Severe sepsis, an example of a complex trait, results from multiple etiologies, such as Gram-positive and Gram-negative bacteria, or fungal infections. The progression to severe sepsis is often mediated by a common biological pathway, with variations unique to specific infectious agents. Therefore, genetic variations within inflammatory mediators involved in the sepsis pathway have been hypothesized to play a role [4,6,7]. However, in addition to genetic factors, host characteristics and pathogen load also influence the phenotype. The relative contribution of host genetic factors in complex traits like severe sepsis would be modest.

Focusing only on the contribution of genetic variation to disease, the exact pattern of genetic variation influencing complex traits is still unclear and several theories have been proposed [8]. One model, termed the common disease-rare variant model, suggests that phenotypic variation in complex traits is due to numerous rare genetic variants at multiple loci, with each variant single handedly causing disease. Although the frequency of each rare variant is low, populations may have several such variants. An example of the common disease-rare variant model includes mutations in the *BRCA1* and *BRCA2* genes, which have been implicated in the susceptibility to breast and ovarian cancer [9]. The frequency

---

ARDS = adult respiratory distress syndrome; IL = interleukin; LD = linkage disequilibrium; SNP = single nucleotide polymorphism; TNF = tumor necrosis factor.

of each of the four mutations within these genes is less than 5%, but more than 80% of subjects with these mutations develop breast cancer.

In contrast, the common disease-common variant model suggests that common variants underlie complex traits. Such variants may be maintained through generations due to some form of balancing selection, where the same genetic variant may be protective for certain diseases and harmful in others. This model may be particularly important in critical illnesses, which often occur due to differences in expression of inflammatory mediators. A robust pro-inflammatory response with TNF and IL-6 release may increase the risk of complications, such as severe sepsis or adult respiratory distress syndrome (ARDS), yet that same response may be critical for an adequate host response to infection. Therefore, genetic variants associated with a pro-inflammatory response could be protective and detrimental under different conditions. An example of balancing selection is the guanine to adenine transition at the +250 site within the lymphotoxin alpha gene, which is associated with increased TNF expression and also with higher risk of severe sepsis but lower risk of prolonged mechanical ventilation after coronary artery bypass graft surgery [4,10]. Complex traits may also occur due to a combination of rare and common variants. Finally, interactions may occur among genes (epistasis) and with environmental factors (gene-environment interactions) to influence the phenotype (Table 1).

Nomenclature: polymorphism, mutation, and SNPs

Nucleotides are the building blocks of DNA and contain one of the following four bases: adenine (A), thymine (T), guanine (G), or cytosine (C). A polymorphism is a common variation in the sequence of DNA among individuals (>1% of the population). Substitution of one of the four base pairs by another base pair is called single nucleotide polymorphism or SNP; for example, a SNP may change the DNA sequence from A*A*TCG to A*G*TCG. Mutations are also heritable changes in the DNA sequence, but have a frequency of <1%. Polymorphisms occur at a rate higher than can be explained by new mutations, suggesting that they may confer some survival advantage. Variable number of tandem repeats is another type of polymorphism, where a particular repetitive sequence is present in different numbers in different individuals. An example of a tandem repeat is the tetranucleotide $(CATT)_n$ repeat within the promoter region of the macrophage inhibitory factor gene, where subjects can have from five to eight repeats [11].

All SNPs are not the same: choosing candidate SNPs

The genes in the human genome account for a very small fraction of the total DNA, and more than 90% of the sequences between genes do not encode any particular product [12]. Variations within DNA are ubiquitous. SNPs

occur every 1,000 base pairs in the human genome, and most SNPs do not lead to a change in protein structure or secretion. When SNPs lead to changes in amino acids they are called non-synonymous or missense SNPs. Some of the non-synonymous SNPs in the coding region may affect protein structure and lead to alterations in phenotype. An example is the G to A coding polymorphism at the +1691 site in the factor V gene of the coagulation cascade [13]. This polymorphism leads to the substitution of an arginine by glutamine at amino acid position 506, which is one of the cleavage sites for activated protein C. Factor V inactivation is delayed because the cleavage site is not present, and leads to a hypercoagulable state.

Similar to non-synonymous SNPs, those in the promoter region are also important. Although they do not affect protein structure, they may affect binding of transcription factors and alter expression of the protein in response to an appropriate stimulus. For example, an insertion/deletion polymorphism, termed 4G/5G, is found 675 base pairs upstream of the trans-criptional initiation site in the plasminogen activator inhibitor-1 gene [14,15]. Although both alleles bind a transcriptional activator, the 5G allele reduces transcription by binding a repressor protein, and is associated with lower circulating plasminogen activator inhibitor-1 concentrations [16,17].

However, most SNPs have no effect on the phenotype because they are either in non-coding regions or they are synonymous SNPs, which are variants that code for the same amino acid. Of the SNPs in the non-coding region, those in the 5′ or 3′ untranslated region are probably more important than those in introns, which are non-coding sequences of DNA that are initially copied into the RNA but cut out of the final RNA transcript. They may play critical roles in the post-transcriptional regulation of gene expression, including modulation of the transport of mRNAs out of the nucleus and stabilization of protein [18]. It is important to understand these distinctions when choosing SNPs during candidate gene analysis for causal variants. In general, promoter region and non-synonymous SNPs are likely to be more important than those in the non-coding region.

SNPs are not necessarily causal: role of genetic markers, linkage disequilibrium, and haplotype blocks

Knowing the causal SNP may often be difficult. Often, we may discover a SNP 'associated' with a specific phenotype, but it is simply a 'marker' rather than the causal variant. This marker is co-inherited along with the causal variant because it tends to be on the same piece of DNA. This phenomenon where two genetic variants are inherited together through generations is called linkage disequilibrium (LD). Several methods can be used to measure LD. Two most commonly used are Lewontin *D*′ and R^2. Both are measures of correlation and expressed on a scale of 0 to 1, with a higher number indicating greater LD or that these SNPs are more

Table 1

Nomenclature and explanation of some terms in genetic epidemiology

Term	Explanation
Allele	Alternative form of gene
Epistasis	The interaction of two or more genes to influence a single phenotype
Exon	Part of the DNA that contains the code for producing the gene's protein
Haplotypes	Closely related variants that are inherited as a unit
Haplotype phase	Determination of which of the two chromosomes a series of markers or SNPs are present on
Intron	A non-coding sequence of DNA that is initially copied into the RNA, but is cut out of the final RNA transcript
Linkage disequilibrium	Non-random association of alleles at linked loci
Mutation	Change in DNA sequence with frequency ≤1%
Non-synonymous or missense SNPs	SNPs that are associated with change in amino acid in the transcribed protein
Polymorphism	Variation in DNA with frequency >1%
Recombination	Maternal and paternal chromosomes pair up and exchange segments of DNA in germ cells
Single nucleotide polymorphism (SNP)	Common, but minute variations in the DNA sequence
Synonymous SNPs	SNPs that are not associated with change in amino acid in the transcribed protein
Tag SNPs	There is a degree of redundancy when SNPs are in linkage disequilibrium with each other within a haplotype. A single or a group of SNPs can be used to identify the entire haplotype block and is called tag SNP
Transmission disequilibrium test	A family based study design using parents-affected child trios. This design tests for an association between a specific allele and disease in the child by testing whether heterozygous parents transmit this allele to affected children more frequently than expected

likely to be inherited together. These measures of LD are statistical measurements in population genetics and do not necessarily imply distance between the two sites. LD maps for SNPs within a single gene are available publicly and provide important insights into choosing marker SNPs for candidate gene analysis.

LD is a powerful tool in genetics. During meiosis, pieces of maternal and paternal DNA are exchanged via recombination. However, markers in LD remain tightly linked and are transmitted through generations as regions of DNA called haplotype blocks. Once an association is determined between a marker and disease, one could focus on the 'block' of DNA to identify the causal polymorphism. These 'blocks' can be identified, or tagged, by one or more polymorphisms on the block. Once a haplotype of interest has been described, additional work can be done to sequence the haplotype and tease out the specific functional polymorphism within the haplotype that appears to cause the phenotype.

Haplotype and haplotype tag SNPs

Haplotyping is a way of describing blocks of DNA with a pattern of alleles. A potential problem in constructing haplotypes from results of genotype alone is that it is often difficult to determine which set of alleles derives from the

paternal chromosome and which set derives from the maternal chromosome. In other words, how are adjacent bases aligned on each chromosome? The specific arrangement of markers on each chromosome within a pair is called haplotype phase. Although phase can be determined by molecular genetic techniques, such methods are expensive. Therefore, statistical software is used to estimate the haplotype frequencies in a population based on genotype data and LD.

Commonly used statistical programs either use iterative likelihood (SAS Genetics, EH Plus) or Bayesian methods (PHASE) to estimate haplotype frequencies in the population. Consider an example of estimation of haplotypes and frequencies of each of these haplotypes in the promoter region of the TNF gene with two SNPs at the -308 and -238 sites (Figure 1). Based on arrangement of these alleles on the maternal and paternal chromosomes, an individual with a GA genotype at both sites could potentially have four different haplotypes, G/G, G/A, A/G, and A/A. Assuming that no LD exists between these sites, the probability of each of these haplotypes is 0.25. However, the estimated probabilities based on LD differ significantly. It is important to emphasize that statistical methods can only estimate probabilities of each haplotype.

Figure 1
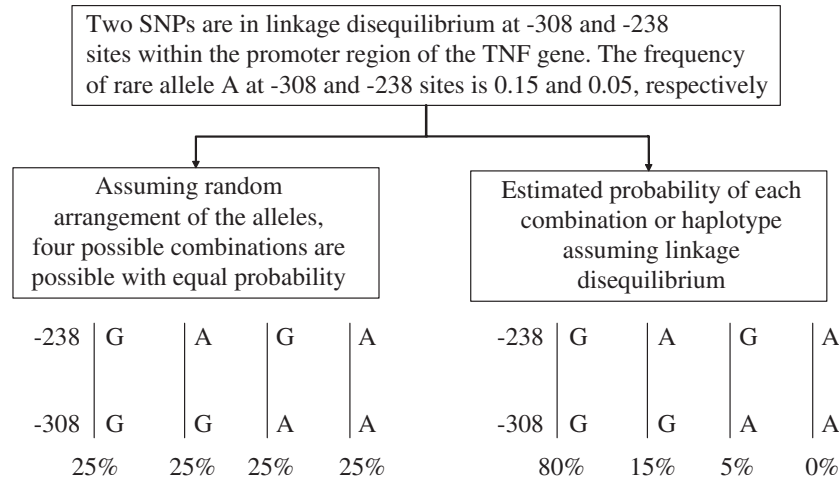
Estimation of haplotype frequencies for two tumor necrosis factor (TNF) single nucleotide polymorphisms (SNPs) at -308 and -238 promoter sites.

The human chromosome is a mosaic of several such haplotype blocks, which are often 11 to 22 kb in size, but may extend longer [19]. Although multiple polymorphisms (SNPs or variable number of tandem repeats) may be present on each haplotype block, only two or three of them are required to identify a particular haplotype. These SNPs are called haplotype tag SNPs, and are often used as genetic markers in gene association studies. Haplotype tag SNPs are an important tool in mapping genetic determinants of disease and, therefore, there is much interest in developing a haplotype map of the entire human genome [20,21].

Study design

Two broad approaches are used to assess the roles of genetic variants in disease: linkage analysis and association studies (Figure 2). Linkage analysis follows meiotic events through families for co-segregation of disease and genetic variants. In contrast to chronic illnesses like diabetes, obtaining an accurate family history about critical illnesses in the past, such as whether a family member developed ARDS after pneumonia, is difficult. Therefore, this approach is less useful in acute illnesses, and has not been used widely in the critically ill. In contrast to linkage analysis, association studies detect association between genetic variants and disease across individuals in large populations. Most association studies are population based, but family based studies using parents-affected child trios (transmission disequilibrium test) can also be conducted. This design tests for an association between a specific allele and disease in the child by testing whether heterozygous parents transmit this allele to affected children more frequently than expected [22].

Gene association studies can be cohort or case-control. Cohort studies are time consuming and expensive to

Figure 2
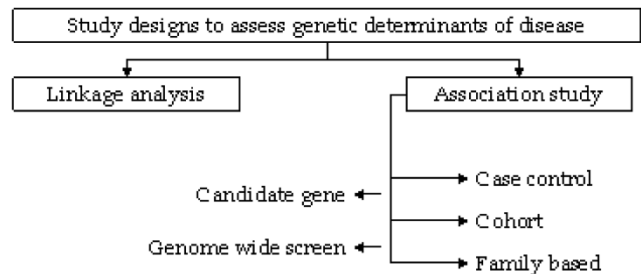
Overview of genetic studies.

conduct, and are impractical for rare diseases, whereas case-control designs can be affected by selection bias or information bias. However, there are study design problems unique to gene-association studies in critical care. A common practice in case control studies is the use of blood bank donors as a control population. For example, consider a case-control design to study genetic variants influencing susceptibility to pneumonia and severe sepsis. The allele frequency in the control population is often driven by subjects who volunteer to participate in the control group. Little information is available on whether individuals in the control group would or would not develop pneumonia when exposed to an adequate pathogen load in the presence of similar non-genetic risk factors for pneumonia susceptibility.

Even assuming that pneumonia does occur uniformly in the controls and cases, it is not known if severe sepsis would then develop among controls. Severe sepsis and other

critical illnesses often occur due to differences in innate immune response. Therefore, while a particular innate immune response like higher TNF production can be protective for pneumonia susceptibility, it might increase risk of severe sepsis. Critical illness occurs in the continuum of a healthy host, who develops infection or trauma, progresses to organ dysfunction or severe sepsis, and death. Taking only the cases at the end of this spectrum, those with established severe sepsis, and comparing them to healthy blood donors could be an entirely spurious process. This association could be confounded by the inciting stimulus that led to severe sepsis.

An inception cohort design is thus a stronger approach. But such studies are time consuming and it is impractical to follow large population based cohorts for long periods, waiting for infections and critical illness to occur. One must, therefore, identify a population at risk, and a single inception cohort may not be able to provide all the answers. One example would be to follow a cohort of elderly subjects for development of pneumonia, while another inception cohort of subjects who present to physicians' offices or emergency rooms with pneumonia could be followed for development of subsequent complications.

Finally, gene-environment interactions are also important to consider in gene association studies. Many interventions in the intensive care unit alter the cytokine cascade, such as strategies to ventilate patients, medications, or surgical techniques. Since differences in expression of proteins involved in the cytokine cascade are hypothesized as candidate genes, interactions between cytokine gene polymorphisms and these interventions would be important.

Candidate gene approach and genome-wide screen

Regardless of the overall study design, one also needs to decide what methodology to use to examine genetic variation. There are two general approaches: genome-wide association studies and candidate gene association studies. Genome-wide association studies are philosophically similar to whole genome linkage analyses, where the investigator does not have an *a priori* idea of the susceptibility locus, but is trying to locate a chromosomal region that is associated with the 'disease' of interest [23]. This approach is hypothesis-generating, and it is technologically intensive and expensive. However, as the cost of genotyping continues to decrease, this methodology becomes more viable. The exact number of SNPs and type of SNPs (all versus non-synonymous SNPs only) to be used for a genome-wide screen is still a matter of debate.

The candidate gene approach examines the role of genetic variation in one or more genes most likely to be involved in the biological pathway. This approach requires an understanding of the biological mechanisms to identify candidate genes and is commonly used because it is technologically non-intensive and relatively inexpensive.

Alternatively, a hybrid approach can be used: a genome-wide screen is used to identify genetic variation spaced throughout the human genome, followed by a candidate gene approach to examine genes within the region of interest.

Phenotype

Accurate definition of phenotype is critical to genetic studies. False positive or false negative results are often due to differences in definitions of phenotypes across studies. Critical illnesses are heterogeneous conditions or syndromes and occur due to a variety of etiologies, each leading to different outcomes. Although clinical definitions of ARDS or severe sepsis are useful diagnostic criteria for clinicians, they may be too expansive for understanding the role of genetic variation. Different sets of genetic markers may underlie susceptibility to ARDS due to infections and trauma [24]. Similarly, genetic variation underlying severe sepsis susceptibility due to different infections may also vary due to interactions between individual organisms and genetic variants.

Statistical issues in gene association studies
Power

Irrespective of study design, it is critical to have sufficient power to detect association. As described previously, the relative risk for critical illness for individual loci would be small, with relative risk ≤2. Sample size estimates for gene association studies are determined by the allele frequency and relative risk of the genetic marker of interest. In general, association studies may be more likely to provide statistical evidence of a disease gene with low relative risks than linkage studies [25]. However, approximately 1,000 cases and 1,000 controls will be required to detect modest relative risks of 1.5 [26]. Larger sample sizes would be necessary for rare alleles (frequency <10%), whereas smaller sample sizes would be required if the relative risks are larger. Numerous statistical tools are available to determine sample sizes required for different levels of significance, for example Quanto [22,27] and Genetic Power Calculator [28,29].

Multiple testing

There is no easy statistical solution to the problem of multiple-testing. If thousands of tests are performed, then there will be many false-positive results. One of the current approaches is to use a false-discovery rate (FDR) statistic to decide what proportion of true positives to false positives is acceptable to the investigator, choose a level of significance based on this proportion, and follow-up on all results that achieve this level of significance [30]. Thus, the first stage of analyses in which multiple-testing is performed is usually considered to be hypothesis-generating, and results of these analyses will contain some false positives. However, follow-up analyses in another population, that is, replication, should differentiate between true-positive and false-positive results.

Increasingly, the use of permutation tests is being advocated to estimate p-values. P-values for genotype or haplotypes are

calculated by random permutations thousands of times. The resulting empirical distribution is used to estimate the p-value for the test-statistic obtained from the actual data. Permutation analyses will account for some of the relatedness among the markers, which are linked if present on the same chromosome. This will remove the dependence of the test-statistic on an underlying distribution. Several statistical packages like R Statistical Computing Environment [31] and SAS Genetics enable the estimation of permutation statistics and FDR.

Replication of genetic studies

The strongest evidence that a particular variant or candidate gene is associated with a trait, and thus may be causal, or in strong LD with a causal variant, is to replicate the result [32]. Replication is defined as doing the analyses in a different population, preferably by different investigators, using different methods to avoid introduction of bias. DeMeo and colleagues [33] recently used linkage analysis to narrow down the candidate genes for chronic obstructive pulmonary disease to chromosome 2q. Using microarray technology on murine and human lung tissue, they identified three genes of interest on chromosome 2q. The associations between these three genes and chronic obstructive pulmonary disease were tested using family based design, and the association with one of the genes, a serine protease inhibitor or SERPINE2, was confirmed in another case-control design using different patient populations from the United States.

Studies have attempted to replicate work in populations of different ethnic origin. For example, a recent report showed that the association between polymorphisms within the selenoprotein S gene with TNF and IL-6 expression in a study of Caucasians was replicated among Mexican families [34]. However, failure to replicate results for a genetic marker in populations of different ethnic origin does not suggest that the results are merely due to type I error. Rather, differences in LD between the genetic marker and the causal variant may lead to different results.

Population admixture

Sub-populations within a population may have a different genetic architecture. Differences in frequency of genetic variants within the population may lead to false positive results. False positive associations between genetic markers and disease can occur due to association of disease with a sub-population, rather than the genetic marker. Self-reported race is used commonly to stratify subjects to avoid ethnic stratification. Population admixture is more common among self-identified African-American subjects compared to those identifying themselves as of Caucasian ethnic origin [35]. Although population admixture does occur in most genetic association studies, the extent to which results would be affected is less clear. Techniques have been developed to detect and correct for population stratification by typing unlinked markers [36-38]. Whether this approach is adequate is controversial [39].

> This article is part of a thematic series on
> *Translational Research*
> edited by John Kellum.
>
> Other articles in the series can be found online at,
> http://ccforum.com/articles/
> theme-series.asp?series=CC_Trans

Conclusion

Genetic association studies will be more valid if study design issues are carefully considered during the planning phase of a study and rigorous statistical methods are used during analysis. There are several challenges to conducting well designed genetic studies in critical care, including recruiting large cohorts to obtain sufficient power, accurately identifying phenotypes, identifying appropriate case and control groups, and choosing a candidate gene or whole genome approach. However, if such considerations are met, one can be cautiously optimistic that genetic association studies may lead to better understanding of biological mechanisms and improve our ability to target therapy in the critically ill.

Competing interests

The authors declare that they have no competing interests.

References

1. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, *et al.*: **Initial sequencing and analysis of the human genome.** *Nature* 2001, **409:**860-921.
2. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, *et al.*: **The sequence of the human genome.** *Science* 2001, **291:**1304-1351.
3. Weiss KM, Terwilliger JD: **How many diseases does it take to map a gene with SNPs?** *Nat Genet* 2000, **26:**151-157.
4. Mira JP, Cariou A, Grall F, Delclaux C, Losser MR, Heshmati F, Cheval C, Monchi M, Teboul JL, Riche F, *et al.*: **Association of TNF2, a TNF-a promoter polymorphism, with septic shock susceptibility and mortality.** *JAMA* 1999, **282:**561-568.
5. Stuber F, Udalova IA, Book M, Drutskaya LN, Kuprash DV, Turetskaya RL, Schade FU, Nedospasov SA: **-308 tumor necrosis factor (TNF) polymorphism is not associated with survival in severe sepsis and is unrelated to lipopolysaccharide inducibility of the human TNF promoter.** *J Inflamm* 1995, **46:** 42-50.
6. Gallagher PM, Lowe G, Fitzgerald T, Bella A, Greene CM, McElvaney NG, O'Neill SJ: **Association of IL-10 polymorphism with severity of illness in community acquired pneumonia.** *Thorax* 2003, **58:**154-156.
7. Waterer GW, Quasney MW, Cantor RM, Wunderink RG: **Septic shock and respiratory failure in community-acquired pneumonia have different TNF polymorphism associations.** *Am J Respir Crit Care Med* 2001, **163:**1599-1604.
8. Zwick ME, Cutler DJ, Chakravarti A: **Patterns of genetic variation in Mendelian and complex traits.** *Annu Rev Genomics Hum Genet* 2000, **1:**387-407.
9. King MC, Marks JH, Mandell JB: **Breast and ovarian cancer risks due to inherited mutations in BRCA1 and BRCA2.** *Science* 2003, **302:**643-646.
10. Yende S, Quasney MW, Tolley E, Zhang Q, Wunderink RG: **Association of tumor necrosis factor gene polymorphisms and prolonged mechanical ventilation after coronary artery bypass surgery.** *Crit Care Med* 2003, **31:**133-140.

11. Donn RP, Shelley E, Ollier WE, Thomson W: **A novel 5′-flanking region polymorphism of macrophage migration inhibitory factor is associated with systemic-onset juvenile idiopathic arthritis.** *Arthritis Rheum* 2001, **44:**1782-1785.

12. Stein LD: **Human genome End of the beginning.** *Nature* 2004, **431:**915-916.

13. Bertina RM, Koeleman BPC, Koster T, Rosendaal FR, Dirven RJ, de Ronde H, van der Velden PA, Reitsma PH: **Mutation in blood coagulation factor V associated with resistance to activated protein C.** *Nature* 1994, **369:**64-67.

14. Dawson SJ, Wiman B, Hamsten A, Green F, Humphries S, Henney AM: **The two allele sequences of a common polymorphism in the promoter of the plasminogen activator inhibitor-1 (PAI-1) gene respond differently to interleukin-1 in HepG2 cells.** *J Biol Chem* 1993, **268:**10739-10745.

15. Eriksson P, Kallin B, 't Hooft FM, Bavenholm P, Hamsten A: **Allele-specific increase in basal transcription of the plasminogen-activator inhibitor 1 gene is associated with myocardial infarction.** *Proc Natl Acad Sci USA* 1995, **92:**1851-1855.

16. Westendorp RG, Hottenga JJ, Slagboom PE: **Variation in plasminogen-activator-inhibitor-1 gene and risk of meningococcal septic shock.** *Lancet* 1999, **354:**561-563.

17. Hermans PW, Hibberd ML, Booy R, Daramola O, Hazelzet JA, de Groot R, Levin M: **4G/5G promoter polymorphism in the plasminogen-activator-inhibitor-1 gene and outcome of meningococcal disease. Meningococcal Research Group.** *Lancet* 1999, **354:**556-560.

18. Mignone F, Gissi C, Liuni S, Pesole G: **Untranslated regions of mRNAs.** *Genome Biol* 2002, **3:**reviews0004.

19. Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, Higgins J, DeFelice M, Lochner A, Faggart M, *et al.*: **The structure of haplotype blocks in the human genome.** *Science* 2002, **296:**2225-2229.

20. The International HapMap Consortium: **A haplotype map of the human genome.** *Nature* 2005, **437:**1299-1320.

21. **International HapMap Project** [www.hapmap.org]

22. Gauderman WJ: **Candidate gene association analysis for a quantitative trait, using parent-offspring trios.** *Genet Epidemiol* 2003, **25:**327-338.

23. Hirschhorn JN, Daly MJ: **Genome-wide association studies for common diseases and complex traits.** *Nat Rev Genet* 2005, **6:**95-108.

24. Lin Z, Pearson C, Chinchilli V, Pietschmann SM, Luo J, Pison U, Floros J: **Polymorphisms of human SP-A, SP-B, and SP-D genes: association of SP-B Thr131Ile with ARDS.** *Clin Genet* 2000, **58:**181-191.

25. Risch NJ: **Searching for genetic determinants in the new millennium.** *Nature* 2000, **405:**847-856.

26. Reich D, Patterson N, Jager PLD, McDonald GJ, Waliszewska A, Tandon A, Lincoln RR, DeLoa C, Fruhan SA, Cabre P, *et al.*: **A whole-genome admixture scan finds a candidate locus for multiple sclerosis susceptibility.** *Nat Genet* 2005, **37:**1113-1118.

27. **Quanto** [http://hydra.usc.edu/GxE/]

28. Purcell S, Cherny SS, Sham PC: **Genetic Power Calculator: design of linkage and association genetic mapping studies of complex traits.** *Bioinformatics* 2003, **19:**149-150.

29. **Genetic Power Calculator** [http://pngu.mgh.harvard.edu/~purcell/gpc/]

30. Hochberg Y, Benjamini Y: **More powerful procedures for multiple significance testing.** *Stat Med* 1990, **9:**811-818.

31. **R Statistical Computing Environment** [http://www.r-project.org/]

32. de Bakker PIW, Yelensky R, Pe'er I, Gabriel SB, Daly MJ, Altshuler D: **Efficiency and power in genetic association studies.** *Nat Genet* 2005, **37:**1217-1223.

33. DeMeo DL, Mariani TJ, Lange C, Srisuma S, Litonjua AA, Celedon JC, Lake SL, Reilly JJ, Chapman HA, Mecham BH, *et al.*: **The SERPINE2 gene is associated with chronic obstructive pulmonary disease.** *Am J Hum Genet* 2006, **78:**253-264.

34. Curran JE, Jowett JBM, Elliott KS, Gao Y, Gluschenko K, Wang J, Azim DMA, Cai G, Mahaney MC, Comuzzie AG, *et al.*: **Genetic variation in selenoprotein S influences inflammatory response.** *Nat Genet* 2005, **37:**1234-1241.

35. Sinha M, Larkin EK, Elston RC, Redline S: **Self-reported race and genetic admixture.** *New Engl J Med* 2006, **354:**421-422.

36. Pritchard JK, Stephens M, Rosenberg NA, Donnelly P: **Association mapping in structured populations.** *Am J Hum Genet* 2000, **67:**170-181.

37. Ardlie KG, Lunetta KL, Seielstad M: **Testing for population subdivision and association in four case-control studies.** *Am J Hum Genet* 2002, **71:**304-311.

38. Freedman ML, Reich D, Penney KL, McDonald GJ, Mignault AA, Patterson N, Gabriel SB, Topol EJ, Smoller JW, Pato CN, *et al.*: **Assessing the impact of population stratification on genetic association studies.** *Nat Genet* 2004, **36:**388-393.

39. Cardon LR, Palmer LJ: **Population stratification and spurious allelic association.** *Lancet* 2003, **361:**598-604.