Genetic Susceptibility to Complex Traits: Moving Towards Informed Analysis of Whole-Genome Screens

Michael R. Green¹,², Emily Camilleri¹, Maher K. Gandhi²,³, Lyn R. Griffiths¹,²*

¹Genomics Research Centre, Griffith Institute of Health & Medical Research, Griffith University, Southport, Queensland, Australia; ²Griffith Medical Research College, a joint initiative of Griffith University and the Queensland Institute of Medical Research, Herston, Queensland, Australia; ³Clinical Immunohaematology Laboratory, Queensland Institute of Medical Research, Herston, Queensland, Australia

Abstract

Susceptibility to complex traits, by definition, involves aetiological polymorphisms at multiple genetic loci combined with variable contributions by environmental factors. However, the approaches taken to identifying genetic loci implicated in susceptibility to complex traits frequently overlooks the compounding contribution of multiple loci in favour of highlighting a single gene solely responsible for predisposition. It is only in a small minority of cases that this has resulted in clear disease heritability associated with polymorphisms in a single gene. More often, this approach has led to an accumulation of single-gene associations with minor contributions to disease susceptibility. As the genomic era advances and genome-wide screens become higher in resolution and throughput, the need for simultaneous consideration of multiple loci is becoming more important. With special reference to non-Hodgkin’s lymphoma (NHL), this chapter will overview the current progress made in elucidating genetic polymorphisms associated with disease susceptibility. We also present novel data from a high-resolution single nucleotide polymorphism (SNP) microarray screen for susceptibility loci that are involved in NHL. Using an ‘informed approach’, the findings are highlighted within the context of cellular pathways, and provide insight and new ideas for methods of analysis for genome-wide screens for susceptibility.

* Corresponding author: l.griffiths@griffith.edu.au
Introduction

Disease susceptibility is conferred by both environment and genetic risk factors. In many cases, the genetic contribution to disease susceptibility is minor in comparison to the environmental risk factors. Furthermore, the genetic contribution to disease generally consists of a low number of alleles that confer high-risk in addition to a high number of alleles that confer low-risk (Figure 1). The aim of many genetic susceptibility studies is to elucidate high-risk alleles that show an outstanding association with a disease phenotype. However, within complex diseases, focusing on determining high-risk alleles and thereby overlooking the alleles with minor contributions to disease risk may overlook the majority of the genetic components of the disease. With the advent of high-resolution single nucleotide polymorphism (SNP) microarrays, that are capable of simultaneously genotyping hundreds of thousands of polymorphic markers, the ability to scan the genome for minor but compounding contributors to disease susceptibility has become achievable. Here, we will provide an overview of the evolution of approaches to determining genetic susceptibility, their application and successes in non-Hodgkin’s Lymphoma (NHL), and our impressions of the future of genome-wide susceptibility scans.

Figure 1: Illustration of the genetic component of disease, showing that the genetic component may be a small part of the overall aetiology, and high-risk alleles are a small part of the genetic basis for susceptibility.
Evolution of Genome-Wide Approaches for Determining Susceptibility

At a time in which knowledge of polymorphisms within the human genome is evolving at an ever increasing rate, genetic studies for determining susceptibility to disease have also been forced to adapt. Family-based linkage studies have been successful in determining genetic loci associated with diseases such as Huntington’s, Alzheimer’s and familial breast cancer. These studies are based upon the co-inheritance of genetic markers (polymorphisms) and phenotypes through families over several generations. However, this approach has not been as successful in determining susceptibility to non-mendelian complex traits due to the inability to detect inheritance patterns within family pedigrees. This has led to the growing popularity of association studies, wherein the contribution of genetic loci to disease susceptibility is determined by comparing allele frequencies of genetic polymorphisms between affected cases and unaffected controls, resulting in higher power and genetic resolution than that obtained from linkage studies [Risch and Merikangas, 1996]. Controls for these studies can either be family-based, or be made up of a cohort of individuals derived from a population. Family-based studies use unaffected relatives (for example siblings or cousins) as controls, or can alternatively employ pseudosiblings formed from the set of genotypes that are not passed down from the parents to an affected offspring [Thomson, 1995]. These studies have the advantage of overcoming population stratification; a phenomenon where ethnic differences in both allele frequency and disease risk can cause an accumulation of type I errors (i.e. false associations). However, the difficulty in recruiting cases and their relatives is a major limiting factor for these studies.

Association studies employing cohorts of a population as case and control subjects are now the most popular format for genome-wide investigations. These studies come in two forms: (i) hypothesis-based investigations, or (ii) hypothesis-
generating investigations. The hypothesis-based investigations are directed
towards identifying associations between a disease and polymorphisms within
candidate genes for which there is usually functional evidence for a role in the
disease in question. This approach has been utilised to determine many genetic
risk factors for NHL (see below), but is limited towards studies of genes implicated
in well-defined etiological mechanisms of the disease. In order to overcome this
limitation, and allow investigation of diseases with varied or poorly-defined
aetiologies, hypothesis-generating approaches can be employed. These studies use
genome-wide screens in order to find differences between case and control
populations, and then generate hypotheses based upon the findings. A platform
allowing high-throughput whole-genome screens for these hypothesis-generating
studies is SNP microarrays.

SNPs are both abundant and spread evenly across the genome, making them
excellent markers for high-resolution scans for genetic susceptibility loci.
Oligonucleotide microarray-based interrogation of these markers allows high-
throughput analysis with ever increasing coverage and decreasing cost. New
generation arrays allow for much higher resolution than that attainable with other
polymorphic markers such as microsatellites, and therefore allow the elucidation
of more narrowly defined loci associated with disease susceptibility. The
interrogation of large numbers of samples, which are needed in order to provide
significant power to associations, can also be further facilitated through the use of
DNA pooling. This method pools genomic DNA from cases and controls and
extrapolates allele frequencies for each SNP using the relative probe fluorescence
intensities from the SNP microarrays. By performing multiple replicates of the
prepared pools, inferred allele frequencies can be remarkably accurate [Bansal et
al., 2002; Docherty et al., 2007], and allow a first-pass screen of the genome to be
performed at low cost. Once allele frequencies have been derived for SNPs,
differences in allele representation between case and control cohorts can be used to provide associations between genotypic variants and disease susceptibility. These polymorphisms may be either directly associated with disease susceptibility due to a functional role in the expression or coding of a gene, or indirectly associated with susceptibility due to linkage disequilibrium with a directly associated allele. Alleles conferring a large relative risk can lead to large differences in allele frequency between case and control cohorts, whereas alleles conferring a small relative risk demonstrate more modest differences between cohorts [Carlson et al., 2004]. Unfortunately, it is the objective of most genome-wide association studies to direct their attention towards alleles that are most differentially represented between case and control cohorts. This effectively overlooks the contribution of low-risk alleles that make up the majority of the genetic component of disease susceptibility.

Genetic Susceptibility to Non-Hodgkin’s Lymphoma

A genetic basis for susceptibility to NHL has become evident due to the strong aggregation of lymphomas in otherwise healthy families [Wang et al., 2007]. The most pronounced risk factor for developing NHL is strong immunodeficiency due to disorders such as Ataxia Telangiectasia (AT), Bloom’s syndrome or Nijmegen breakage syndrome (NBS). In these diseases, susceptibility to NHL is conferred by defects within pathways that respond to DNA damage [Tran et al., 2008]. This observation has lead to an accumulation of hypothesis-based studies that have found numerous associations between NHL and polymorphisms in the DNA repair system. Furthermore, susceptibility has also been linked to loci associated with folate and one-carbon metabolism, as well as immune-responses.
Variants within DNA Repair Genes

Genes involved in all facets of DNA-repair have had variable associations with disease susceptibility. This includes genes involved in lymphocyte developmental processes such as the recombination of immunoglobulin and T-cell receptor variable, diversity and joining regions, and the associated non-homologous end-joining repair pathway (RAG1 [Hill et al., 2006], LIG4 [Hill et al., 2006] and H2AFX [Novik et al., 2007]). However, genes involved in other mechanisms of DNA repair that are not necessary for lymphocyte development have also been associated with susceptibility, such as homologous recombination (BRCA1 [Shen et al., 2006], BRCA2 [Hill et al., 2006], WRN [Hill et al., 2006; Shen et al., 2006] and XRCC3 [Ekstrom-Smedby et al., 2006a; Shen et al., 2006]), nucleotide excision repair (ERCC2 [Ekstrom-Smedby et al., 2006a], ERCC5 [Shen et al., 2006]), base excision repair (XRCC1 [Ekstrom-Smedby et al., 2006a; Shen et al., 2007]), direct repair (MGMT [Hill et al., 2006; Shen et al., 2007]), and mismatch repair (MSH2 [Paz-y-mino et al., 2003; Hishida et al., 2003a]).

Variants within Folate Related Genes

The folate and one-carbon metabolism pathway is also linked to DNA-repair through the role of 5,10-methylenetetrahydrofolate reductase (MTHFR) in catalyzing the dissociation of thymine-thymine dimers. However, more important for lymphoma risk is the role if this pathway in regulating DNA methylation by affecting the amount of available methionine, and in controlling the balance of deoxynucleotide synthesis, particularly uracil which can lead DSB formation when misincorporated into DNA [Skibola et al., 2004]. Several genes within this pathway have been associated with lymphoma risk, including MTHFR [Matsuo et al., 2001; Matsuo et al., 2004; Niclot et al., 2006; Rudd et al., 2004], methionine synthase [Matsuo et al., 2001; Lincz et al., 2003; Niclot et al., 2006], serine hydroxymethyltransferase [Hishida et al., 2003b], thymidylate synthase [Hishida
et al., 2003b; Skibola et al., 2004; Niclot et al., 2006], and folylpolyglutamate synthase [Lim et al., 2007]. However, these associations are heavily confounded by variation in the dietary intake of folate and methionine and multiple co-factors for their metabolism [Lim et al., 2007].

**Immune Related Gene Variants**

Although the association between immune-gene polymorphisms and lymphoma in healthy individuals is not as pronounced as the association with immunodeficiency, there have been several findings linking these polymorphisms with both lymphoma susceptibility and disease outcome. For example, in Hodgkin’s lymphoma there are strong links between polymorphisms with the human leukocytes antigen (HLA) class I and class II regions due to their role in mediating responses to the causative pathogen, Epstein-Barr virus (EBV). Although EBV does not play as prominent of a role in NHL aetiology, these same polymorphisms have also been associated with NHL susceptibility [Houlston et al., 2003; Machulla et al., 2001]. A polymorphism within the TNFα gene, localized to the HLA class III gene cluster, has also been associated with lymphoma risk [Rothman et al., 2005], as have the genes for interleukin-10 (IL-10) [Guzowski et al., 2005; Rothman et al., 2005] and the alpha sub-unit of the IL-10 receptor (IL-10RA) [Nieters et al., 2006] which reduce TNFα expression. Interestingly, a recent study by Wang et al. (2007b) has provided evidence that lymphoma predisposition associated with autoimmune disease is restricted to those individuals carrying a variant allele of either of the implicated TNFα or IL-10 polymorphisms. Furthermore, TNFα and IL-10 gene polymorphisms have also been linked with patient outcome, demonstrating that their role is not restricted to disease genesis [Warzocha et al., 1998; Lech-Maranda et al., 2004].
Inconsistencies in Previous Associations

Despite the accumulation of positive associations with NHL, many of these investigations were hypothesis-based and their replication is disappointingly low (Table 1). This is contributed by factors associated with study design, including low statistical power, multiple hypothesis testing and population sub-structure; but it is also a factor of gene-environment and gene-gene interactions. For example, as well as the gene-environment interactions previously highlighted between HLA polymorphisms and EBV infection, the association between XRCC3 polymorphisms and the risk of developing Follicular Lymphoma are exacerbated in individuals that smoke [Ekstrom-Smedby et al., 2006]. Gene-gene interactions have also been shown in NHL between the DNA repair genes RAG1 and LIG4 [Hill et al., 2006], as well as the TYMS, MTR and MTHFR genes that function in folate metabolism; associations with two polymorphisms within the TYMS only gained significance when analysed in tandem with polymorphisms in the MTR and MTHFR genes [Niclot et al., 2006].

Table 1: Examples of contradictory findings regarding genetic susceptibility in NHL.

<table>
<thead>
<tr>
<th>Gene</th>
<th>Positive Associations</th>
<th>Negative Associations</th>
</tr>
</thead>
<tbody>
<tr>
<td>RAG1</td>
<td>Hill et al., 2006</td>
<td>Shen et al., 2007</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Scott et al., 2007</td>
</tr>
<tr>
<td>BRCA2</td>
<td>Hill et al., 2006</td>
<td>Scott et al., 2007</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>XRCC1</td>
<td>Hill et al., 2006</td>
<td>Ekstrom-Smedby et al., 2006</td>
</tr>
<tr>
<td></td>
<td>Shen et al., 2006</td>
<td></td>
</tr>
<tr>
<td>MSH2</td>
<td>Paz-y-mino et al., 2003</td>
<td>Hishida et al., 2003a</td>
</tr>
<tr>
<td>TYMS</td>
<td>Skibola et al., 2004</td>
<td>Lim et al., 2007</td>
</tr>
<tr>
<td></td>
<td>Niclot et al., 2006</td>
<td></td>
</tr>
<tr>
<td>SHMT1</td>
<td>Hishida et al., 2003b</td>
<td>Niclot et al., 2006</td>
</tr>
<tr>
<td>MTHFR</td>
<td>Skibola et al., 2004</td>
<td>Lim et al., 2007</td>
</tr>
<tr>
<td></td>
<td>Matsuo et al., 2001</td>
<td>Gemmati et al., 2004</td>
</tr>
<tr>
<td></td>
<td>Niclot et al., 2006</td>
<td>Lincz et al., 2003</td>
</tr>
</tbody>
</table>
Informed Analysis of Genome-Wide SNP Microarray Data

The large number of polymorphisms associated with a moderately increased risk of developing NHL, as well as the large amount of inconsistency among independent studies, should demonstrate that complex disorders such as NHL are the result of susceptibility alleles within multiple genes that each have a minor contribution to disease onset. That is, these alleles may be neither necessary nor sufficient to individually cause disease, but may act in concert with other alleles to cause disease when a threshold of susceptibility is reached [Johnson and Todd, 2000]. Therefore, modern genetic mapping of complex diseases should allow for simultaneous consideration of multiple unlinked loci [Cox et al., 1999].

The consideration of multiple loci in microarray data has been addressed by the development of Gene Set Enrichment Analysis (GSEA) platforms which allow the interpretation of aberrations within the context of cell signalling pathways or gene ontology [Subramanian et al., 2005]. However, this approach thus far has been restricted to gene expression microarray analysis. Therefore, the wealth of information provided by genome-wide SNP microarray studies has been only relatively crudely interpreted in terms of deriving genetic aetiology. To address this issue and look at a better way of defining genetic contributors to disease, here we describe a simple method for GSEA of SNP microarray data. In order to illustrate the potential of this approach to analyse whole-genome screens for genetic susceptibility, we have used an example data-set from Affymetrix 250K Sty SNP microarrays performed on pooled genomic DNA of NHL cases (n=100) and controls (n=100). These data will allow us to provide an example of how whole-genome association study results can be interpreted within the context of cellular pathways in order to elucidate genes with a compounding contribution to disease susceptibility.
Methodology for Informed Analysis of Genome-Wide Susceptibility Screens

This method employs the calculation relative allele signal (RAS) values from probe intensity data of cases and controls, followed by the application of thresholds to find informative SNPs, and the use of a web-based GSEA platform to derive meaningful cellular-pathway based associations. Probe intensity data are first cleaned by removing SNPs with failed calls in any of the first 3 quartets in order to eliminate outliers. The RAS values and delta-RAS values (ΔRAS), representing the difference in allele frequency between the case and control pools, are then calculated for every SNP. ΔRAS values fit a normal distribution (Figure 2), allowing simple thresholds to be applied to the tails in order to define informative SNPs. The selection of informative SNPs is therefore based upon the relative distribution of their ΔRAS values compared to other markers, rather than statistical stringencies. This allows the selection of low-risk as well as high-risk alleles in order to provide a platform for comprehensive analysis of compounding contributors to disease susceptibility.
Figure 2: Delta values generated from the differences in case and control allele frequencies showing a normal distribution to which simple thresholds can be applied.

Annotation data accompanying the Affymetrix SNP arrays are then used in order to determine genes associated with informative SNPs. It is this level of the analysis at which a threshold is applied in order to control for type I errors. The use of Affymetrix SNP microarrays in this method allows for the presence of multiple polymorphisms associated with most genes. Utilising only those genes with multiple low-stringency associations allows the elimination of numerous false-positive associations. Genes of interest (GOI) are therefore defined as those with multiple ≥3) informative SNPs showing association with the trait. GOIs elucidated in our data included genes with well-defined roles in NHL pathogenesis, such as BCL2 [Hanada et al., 1993; Kramer et al., 1998], FOXP1 [Barrans et al., 2005; Brown et al., 2005; Brown et al., 2007] and SWAP70 [Shipp et al., 2002]. There were also a host of cell-cycle regulatory genes (eg. CCND1,
CCND2, CDK6 and CDK8), genes involved in cytokine signalling (eg. IL6, IL12B, CXCR4, and IL10RA), DNA repair genes (eg. MGMT, ERCC4, MSH2, RAG2 and XRCC4), growth factors and their receptors (eg. FGF1, FGFR2, VEGFC, and PDGFC), and well as genes involved in signal transduction from growth factor and interleukin receptors (eg. GRB2, FYN and JAK2).

In order to make sense of these long lists of GOIs, they can be organised by cell signalling pathways using a freely available web-based GSEA platforms such as WebGestalt [Zhang et al., 2005]. These utilities group candidate genes based upon pre-defined associations with KEGG pathways, BioCarta pathways or gene ontology. GOI resident in the same cellular pathways may have a compounding contribution to disease susceptibility and therefore should be considered in unison in association studies. However, there may also be interaction between pathways due to pathway crosstalk or pathways containing common constituents. These pathway associations can therefore be represented graphically using nodes consisting of cellular pathways, and straight edges linking nodes that are mutually dependent due to crosstalk or common constituents. Pathways that are central to disease susceptibility also fall centrally within graphical representations of the pathway data. For example, using this method, we associated several cell signalling pathways with susceptibility to NHL, each containing between 2 and 27 mutually inclusive constituents. Several of these pathways have important roles in the cellular physiology of B-cells, the cell-of-origin for the subtypes of NHL investigated here.

Results of Pilot Study into NHL Susceptibility

Visualisation of GSEA results by linking pathways by common components elucidated three central pathways involved in cytokine signalling and the transduction of signals through the MAPK and JAK-STAT pathways. The
transduction of signals from cytokine receptors is critical for the normal development of lymphocytes. But these normal cellular processes have also been described to have a pathogenic role in NHL. The expression of several cytokines has been linked with NHL pathogenesis [Fayad et al., 2001; Hulkkonen et al., 2000; Jones et al., 2002], including IL6 which was highlighted as a candidate in this investigation and contains polymorphisms within its 5’ flanking region that alter its expression [Hulkkonen et al., 2000]. Signals from cytokine receptor are transduced through the MAPK and JAK-STAT pathways resulting in the activation of transcription factors (Figure 4). This resultant gene transcription can result in the initiation of processes such as proliferation or inhibition of processes such as apoptosis [Imada and Leonard, 2000; Treisman, 1996].

Figure 4: An example of signal transduction from a cytokine receptor, in this case the IL6 receptor, through the JAK-STAT (left) and MAPK (right) signalling pathways.
The interaction between 50 GOIs within the three pathways presents a situation where aberrations at different levels of the signalling cascade can have a compounding effect on NHL susceptibility. Therefore, the combinations of low-risk polymorphisms that effect the function or expression of cytokines, cytokine receptors, scaffolding proteins, signal transduction molecules, transcription factors and inducers or inhibitors thereof, could result in an accumulated high-risk of NHL. The statistical analysis of these multi-locus models is beyond the scope of this chapter. However, it should be noted that the formation of interaction networks within and between the central pathways resembles that of Bayesian networks [Rodin and Boerwinkle, 2005]. Using Bayesian model averaging, it is therefore possible to select a model or set of models that explain the complex relationship between genetic risk factors [Fridley, 2008]. These analyses can be performed using the R library RMA (http://cran.r-project.org/src/contrib/descriptions/BMA.html) and are reviewed in Hoeting et al. (1999).

Conclusion

Studies to derive genetic features associated with susceptibility to NHL have generated a plethora of associations accompanied by numerous inconsistencies. These studies have employed strict statistical stringencies to their investigations in the hope of deriving a single polymorphism or small number of polymorphisms associated with a high risk of NHL. However, the cryptic inheritance patterns of this disease highlight the complexity of its genetic basis, lending proof to the hypothesis that susceptibility is conferred by a large number of interacting and compounding low-risk alleles. In order to address this and provide a means of elucidating interacting low-risk alleles, we have described a method of informed gene pathway analysis for genome-wide susceptibility screens. Using this approach within a pilot study we have implicated 50 genes overlapping three cell signalling pathways that have a plausible role in NHL disease etiology. This provides support
for the use of our method of informed analysis for the elucidation of compounding contributors to susceptibility of complex traits such as NHL.

References:


