

Professor Jotun Hein and Professor William Cookson  
**Case for Support BBSRC Grant Application September 2005**

“Integrative Analysis of the Genetic Factors behind Asthma and Atopic Dermatitis”

**Part I: Previous Research Track Record**

**Principal Investigator: Professor Jotun Hein**

I moved to the Department of Statistics at Oxford University in September 2001, coming from Aarhus University, where I was director of BiRC (Bioinformatics Research Center). Since I have come to Oxford I have worked on a series of issues. Most of my work is focused on developing methodologies that can analyze data arising in molecular genetics and genomics with a comparative/evolutionary focus. The work relevant to this proposal is especially Molecular Population Genetics and Comparative Genomics.

In *population genetics*, especially with Carsten Wiuf, Thomas Mailund and Mikkel Schierup from Aarhus University we work on the coalescent based methods of association mapping. We (Hein, Schierup and Wiuf) have published a 300 page book on molecular population genetics titled “Gene Genealogies, Sequence Variation and Evolution” Oxford University Press, and are presently developing a tutorial in association mapping that we hope to publish as a booklet in 2006 and are also involved in a very large EU collaboration (Holland, Denmark, Iceland and UK) to find susceptibility genes for breast and prostate cancer.

In *comparative genomics*, the most fundamental investigation is to find genes in a pair of aligned genomes. However, this can immediately be extended to many sequences, unaligned sequences, allowing for alternative splicing, RNA gene finding and a stronger focus on specifically motivated biological problems. Stephen McCauley (Dphil student) works on extending these models to annotating large sets of viral genomes and Ben Holtom (Dphil student) is working on regulatory signals in collaboration with Harwell (John Hancock) in mouse.

Bioinformatics is a field strongly in demand and we have contributed to the development of this field. Other research areas we are actively involved in include statistical alignment, RNA structure and protein coding constraints, haplotype block definition via reconstructed evolutionary histories and pathogen analysis. Besides being an active research group, we do much to increase awareness of this field in terms of public courses, seminars and organize lecture series. Additionally, we teach a part time MSc in Bioinformatics and thus educate researcher in this field. Presently, I have 5 D.Phil. students (2 co-supervised with MRC Harwell and Wellcome Trust Centre for Human Genetics) and my group continuously supervises a series of student projects. 2-3 postdocs are working on population related issues, including association mapping. 2-3 postdocs work on models of molecular evolution relevant for comparative genomics, such as alignment and signal finding. 1 student is already co-supervised by **WC** and **JT** (Lizhong Hao) and will be joined by another co-supervised student in the autumn 2006 (Joanna Davies).

**Co-Investigator: Professor WOC Cookson**

I trained in medicine at the University of Western Australia and subsequently qualified as a specialist in internal and respiratory medicine. Whilst I was in Australia I developed an interest in epidemiological research, and an investigation of asbestosis in the miners of Wittenoom Gorge was the basis of my MD thesis. At the same time I carried out investigations of the epidemiology of asthma. On my return to the United Kingdom in 1984 Julian Hopkin and I set up a programme of research into the genetics of asthma. This research was the basis of my DPhil degree in Oxford. Over the past 18 years my group and I have been fortunate to have found many of the established genetic effects on asthma, through a combination of positional cloning and candidate gene studies. In 2003 we published the positional cloning of two previously unknown genes influencing the disease.

These large studies required 8 years of work and the participation of over 30 scientists in an international partnership. We have also shown that the combination of eczema and asthma is a different disease to asthma alone, and have also made significant advances in finding the genes causing infantile eczema. In the past five years my group have had eight publications in journals with impact factors approximating 30, and two other publications with individual impact factors greater than 30. We have generated three patents in this time to add to the five we have previously registered. Dr Miriam Moffatt and I have founded a spin-out company with the University of Oxford in order to develop these findings into effective therapeutics. The Wellcome Trust holds equity in the company.

Currently I am a Professor of Human Genetics at Imperial College in London and jointly with Miriam Moffatt, who is Reader in Human Genetics at Imperial College, I supervise a group that contains four post-doctoral scientists, four DPhil students, and a Clinician Research Fellow. The group have recently moved from the Wellcome Trust Centre for Human Genetics in Oxford to the National Heart and Lung Institute (NHLI). The NHLI is part of Imperial College, and is the premier institution in Europe for the study of respiratory and allergic disease.

The findings from my group are changing the understanding of the aetiology of asthma and AD, and have led to some exciting new possibilities for the prevention and treatment of these diseases. As a consequence of this work I am an acknowledged international leader in asthma and AD research and in the study of complex genetic diseases.

**Co-investigator: Dr Chris Holmes**

I moved to Oxford to become a lecturer in Statistical Genetics in February 2004, having previously spent 3 years as a lecturer in Statistics in the department of Mathematics at Imperial College London. During and subsequently following my PhD in 1999 I have been working on Bayesian statistical modelling techniques for non-linear complex systems. This includes methodological development in Machine Learning and Stochastic Computation, in particular novel Markov Chain Monte Carlo algorithms.

I am currently Honorary Secretary of the Research Committee of the Royal Statistical Society (RSS). In 2003 I was awarded the biennial Research Prize by the RSS for my work on Bayesian nonlinear regression models. I serve as Honorary Editor of the journal of Evolutionary Bioinformatics. I am co-organizer of a four week research program on "Bayesian nonparametric regression" to be held at the Isaac Newton Institute for Mathematical Sciences in Cambridge during July 2007.

My research group is based at the Oxford Centre for Gene Function and currently contains 4 postdocs and 6 DPhil students many of whom work on joint projects involving collaboration with investigators at the Pathogen Group in Zoology; the Wellcome Trust Centre for Human Genetics; the Centre for Statistics in Medicine and; the Broad Institute, MIT. Two of the postdocs are working on a highly related project in integrative genomics for biomarker discovery, as part of the EU MolPAGE consortia. The other two postdocs work on novel Bayesian methodology for pattern recognition models including non-linear methods and population based Markov Chain Monte Carlo algorithms for statistical computing. Of the six DPhil students, one works on new Bayesian approaches to mapping QTL jointly with Richard Mott at the WTCHG; one works on integrative genomics combining gene expression data with clinical markers for prognostic forecasting, joint with Doug Altman at the CSM; one works on novel Bayesian clustering models for de novo motif finding of regulatory elements; one works on models for SNP calling, in a joint project with the Broad Institute at MIT; one has just started working on copy number variants and their functional consequence in the mouse, and one has just started working on epidemiological models of sequence evolution in pathogens, incorporating space-time data in viral genealogies.

## Case for Support BBSRC Grant Application September 2005

“Integrative Analysis of the Genetic Factors behind Asthma and Atopic Dermatitis”

### **Part I: Research Proposal**

#### **Background**

##### **A Introduction of topic of research and its academic and wider context**

Asthma is the most common disease of childhood, and affects one child in seven in the United Kingdom. Atopic Dermatitis (AD, eczema) affects similar numbers of children. About 60% of children with severe AD will have concomitant asthma. Treatments for both diseases are unsatisfactory. Abandonment of orthodox medical therapy for AD is common in many families who have children with the disease.

Susceptibility to asthma and AD is strongly familial and has an established genetic component<sup>1</sup>. Recent progress in disease gene identification has highlighted the importance of epithelial innate immune mechanisms in mediating susceptibility and severity of asthma and AD<sup>1</sup>. Epithelial innate immune mechanisms have also been implicated in several other human diseases with large impacts on society including inflammatory bowel disease and infections of the lung and skin such as Tuberculosis and Leprosy. Epithelial cells, such as keratinocytes from the skin and airway epithelial cells from the lung are immunologically very active. Mechanisms utilised by these cells include the recognition of external and internal pathogens through pattern recognition receptors, active and passive barrier defences and antimicrobial defences, and signalling molecules to engage other inflammatory and immune cells. It is also likely that epithelial cells secrete homeostatic factors, which subdue inflammatory responses in the absence of a danger and damage.

A deeper understanding of asthma and AD is essential for the development of satisfactory therapies and disease risk management. A comprehensive study of epithelial cell molecular biology is required to achieve this. The research group of Cookson and Moffatt have been completing a high-throughput genomics programme to identify genetic susceptibility alleles for asthma and AD and to characterise the transcriptome of epithelial cells in a variety of clinically relevant conditions. The substantial quantities of data covering gene expression, genetic variation and disease phenotype (described in detail below) that these studies and others like it for other human disease have generated has precipitated a rapidly developing and dynamic research area focussed on the development of mathematical and statistical methods to exploit the information contained in multiple data sources.

**The size of these data sets and the nature of observations presents an unprecedented opportunity to uncover the genetic systems that co-vary with disease, by moving beyond isolated analysis of individual data.** The data sets record genetic association, genetic variation and gene expression all within disease and disease-related contexts. This makes possible the application of mathematical and statistical methods to perform integrated analyses and generate joint models of genetic variation, gene expression and co-regulation across disease states.

This proposal supports the acquisition of multi-disciplinary expertise to develop an integrative analytical framework for the analysis of the Cookson and Moffatt data sets. This framework will incorporate both novel and established analytical methods to generate experimentally testable hypotheses. This type of multiple data study will become increasingly common in the future and the methods we develop will be of general utility to the research community. We will therefore produce a freely available software analysis package incorporating our integrative methods.

The data collection for the Cookson and Moffatt studies has been funded by the Wellcome Trust, the Medical Research Council and the European Union. This funding has covered the collection of families, the culture of cells for RNA extraction and for genotyping and gene expression studies. The greatest part of the funding has been for reagent costs to produce the billion genotypes and the 1500 expression arrays described below. Simple data storage and bioinformatics is provided through core funding at the Wellcome Trust Centre for Human Genetics. The analysis of data so far funded extends to conventional analyses of linkage and association through the existing collaborative network of Cookson and Moffatt. The principal external collaborators are Mark Lathrop in Paris and Goncalo Abecasis in Michigan. Abecasis will provide statistical tools to access the Hapmap project. No provision in any application has yet been made to carry on the type of innovative investigation that this high dimensional data set can support. The development of statistical tools to analyse this depth of data is extremely timely, as many other groups world wide begin to utilise the technology and information that are now

available following the successful completion of sequencing of the human genome. The integrative approach to the data will inevitably shed new light on the understanding of the pathogenesis of asthma and AD, and will naturally lead to a systemic investigation of the systems biology of epithelial inflammation in a variety of different disorders.

## **B The data sets**

**1. Global gene expression of stimulated epithelial cells:** The response of keratinocytes and airway epithelial cells to immunogenic substances such as bacterial proteins and whole bacteria has been investigated extensively by the Cookson & Moffatt group. In particular, gene expression measurements with multiple biological replicates have been recorded in a time series after stimulation or differentiation. Affymetrix Human Genome arrays were used to measure expression of approximately 45,000 human transcript elements in approximately 24,000 genes.

**2. eQTL<sup>19</sup> mapping dataset** of 1000 subjects (600 sibpairs). Gene expression levels in EBV transformed cell lines from children with asthma and AD and their siblings have been profiled under controlled conditions using Affymetrix Human Genome GeneChips.

**3. Whole genome association study** of 3000 subjects in nuclear families identified through probands with asthma or AD. The families include 910 sibpairs selected to provide extreme trait distributions for the optimal analysis of association. Typing of 100,000 SNPs is being carried out in all family members and an additional 350,000 SNPs are being typed in probands and their siblings. This dataset contains all the subjects from the eQTL dataset described above. It will be analysed in collaboration with Abecassis.

## **C Aims of the analysis (see Figure 1)**

The guiding aim of the analysis is to identify systems of candidate genes (CGs) and regulatory elements (REs) involved in asthma and AD biology through the integration of complementary data sets. The main aims, outlined in Figure 1, are:

- 1) Identification of CGs and REs that mediate **epithelial cell** biology, specifically immune mechanisms.
- 2) Identification of candidate **genetic** variation for asthma and AD susceptibility.
- 3) **Integration** of evidence from above methods to generate testable disease candidates and pathways.

### **C1) Epithelial cell genomics – Dataset 1**

**1.1** Transcripts of differential abundance will be identified and distilled into expression clusters (ec) of shared expression profiles.

**1.2** Within clusters, the presence of shared regulatory and functional themes will be investigated, including shared or correlated chromosomal physical position, regulatory factor binding sites and gene ontology terms.

### **C2) Genetic Variants for Asthma and Eczema susceptibility – Dataset 2 & 3**

**2.1** Incorporation of potential disease loci (PDLs) from association mapping approaches with quantitative trait loci mapping for expression observations in both affected and unaffected individuals.

**2.3** Characterisation of mapped loci to identify: *cis* and *trans* correlations and one-to-many, many-to-one associations between genetic variation and expression. This investigation will specifically aim to identify elements of the eQTL profile that vary with disease status and hence are likely due to the presence of a genetic variant associated with disease.

### **C3) Integration – Dataset 1, 2 & 3**

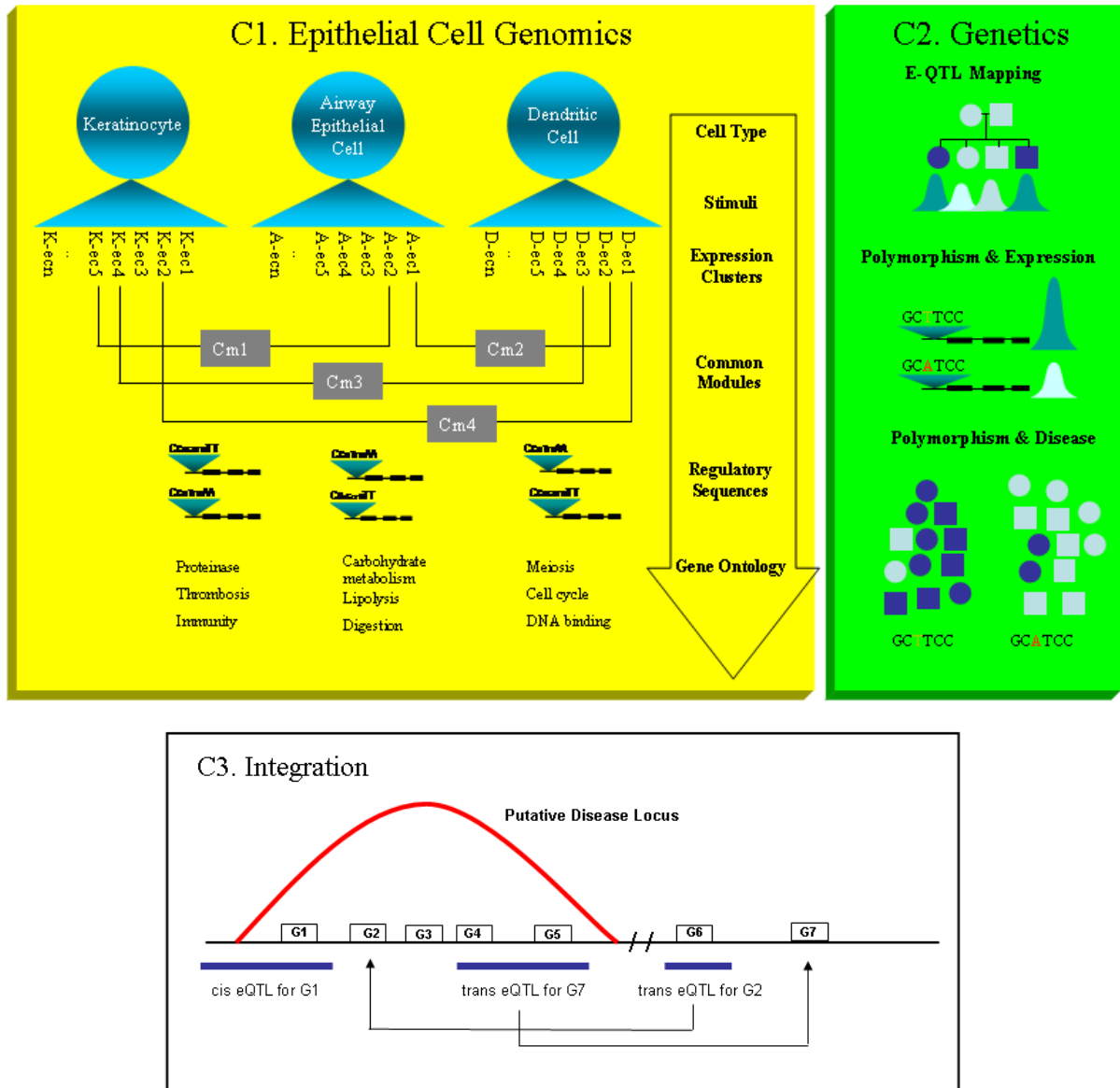
Support from all three data sets will be used to present robust candidates for experimental investigation. The major forms of support that will be considered are as follows:

3.1 Candidates whose expression is implicated in epithelial cell biology and/or disease and located within a PDL.

3.2 Candidates whose expression is implicated in epithelial cell biology and/or disease and whose eQTL is located within a PDL (for example G7 in Figure1).

3.3 CGs and REs that contribute to models of second order interactions between PDLs and eQTLs via epistatic or regulatory network interactions (for example G6 in Figure1).

3.4 CGs and REs associated with epithelial cell biology and show disease related expression.



**Figure 1: Schema of analytical framework.** C1 Epithelial Cell Genomics aims to identify robust clusters for which common regulatory themes and gene ontology can be identified. C2 Genetics provides the added information of genetic variation associated both with disease and with the mediation of gene expression. C3 Integration presents a scenario in which the eQTLs and disease associations could be combined to inform on complex regulatory interactions.

## **D. Methodology**

The methodology applied will have three main components: (a) Disease Association and eQTL Mapping, which will assist in defining regions of the genome that harbour CGs and REs that are relevant for a more detailed model of asthma and AD; (b) Expression analysis, which will assist in the delineation of underlying functional interactions leading to gene expression indicative of epithelial cell biology, asthma and AD; (c) Regulatory Signal models and their variation with disease status will incorporate information from comparative genomics and models of regulatory element interaction to infer shared regulatory machinery within expression clusters; (d) Integration and Modelling of supporting evidence from all of the above analyses towards a cohesive understanding of the complex interactions between PDLs, genetic variation and expression and how these interactions vary across disease states and differing epithelial cell states. A network or interaction model will be generated using methods of graphical modelling with both inhouse data and public databases to propose predictive models for epithelial cells and characterise critical molecular interactions within asthma and AD biology. Finally, supporting and extending methodologies from above will contribute to (E) Future Directions of the study and include interfacing and data exchange with contemporary public databases.

### **D(a) Disease Association and eQTL Mapping**

Mapping the human genome for regions and positions that are responsible for disease susceptibility and differential gene expression is central to this project. Both expression levels and disease status can be considered as phenotypic traits and observing these together with segregating markers allows expression quantitative trait loci (eQTL) and disease loci (PDL) to be found. To date the success of molecular profiling technologies and association mapping approaches to identify genes underlying complex disease has been limited. Association mapping approaches often lack the resolution to unambiguously identify a reasonable number of potential candidates for further study. Advanced multipoint coalescent based methods allows refinement of this<sup>2,3</sup> and recent advances in models allow improved handling of population structure and complex inheritance<sup>4</sup>. The strength of integrating the data sets as proposed here is that the resolution of a whole genome association study is extended and improved by the observations of gene expression throughout putative disease loci<sup>18</sup>. Further, within the disease context of interest these expression observations can be mapped to genetic loci (eQTL analysis) thus informing models of both variation of expression and variation of the regulation of expression across genome. Genetic variation directly mediating expression can be mapped and compared to regions of disease association to identify SNPs that play a role in disease. This integration provides a powerful extension of the molecular profiling approach.

The analysis of the association mapping dataset will be performed in collaboration with Goncalo Abecasis and Mark Lathrop and will be completed as specified in previous funding. Completion of this analysis is not part of this proposal. We seek to determine which heritable eQTLs determine risk for AD or asthma and to build statistical models that predict the risk of the disease from this data. We will determine the role of genetic polymorphism in modulating expression levels using modified QTL mapping techniques. The combined modelling of eQTL with disease phenotype and its highly dimensional nature will require the development of novel statistical techniques.

Automated strategies for detecting eQTLs will be adapted from standard QTL mapping methods. Existing packages include QTL Cartographer<sup>5</sup>, WebQTL<sup>6</sup>, Merlin<sup>7</sup> and QTDT<sup>8</sup>. Standard QTL methods were originally designed to consider a small number (relative to thousands of gene expressions) of quantitative phenotypes and are typically based on a generalised least squares approach. The extra dimensionality of the dataset considered here requires modification of standard methods for computational efficiency and analytical power. Possible extensions include the use of parental observations and models of gene-gene interactions. Novel methods for significance testing will be investigated since obtaining an empirical significance threshold for each gene is highly computational<sup>9</sup>. The major focus of the eQTL methodology will be the integration of genetic variation, expression, disease association and phenotype to understand the correlation structure and interactions underlying disease (refer section D(d)).

### **D(b) Expression Analysis:**

The aim of the expression analysis is the application of analytical approaches that sufficiently capture the critical patterns of expression across the experimental design. Two different expression data sets are considered here: time series of epithelial cell biology and gene expression of a complex disease sibpair cohort. Each dataset offers unique challenges and opportunities to capture greater understanding of the expression variation that is not afforded by standard analytical methods such as significance analysis of microarrays (SAM)<sup>10</sup> and Wilcoxon rank tests. These standard methods are extremely useful for the identification of robust changes, they do not consider the complex relationships between genes in several ways. For example, time series data sets potentially capture relationships and

dependencies of gene expression within and between time points which may suggest causative co-regulation. These dependencies and interactions could be better uncovered using statistical modelling approaches such as Bayesian model based methods that aim to identify co-expressed clusters of genes under a model of temporal dependence between observations, that is utilising gene expression measures in time to better judge cluster membership<sup>11,12</sup>.

Secondly, the asthma and AD expression dataset of sibpairs inherently contains underlying structures of shared genetic disease risk. It is likely that this shared risk across affected and unaffected sibpairs will confound standard analytical approaches to identify transcripts that are necessary but not sufficient for disease onset and transcripts that contribute to disease with small effect. A more complex model could be derived to accommodate this aspect of the dataset. Further information derived from SNP genotyping could be used to gain an understanding of the degree of genetic sharing for each sibpair at each loci to enhance the model. To generate novel analytical methods to address these issues, simulated data sets will be generated based on known inheritance models for complex disease. The ability of methods to detect associated gene expression profiles under conditions of reduced penetrance, genetic heterogeneity and common, small effect genetic variants will be assessed.

#### **D(c) Regulatory Signals:**

Co-regulation of genes via shared transcriptional networks provides the basis for context-dependent gene expression, an understanding of which is vital to the understanding of disease etiology and disease progression. In particular, transcription factors (TF) and their transcription factor binding sites (TFBS) provide a key component in the understanding of how co-regulation is achieved.

Recently, a number of key papers<sup>13,14,15</sup> have demonstrated the ability to identify upstream TFBS using advanced bioinformatic techniques, by searching for motifs (common patterns in the binding sites) which can be related to measurements of differential expression obtained via microarrays<sup>13,14</sup>. Importantly our proposal will extend and improve upon the existing methods<sup>13,14,15</sup> in a number of key respects

- We shall develop joint statistical models that simultaneously characterizes multiple TFBS *and* learn their functional role in co-regulation of gene expression within the context of a disease phenotype.
- We shall use Bayesian nonlinear machine learning techniques<sup>16</sup> to infer multiple TF interaction involved in co-regulation, making use of both the order and relative location of the TFBS.
- We shall incorporate *semi-supervised* learning methods to help infer such structures.

Existing methods that combine sequence and expression data tend to use a two-stage learning process whereby initially gene expression data is pre-processed into clusters (or ranked by differential expression) and subsequent TFBS are found by searching for common motifs<sup>13,14</sup>, treating the clusters as fixed. We intend to integrate these two stages within a joint Bayesian statistical model of TFBS and their nonlinear interactions predicting co-expressed clusters of gene profiles and their variation with disease status. This joint borrowing of strength of the two sets of observables (sequence data and expression data) should lead to more coherent clusters and more powerful detection. Proof of principle of the advantages of joint modelling is indicated by ref(17), though they do not use sequence data or explicitly search for TFBS. Furthermore we intend to investigate whether the order and relative location of identified TFBS helps in learning about their interaction and functional role in co-regulation. This extra information may offer potential gains in inferring their functional inter-dependencies which has not been previously considered.

Our approach then will simultaneously infer potential clusters of co-expressed gene profiles and the multiple TFBS that co-regulate them and how these associate with disease status. We shall treat the number of gene expression clusters, the number of their regulatory TFBS and the nature of the dependence structure as unknown. Using Bayesian probability models and simulation methods such as Markov chain Monte Carlo (MCMC) we aim to infer a joint structure on TFBS location and their role in co-regulation of the expression data. The research will compare a number of motif models including generative methods, such as multinomial-Dirichlet distributions<sup>15</sup> and diagnostic approaches, such as logistic regression<sup>16</sup> within a semi-supervised approach. Semi-supervised learning is a technique which allows for a combination of labeled and unlabeled samples. In our case the labeled samples would relate to known up-stream motifs (from known gene TFs in the literature), which help in the discovery of unknown (unlabeled) data. To model the functional dependence we shall explore machine learning methods<sup>16</sup>, such as decision tree methods to predict the co-expressed gene profiles. As part of this study and in (E) Future work, see below, we will investigate the benefit of using comparative genomics in helping to locate and characterise the regulatory elements and signals.

#### **D(d) Integration and Modelling to infer regulatory systems co-varying with disease status**

The inference of regulatory networks in mammalian systems has proved to be a difficult and complex problem. This proposal will focus on a much smaller regulatory neighbourhood within a disease context around a well supported set

of candidate genes for the phenotype of interest. This greatly reduces the dimensionality and complexity of the problem to initially mapping first and second order local interactions. We will combine the information and analysis from methods D(a)-D(c) in order to characterise the changing nature of regulation with disease state. This will necessitate novel statistical methods. It is difficult to envisage a priori the methods best suited to this task: though it is highly likely to involve constructing small stochastic networks of co-regulation, incorporating the genetic variants associated with disease, which can evolve with phenotype.

Schadt et al., (2005) presented a multi-step procedure to identify genetic variants that were causal to gene expression variation associated with a phenotypic trait. We will extend and modify this approach with utilisation of the information provided from disease association studies, expression time profiles of epithelial cell biology and the regulatory modules highlighted from D(c). While the objectives of Schadt et al are similar to ours, our data structures are much richer and more complex and this will necessitate different approaches to be adopted. For instance, our principal phenotypic trait is qualitative (diseased/non-diseased) and this allows us to consider the problem of modelling changes of eQTL patterns given a change in phenotype, as opposed to modelling the phenotype given the eQTL data (as in Schadt et al). Moreover, the Schadt et al study considered 139 micro-satellites markers on a population of 111 F2 mice. We will be developing methods for analysing data on a much denser map of at least 100,000 SNP markers on hundreds of nuclear families from a human population. This will provide substantially greater information regarding the interaction of genetic variation with both phenotype and expression, and much greater challenges regarding multiple testing and accounting for population structure. We will utilise Bayesian probability models to integrate the data sets and characterise the levels of uncertainty associated with any one model.

## **E The Future**

We are critically aware that the field of statistical bioinformatics is changing at pace. The present description focus on 3 data sources (SNPs, phenotypes (expression levels and disease status) and genomes), but now and increasingly in the future additional data sources and modelling techniques will be relevant. These will have to be handled 3-4 years from now on a large scale. In the latter stages of the project we aim to assess the follow future directions and depending on the speed of progress made in D(a)-D(d) seek to investigate one or more of the following themes

- *Networks*: It is likely that as the field of network inference progresses regulatory networks in other biological contexts and organisms will become available and it would be very informative to place and compare inferences generated from this work within other networks. Methods for network comparison are a rapidly developing research area.
- *Protein interaction data*: There is a growing body of protein-interaction data and this data is a useful extension to inferences of functional interaction between disease gene candidates and co-expressed genes.
- *Ontologies for Functional Annotation*: This project will lead to a small subset of genes of interest for asthma and AD.. Ontologies are key in making automated and vocabulary controlled statements about function and it will be interesting to interface the analytical framework presented in the proposal with contemporary advances in gene ontology methodology.
- *Comparative Genomics*: The availability of a large number of mammalian genomes allows further characterisation of regulatory signals by observing their conservation or mode of evolution. This assumes some knowledge of how regulatory signals evolve. This can be modeled, where not only a sequence has been observed with its expression patterns, but additionally information regarding how surrounding regions of the sequence evolves. Given the large number of genomes, this may prove a very important contribution.

## **F Software Development**

This project involves the development of methodologies covering expression analysis, gene mapping and comparative genomics. The methods are continuously applied to a large state-of-the art data set. Given the generality of many of the problems, these algorithms and models will be incorporated into a programming package of growing capability. A software package that provides an integrated framework for the analysis of these datasets is currently not available. As the capability to generate large-scale genetic and genomics datasets becomes more widely available there is an increasing need for an integrated suite of algorithms to fully exploit the power of these datasets. We feel that because of access to a pioneering example of this data and joint expertise of Cookson and Moffatt, Hein, Holmes and Taylor we are ideally placed to contribute such a package. We will make this user-friendly and generally available. We propose to adopt a software framework similar to that employed by The Institute for Genomic Research (TIGR) in their software suite for analysis of expression data, TM4 (<http://www.tm4.org>). The TM4 software suite has a modular structure containing 4 component tools for specialised computational tasks. They are developed in Java and C/C++.

The software developed in this proposal will need to be optimised for the efficient computation of large data sets and will consist of several component tools which will interface to provide seamless data transfer:

- Input parser: This module will contain tools to parse input data consisting of individuals with associated phenotype and molecular data (e.g. SNP genotypes and expression levels under different criteria). Public database interfaces will allow translation of expression and genotype probe identifiers where necessary.
- Pedigree, Association and eQTL Analysis: This module will provide association mappings for phenotypic and expression traits. Graphical representation, subsetting and cross-referencing of multiple mappings will be accommodated. For example, the visualisation of coincident phenotype and eQTL mappings. These visualisations will interface with a genome browser (Ensembl or UCSC) to permit visualisation of the functional genetic elements and conservation profiles within mapped loci.
- Expression profiling: There is currently an expanding body of expression profiling tools in both the BioConductor project ([www.bioconductor.org](http://www.bioconductor.org)) and TM4 software databases. We plan to submit any novel tools generated in this proposal to BioConductor and provide an interface within the proposed software to R. This allows maximum flexibility and prevents duplication of effort.
- Regulatory Signals: This module would contain currently available tools for the identification of regulatory elements, new tools developed in this proposal and permit interface with output from other tools in the software suite (e.g. eQTL analysis) for the identification of key regulatory features of interest. The module would also contain multiple interfaces with sequence, genome and regulatory element databases.
- Integration tools: The tools required in this model will become clearer as the project progresses. It is likely that several statistical modelling tools with clear graphical representation tools will be required. This will likely involve graphical models and algorithms for the probabilistic inference of interactions in the form of edges between nodes of local networks of genes, regulatory elements and observed phenotypic traits.

## **G Timeliness**

The large amounts of data from a series of throughput technologies coupled to computer intensive statistical modelling are allowing predictive modelling to be inferred for well-defined biological systems. This kind of ambitious integrative data analysis will dominate biomedical research in the coming years. This project proposed here combines the data already being generated at the WCHG by the Cookson group with modelling expertise of Jotun Hein and Chris Holmes, to model cell types relevant to the understanding of Asthma and Atopic Dermatitis and its pathogenesis. Since this approach is transferable to the analysis of other diseases the value of this analysis effort is strongly augmented by the production of generally available analysis software.

## **H Milestones and management of the project**

Our project has two main tasks: (i) an integrated analysis of multi-level data relevant for asthma and AD; (ii) the development of methods and software that can perform this analysis and similar analysis for other disease types. The postdoc (PD) and scientific programmer (SP) will be placed in Oxford Centre for Gene Mapping and under daily supervision of JH and CH. Additionally, JT will participate in one weekly meeting. Finally, there will be a monthly meeting with WC as well, that will alternate between Oxford and Imperial where progress and problems will be discussed. Jotun Hein will supervise the project generally, Chris Holmes with focus on statistical matters, while Jennifer Taylor and Bill Cookson will give biological expertise and data interpretation. The PD, who will have statistical expertise, and SP, with computer science expertise (algorithms, software engineering and interfaces) will continuously work together on the same tasks and establish a division of labour, where both have some engagement in all aspects, but the PD will lead in statistics and biological interpretation, while the SP will lead in implementation and algorithms.

Year 1-Months 1-3: PD+SP reads biological literature and collect relevant existing programs for disease associations, that are tested on existing data. Disease Association Mapping and eQTLs,

Year 1 Months 4-6: PD+SP plans in more detail the overall structure of the package to be developed and the algorithms to be used. PD+SP make simple simulation models that can generate artificial data involving all facets of the analysis (genetic variation, regulatory signals and expression, genomes from related species). This is necessarily simplistic, as full details of complex aspects such as regulation are unknown, but can become increasingly realistic as the project proceeds.

Year 1 Months 7-12: Simple analysis programs are developed that can be continuously tested on simulated data. Pipe line based on modifications of existing programs created for: i) SNP analysis with respect to disease susceptibility and expression level, ii) regulatory signals from multiple genomes, iii) expression analysis.

Year 2 Months 13-18: Basic analysis based on first newly developed methods.

Year 2 Months 19-24: Software development and testing

Year 3: Large scale analysis of all existing data and incorporation of external databases, software optimisation.

Year 4: Major publication activity. Finalizing software and incorporation of other data types such as protein interactions and more refined models of networks.

### **Justification of resources**

Asking for 2 full salaries for 4 years reflects the high level of ambition of this project. Software development, computational intensity and the extensive implementation necessitates the SP. The 4 years instead of for instance 3 is due to the novelty of the proposal and hence the extensive training that the PD +SP will experience so that their full value is first realized after some time. The bulk of the resources are allocated in the PD +SP salaries that will be in Oxford Centre for Gene Function in the Bioinformatics and genome analysis group. Researchers in computational biology and bioinformatics are in strong demand and hard to attract, we have therefore to offer a very good salary. A high level of computing is clearly required. A suitable choice would be something similar to a Dell Precision M70 Advanced, upgraded to 2MB of memory (£1820). Once the project is running, a work station will also be required for larger computations. A suitable choice would be something similar to a Dell Precision 470 Advanced, upgraded to 3GB of memory, DVD+-RW drive and dual Xeon 3GHz processors (£2185). It can be anticipated that one or more licences for mathematical software (MatLab or Mathematica), and reference books, will be essential for the statistical modelling aspects of the project. We will need service support from the Statistics Department and in particular when distributing software a www-page and ISP facilities are required and hence it is necessary to support (10%) a computer officer. The postdoc and the scientific programmer will be expected to publish at and attend international conferences on bioinformatics (e.g. RECOMB or ISMB) as well as asthma (e.g. the American Thoracic Society meeting), these will of course incur travel and conference attendance costs for them and for the PI and CIs on the grant.

### **References:**

1. **Cookson, W.** (2004). The immunogenetics of asthma and eczema: a new focus on the epithelium. *Nat. Rev. Immunol.* Dec; 4(12): 978-88.
2. Morris AP, Whittaker JC, Balding DJ. (2002) Fine-scale mapping of disease loci via shattered coalescent modelling of genealogies. *Am J Hum Genet.* Mar;70(3):686-707
3. **Hein, JJ,** Schierup, MH and C.H. Wiuf (2005) "Gene Genealogies, Variation and Evolution" Oxford University Press.
- 4 Marchini J, Donnelly P, and Cardon LR (2005) Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nat. Genet.* Apr;37(4):413-7.
5. Basten C.J., Weir, B.S. and Z.-B. Zeng. (2001). QTL Cartographer, Version 1.15. Department of Statistics, North Carolina State University, Raleigh, NC.
6. Chesler EJ, Lu L, Wang J, Williams RW and Manly KF. (2004). WebQTL: rapid exploratory analysis of gene expression and genetic networks for brain and behaviour. *Nat. Neurosci.* May; 7(5):485-6.
7. Abecasis GR, Cherny SS, Cookson WO and Cardon LR. (2002). Merlin—rapid analysis of dense genetic maps using sparse gene flow trees. *Nat. Genet.* Jan;30(1):97-101.
8. Abecasis GR, Cardon LR, Cookson WO (2000). A general test of association for quantitative traits in nuclear families. *Am J Hum Genet.* 66:279-292.
9. Carlborg O, De Koning DJ, Manly KF, Chesler E, Williams RW and Haley CS. Methodological aspects of the genetic dissection of gene expression. (2005) *Bioinformatics* May 15;21(10):2383-93.
10. Tusher VG, Tibshirani R, Chu G. (2001) Significance analysis of microarrays applied to the ionizing radiation response. *PNAS.* Apr 24;98(9):5116-21.
11. Heard, N. A., **Holmes, C. C.**, Stephens, D. A., Hand, D. J. and Dimopoulos, G. (2005) Bayesian Co-clustering of Anopheles Gene Expression Time Series: A Study of Immune Defense Response To Multiple Experimental Challenges. *Proceedings of the National Academy of Science USA*, 102, 47, 16939-16944.
12. Heard, N. A., **Holmes, C. C.** and Stephens, D. A. (2005) *J. Am. Stat. Assoc.* (in press).
13. Beer MA, Tavazoie S. Predicting gene expression from sequence. (2004) *Cell.* Apr 16;117(2):185-98
14. Conlon EM, Liu XS, Lieb JD, Liu JS. Integrating regulatory motif discovery and genome-wide expression analysis. (2003) *PNAS*, 100(6):3339-44.
15. Zhou, Q. and Wong, W.H. CisModule: De Nova Discovery of Cis-Regulatory Modules by Hierarchical Mixture Modeling. (2004) *PNAS. USA*, 101: 12114-12119.
16. Denison D, Holmes C, Mallick B, Smith AFM. "Bayesian methods for nonlinear classification and regression", (2002) Wiley: Chichester.
17. Segal E, Shapira M, Regev A, Pe'er D, Botstein D, Koller D & Friedman N. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. (2003) *Nat. Genet.* 34, 166-176
18. Schadt EE, Lamb J, Yang X, Zhu J, Edwards S, Guhathakurta K et al., an integrative genomics approach to infer causal associations between gene expression and disease. (2005) *Nat. Genet.* Jul;37(7):710-7.
19. Morley M, Cliona M. Molony, Weber TM, Devlin JL, Ewens KG, Spielman RS and Cheung VG. (2004) Genetic analysis of genome-wide variation in human gene expression. *Nature* Aug 12;430(7001):743-7.

