Genomic approaches for the identification of risk loci for Rheumatoid Arthritis

Antonio Julià Cano

Grup de Recerca de Reumatologia

Institut de Recerca Hospital Universitari Vall d'Hebron

A thesis submitted for the degree of

Doctor of Phylosophy

January 2010

To my parents,

Acknowledgements

First of all, I would like to express my gratitude to my advisor, Sara Marsal, who gave me complete freedom to develop my scientific interests yet reminding me to keep focused on important biomedical problems. Her incomparable dedication to the study of human disease has been a constant reference and motivation throughout these years.

These have been years of intense work in which many important people have contributed to my formation. I would like to thank them in historical order:

- Dominique Gallardo and Francisco Vidal for teaching me the importance of rigor in the lab bench.
- Jerry Lanchbury from the Guy's and King's college (London, UK) for his helpful advice in the study of the complex genetics of Rheumatoid Arthritis.
- Joaquim Ariño from the Universitat Autònoma de Barcelona for his fundamental support and help in the first years of our research group.
- Anna Bassols and Laia Miquel Serra from the Universitat Autònoma de Barcelona for their invaluable help and knowledge on human tissue cultures.
- Ioannis Ragoussis from the Wellcome Trust Human Genetics Centre (Oxford, UK) for providing an in-depth training in the most recent genomic technnologies.

- Jason Moore and Marylyn Ritchie from the Vanderbilt University (Nashville, USA) for their close collaboration and discussions on the multiple aspects of epistasis.
- Elisabeth Vilella from the Hospital Universitari Sant Joan de Reus for sharing her pioneering expertise in DNA Biobanking.
- Jesús Tornero (Hospital Universitario de Guadalajara), Javier Ballina (Hospital Universitario de Asturias), Juan de Dios Cañete (Hospital Clínic de Barcelona) and Alejandro Balsa (Hospital La Paz de Madrid) for their willingess to collaborate.
- Simó Schwartz and Toni Andreu from the Hospital Univeristari Vall d'Hebron for their support in the difficult moments of a growing research group.
- Jaume Bertranpetit from the Universitat Pompeu Fabra for his open mindedness and for making human genetics a better science in our country.
- Alberto Orfao and Andres García from the Banco Nacional de ADN for their helpful collaboration and for raising the standards of human sample Biobanking in Spain.
- Josep Lluis Gelpí, Jordi Camps and in general the Barcelona Supercomuting Centre Life Sciences team, for their invaluable technical support in high performance computing.
- David Clayton from the Cambridge Institute for Medical Research (Cambridge, UK) for the instructive discussions and advice on the fast-evolving world of GWAS analysis.
- The clinical staff from the Hospital Universitari Vall d'Hebron from which I will always remember their dedication and implication.

Finally, I would like to thank all the patients that have collaborated in this biomedical research work. If any of this or other future work helps to improve their lifes, it will be the best of all rewards.

Abstract

Rheumatoid Arthritis (RA) is one of the most prevalent autoimmune diseases in the world and is characterized by the chronic inflammation of the synovial joints. The origin of the disease is unknown but it is actually accepted that it is caused by the complex interaction of a genetic susceptibility background and environmental factors. To date, the characterization of the genetic architecture of RA is far from complete. In the present work we will use the power of two distinct genomic approaches to identify new candidate genes for the susceptibility to RA.

In the first genomic approach, we have used gene expression microarrays to characterize the *in vitro* transcriptional response of the synovial fibroblast (SF) to the stimulation with RA synovial fluid. Using a reverse engineering approach, we have inferred the main transcriptional regulatory network that governs the response to this complex proinflammatory stimulus. We have then studied the genes in this regulatory network as risk factors for RA susceptibility using a case-control approach. We have analyzed the association of each gene with disease independently, but we have also analyzed the presence of higher order interactions associated with disease risk (i.e. epistasis) using the Multifactor Dimensionality Reduction method.

In the second genomic approach, we have used whole genome genotyping microarrays targeting more than 300,000 SNPs (Single Nucleotide Polymorphisms) markers to perform a Genome-wide Association Study (GWAS) in RA. In order to increase the statistical power of our study we have implemented a liability-based design. We have

subsequently validated those loci showing highest evidence of association using an independent replication cohort. Also, in order to integrate our findings with the evidence of previous GWAS in RA, we have determined those genomic loci showing increased clustering of signals between studies. Finally, we have performed an exhaustive genome-wide analysis of the two-way epistatic interactions associated with RA applying parallel computation.

Using the SF *in vitro* stimulation model we have identified n = 157 genes significantly associated with the response to RA proinflammatory stimulus. Within this set of differentially expressed genes there are genes that have been clearly associated to RA pathophisiology but also new genes not previously linked to this disease. From the differential expression data we have been able to identify a 13 gene Nuclear Factor kappa-Beta (NF-kB) transcriptional regulatory network, as the key transcriptional regulatory force in this RA SF model. Whilst several of the genes in the network showed nominal association to disease, we have identified a significant epistatic interaction between *interleukin 6* (*IL6*) and *interleukin 4 induced 1* (*IL4I1*) genes.

In the GWAS approach we have identified several candidate genes for RA, advanced RA and chronic arthritis risk. Using an independent replication dataset we have found an intronic SNP in *Kruppel-Like Factor 12* (*KLF12*) gene as the most strongly associated SNP with RA. The meta-analysis with previous GWAS results has also identified several genomic regions -including *KLF12* locus- that are likely to harbour new risk variants for RA. In the genome-wide epistasis analysis we have found a number of SNP pairs associated with RA with a significance close to the conservative multiple test correction threshold. Also, we have found that two-way interactions including the HLA region, the strongest main effect in RA, are ranked secondarily to many other potentially interacting loci, thus suggesting a minor role for this locus in the epistatic susceptibility to disease.

The two alternative genomic approaches we present in this work have identified a group of new loci which are likely to be associated with the risk to RA. This group of candidate loci should be now validated in independent populations to confirm their implication in RA susceptibility.

Contents

List of Figures

Nomenclature

λ Genomic Inflation Factor

λ_R Relative Risk

ACR American College of Rheumatology

AMD Acute Macular Degeneration

APC Antigen Presenting Cell

CCP Cyclic Citrullinated Peptide

CFH Complement Factor H

CNV Copy Number Variant

DNA Deoxyribonucleic Acid

ECM Extracellular Matrix

$GWAS$ Genomewide Association Study

HLA Human Leukocyte Antigen

$IL1\beta$ Interleukin 1 Beta

$IL4I1$ Interleukin 4 Induced 1

$IL6$ Interleukin 6

$IMID$ Immune–Medited Inflammatory Disease

$KLF12$ Kruppel-Like Factor 12

LD Linkage Disequilibrium

LOD Logarithm of Odds

MDR Multifactor Dimensionality Reduction

$mRNA$ messenger Ribonucleic Acid

$NFkB$ Nuclear Factor kappa Beta

OR Odds Ratio

PCA Principal Components Analysis

$PTPN22$ Protein Phosphatase 22

QC Quality Control

RA Rheumatoid Arthritis

RF Rheumatoid Factor

RNA Ribonucleic Acid

SF Synovial Fibroblast

SNP Single Nucleotide Polymorphism

TCR T Cell Receptor

$TNF\alpha$ Tumor Necrosis Factor Alpha

$Treg$ Regulatory T cell

$WTCCC$ Wellcome Trust Case Control Consortium

Chapter 1

Introduction

"Why am I sick?" As soon as we have consciousness of our self we start asking this question. People want to know the reason why their body is being harmed and they feel unhappy. For many centuries, humanity has devoted much effort in trying to understand this intrinsic aspect of life. One fundamental step in this quest was the discovery that certain microscopic organisms can enter our body and attack us. This was a simple answer to the disease origin question: it's the "invaders". Fortunately, for many of these minute creatures we have been able to find treatments that can fight them and reduce its impact on human life. However, there are still some microorganisms like the human immunodeficiency virus (HIV) which we still cannot defeat and that are a constant reminder that we should never underestimate the power of nature, but also that we should never overestimate our capacity to overcome threats.

1.1 Genes as disease causing agents

By the end of the nineteenth century, a new area of research appeared. For centuries, health professionals had noticed that, for a particular disease, some families carried more diseased individuals than others. Therefore, there needed be something that parents passed on to their offspring that perpetuated the manifestation of the disease. Using a robust methodological approach, Gregor Mendel identified in 1866 the rules of heredity, practically 100 years before the chemical structure of the hereditary material was resolved by James Watson and

Francis Crick. Once the hereditary molecule was characterized, a second cause of disease emerged: the mutation of the DNA sequence. Genes give rise to proteins, and proteins are the workhorse of the cells that build up our body; thus, from a mechanistic point of view, it seemed clear that a faulty protein due to a mutation in its encoding DNA should negatively influence those biological functions in which it participates and, henceforth, it could end up causing a disease. To date, more than 1,500 human diseases caused by mutations in a gene have been already characterized (Peltonen & McKusick, 2001).

1.2 Complexity in disease causality

For many diseases however, neither an infectious agent nor a causal mutation could be identified. And many of these diseases are quite prevalent in human populations. Thus, by the end of the twentieth century, the research on human diseases needed to broaden its view of causality and started to embrace complexity. The introduction of technologies that can analyze thousands of biological variables in parallel (i.e. high throughput technologies) has been crucial in this transition. With these genomic technologies, researchers are now able to study the organism taking into account all its elements and not just a limited set. This way of performing research is conceptually new in biomedical research and it implies profound conceptual changes; one of them is that we are less attached to our subjective knowledge of the disease process. Deductions from the whole set of transcripts, for example, are more powerful because they are more objective. But also this analytical approaches offer a new possibility: to use this information in conjuction to obtain a complete characterization of the normal and pathological states. For this objective a new area of science, systems biology, is now rapidly emerging.

The new view on biology that genomic technologies are providing is having a profound impact on how we view diseases. Individuals are now seen as particular "mosaics" of genetic variations which can have differential behaviours according to the environmental factors with which their interact throughout their life. What we previously defined as a single disease according to a set of signs and symptoms, it may now be described as a large heterogeneous collection of molecular

variations. In the present dissertation we will use the power of genomic technologies to identify new genetic variations associated with the risk to develop RA, a complex and heterogeneous disease.

First strategy: genomics of expression and complexity

In the first strategy we have used gene expression microarrays to identify the set of genes that are associated to the response to a pathogenic insult. Using a bioinformatic approach called reverse engineering we have estimated the principal network of genes that govern this gene expression response. Once we have identified the main regulatory network, we have studied the association of genetic variation within these genes with the susceptibility to develop the disease. Given that we have evidence that the gene products of these genes belong to a common regulatory pathway, we have also studied the presence of epistasis associated with disease risk.

Second strategy: genetic genomics and complexity

In the second strategy we have used whole genome genotyping microarrays to directly identify SNPs associated to the risk to develop the disease. These microarrays are built upon tagSNPs which are highly informative SNPs that cover most of the common variation in the genome. Using this approach we have objectively assessed which genetic variations are more relevant to disease susceptibility. In order to increase the power of the approach, we have introduced an epidemiological design that maximizes the difference in liability between individuals. We have compared our results with the results of previous genome-wide scans for the same disease. Finally, we have conducted a genome-wide exploration to identify two-way epistatic interactions associated with disease risk.

Summary of the Contributions

- Identification of the genes significantly associated with the SF response to the RA proinflammatory environment.
- Identification of NF-kB as the main transcription factor governing the response to RA synovial fluid.
- Characterization of the main transcriptional regulatory network associated with the SF response to the RA environment. Under RA synovial fluid complex stimulus, transcription factor NF-kB upregulates 11 genes and downregulates 2 genes. While some of these genes are known to be regulated by NF-kB, there are other genes that had never been previously associated with the regulation by this transcriptional factor.
- Identification of a significant two–way SNP interaction associated with a high risk to develop RA. Multifactor Dimensionality Reduction epistasis analysis on the promoter SNPs of the 13 corregulated genes of the SF network revealed that the interaction between *IL4I1* and *IL6* genes was better at predicting disease status than any other higher level combination.
- First genome-wide association study performed in the Spanish population: identification of a low impact of genetic structure albeit a predominant West to East trend different from other European populations.
- Third GWAS study on RA: identification of a new gene for susceptibility to RA. SNP rs1324913 located in the first intron of transcription factor *KLF12* is associated to RA risk in a reproducible dominant model.
- Meta-analysis with previous GWASs: identification of genetic regions that show increased probability to harbour a susceptibility variant for RA.
- First exhaustive genome-wide scan for epistasis performed for a complex disease. The results for the top ranking SNP pairs are promisingly close to multiple test corrected significance. Interactions of variations in the HLA region with other non-HLA loci rank sensibly lower compared to other genomic epistatic interactions.

Chapter 2

Rheumatoid Arthritis

2.1 Epidemiological Perspective

Rheumatoid Arthritis is a chronic inflammatory disease that affects the synovial joints. It is one of the most frequent autoimmune diseases in the world with a worldwide prevalence of approximately 1%. To date, the most exhaustive epidemiological survey performed in Spain (Carmona *et al.*, 2002) estimated a 0.5% (0.25–0.8, 95% CI) prevalence. This value is similar to other Mediterranean countries like France (0.6%), Italy (0.3%) or Greece (0.3–0.7%), but relatively lower than northern European countries like the UK (0.8–1.1%), Finland (0.8%) or Sweden (0.5–0.9%). At the two prevalence extremes we find a relatively large number of cases in native–American populations like the Chippewa (6.8%) (Harvey *et al.*, 1981) (Harvey, Lotze et al. 1981) and Pima (5.3%) (Del Puente *et al.*, 1989) Indians and the lack of evidence from sub–Saharan populations like the Nigerians (Silman *et al.*, 1993b). In the European countries, incidence estimates follow a similar trend to the prevalence estimates, with Northern countries having a relatively higher occurrence of disease than Southern countries (0.44–0.8% vs. 0.31–0.5%, respectively).

In all countries examined, RA is characterized by an increased frequency in females compared to males, in an approximate 3:1 ratio. This differential susceptibility is common also for several other autoimmune diseases (Figure 2.1).

The aetiology of RA is still unknown but there is clear evidence that there is a complex genetic architecture that increases the risk to develop it. Environmental

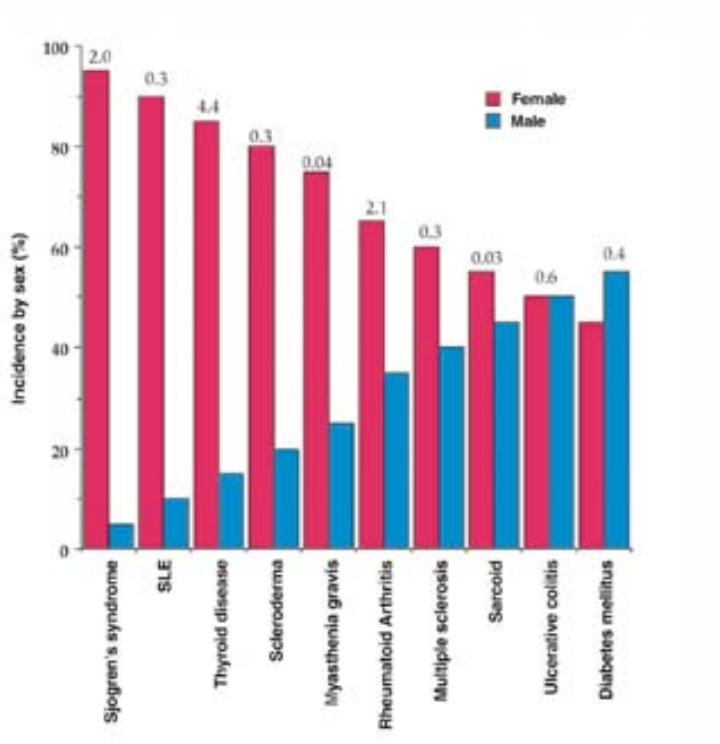


Figure 2.1: Incidence of different autoimmune diseases according to sex. Taken from Whitacre (Whitacre, 2001).The numbers above the bars refer to the total number of disease cases (x 1,000,000) in the USA.

factors have also an important role as disease triggers but, apart from cigarette smoking (Odds Ratio (OR)∼12 in monozygotic twins) (Silman *et al.*, 1996), there is no other environmental factor that has been consistently associated with the risk to RA. More specifically, there is yet no strong evidence in favour of an implication of any microorganism in the generation of the disease. For all these reasons, RA is actually classified as a common complex disease.

2.2 Pathophysiology of the disease

In order to understand the pathological processes that occur in RA, we must first have a clear picture of the target organ of this chronic inflammatory disease: the synovial joint.

2.2.1 The synovial joint

Evolution has brought to many species the ability to move. Being able to change the spatial localization has important advantages: it increases the probability to find necessary resources like food and it can also be a useful defence mechanism. In higher vertebrates like humans, movement is performed by the combination of internal rigid structures (i.e. bones) and soft contractile structures (i.e. muscles and tendons). The contractions and distensions of muscles allow the relative movement of the skeletal bones at specific angles that allow the execution of multiple necessary tasks like walking, grabbing tools, etc. Whilst some joints like the skull synarthrosis or the vertebral ampharthrosys are attached by intermediate connective tissue, the synovial joints do not have any rigid binding. Instead synovial joints have two opposing cartilage surfaces that interact through a viscous lubricating liquid called synovial fluid. This is due to the necessity to perform big angles of trajectory like rotation or flexion.

Embrionary development of the synovial joint

To better understand the synovial joint architecture, it is useful to revise its embrionary development. The musculoskeletal system in its whole originates from the mesoderm. Early in the development, part of this embryonic sheath differentiates into a condensed group of cells called pre–chondrogenic tissue (Spitz & Duboule, 2001). This cartilaginous tissue will rapidly expand and ramify, giving rise in very few weeks to a scaled version of the adult skeleton. Gradually, through a process called endochondral ossification, the cartilage tissue will be replaced by bony tissue. This transition is promoted by the growing vascular system which gradually infiltrates the embryonic cartilage. Through the new arteries and capillaries, monocytic cells will arrive and start to degrade the cartilaginous extracellular matrix (ECM) through the secretion of large amounts of metalloproteinases. In parallel, the resident fibroblasts –called osteoblasts–will start secreting collagen type I fibers and hydroxyapatite–like calcium phosphate to replace the previous ECM (Bueno & Glowacki, 2009). This new matrix will have the same composition as the adult bone and, therefore, it will already provide the embryo the structural properties that are characteristic of this tissue.

In parallel to the ossification process, certain groups of cells located in very specific areas of the embryonic cartilage called interzones will initiate an alternative differentiation program (Figure 2.2). They will start by losing their ability to synthesize ECM and they will gradually acquire a flattened morphology. These interzone cells will finally die by apoptosis and will give rise to a tissue cavity that will be the future intraarticular space (Francis– West *et al.*, 1999). At each side of this cavity, two dense populations of cells will then multiply and originate the opposing joint cartilage tissues. Surrounding the intraarticular space and the former cartilage tissues, a group of prechondrogenic cells will differentiate into the articular capsule and the tendon tissues that will finally bind to the adjacent muscular tissues. Linning the inner surface of the articular capsule, a slim but highly vascularized connective tissue will begin to form. This tissue will be responsible for the synthesis of a viscous substance that will fill the intraarticular space and lubricate the opposing cartilages. Since this fluid macroscopically resembles the white of an egg (in Latin *syn ovia*) it was first called synovial fluid, and the producing tissue the synovial membrane.

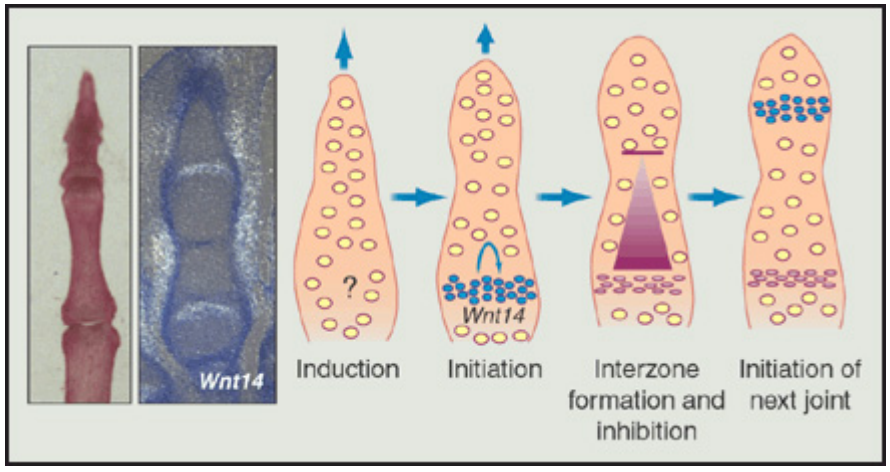


Figure 2.2: Articulating joint formation. Taken from Spitz and Douboule (Spitz & Duboule, 2001). Schematic representation of the formation of the embryonic chick limb. The formation of joints in the embryo is a sequential process involving many yet unknown mechanisms.

Fibroblasts are the main cellular type in the synovial joint

The synovial joint is composed by three types of connective tissue: cartilage, subchondral bone and the synovial membrane. All connective tissues in the human body have one common cellular type: the fibroblast cell. Fibroblasts are responsible for most of the ECM synthesis that is characteristic of these tissues. Given that they are more easily obtained and cultured than other human cell types they have been much used in *in vitro* studies. Until recently, however, fibroblasts were thought to be a rather homogeneous and "dull" cell, carrying out just mere structural functions in the organism. Recently, new evidence has proportioned a more sophisticated view on the roles of fibroblasts. A comprehensive study of the transcriptional profiles of a large panel of human fibroblasts using microarray technology (Chang *et al.*, 2002) demonstrated that this cell type is in fact a very heterogeneous family. The consequences of this discovery are very important at many levels. It means, for example, that experimental results obtained with one fibroblast type need not be reproducible in another fibroblast type, and consequently it raises the importance of carefully selecting the experimental cell type. There is also increasing evidence that fibroblasts have fundamental roles in immunity and in the generation of malignant tissues. For example, the development of the extensive immune cell repertoir requires the interaction with resident stromal cells in the bone marrow (Wilson & Trumpp, 2006). Also, there is increasing evidence that fibroblasts residing at the different connective tissues in the organism are fundamental for the signal cross–talk that takes place during an immune response (Pierer *et al.*, 2004). In cancer, there is also evidence that the tumor progression is less an absolutely autonomous process and it crucially depends on the interaction with the cells of the surrounding connective tissue (Elenbaas & Weinberg, 2001).

Cartilage fibroblasts, the chondrocytes, are the unique cell type present in this tissue. The cartilage has no vascular system and, therefore, the metabolism of chondrocytes depends on the diffusion of nutrients from the synovial fluid. The chondrogenic progenitors are placed in the basal zone, limiting with the subchondral bone. They are flat cells that secrete very little ECM but, as they differentiate into mature chondrocytes, they start acquiring a rounded shape and producing high amounts of ECM, mainly, collagen II fibers and aggrecan. The composition

of this matrix is essential for the functions of the cartilage: minimal friction and maximal resilience. The negative charges of the aggrecan proteoglycan attract large amounts of water molecules which provide the tissue the necessary levels of compressibility that are associated with the articular movement and scaffolding under gravity. In RA, the chronic inflammation state leads to the destruction of this tissue by different mechanisms, including synovial fibroblast invasion, increased synthesis of matrix metalloproteinases (Page-McCaw *et al.*, 2007) and by the cytokine induction of chondrocyte apoptosis (McInnes & Schett, 2007) (Figure 2.3). The loss of cartilage in RA causes the friction between the adjacent bone tissues, which gradually leads to further structural damage, loss of function and pain.

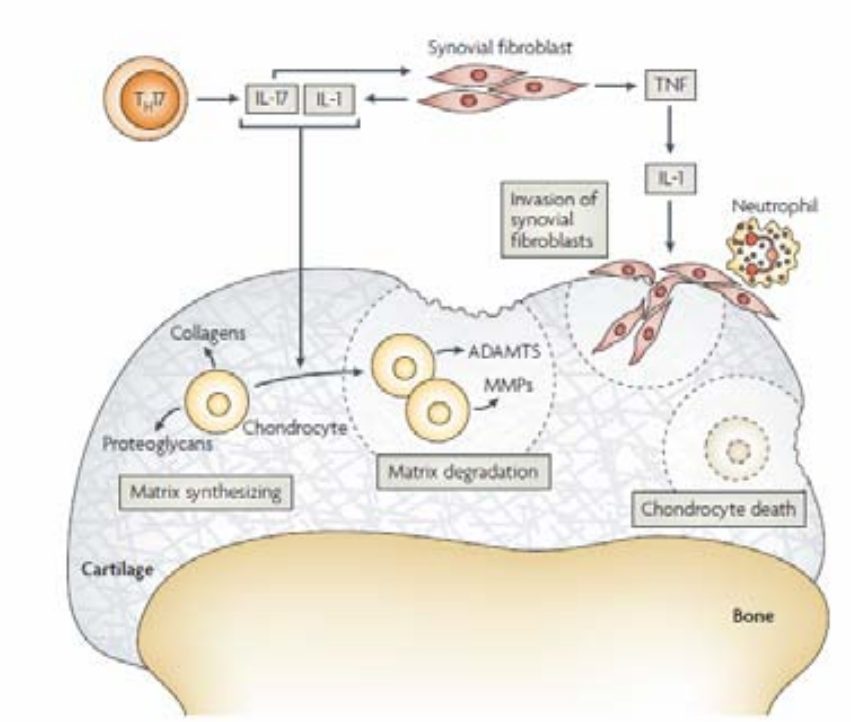


Figure 2.3: Pathways regulating chondrocyte activation and cartilage degradation in rheumatoid arthritis. Taken from McInnes and Schett (McInnes & Schett, 2007). Cartilage tissue degradation in RA is a multistep process based on the simultaneous activation of different cell types.

The process of endochondral ossification that takes place during the embri-

onary development removes almost all the cartilage tissue from the skeleton. Only the distal regions of the bones, the epiphyses, will retain this type of tissue up to the adult phase of the organism. Specialized macrophages called osteoclasts will destroy the cartilage whilst tissue–specific fibroblasts, the osteoblasts, will synthesize the new ECM. This cellular duet will remain active throughout all the life of the organism maintaining a constant renewal of the tissue. Interestingly, when an adult bone is fractured, the healing process recalls the embrionary process: first chondrogenic precursors fill the lesion with cartilage tissue which is subsequently replaced by new bone tissue through the osteoblast–osteoclast system (Page-McCaw *et al.*, 2007). In RA, the increased production of proinflammatory cytokines like RANKL, TNFα, IL1β or IL17 breaks the balance of this system in favour of an increased number of active osteoclasts (Kong *et al.*, 1999). This cytokine–mediated osteoclastogenesis is going to be responsible for the observed subchondral bone erosion characteristic of RA patients (Figure 2.4). In some extreme cases, the bone can actually disappear which, in the case of the cervical synovial joints, could lead to paralysis and death.

Figure 2.4: Typical subchondral bone lesions in the hand joints of RA patients. The osteoclasts front, promoted by proinflammatory cytokines, gradually erodes the bone and forms cavities (left image, red arrow) which are filled by the soft proinflammatory tissue (i.e. pannus). If the chronic inflammation is not controlled, this will finally destroy the subchondral bone leading to joint luxation (right image, asterisk) or fusion.

Similar to the bone, the synovial tissue is also composed by fibroblasts (also

called "type B synoviocytes") and macrophages (also called "type A synoviocytes"). These two cell types concentrate in the intimal layer of the synovial membrane, facing the intraarticular space filled with synovial fluid (Figure 2.5). In this way, synovial fibroblasts can more easily refill the intraarticular space with hyaluronic acid and lubricin, the molecules responsible for the viscous and lubricating properties of the synovial fluid (Khurana, 2009). Also, macrophages can capture and remove debris from this fluid matrix, thus maintaining the desired physicochemical properties of the synovial fluid. In its basal layer, the synovial membrane is rich with vascularization. The diffusion of plasma from the synovial vascular network is essential for the maintenance of the metabolic necessities of the articular cartilage.

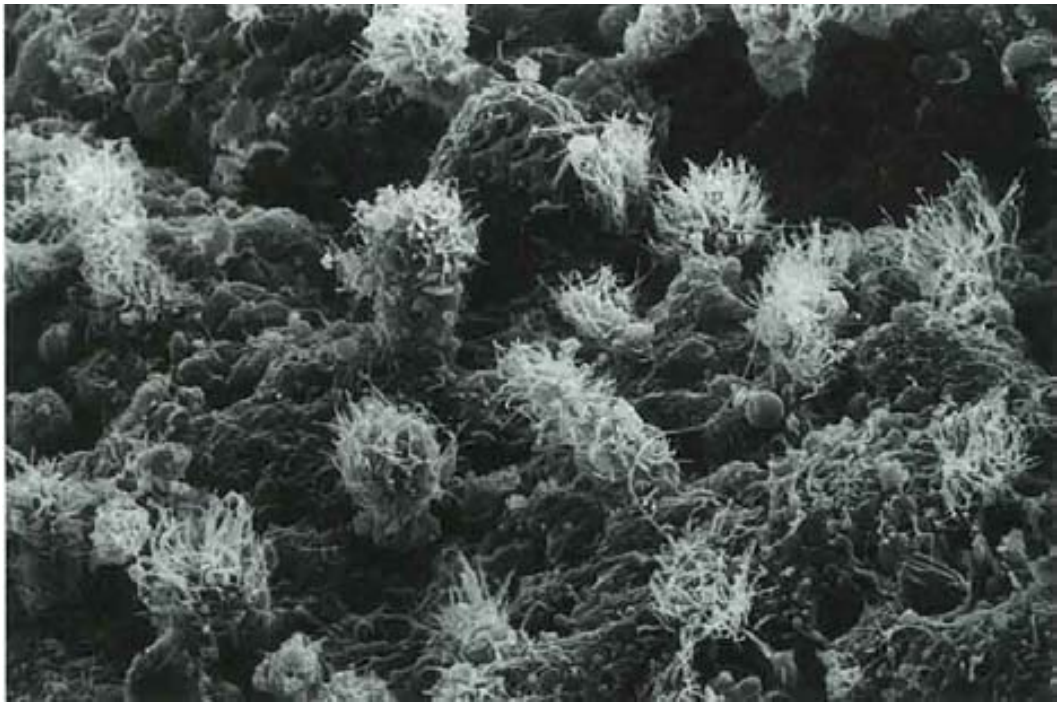


Figure 2.5: Electron micrography of the synovial membrane. (Taken from Shikichi et al. (Shikichi, Kitamura et al. 1999)). Synovial fibroblasts secrete large amounts of hyaluronic acid and lubricating proteins into the synovial fluid. In the image, synovial fibroblasts of the horse project microvilli into the intraarticular cavity, whereas synovial macrophages have a spherical shape and have lamellipodia.

The synovial membrane, a rather acellular tissue in normal conditions, be-

comes hyperplasic in RA. Whilst the intimal layer is still composed by fibroblasts and macrophages, the subintimal zone is colonized by multiple elements of the innate and acquired immune response including macrophages (Firestein and Zvaifler 1997), mast cells (Woolley, 2003), CD4+ T lymphocytes (Kremer *et al.*, 2003), CD8+ T lymphocytes (Haworth *et al.*, 2008), Natural Killer cells (Perricone *et al.*, 2008), B lymphocytes and plasma cells (Edwards *et al.*, 2004). All these cell types interact through multiple cytokine networks that lead to the chronic state of inflammation characteristic of RA. To date, more than 20 different cytokines are considered to be "key" in the pathogenesis of RA (McInnes & Schett, 2007), reflecting the complex nature of the disease but also the level of ignorance that we still have on the mechanisms underlying this pathology.

From all the cell types in the RA synovial membrane, synovial fibroblasts (SFs) seem to have a leading role in joint destruction (Meyer *et al.*, 2006). They display a tumor–like phenotype (Firestein, 1996) with the capacity to attach to the neighbouring cartilage and deeply invade its ECM (Huber *et al.*, 2006). SFs synthesize large amounts of TNFα, IL1β and IL6 proinflammatory cytokines which induce the production of ECM degrading enzymes like matrix metalloproteinases, aggrecanases and cathepsins (Pap *et al.*, 2000). Also, they promote the migration, activation and survival of T and B lymphocytes into the synovial membrane by expressing the required chemokines and cytokines (i.e. TNF, IL16, IL15, BAFF and IL7). RANKL, the main cytokine promoting osteoclastogenesis, is produced by T cells in the synovial membrane that have been activated by SFs expressing the CD40 receptor ligand (Lee *et al.*, 2006). Finally, SFs promote angiogenesis in the synovial membrane by the production of vascular growth factors like VEGF, bFGF, oncostatin M and IL18 (Firestein, 1999). For all these reasons, the characterization of the molecular pathways and regulatory mechanisms of the SF in RA will be essential to understand the initiation and progression of this disease.

Once we have viewed the elements of the disease scenario we will now describe the process of constitution of RA as a distinct clinical entity. We will then revise which are the different complex immunopathological features of RA and how they were characterized.

2.3 Rheumatoid Arthritis as a clinical entity

The first step to find a solution to a problem is to have a clear definition of the problem. Hippocrates (Greece, V–IV centuries B.C.) was aware of this notion and applied with unprecedented rigour to the problem of human healing. In his view of health, which prevailed in occidental medicine up to the seventeenth century, there was an equilibrium between the four "fundamental fluxes" ("*rheuma*" in Greek) that constitute the human body. If one such fluxes accumulated in an organ, then disease developed. In the middle ages, for example, most articular diseases were commonly classified as gutta (drop in Latin), exemplifying the hippocratic notion of the abnormal accumulation of fluids in the joint.

In the seventeenth century, the French physician Guillaume de Baillou was the first to use the term "rheumatism" to specifically describe the accumulation of harmful "fluxes" in the joints. Some years later, Thomas Syndenham, considered the "English Hyppocrates" for its observational capacity and methodological rigour, differentiated two forms of arthritis, one chronic and destructive which he called "rheumatic gout" (i.e. the equivalent of RA) and another acute that generally occurred in young individuals which he called "rheumatic fever". From this time onwards, RA was successively reclassified under different names like "chronic rheumatism", "rheumalgia" or "scorbutic rheumatism" which added little more than confusion. It was finally the English physician Alfred Garrod who in his 1859 medical treatise used for the first time the term "Rheumatoid Arthritis" to design and define this clinical entity (Storey, 2001). However, it was not until 1922 that the English National Health Services accepted the use of this clinical definition. Why did it take more than 60 years to accept RA as disease entity on its own? This delay was most probably due to two fundamental aspects of RA: the lack of a specific clinical diagnostic test and the high level of clinical heterogeneity. These two aspects are still present in RA and they are fundamental to understand the difficulties associated with this heterogeneous disease including the lack of an early diagnosis or the lack of replicability of many research findings. The latter was and still is a matter of much concern between rheumatology specialists. For this reason, in 1957 the American College of Rheumatology (ACR) established the first standardized diagnostic system for RA (Ropes *et al.*, 1957). It was based

on a set of 11 clinical criteria which were used to categorize patients either as "definitive" RA (6 $\geq$ criteria), "probable" RA (3 to 5 criteria) or "possible" RA (using other aspects). Only one year later, a fourth category representing the most specific form of phenotype was included: "classic" RA, where $\geq$7 criteria had to be fulfilled. The ACR 1957 diagnostic criteria for RA prevailed in the clinical practice of Rheumatology for the next 30 years. Its extended use greatly improved the acquisition and diffusion of knowledge in the clinical research of RA and, consequently, patients benefited from an earlier diagnostic as well as a better targeted treatment. However, its extended use also showed several weaknesses. For this reason, the ACR published in 1987 a reformulation of the diagnostic system of RA (Arnett *et al.*, 1988). This new system was now constituted by 7 criteria from which a patient had to fulfil at least 4 to be diagnosed as RA (Figure 2.6). This time, no intermediate qualifications of the disease were considered.

Criterion	Definition
1. Morning stiffness	Morning stiffness in and around the joints, lasting at least 1 hour before maximal improvement
2. Arthritis of 3 or more joint areas	At least 3 joint areas simultaneously have had soft tissue swelling or fluid (not bony overgrowth alone) observed by a physician. The 14 possible areas are right or left PIP, MCP, wrist, elbow, knee, ankle, and MTP joints
3. Arthritis of hand joints	At least 1 area swollen (as defined above) in a wrist, MCP, or PIP joint
4. Symmetric arthritis	Simultaneous involvement of the same joint areas (as defined in 2) on both sides of the body (bilateral involvement of PIPs, MCPs, or MTPs is acceptable without absolute symmetry)
5. Rheumatoid nodules	Subcutaneous nodules, over bony prominences, or extensor surfaces, or in juxtaarticular regions, observed by a physician
6. Serum rheumatoid factor	Demonstration of abnormal amounts of serum rheumatoid factor by any method for which the result has been positive in <5% of normal control subjects
7. Radiographic changes	Radiographic changes typical of rheumatoid arthritis on posteroanterior hand and wrist radiographs, which must include erosions or unequivocal bony decalcification localized in or most marked adjacent to the involved joints (osteoarthritis changes alone do not qualify)

* For classification purposes, a patient shall be said to have rheumatoid arthritis if he/she has satisfied at least 4 of these 7 criteria. Criteria 1 through 4 must have been present for at least 6 weeks. Patients with 2 clinical diagnoses are not excluded. Designation as classic, definite, or probable rheumatoid arthritis is *not* to be made. See Table 3 for definitions of abbreviations.

Figure 2.6: Table of 1987 revised ACR criteria. Excerpt from the original 1988 article in which the American Rheumatology Association committee defined the new diagnostic criteria (Arnett *et al.*, 1988).

2.4 Rheumatoid Arthritis: immunopathology

Rheumatoid Arthritis is a disease where the immune system is chronically activated in the synovial joints. Thus, the advances in the knowledge of the pathophysiological mechanisms of RA have been inevitably linked with the progress in the understanding of immunology. One crucial step in immunology's timeline

was the identification at the end of the nineteenth century of blood as the essential tissue for immune activity by Emil von Behring and Shibasaburo Kitasato (Kantha, 1992). Behring and Kitasato formulated the humoral theory of immunity which states that there are elements in the human serum that are able to bind bacterial toxins and cancel their harmful action. Some years later, Jules Bordet showed that this defence system could be subdivided into two subtypes according to the resistance to heating. Based on this categorization, Paul Ehrlich hypothesized that the thermorresistant component was the product of a specific cellular recognition system which he called "antibody". According to Ehrlich, the antibody would have a double function: to bind and neutralize the invasive agent (called "antigen") and to bind to the thermolabile component of the serum. The latter process, he devised, would be a mechanism to increase the neutralizing efficacy of the antibody and consequently he called the thermolabile substance "complement".

2.4.1 The discovery of the Rheumatoid Factor and autoimmunity

The recognition of the serum as an essential executor of the immune response stimulated the study and identification of different serum reactivity processes associated to human diseases. Two months before the Nazi occupation of Norway during the Second World War, Norwegian researcher Eric Waaler published a manuscript (Waaler, 1940) in which he described the distinct capacity of serum from RA patients to agglutinate sheep red blood cells compared to control individuals. Later, American immunologist Noel Rose rediscovered the same agglutinating factor and concluded that it could be essential in the diagnostic of the disease (Rose *et al.*, 1948). A few years later, Waaler's and Rose's factor was called "Rheumatoid Factor" (RF) and it was not until 1957 that it was characterized as an anti–IgG antibody.

An antibody binding to other antibodies? It was clear that some new biological process was being discovered which seemed to confront to the basic principle of immunity, that is, the discrimination of "self" from "non–self". Up to that moment, immunology researchers had been working with the conviction that

there was a principle of no autoaggresssion (or "horror autotoxicus" as Ehrlich described it). However, there was increasing evidence that in many human diseases the activation of immune system was not associated to the presence of an invasive microorganism. Finally, the works of Roitt and Rose at the end of the 50's clearly demonstrated the existence of autoimmune-type reactions in the human pathology. The first author observed that mixing serum of patients with Hashimoto's disease (an autoimmune disease of the thyroid gland) with purified thyroglobulin, he could observe red blood cell precipitates, clearly demonstrating the presence of autoantibodies against this thyroid–specific protein in these patients (Campbell *et al.*, 1956). Rose also studied Hashimoto's disease but, instead, he showed that rabbits developed autoimmune thyroiditis after they were immunized against their own thyroglobulin protein (Rose *et al.*, 1965). In an attempt to formalize this new aetiology phenomenon, German immunologist Ernest Witebsky published in 1957 a group of 3 postulates defining autoimmune diseases (Rose & Bona, 1993) (Table 3.1).

Table 3.1 Witebsky's postulates of autoimmunity

- An autoimmune reaction is identified in the form of autoantibody or cell-mediated immune reaction
- The corresponding antigen is known
- An analogous response causes a similar disease in experimental animals

However, as Waaler already noted in his 1940 manuscript, not all RA patients were positive for RF. Also, patients with other diseases could produce this autoantibody in their sera. Therefore, RA did not fall into Witebsky's established definition of autoimmune disease.

2.4.2 HLA association with RA

One key aspect in the understanding of the immune response was the identification of the clonal selection process. This mechanism discovered by Francis Burnett in 1959 and later demonstrated by James Gowans, describes the principles by which the cells of the acquired immunity, the lymphocytes, operate (Rajewsky, 1996). The clonal selection theory can be divided into several phases: first, there is an initial repertoire of lymphocytes that are able to recognize all possible 3–Dimension molecular structures. In the first stages of the life of the organism, those lymphocytes that recognize "self" molecular structures are eliminated. The surviving lymphocytes will then migrate to the periphery of the body where they will wait for the antigen to appear. When a microorganism comes to scene, those lymphocytes that are able to recognize its 3D structures (i.e. antigens) will proliferate. Finally, these proliferating lymphocytes will give rise to two different subpopulations of cells: the effector cells which will direct antigen elimination and eventually will disappear, and memory cells that will be stored in the organism so that faster and stronger immune responses are elicited in future infections of this same microorganism.

In 1974, Rolf Zinkernagel and Peter Doherty showed that the last phase of the clonal selection theory, the lymphocyte recognition of the antigen, was not performed directly on the infecting microorganism (Zinkernagel & Doherty, 1974). Instead, the microorganism had first to be processed by other types of cells (i.e. Antigen Presenting Cells or APCs) and then exposed to lymphocytes by a specific group of proteins called the Major Histocompatibility Complex in mice or the Human Leukocyte Antigen (HLA) in humans. Once discovered this HLA–restricted lymphocyte recognition mechanism, researchers started asking themselves if this system would be the key to the origin of autoimmune diseases. In 1976, American physician Peter Stastny saw that peripheral blood mononuclear cells (PBMCs) isolated from RA patients cocultured with PBMCs from control individuals had normal growth rates but when they were cultured together with PBMCs of other RA patients they had very poor proliferation levels (Stastny, 1976). Importantly, Stastny saw that this lack of stimulation had a strong correlation with their HLA serotype. Although HLA typing at that time could only be performed at the

protein level, it was the first strong evidence that there was genetic variation associated with RA.

The association of genetic variants to disease risk can be a very useful way to discover fundamental aspects of the disease aetiology. In some cases, however, the identification of the precise role of the genetic variation with the disease predisposition can be very challenging. One such example is the association of HLA variability with RA. More than 30 years after the association of the HLA locus with RA and still there is no convincing explanation of its implication in the disease. Initially it was thought that, similar to other autoimmune processes like Hashimoto's thyroiditis, there should be an organ–specific antigen which could raise an immune reaction in predisposed individuals (Gregersen *et al.*, 1987; Verheijden *et al.*, 1997). Since no clear arthritogenic antigen has been identified, many other alternative hypotheses for the implication of the HLA have been formulated but, to date, there is not enough evidence for any of them (Firestein, 2003). Nonetheless, the recognition of a T cell mediated effect has been very useful in the development of efficacious therapies in RA. In particular, the study of Collagen Induced Arthritis (CIA) (a T cell model of RA in mice) showed that the blockade of the TNFα cytokine was very efficient reducing the inflammation and joint damage (Williams *et al.*, 1992). This evidence was the firm rationale from which therapies blocking systemic TNFα started to be evaluated in RA patients in the mid 90's (Lipsky *et al.*, 2000). Actually, TNFα blocking therapies have proven a major success in RA treatment and are a clear landmark in the treatment of chronic inflammatory diseases (Smolen *et al.*, 2007).

2.4.3 Cytokines and new T cell subtypes associated with RA

Multicellular organisation is possible through the existence of powerful communication networks between cells. The nervous system can communicate through long distances through axons, and the endocrine system, through mediators that travel in the blood can reach distant targets. However, there is a short–range communication system that is fundamental for multicellular cross-talk and, especially for immune cell communication. These messengers are small molecular

weight proteins called cytokines, although previously they have also been known as interleukins, interferons or colony–stimulating factors (CSFs) The development of essential molecular biology techniques like DNA cloning during the 70s contributed to a boom in the identification of multiple cytokines during the 80s that is still ongoing. The discovery of these diverse communication systems has been extraordinarily influential in our understanding of many pathological processes. Importantly, they have significantly improved the treatment of many common diseases like cancer (i.e. haematopoietic reconstitution with CSFs after chemotherapy) or autoimmune diseases (i.e. prevention of bone erosion in RA with TNFα blocking agents).

One of the first consequences of the knowledge of the diverse cytokine repertoire was the acknowledgement of distinct CD4+ lymphocyte (T–Helper) subsets according to their cytokine profile. In 1986, Robert Coffman and Tim Mossman demonstrated that there were two major CD4+ subtypes: cells expressing interferon gamma (IFNγ), IL2 and TNFα which they called T_H1, and cells expressing IL4, IL5 and IL10 which they called T_H2 (Cherwinski *et al.*, 1987). T_H1 cells are fundamental in the defence against virus and bacteria since they are able to activate macrophages, cytotoxic (CD8+) T cells, and can induce the production of specific antibodies that neutralize (i.e. opsonize) these infecting agents. T_H2 cells are key elements in the defence against parasitic and mucosal infections since they are able to induce B cells to produce large quantities of antibodies specific for this type of microorganisms. In the course of an infection, dendritic cells migrate to the secondary lymphoid organs where they present the processed antigens to the naïve T cells through the HLA system. Once the CD4+ cells specific for those antigens have activated, they proliferate and migrate to the infection focus where, according to the local profile of cytokines, they differentiate into the T_H1 or the T_H2 phenotype. The T_H1/T_H2 theory was a key intellectual breakthrough not only for the definition of new subphenotypes but also because it introduced the notion that cells of one T cell subtype can inhibit the formation of cells of the other subtype (Coffman, 2006).

On the basis of studies in rodent models, RA was initially though to be a T_H1–cell mediated disease (Courtenay *et al.*, 1980; Remmers *et al.*, 2007). However, like in many other autoimmune diseases, the evidence obtained from human

samples did not fit the established T_H1/T_H2 model. Although high levels of TNFα cytokine were detected in RA synovial biopsies, only small amounts of IL2 or IFNγ could be found. Recent findings, however, are broadening our vision on the CD4+ subtype paradigm. In 1995, Sakaguchi and colleagues (Sakaguchi *et al.*, 1995) showed that a particular subset of CD4+ lymphocytes was essential to prevent the development of autoimmune processes in mice, including arthritis. These cells express the IL2 receptor (i.e. CD25) and are called T regulators or, more commonly, Tregs. Tregs have been also found in humans (Misra *et al.*, 2004) and, obviously are now an objective of intense research in many autoimmune diseases including RA (Ehrenstein *et al.*, 2004). Its implication in RA aetiology however is still not clear. Several studies, for example, have reported an enrichment of Tregs in the synovial fluid of RA patients (Raghavan *et al.*, 2009) but there is notable controversy regarding the frequency of CD4+CD25+ Tregs in the peripheral blood (Sarkar & Fox, 2008). Some authors have suggested that Tregs in RA could have lost their suppressive capacity although several other researchers show different evidence. More promising it could be the link between Tregs and its relation to the response to anti–TNFα treatment. In this regard, we have shown (Julià *et al.*, 2009b) that patients that don't respond to TNFα blockade therapy tend to have lower CD4+CD25+ Tregs at baseline compared to those patients that will respond. In this sense, we hypothesize that the reduction of the high TNFα levels in RA will be beneficial only to those patients with sufficient baseline Treg cells to suppress the autoimmune activity.

More recently, a new CD4+ T cell subtype has been identified that could also be crucial to understand the complex immunophatology of RA. IL17 was identified in 1995 as a T cell derived cytokine that promotes inflammation and neutrophil activation (Yao *et al.*, 1995); however it was not until 2005 that the cell source of this proinflammatory cytokine was identified in mouse (Harrington *et al.*, 2005). The IL17 producing lymphocytes were consequently named T_H17 and their role has been in various situations associated with inflammation and ECM destruction. T_H17 cells are characterized by the expression of the proinflammatory cytokines IL17A, IL17F and IL22. IL17A has been found in the synovial membrane (Chabaud *et al.*, 1999) and the synovial fluid (Kotake *et al.*, 1999) of RA patients and it is associated with disease severity (Kirkham *et al.*,

2006). Its deficiency has also been found to have profound antiarthritic effects in mice (Lubberts *et al.*, 2004; Nakae *et al.*, 2003). A recent line of research is working on the hypothesis that the increased IL1β and IL6 production from synovial fibroblasts observed in RA would be the responsible for the shift of CD4+ T cells towards this subtype. Once sufficient numbers of T_H17 are been generated, the inflammatory process may then become autonomous with IL17 dominance and independence of TNFα or IL1β activation (van den Berg & Miossec, 2009).

The discovery of new pieces of the complex immunology puzzle is still ongoing. For example, apart from the Treg and T_H17 cells, there are new CD4+ subtypes that are being characterized like the T follicular helper cells, the T_H22 cells and the T_H9 cells (Bluestone *et al.*, 2009). Will these new cell types be also fundamental in the aetiology of RA? How will they fit into the complex scenario of activated fibroblasts, antibody production and chronic lymphocyte infiltration of RA? It is clear that this and many more fundamental answers will need the detailed study of multiple molecular and cellular mechanisms. Perhaps, the identification of genetic variation associated to disease risk will be fundamental to disentangle the key immunological factors that cause RA.

Chapter 3

Genetics and Genomics

3.1 The concept of heredity and Mendel's laws

Heredity is a relatively modern concept in human history. Hippocrates, for example, conceived that both parents produced a seed ("semen" in Greek), that intermingled to produce the embryo. However, he could not find a satisfactory explanation of why certain traits appeared in the offspring and others did not. For more than 1,500 years, researchers weren't able to find any consistency in their observations and, therefore, the heredity concept wasn't developed. In the seventeenth century, the introduction of theory, dissection and experimentation into the study of biological phenomena lead to important advances in the understanding of the generation of life, namely, the identification that egg cells in female organisms and spermatozoa in male organisms as essential elements for reproduction. Still, however, nobody accounted for the underlying heredity of traits. The definite advance came thanks to the industrial revolution in the eighteenth and nineteenth centuries. This key period in Europe's history started in the United Kingdom and, a part from other important advances, it promoted a major improvement in agricultural production. Robert Blakewell, a British agriculturalist, was the first to consciously select and cross specimens with the best productive qualities from around the UK to generate better livestock, that is, he discovered selective breeding (Cobb, 2006).

In the nineteenth century, Brno (Czech Repuplic) was considered one of the European capitals of textile and sheep breeding. As such, there was an inten-

sive intellectual life around the problem of breeding and trait production in the descendants. It is under this particular environment that the Czech monk Gregor Mendel was stimulated to produce his fundamental studies on the heredity of traits. In his 1866 article *"Versuche über Pflantzen–Hybriden"* on plant breeding, Mendel discovered two fundamental aspects of heredity:

- Organisms have two factors that give rise to any particular trait, one coming from each parent. When organisms have offspring, any of the two possible factors is chosen randomly and transmitted (i.e. segregated) to them.
- The transmission of a factor from one parent occurs independently of the transmission of the factor of the other parent, independently for each offspring and independently of each trait.

These two key observations are considered the birth of genetics and are known as the "segregation" and "independent segregation" laws of genetics, respectively. To arrive to these conclusions Mendel had performed an exemplar approach to the study of heredity. First he used a model organism that was easy to grow and to selectively reproduce: the pea plant. Second, he registered the results on the crossing of more than 10,000 of such plants, accurately annotating the phenotypic features observable at each generation. Third, he used homogeneous breeds, that is, plants that systematically expressed the same type of trait to precisely observe the effects of crossing in the next generations. Using these approaches he finally was able to confirm that the origin or genesis of the traits needed to be binomial.

Unfortunately for him, Mendel's work did not have impact during his lifetime. It was in 1906, during the English Royal Horticultural Society conference, when his findings were finally acknowledged. During this conference, Cambridge biologist William Bateson, who is the responsible for Mendel's rediscovery, used for the first time the term "genetics" to describe this new phenomenon. Apart from disseminating Mendel's research, Bateson itself made also substantial contributions to genetics. Using the Mendel's research methodology he studied the

inheritance of traits in pea and chicken dihybrids, that is, hybrids (i.e. heterozygotes in modern notation) for two different traits. He found that, for certain combinations of two traits, the observed frequencies in the offspring were different from the predicted frequencies according to Mendel's laws. He saw, however, that some of these deviations could still be explained by the heredity lawss if the existence of a more complex mechanism was assumed. Bateson found that some observed inheritance patterns could be explained by the presence of a masking effect of one trait upon another. He termed this effect "epistasis" (in Greek "to stand upon").

3.2 Hardy–Weinberg, the chromosomal view of heredity and Fisher

The rediscovery of Mendel's laws posed new questions in the characterization of heredity. One of these questions was to understand the consequences of the segregation law in a population with the absence of perturbing selection forces. German pysician Wilhem Weinberg and British mathematician Godfrey Hardy independently answered to this problem: they demonstrated that no matter what genotype frequencies are in a population, these will always reach a stable frequency with only one generation of random mating. This principle was called the "law of pancmictic equilibrium" but now is more commonly known as the Hardy–Weinberg equilibrium. The identification of this mathematical relation has had profound implications in the genetic study of populations as well as practical utilities in association studies of complex traits (Balding, 2001).

In the study of the inheritance of dihybrid crossings William Bateson together with Richard Punnet concluded that, for some traits, there needed be some sort of physical coupling between them. This hypothesis was confirmed some years later by the heredity studies of American geneticist Thomas Morgan using Drosophila as a model organism. Morgan, who initially was critical with the chromosomal theory of heredity, ended up suggesting that those traits showing coupling should be located in the same pair of homologous chromosomes. In order to account for the observation of non–parental combinations in the offspring (i.e. recombinants) he proposed that when two homologous chromosomes paired during meiosis, they

could occasionally exchange parts. Using these principles, one of Morgan's students, Alfred Sturtevant, built up in 1913 the first chromosomal map of an organism (Sturtevant, 1913). For this purpose, he used the estimated number of cross–overs as an indirect measure of physical distance within the chromosome.

Charles Darwin's evolutionary theory, published in 1859, is one of the most important milestones in human knowledge. In this manuscript, Darwin described the principle of natural selection which elegantly combines the struggle for life, heritable variation and differential reproduction. Stemming from these discoveries, most biological researchers started to study the heredity of quantitative traits. However, even after the rediscovery of Mendel's laws, Darwinians were reluctant to accept a particulate view of heredity that seemed to clash with the principle that selection acted on variations in quantitative characters. It was not until 1918 that the English statistician, Ronald Fisher, finally reconciled both views in his work *"The Correlation between Relatives on the Supposition of Mendelian Inheritance"*. Fisher showed that the apparent continuous variability observed for many traits could be explained by the cumulative effects of many Mendelian inherited factors acting together in an additive manner. Notably, in his 1918 paper, Fisher accounted for those particular genetic models in which the cumulative effects of genes (which he called Mendelian factors) that do not follow the additive model:

"(...) *A similar deviation from the addition of superimposed effects may occur between different Mendelian factors. We may use the term Epistacy to describe such deviation, which although potentially more complicated, has similar statistical effects to dominance.*"

Thus, whilst Bateson used the term "epistasis" to describe a masking or reversing effect between two genes, Fisher used the term "epistacy" to describe the statistical interaction between two genes. The different meaning between both terms (qualitative vs. quantitative) and the fact that just the term epistasis prevailed, has generated some confusion between statistical and biological researchers (Cordell, 2002).

3.3 DNA and the basic dogma of biology

During the industrial revolution, chemistry had been rapidly evolving. In 1869, German physician Friedrich Miescher had already isolated an organic acid with high phoshporous content from cell nuclei which he called "nuclein". Some years later, the also German physician Albrecth Kossel elucidated the chemical nature of both DNA and RNA which he called "thymonucleic" and "yeast nucleic" acids, respectively. By then it was known that the nucleic acid molecule was made of two purines (adenine and guanine), two pyrimidines (cytosine and thymidine in DNA or uracil in RNA), a phosphate group and a pentose sugar. In 1909, chemist Phoebus Levene identified the pentose sugar in the "yeast nucleic acid" to be a ribose, and so this substance was now called ribonucleic acid (RNA). However, it took Levene's group 20 more years of research identify the 2'–deoxyrribose as the conforming pentose of the "thymonucleic acid", which was then called deoxyrribonucleic acid (DNA).

In his 1944 book "What's Life" Physicist Erwin Schrödinger suggested that an "aperiodic crystal" should be the basis of hereditary information. This same year, Oswald Avery and coworkers proved that the transforming principle –the chemical principal that is able to transform a non–virulent Pneumococcus strain into a virulent form–was DNA (Avery *et al.*, 1944). Boosted by this discovery Erwin Chargaff performed a detailed chemical analysis on DNA and identified that its nucleotide composition followed some precise rules, that is, there is the same quantity of Thymidines as Adenines as there is the same quantity of Guanines as Cytosines (Vischer & Chargaff, 1948). Therefore, the last step to characterize the "molecule of life" was to determine its 3–Dimensional chemical structure. Finally, molecular biologists James Watson and Francis Crick solved in 1953 this fundamental stereochemistry problem: the DNA has a double–helical structure (Watson & Crick, 1953). With the knowledge of the structure of DNA, research rapidly moved into "cracking" this information encoding system. Francis Crick, Sydney Brenner and other authors were responsible for its final characterization. First, they discovered that RNA was the "messenger molecule" that linked DNA with the protein product. Second, they concluded that the chemical code embodied in a gene consisted in non–overlapping groups of 3 DNA bases (i.e. codons).

By 1966, the codons of all 20 aminoacids necessary for life had been identified. With this discoveries, the central dogma of biology (i.e. the transmission of genetic information from DNA to RNA and from RNA to protein) had been finally established (Crick, 1970).

Once the genetic code was cracked, the DNA molecule became central to biological research. Alan Maxam and Walter Gilbert provided in 1977 the first method for determining the sequence of DNA (Maxam & Gilbert, 1977), to be later superseded by Frederick Sanger's chain–terminating dideoxy method (Sanger *et al.*, 1977). Another fundamental technical breakthrough in the characterization of the DNA sequence was the development in 1983 of the Polymerase Chain Reaction by American biochemist Kary Mullis (Saiki *et al.*, 1985). This method, which allows the exponential amplification of the DNA sequence of interest, has been a crucial to speed the characterization of the DNA sequences of thousands of species. This way, in 1986 the scientific World was confident enough to start one of the most ambitious projects of humanity: the determination of the complete sequence of the Human Genome (DeLisi, 2008).

3.4 Heritability, linkage and linkage disequilibrium

In the 50s epidemiologists were aware that epidemiology, the study of the determinants of disease in populations, had yet not approached genetics. Those diseases following Mendelian inheritance patterns were rare and, perhaps, less "interesting" from an epidemiological perspective. However, with the knowledge of the DNA sequence and the increasing number of available markers, epidemiologists started the difficult quest to study more prevalent diseases but with more complex genetic background. It could be said that genetic epidemiology, as such, began as a new discipline at the end of the 70s with the publication of the first book entirely devoted to it (Morton & Chung, 1978). One of the first objectives that genetic epidemiologists tackled was the determination of the heritability of complex traits. Heritability is a descriptive statistic that refers to the proportion of phenotypic variability of any particular trait that can be accounted for genetic variability amongst individuals (Visscher *et al.*, 2008). Francis Galton, who

was cousin of Charles Darwin and who was interested in separating the "nature" vs. "nurture" from certain human developmental and human traits, pioneered in 1875 the use of twins to evaluate the relative importance of genetic factors in a trait: if a trait was due to genetics, twins under equal nurture should manifest it equally. Some years later, Weinberg realized from his medical student work at the obstetrics clinic that there had to be two kinds of twins: those becoming from a same "egg" (monozygotic) and those coming from two eggs (dizygotic). Therefore, the comparison of the correlation of a trait between monozygotic twins against the correlation between dizygotic twins should provide a good estimate of the size of the genetic effects influencing it.

After the implication of genetic factors in complex diseases it became then necessary to find the gene(s) underlying this risk. However, it took 40 years after Sturtevant maps in Drosophila to characterize the first autosomal linkage in humans. In 1951 Jan Mohr, a Norwegian physician, determined the presence of linkage between the Lutheran and Lewis blood groups (Mohr, 1951). Soon afterwards, American geneticist Newton Morton developed a fundamental statistical method for the detection of linkage in families: the LOD score. Morton himself, used the LOD score to identify one year later the presence of linkage between the gene causing elliptocytosis –a rare monogenic disease affecting red blood cell shape–and the Rh blood type genes (Morton, 1956).

From this moment onwards, the interest in identifying linkage for human diseases grew exponentially, with medical researchers collecting large pedigrees showing high frequency of a disease and trying to link them to the growing number of available enzymes and proteins whose chromosomal location had already been estimated. It was not until 1979, when Solomon and Bodmer used the fragment length polymorphism (RFLP) technique, that DNA was directly used to study linkage of human traits (Solomon & Bodmer, 1979). In this technique, the variations in the DNA sequence become differential targets for bacterial restriction enzymes and, therefore, the digestion fragments can be used as markers to study the correlation with the disease inheritance in a family (Donis-Keller *et al.*, 1987). However, the definite impulse to the study of linkage with familial traits came with the invention of the Polymerase Chain Reaction (PCR) technology. With this technology a new type of DNA polymorphism in the human genome could

be discovered: the microsatellite. Microsatellites are tandem repetitions of short DNA sequences and they presented several advantages: they are ubiquitous in the genome and they have a high degree of polymorphism which makes them much more informative than the previous RFLPs (Weissenbach *et al.*, 1992). With only 400–500 microsatellite markers it was now possible to scan the whole genome of pedigrees to search for regions showing linkage with disease. The drawback of this approach is that, since linkage mapping depends on the number of recombination events –and the number of recombinatorial events per family is of necessity low–, the resolution was quite low (i.e. in the order of several megabases). Nonetheless, this approach had a tremendous success in identifying the genes responsible for many monogenic diseases. The first gene to be mapped using this approach was the gene responsible for Huntington's disease, and was discovered by analyzing the linkage patterns in a large pedigree from a Venezuelan town which had a very high incidence of the disease (Gusella *et al.*, 1983).

The family–based linkage approach showed to be very useful to identify those genes responsible for diseases following Mendelian patterns. However, for those diseases not following Mendel's inheritance rules, the linkage approach seemed to be rather inefficient (Julià & Marsal, 2003). After much frustrated efforts to characterize genes for susceptibility for many common diseases using linkage, it started to become clear that this group of diseases had to have a more complex origin. In this new scenario, complex diseases are not considered to be "caused" by mutations in one gene but instead they are "induced" by a genetic risk background in combination with multiple triggering environmental factors (Cordell & Clayton, 2005). This genetic risk background would be composed by multiple DNA variants at different loci, each adding a small increase in the penetrance upon the phenotype. If this low penetrance variants escape linkage studies, how will we be able to detect them? Eric Lander, a mathematician working at the Whitehead Institute for Biomedical Research in Boston, described in 1996 what he thought it would be the research on human traits after the human genome sequence was available (Lander, 1996). In this illuminating paper entitled *"The New Genomics: Global Views of Biology"*, Lander hypothesized that common diseases were caused by genetic variations of modest effect but, importantly, that these variations should be common in the general population, a theory known

as the "Common Disease–Common Variant" hypothesis. It seemed that genetic susceptibility to common diseases would not be characterized by using linkage on large pedigrees but instead, by using association in case and control samples obtained from the general population.

Lander's vision on the characterization of the genetic basis of human variability was largely based in the increased acceptance of linkage disequilibrium as a powerful analysis tool. Linkage disequilibrium (LD) is the non–random association of alleles within a population, and was first described by American biologist Richard Lewontin in 1960 when studying the dynamics of polymorphisms in different organisms (Lewontin & Kojima, 1960). LD stems from the assumption of a "founder" effect, that is, the spread of a genetic variation in a population from one first individual carrying the mutation (i.e. the founder) after a large number of generations. After hundreds or thousands of successive chromosomal recombinations, only those alleles that are very close to the original mutation will remain highly correlated with it (i.e. in high LD). Therefore, if one finds a polymorphism associated with a disease using a case and control cohort from a population, he can be certain that the causal genetic variation is very close (Cardon & Bell, 2001). Although this concept had been originally proposed by Lander and Botstein in 1986 (Lander & Botstein, 1986), it was not until 1994 that a study on a human disease exploited it to refine the location of the causal gene. In their influential 1996 paper "*The Future of Genetic Studies of Complex Human Diseases*" Risch and Merikangas showed that linkage studies were underpowered to find such genes and that association studies exploiting LD would have more statistical power to identify risk loci (Risch & Merikangas, 1996).

3.4.1 Single Nucleotide Polymorphisms and the development of microarray technology

It became apparent that the primary limitation for conducting genome-wide association analyses was not a statistical one but a technological one. A large number of genetic markers had to be first identified and then genotyped in a cost–effective way. The existence of single nucleotide polymorphisms (SNPs) in the genome had been known since the introduction of the RFLP technique in early 80's. With

time, it was seen that SNPs have very useful properties for association studies: they are highly ubiquitous (1 SNP per 1,000 kb), unlike microsatellites they are highly stable markers (i.e. they "mutate" less throughout generations) and they have greater potential for automation. SNPs had therefore a great potential to become a good biomarker for human complex traits, including diagnosis, prognosis and the response to treatment. For all these reasons they became an interesting target not only for research institutions but also for many pharmaceutical and technological companies. In 1999 the SNP Consortium, a collaboration between private companies and public research institutions (Masood, 1999) was launched with the initial objective to discover 300,000 SNPs from the yet incomplete human genome sequence. The final results exceeded the initial expectations and, in just 2 years 1.4 million SNPs had been already released into the public domain (Sachidanandam, Weissman et al. 2001).

In 2001 Lander and coworkers clearly demonstrated that LD in human populations is not monotonic and, instead, the genome is made up of regions with variable levels of LD (Daly *et al.*, 2001). In those genomic regions with high LD, alleles are so much correlated between them that they can be considered as single blocks or, more commonly, haplotype blocks. The identification of the block–like nature of the human genome was soon viewed as a useful means to increase the analytic power in population–based association studies. If one knew *a priori* the LD pattern of a region of interest, then it would be possible capture most of the genetic variation with just a reduced set of highly informative SNPs (i.e. tagSNPs), and avoid the expense of genotyping redundant markers. In order to characterize this type of variation in the whole genome, an international consortium called the HumanHap project was launched in 2002 (IHC, 2003). Using a collection of individuals from four human populations of very different ancestry (European Caucasians, Japanese, Chinese and Nigerian Yorubans), this project identified thousands of tagSNPs that allowed more powered genetic association studies. Although the first phase of this ambitious project was completed in 2005 (Altshuler *et al.*, 2005), it has successively been extended with the inclusion of new human populations as well as an increased coverage of the genome SNP content largely due to the introduction of high–throughput genotyping technologies

One of the first approaches to increase the genotyping throughput were DNA microarrays. Based on the same principles of Edwin Southern's 1973 "*Southern blot*", microarrays are solid surfaces on which specific DNA sequences are fixed (Southern, 1975). These fragments are then used as probes to capture the complementary DNA chain from a complex mixture (i.e. genomic DNA from an individual, mRNA from a tissue reverse–transcribed to cDNA, ...); after hybridization the non–complementary DNA is washed away and the level of hybridization is then measured, generally, through fluorescence detection. However, the first widely used genomic application of DNA microarrays was not SNP genotyping but mRNA gene expression. Pat Brown's lab in Stanford had devised in 1995 a self–made robot to print DNA probes in a solid support (i.e. a glass slide) and, with these so–called "academic" microarrays, they could characterize the mRNA expression patterns of any cell type or tissue (Schena *et al.*, 1995). Closely, Santa Clara based biotechnological company Affymetrix also developed a microarray devise using approaches derived from the manufacturing of computer processors (Wang *et al.*, 1998). Microarrays were able to interrogate thousands of different DNA sequences and it was just a matter of time that this high throughput would scale up to hundreds of thousands or even millions of probes. Thus, with the completion of the Human Genome Project in 2001, the foundations for Genome-wide Association Studies (GWAS) were laid (Lander *et al.*, 2001).

On February 2007 the first comprehensive GWAS study on a complex disease was published (Sladek *et al.*, 2007). It took 6 years from the Human Genome Project completion and the development of extremely high–throughput technologies to attain this milestone. To date, more than 250 genetic variants associated with complex polygenic traits have been identified using GWAS, confirming Risch 1996 predictions on the high power of this approach (Risch & Merikangas, 1996). Before diving into the impact of the GWAS approach in RA, we will first revise the study of the genetic basis of this disease in the "pre–genomic" era.

3.5 Genetics of Rheumatoid Arthritis

3.5.1 Heritability of RA

As soon as RA started to be a distinct clinical entity, the evidence of an inherited susceptibility component became apparent. In his description of "Chronical Rheumatism" included in the 1806 treatise of Medicine, the English physician William Heberden concluded with a significant phrase "Is this not in some degree hereditary?". Although there was empirical evidence, the lack of a standardized definition of RA, hindered the robust assessment of heritability in RA. In 1928 the German physician J Kroner published the first study reporting an outstanding presence of RA in a single family, where 4 generations of women had developed the disease (Lawrence, 1970). With the gradual approach of epidemiologists to the study of the inherited component in complex diseases, more rigorous approaches were introduced. In 1950, Lewis–Fanning used a predecessor of the actual familial risk statistic to estimate that first order relatives of RA patients had 2 times more probability to develop the disease than matched healthy controls.

During the 50s and 60s several epidemiological initiatives in Europe and the United States began to collect large cohorts of individuals. Using these detailed collections it was now possible to obtain more robust estimates of the degree of heritability in RA. The nationwide Finnish Twin Cohort, for example, collected 4,137 monozygotic and 9,162 dyzigotic twins, from which estimates of 12.3% and 3.5% concordance for RA could be obtained, respectively (Aho *et al.*, 1986). In 1993, a large cohort from the UK, the Arthritis Research Campaign, reported similar estimates of twin concordance: 15.4% monozygotic concordance vs. 3.6% dizygotic concordance (Silman *et al.*, 1993a). However, although the comparison of concordance rates between twins can inform about the presence of a heritable component, it is known to be affected by the prevalence of the disease in the population. For this reason, Fisher's variance components approach to the measure of heritability is actually preferred. Using the data from the two previous European cohorts, the proportion of liability to RA accounted for genetic factors was estimated to be of 60% (53% and 65% heritabilities for Finnish and UK cohorts, respectively) (MacGregor *et al.*, 2000).

3.5.2 HLA genetic association with RA

The findings of the association of HLA serotypes and RA performed by Stastny (Stastny, 1976), did not come unnoticed by geneticists. With the introduction of the recombinant DNA and sequencing techniques during the 70s, the genes from the HLA region were progressively characterized. The extended HLA region as it is now known, covers a total of 7.6 megabases on the short arm of chromosome 6 and harbours approximately 250 protein codifying genes (Horton *et al.*, 2004). One of such genes, *HLA–DRB1*, belongs to the immunoglobulin superfamily which mediates antigen recognition in the acquired immunity. *HLA–DRB1* codifies for the Beta–chain of the heterodimer protein that antigen presenting cells use to present exogenous peptides to lymphocytes. In 1987, the American rheumatologist Peter Gregersen, identified an aminoacidic pattern in this protein that was more prevalent in RA patients than in controls (Gregersen *et al.*, 1987). He hypothesized that the structural variation of this "shared epitope" would influence the interaction of the HLA molecule with the T cell receptor and promote the appearance of RA, either by shaping the thymic selection of lymphocytes or by modulating the antigen presentation during the immune response. To date, however, there is no consistent evidence supporting either possibility, as there is no evidence supporting any other hypothesis (Firestein, 2003).

3.5.3 The determination of the non–HLA risk component: linkage and association

Once the association between the HLA locus and the RA risk had been confirmed, the next question was: are there other genomic regions that contribute to RA susceptibility? For the robust estimation of the contribution of genetic variation in complex diseases, genetic epidemiologist Neil Risch proposed the Relative Risk statistic or λR (Risch, 1990). With this statistic the risk of having a disease in relatives is compared with the probability of having the disease in the general population (i.e. the prevalence). λ_R started to become extensively used to characterize the genetic contribution of genetic loci to complex traits (Risch, 1987). In adult onset diseases like RA, λ_R was generally calculated using the siblings of the affected proband (i.e. λ_{sib}). Using this type of samples, λ_{sib} in RA has

been estimated to lie between 5 and 10 (Julià & Marsal, 2003). From this global estimation, the fraction attributable to the HLA region is around 30% (Cornelis *et al.*, 1998; Deighton *et al.*, 1989; Rigby *et al.*, 1993; Shiozawa *et al.*, 1998). This means that, more than 50% of the genetic variation associated with the risk to RA remained to be discovered.

As soon as the first RFLP maps were available, linkage studies started to proliferate. However, it was not until microsatellites were characterized and faster genotyping techniques like capillary sequencing were developed that the boom of linkage studies in complex traits really began. The first whole genome linkage study to be performed in RA was done in 1998 by Cornelis and coworkers (Cornelis *et al.*, 1998). In this linkage scan, they studied the transmission of 309 microsatellite markers in 97 nuclear families with 114 affected sib pairs. Only the HLA region withstood the multiple test correction considered for these studies (LOD score >3), although other genomic regions showed nominal evidence of linkage (LOD score <3–1). This same year, a similar study in Japanese families (n = 41 affected sib–pairs, 358 microsatellite markers) identified three genomic regions (chromosomes 1, 8 and X) with significant LOD scores but none overlapped with the previous linkage scan (Shiozawa *et al.*, 1998). In 2001, Gregersen and collaborators performed the first linkage scan in North American families (n = 257, n = 301 affected sibling pairs, n = 379 microsatellite markers) and found high linkage for the HLA region as well as significative evidence in chromosomes 1, 4, 12, 16 and 17 (Jawaheer *et al.*, 2001) (Figure 3.1). From these, only the genomic region in chromosome 16 seemed to overlap with the previous study in European families, and there was no overlap with the Japanese families' results. Using the affected sibling samples from the Arthritis Research Campaign (n = 252, n = 182 families, n = 365 microsatellite markers), a genome linkage scan was also performed in UK families (MacKay *et al.*, 2002). Similar to the North American scan, the HLA region was strongly linked to RA susceptibility and there were several other genomic regions showing suggestive evidence of linkage. In this case, there seemed to be an increased degree of overlap with other nominal peaks of the two previous linkage scans on Caucasian samples.

Given the lack of a strong non–HLA signal and the weak evidence of secondary peaks, genetic researchers started to question the merits of the linkage–based ap-

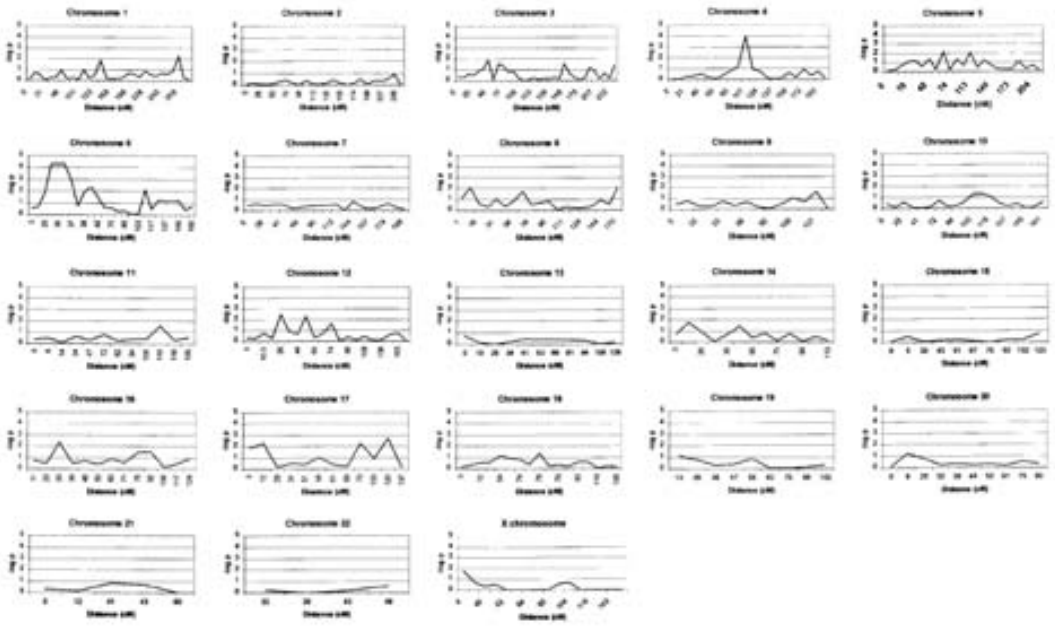


Figure 3.1: Whole genome LOD score results in 257 North American families. (Taken from Jawaheer et al. (Jawaheer *et al.*, 2001)). The strongest linkage signal is found in the short arm of chromosome 6, where the HLA locus resides (only the partial linkage peak is shown). Apart from the HLA region, only one other region in chromosome 4 shows significant linkage at the genome-wide level (LOD score >3).

proach in the identification of new genes for RA (Jawaheer & Gregersen, 2002). Even for regions showing some degree of overlapping between studies, it was certainly daunting to try to find the causal gene. The genomic regions "linked" to disease by this approach encompassed several megabases; in order to better map the location of the causal gene, either more family samples had to be recruited (which was generally unfeasible due to the late age of onset of the disease) or a case–control association study was implemented. In many cases, the later approach was generally discontinued by the lack of replicability. There is, however, one important exception to this lack of success which was the identification of *PTPN22* as a candidate risk locus for RA. Peter Gregersen, in collaboration with other public and private partners (including Craig Venter's Celera company), performed a case–control study of candidate genes for RA, either by family linkage analysis or by direct implication in the disease pathophysiology (Begovich *et al.*,

2004). In this case, they restricted the study only on functional SNPs, that is, exonic SNPs whose variation introduced either aminoacid substitutions or translation stop codons (Botstein & Risch, 2003). From all 87 functional SNPs finally tested, they identified a SNP codifying the change of Arginine to Tryptophan at position 620 of Protein Phosphatase 22 as strongly associated to RA susceptibility. This SNP had been because of the fact that it lay within the chromosome 1 linked region the previous scan in North American families. It can be said, however, that they had certain luck to find such association because the SNP was notably far from the original linkage peak (>9 megabases) and, furthermore, it was shown not to be the main responsible for the observed linkage singal.

The introduction of more sophisticated cellular and molecular biology techniques lead to important advances in the characterization of the pathophysiological processes in complex diseases. Genetic researchers were aware of such progresses and exploited this knowledge to perform an alternative search for candidate genes to test for association. Although linkage mapping also produced candidate genes, it was common to call the biological knowledge approach the "candidate gene" approach. In 2003 our research group performed a case–control association study using the Corticotrophin Releasing Hormone (*CRH*) gene as a candidate for RA susceptibility. It had been previously shown both in RA patients and in rodent models of the disease, that there was an evident deregulation of the hypothalamo–hypophysary axis, from which the CRH is a key hormone (Chikanza *et al.*, 1992; Sternberg *et al.*, 1989a,b). Following this evidence, in 2000 a research group from the Guy's King's and St. Thomas' School of Medicine (London, UK) tested for association in the *CRH* locus and found a significative association (Fife *et al.*, 2000). In collaboration with this group, we decided to provide the first replication study of this locus in an independent population. However, although having a similar statistical power to detect association we found no evidence of significant association between the *CRH* locus and RA in the Spanish population (Julià *et al.*, 2003, 2004). What could explain the discrepancy? Was our study a false negative finding or was it the English a false positive? Or were they both true positives? Under these circumstances, which have been very common in the study of complex diseases (Hirschhorn *et al.*, 2002), genetic researchers generally tend to be conservative in their reasoning and follow "Occam's razor" principle

(that is, to choose the simplest explanation from all possible explanations). In the case of genetic association studies of complex diseases, this has generally meant to regard the lack of replication in independent populations, as a proof of the lack of association with the disease.

3.6 GWAS studies in RA

The introduction of the first genome-wide genotyping technologies in 2006 lead to an explosion of GWAS studies on common disease traits that is still ongoing (Figure 3.2). RA was one of the first complex diseases to be analyzed using this approach. In August 2007, the Wellcome Trust Case Control Consortium (WTCCC) in UK published a combined GWAS approach with seven common diseases including RA (WTCC, 2007). For each disease they collected approximately 2,000 cases which were compared to a common 3,000 control cohort. Apart from providing new candidate loci for this group of diseases, the WTCCC study is now considered a fundamental milestone in the GWAS approach, providing the reference methodology for most subsequent studies. These methods include varied quality control analyses, genotyping algorithms, the characterization of genetic variation associated to geographic ancestry as well as several other statistical approaches now idiosyncratic for GWAS.

Regarding RA, the WTCCC study showed unequivocal association to the *HLA–DRB1* and the *PTPN22* loci ($P = 2.6$ e–27 and $P = 4.9$ e–26, respectively). From all studied diseases, it was the only one to have a marker differentially associated to RA, according to sex. In particular, the sex-stratified analysis identified a SNP in chromosome 17 to be highly associated in females but not in males, which was certainly appealing because of the strong sex bias that is characteristic of RA. Within the so–called "moderate" range of association (P = 1e–7 to 1e–5), nine new loci were associated with the risk to develop RA. Compared to posterior GWAS studies, however, the WTCCC did not provide a replication study of the most significant loci (i.e. one-stage design). Thus, the robustness of these associations would remain questioned until other studies provided further confirming evidence.

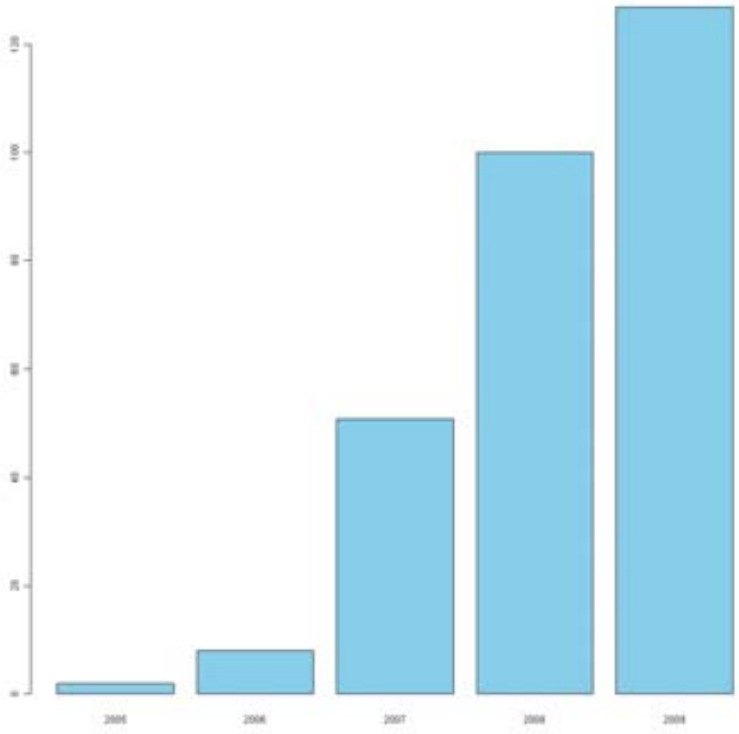


Figure 3.2: Number of GWAS studies on disease traits. Number of different studies reporting SNP–trait associations with P <1e–5, according to the National Human Genome Research Institute Catalog (http://www.genome.gov/gwastudies/)

Only one month after the publication of the WTCCC study, Peter Gregersen and collaborators identified a SNP between *TRAF1* and *C5* genes in chromosome 9 to be associated with RA (Plenge *et al.*, 2007b). Compared to the WTCCC GWAS, however, in this study only RA patients that were positive to anti–cyclic citrullinated peptide antibodies (anti–CCP) were used. The RA cohort was made up of North American patients (n = 908) and Swedish patients(n = 485). The integration of both cohorts was essential in choosing SNPs for replication since, separately, none reached multiple–test significance. Compared to the WTCCC study, replication cohorts -North American and Swedish, as well- were used to identify true positive associations from statistical fluctuations (i.e. two-stage design). It is important to note, however, that whilst the North American repli-

cation samples (n = 676) showed a clear association with the *TRAF1–C5*, the Swedish replication cohort (n = 568) did not show any significant association. Also, looking back at the association signals in the WTCCC for this region, they did not find any evidence of association even having genotyped a SNP in complete LD with the associated SNP. Together, these results evidence the difficulties of replicating low penetrance loci in complex diseases even for relatively large cohorts.

In November 2007, the same UK and USA research groups that participated in the previous GWAS, simultaneously published an association of the *TNFAIP3–OLIG3* locus with RA (Plenge *et al.*, 2007a; Thomson *et al.*, 2007). The *TNFAIP3–OLIG3* locus is located in the long arm of chromosome 6 and it was one of the nine "moderately associated" loci identified in the WTCCC study. With a replication cohort of 1,860 RA patients and 2,938 controls, the UK group only validated this candidate SNP from all nine tested. In the USA study, they used a different analysis strategy. First they performed a GWAS study with a moderate sample size (n = 397 RA patients -not included in the previous GWAS- and n =1,211 controls). Then, they tried to validate the top 90 SNPs showing the highest statistical evidence in a very large cohort (n = 5,063 cases and n = 3,849 controls, including individuals from the *TRAF1–C5* study). Interestingly, from this large list of candidate SNPs, only the only the SNP in the *TNFAIP3–OLIG3* locus (ranking in the 77^{th} position showed a significant P value in the replication analysis. This means that 89 GWAS candidate SNPs were in fact false positive associations.

After the GWAS studies in RA in the UK, North American and Swedish populations, we published in 2008 a genome-wide search for candidate loci in RA in the Spanish population (Julià *et al.*, 2008). Using a two–stage design, we identified a set of loci which were associated to the risk of RA. From these, a SNP in the *KLF12* gene showed the strongest association to RA susceptibility. In Chapter 5 we present this work, describing in detail the rationale of the study design and the different methodological approaches that we implemented.

Two months after our GWAS publication, the UK and USA groups published a collaboration study in RA. In this study they performed a meta-analysis with the data generated in the previous GWAS (total: n = 3,393 cases, n = 12,462 controls)

from which they selected the 31 most associated SNPs for validation in a large replication cohort of autoantibody–positive (i.e. RF positive and/or anti-CCP positive) RA patients (n = 3,929 *vs.* n = 5,807 matched controls) (Raychaudhuri *et al.*, 2008). They significantly replicated the SNPs in the *CD40* and *CCL21* loci, and found suggestive evidence of association for the *MMEL1–TNFRSF14*, *CDK6*, *PRKCQ*, *KIF5A–PIP4K2C* loci. Independently, the UK research group also performed a replication study on the 49 most significant SNPs from the original WTCCC study (*P* values from 1e–4 to 1e–5) (Barton *et al.*, 2008). Contrary to the meta-analysis approach, they did not select the replication RA patients based on their autoantibody status (n = 4,106 RA patients, n = 11,238 controls). However, two of the 3 positively replicated loci were common with the meta-analysis: kinesin *KIF5A* and *PRKCQ* genes. The third replicated locus was the IL2 receptor *IL2RB* locus in chromosome 22. Noticeably, these three loci did not show differential association effects according to the antibody status. This is an important finding since it suggests the existence a common genetic risk core in RA patients irrespective of the presence of autoantibodies.

In 2009 two new GWAS studies in RA have been published. In June 2009, Peter Gregersen and collaborators identified an association of the *REL* locus with the risk to RA. They used a two–stage association design: in the GWAS stage they expanded the cohort previously used to identify the *TRAF1–C5* locus up to a final number of 2,418 patiens and 4,504 controls, and in the replication stage they used a total of 2,604 RA patients and 2,882 controls. Following the trend of the previous approaches, they used cohorts with a highly predominant seropositivity (i.e. either CCP autoantibodies or RF). In November 2009, the same research group published significant associations at the *CD28*, *PRDM1* and *CD2/CD58* loci (Raychaudhuri *et al.*, 2009b). In this case, however, they used a completely different approach to select the candidate genes for replication. From the previous GWAS it had become increasingly evident that there were functional commonalities between the associated genes. For example, there were genes associated with T cell activation by APCs (*HLA–DRB1* and *PTPN22*) and genes from the NF–kB pathway (*CD40*, *TRAF1*, *TNFSF14*, *TNFAIP3* and *REL*). Based on this evidence, it seemed very likely that part of the yet unidentified risk loci for RA should also belong to these genetic pathways. Their relatively

low penetrance, however, would have "buried" them within the group of nominally associated genes (P <0.05) in the previous GWASs and, therefore, they wouldn't have been recognized as suitable candidate genes. To rescue this group of functionally related candidate genes, genetic researchers from the Broad Institute (Massachussets, USA) recently built a data mining tool called "GRAIL" (Raychaudhuri *et al.*, 2009a). Briefly, GRAIL uses the list of robustly replicated loci and the list of nominally associated loci to search for relationships between them based on previously published evidence. In this study, the list of established loci consisted on n = 16 SNPs, whereas the list of candidate loci consisted on the group of n = 179 SNPs that showed nominal association in the previous UK-USA meta-analysis (Raychaudhuri *et al.*, 2008). After applying GRAIL's text mining approach they found significant connectivities with 22 new candidate loci. Using a very large sample size of 7,957 autoantibody-positive RA patients and 12,462 controls, they could statistically validate the three mentioned loci (i.e. *CD28*, *PRDM1* and *CD2/CD58*), as well as suggesting further candidate genes.

3.7 Complexity component in RA

The GWAS approach has been a great success in identifying new loci associate to RA risk. Before the "GWAS era", only two loci had been robustly associated with the disease susceptibility, and now more than twenty loci have been robustly associated with RA. These genes have provided new insights into the disease pathophysiology as well as characterized new subgroups of patients. However, there is still a large fraction of the genetic architecture that is not captured by this main effect analysis approach. The total percent variance explained by all known non–HLA common genetic variants is approximately 4%. Considering that around 60% of RA risk is thought to be genetic, and one–third comes from the HLA locus, this indicates than more than half of the genetic variation associated with still needs to be elucidated. It would seem that we are reaching a point reminding Zeno's paradox of "Aquiles and the tortoise": larger and larger sample sizes are being used to discover smaller and smaller genetic effects. Following this reasoning, this should lead to an infinite sum of risk loci which is, evidently,

impossible. Thus, what type of genetic component remains to be identified and how we will be able to detect it?

Common diseases are considered to be complex for their large number of aetiological factors from which we only know a portion. However, complexity can manifest in other different forms (Thornton-Wells *et al.*, 2004). Together these complex causes can be subdivided into two groups: heterogeneity and interactions. Heterogeneity happens when there are multiple independent factors that can cause the disease but also when there are multiple phenotypic variables. In the first case, we can observe heterogeneity when several alleles from the same locus can cause the disease sometimes with opposite effects (i.e. allelic heterogeneity or also pleiotropy), when several independent loci can cause the disease (i.e. locus heterogeneity) or when genetic and non–genetic factors can give rise to the same disease (i.e. phenocopies). In the second case, RA is an exemplar case of a heterogeneous disease with patients showing a wide range of phenotypic manifestations although all included within the same clinical entity. Interactions appear when the effect of two or more factors upon the phenotype cannot be predicted by the independent effects of each factor. Interactions occur between genes and environment but also between genes themselves. Thus, it is evident that the analysis of interactions should be taken into consideration, if we want to complete the architecture of all common complex diseases. This was already true before the GWAS approach but now it is compulsory after the practical exhaustion of the main effects approach for the analysis of complex diseases (Moore & Williams, 2009).

In the present dissertation, we present two genetic studies which use alternative strategies to confront the complexity of RA. In Chapter 4 we use a whole genome gene expression analysis on synovial fibroblasts and a sophisticated bioinformatic analysis to identify the principal regulatory network of the response to the complex proinflammatory insult in RA. We then use this new set of candidate genes to study the presence of gene–gene interactions associated with the risk to RA using the Multifactor Dimensionality Reduction algorithm. In Chapter 5 we use a RA liability–based design and a whole genome association approach to identify new genetic variants associated with RA. We perform main effect analysis and also we use, for the first time, a genome-wide approach for the identification

of epistatic interactions in RA. Finally, in Chapter 6 we discuss the results from both these studies in the context of the actual knowledge of RA pathophysiology and susceptibility mechanisms.

Chapter 4

Genomics of expression and complexity

4.1 The failure of the candidate gene approach

Candidate genes for susceptibility risk have been traditionally selected on the basis of biological hypotheses or the location of the candidate in a previously determined region of linkage. In the latter case, we have seen how the scarce overlap between the different linkage studies in RA families discouraged the continuation of the positional–cloning approach. In the former case, researchers are indeed subject to the partial knowledge of the pathophysiological process of a disease (Hirschhorn & Daly, 2005). By 2005, more than 200 non–HLA loci had been studied in relation to RA using this approach (Plenge *et al.*, 2005) from which less than 10 showed some level of acceptable association (i.e. defined as a significance value of P <0.001 in one study or P <0.05 in two or more studies). As Risch and Merikangas had envisioned (Risch & Merikangas, 1996), the LD–based approach should have more power to detect loci associated to complex traits so, perhaps, it was the way we were choosing our candidates that was inappropriate.

4.2 Microarray analyses to guide candidate gene selection in RA

Reasonably, most of the candidate genes evaluated in RA have been genes associated with the HLA antigen presenting process (Marsal *et al.*, 1994) (i.e. *TAP*

genes, *TCR* genes, *CTLA4* ...) and with those genes involved in the cytokine signalling system (i.e. messenger proteins, cell receptors and downstream signalling effectors) that were found to be overexpressed in RA synovial samples. However, the introduction of the microarray technology in biomedical research at the beginning of the twenty-first century has opened new ways to the selection of disease candidate genes (Gregersen, 1999). Gene expression microarrays allow, for the first time, to analyze the expression levels of all the genes in the genome in a particular organism, tissue or cell type (i.e. the transcriptome). With the adequate design one should be able to identify which genes are actually being modulated in each particular pathological process and thus, identify a set of strong candidates for association with disease risk.

4.2.1 Transcriptional factors are the master regulators of gene expression activity

Cellular life depends on the correct response to varying internal and external stimuli. But, how do cells manage to appropriately regulate their gene expression levels? The key regulating cellular elements are transcription factors. Transcription factors are intracellular proteins that under expression and/or activation bind to specific sequences in the genome and modulate the levels of mRNA transcription of genes. The most recent estimates indicate that approximately 1,391 loci –around 6% of the human protein–coding genes–are transcription factors (Vaquerizas *et al.*, 2009). Together, the transcription factor and the set of genes that it regulates define a transcription regulatory network. Therefore, if a disease is caused by the inappropriate functioning of a cell type/s, it is essential that we identify which are the transcription regulatory network/s that are implicated.

4.2.2 The synovial membrane fibroblasts are fundamental to RA

Synovial fibroblasts (SFs) are a key cell type in RA pathophysiology. Together with macrophages, they are present in high densities in the forefront of the RA synovial membrane (i.e. the pannus), where they actively produce proinflammatory cytokines like TNFα and IL1β as well as ECM–degrading enzymes like

matrix metalloproteinases (Pap *et al.*, 2000; Ritchlin, 2000). In RA, this active front progressively invades the neighbouring cartilage and bone, leading to joint structural damage and loss of function.

4.2.3 Altered behaviour of RA SFs

Culture studies and evidence from animal models indicate that this is due to the acquisition by the SF of a "transformed" phenotype (Firestein, 1996). For example, several authors have identified a loss of contact inhibition in cultured SFs which is reminiscent of neoplastic activity, although the possibility of this being a pseudo–cancerous state has been rather inconclusive (Davis, 2003; Pap *et al.*, 2000). Another important evidence of this transformed activity comes from the Severe Combined Immunodeficiency (SCID) mouse coimplantation model of RA (Muller– Ladner *et al.*, 1996). This animal model harbours a mutation that prevents the formation of B and T lymphocyte lineages and thus, it is appropriate to perform *in vivo* tests of human tissue biopsies without the interference of the immune response of the host. In the RA SCID model, SFs isolated from RA patients are engrafted together with normal human cartilage under the renal capsule or the skin of the animal. After several days, it can be clearly seen how RA SFs have destroyed and invaded the coengrafted cartilage tissue, confirming the preservation of their altered behaviour.

4.3 Objectives of the study

The present study was designed to try to answer the next questions:

- What is the principal transcriptional regulatory network that governs the SF activity in RA?
- Is the genetic variation of the genes within this network associated with the susceptibility to RA?

4.4 The study design

4.4.1 The synovial fluid *in vitro* challenge will reveal the main transcriptional regulatory network of RA SFs

The first important aspect in this design was how to capture the relevant gene expression pattern from RA SFs. Synovial membrane biopsies offer a direct *in vivo* signature associated with the disease process; however, they are complex mixtures of different cell types other than SFs and, thus, it would be difficult, if not impossible, to attribute the observed gene expression changes only to the activity of SFs. Another possibility is to isolate SFs from a synovial membrane biopsy, culture them and analyze their gene expression pattern. Previous microarray studies had shown that fibroblasts retain their gene expression pattern even after several rounds of culture so, in this sense, it seemed a safe model from which to obtain valid conclusions. The next difficulty was to find a way that could evidence the transcriptional regulatory network that the SF used in the RA environment. Previously, other authors had studies single cytokine stimulations to study the gene expression patterns of RA SFs (Pierer *et al.*, 2004; Zhang *et al.*, 2004). Whilst these models can be useful to identify particular mechanisms associated to the response to each cytokine, they are clearly an oversimplification of the transcriptional program of the cell. Furthermore, it is possible that by not taking into account the *in vivo* complexity, we could be observing highly misleading results. Thus, if we want to see the activity of a SF in a real complex scenario, shouldn't we try to mimic this complex scenario as much as possible? The synovial fluid of RA patients is a complex mixture from the secretions of all the immune and non–immune cell types that populate the synovial joint (Distler *et al.*, 2005). As such, it acts as a channel for information flow between different cell types in the joint, including SFs. Furthermore, most SFs lie in the intimal face of the synovial membrane, in direct contact with the fluid. Most likely, an important amount of the external inputs that affect SF functionality will be coming from the elements diffusing in the synovial fluid. For these reasons, in order to capture the relevant transcriptional regulatory network in RA, we have performed an *in vitro* stimulation of SF using synovial fluid obtained from a joint of a RA patient showing high inflammatory activity.

It could be argued, however, that the synovial fluid sample from one individual does not necessarily need to represent the complex proinflammatory pattern from all RA synovial fluids. To clarify this, we performed a multiplex analysis of several cytokines using a protein microarray on synovial fluids from activated joints from other 5 RA patients. What we could observe is that, although there is some variation in the absolute concentration of these cytokines, the relative quantities between these cytokines is highly constant (Figure 4.1).

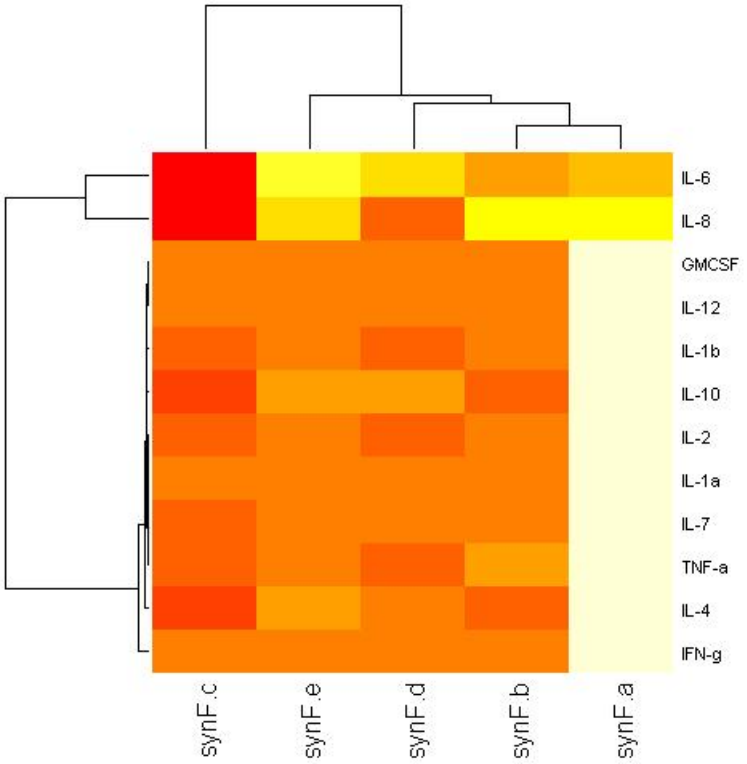


Figure 4.1: Clustering and heatmap of the cytokine concentrations in the synovial fluid of five patients having active RA. It can be seen that, except for 1 patient (synF.c) the correlation between the cytokine profiles is very high (average Pearson's product–moment correlation >0.94). The colour gradient denotes cytokine levels (red: low expression, white: high expression).

4.4.2 The selection of the microarray technology for genomic expression analysis

At the time the study was conceived, new commercial microarray platforms appeared. Following the success of so–called "academic" cDNA microarrays (Schena *et al.*, 1996) and Affymetrix high density synthetic oligonucleotide arrays in the early 2000's, other biotechnological companies proposed new microarray devices. One of these companies was UK based Amersham company, which had recently been acquired by the American multinational General Electric. We decided to use this platform for several reasons. First, cDNA microarrays although proving useful in many cancer studies, are associated with a considerable amount of technical problems. Although more expensive, commercial microarrays provide higher technical reproducibility which, in the case of studies where the biological source is limited (i.e. most studies with human samples, excluding neoplastic cells), it is a major advantage. Second, Affymetrix commercial microarrays, although highly reproducible, could be further improved. The principal reason was that, due to the characteristics of their manufacturing process (i.e. photolithography on silicon wafers) (Lipshutz *et al.*, 1999) , only DNA probes up to 25–mers (i.e. 25 bases of DNA) could be fixed on the array. It was shown, however, that the optimal specificity and sensibility parameters for DNA hybridization in microarrays could be reached with larger probes. This fact was exploited by several competing industries, including Amersham, who developed their own technology to synthesize gene expression microarrays with 50–mer probes. Given this evident technological superiority, we chose to use Amersham's CodeLink microarrays to evaluate the transcriptional response to synovial fluid stimulation in SFs.

4.4.3 R: open–source bioinformatic analysis toolbox

Gene expression microarrays introduced two fundamental aspects in biomedical data analysis which had been rarely seen previously. The first is a computational one: gene expression microarrays are able to interrogate tens of thousands of different transcripts per sample, each transcript generating a single data point. This high dimension of data means that typical Excel–type spreadsheets and statistical software are no longer practical for microarray data handling and

analysis. However, whilst different information technology companies developed user–friendly applications for this specific objective, an open–source initiative had a big impact: the R–language. "R" is an object–oriented statistical programming language initially built as an open–source version of the commercial statistical language "S" by Ross Ihaka and Robert Gentleman (Ihaka & Gentleman, 1996). The R–language had many aspects that favoured its use in the microarray arena: it was fast with high dimensional data, it allowed functional programming (i.e. using directly many mathematical functions), it had strong data and model visualization capabilities and, above all, it was open–source which means it was free to download and its source code available for anyone to inspect and modify. Soon, "R" created a large user and developer community in many different aspects involving statistical analysis. Without a doubt, one of the most productive ones was the analysis of gene expression microarray data. It became so popular that in 2001 it finally springed out as a specific computational biology and bioinformatics project called Bioconductor (Gentleman *et al.*, 2004). Thus, in the present project we used the R statistical language to perform several microarray data analysis steps including quality control, normalization and differential expression analysis. However, it is important to note that, at that time, R packages for microarray analyses had been developed either for two–colour cDNA microarrays or single–colour Affymetrix microarrays. Amersham microarrays were conceptually more close to the probe printing process of cDNA microarrays but were single–coloured like Affymetrix. For this reason, none of the available R packages suited the recently appeared CodeLink microarrays. One of the technical challenges in this study was that we had to program our own version of the quality control, normalization and differential analysis steps for this platform.

4.4.4 Mining the genomic data: Gene Ontologies

The comparison between the gene expression patterns between synovial fluid treated SFs and non–treated SFs helped us determine the relevant set of genes that we were looking for. One common first approach to characterize long lists of differentially expressed genes is to perform Gene Ontology analysis. Gene Ontologies are a controlled vocabulary that is used to define and characterize genes

according to the knowledge that we actually have from their protein product. Gene Ontologies are defined in a hierarchical manner stemming from 3 terms which are biological process (i.e. what is their biological role), cellular component (i.e. where are they localized), and molecular function (i.e. what particular biochemical activity they perform). In order to determine the significantly over-represented Gene Ontologies in the differentially expressed genes, we used Tim Beissbarth's webserver analysis tool GOstat (Beissbarth & Speed, 2004). Using the set of over and underexpressed genes, GOstat searches for their corresponding Gene Ontology definitions from the GO database (Consortium, 2001), and uses the Fisher's exact test to determine if the differences between the two groups are statistically significant. Finally, since the chance of calling a false positive increases with the number of null hypothesis tested simultaneously (i.e. the multiple–test problem), we chose to correct the nominal significance values using Bonferroni's method.

4.4.5 Inference of RA SF transcriptional regulatory network: reverse engineering

We have found the elements of the network, the differentially expressed genes, however we have not identified the network system that relates them. How can we infer the system from its constituent parts? A similar situation had already been found by hardware engineers during the 80s and 90s, when they faced the technical problem of having to clone a complex hardware system that 1) they had not developed and 2) the only available information from the system was the one obtained from separately examining the parts. They called the process of building a system out of the study of its constituent parts "reverse engineering". As soon as the first gene expression microarray datasets were produced, many bioinformatics researchers started to search for methods to "reverse engineer" gene networks from the gene expression measures. Whilst several methods inferred global networks in the purest mathematical sense (Basso *et al.*, 2005), other methods proposed the use of the growing biological knowledge on promoter and transcription factors to build more specific transcriptional regulatory networks.

One of the first methods to implement the latter approach was CARRIE or Computational Ascertainment of Regulatory Relationships Inferred from Expression. This method was developed in Zhiping Weng's bioinformatics Lab in Boston University and published in 2004 (Haverty *et al.*, 2004). Briefly, the method works as follows: once we have defined a group of genes that have significant differential expression, CARRIE starts by determining an equally sized set of genes which have the least expression changes (i.e. genes that are expressed in SFs but that do not vary between treatment conditions). Using the promoter sequences from both sets of genes and the Position Specific Scoring Matrix (PSSM) associated with each transcription factor, CARRIE calculates a binding score. With the empirical distribution of these scores in the negative set, CARRIE determines the score threshold at which binding in the positive score can be deemed significant. Finally, using the binomial probability distribution function, CARRIE calculates 1) the probability that a particular transcription factor binds to the sequence and 2) the probability that a specific transcription factor regulates the entire positive set of differentially expressed genes. If our SF *in vitro* model uses a specific transcription factor, this should appear as statistically significant in the last computational step. At that time, CARRIE's reverse engineering approach had been thoroughly validated in well–known yeast activation models, but in humans, it had only been tested in the human skin fibroblast response to serum stimulation (Haverty *et al.*, 2004). Thus, our study was the first time this method was applied to infer the transcriptional regulatory network on a disease–related design.

4.4.6 Analysis of epistatic interactions in the main transcriptional regulatory network responding to synovial fluid

Once we have identified the transcriptional regulatory network that governs SF response to RA synovial fluid we can answer the second question: is the genetic variation in these genes associated with the risk to develop RA? Following the general approach, genetic markers from each locus could be selected, genotyped in a case–control study design and examined for independent effects. However,

having identified that such genes belong to a common regulatory network, the assumption of independence of effects upon phenotype cannot reasonably hold. If we accept that genomes tend to "buffer" negative genetic variations (i.e. canalization), wouldn't it be more plausible that the variation associated with disease is a combination of elements from this network rather than either one alone? As Bateson already proposed at the beginning in the twentieth century, gene–gene interaction effects (i.e. epistasis) should be accounted for, if we really want to understand the complexities of genotype–phenotype relations.

4.4.7 High–level interactions analysis: Multifactor Dimensionality Reduction

The identified SF transcriptional regulatory network was composed of 13 corregulated genes. Thus combinations between any of these 13 genes could give rise to the fibroblast altered activity. Perhaps it was the genotypic combinations between 3 genes that were harmful. Or maybe more. We had to find a method that would help us search through all possible gene combinations those that were associated with the disease. Logistic regression, the generalization of linear regression to binary (case–control) data, is well suited to detect low–level interactions but fails when the number of predictor variables (i.e. genetic markers) is high (Ritchie *et al.*, 2001). The task of identifying relevant patterns amongst a large group of possible explanatory variables seems to be better conducted by data mining methods. Data mining can be defined as the science of looking for patterns within large datasets using computers. One of the first authors to use data mining strategies for the problem of identifying high–level genotypic combinations associated with phenotype traits was Matthew Nelson. He developed a method called "Combinatorial Partitioning Method" (CPM) to identify sets of partitions of multi–locus genotypes that can predict the variability of a quantitative trait (Nelson *et al.*, 2001). First, for each combination of "n" loci, the CPM determines the number of individuals having one particular genetic combination (i.e. 9 possible combinations in 2–loci combinations, 27 possible combinations in 3–loci combinations, etc.). In the second step, the genotype combinations are partitioned into k disjoint sets according to an objective function that maximizes

the proportion of phenotypic variance explained. Finally, in order to evaluate the consistency of the combinatorial models, the full process is embedded in a cross validation structure. Cross–validation is a data mining technique that consists in the division of the sample into nonoverlapping subsamples of training and testing data. For each partition, the genetic model is built using the information from the training data and subsequently validated in the testing data. We can then use the average performance of the predictions in all subsamples as a measure of the accuracy of the model. The power of this method stems from the fact that the training process is performed without using the information from the testing set, thus giving an unbiased measure of the association of the model with the trait. In the present study, we have used a method derived from Nelson's CPM called Multifactor Dimensionality Reduction (MDR) (Ritchie *et al.*, 2001). This method was devised by Prof. Jason Moore from Dartmouth Medical Centre (New Hampshire, US) for the determination of relevant high order interactions between genotypes and binary traits. In collaboration with Prof. J Moore, we therefore used MDR to determine the presence of epistatic interactions in the identified NF–kB transcriptional regulatory network associated with RA.

4.4.8 Hypernormal controls: epidemiological strategy to increase statistical power

Up to the association analysis of the SF transcriptional regulatory network genes we had addressed several levels of complexity. However, even if a genetic variation for RA risk existed in these genes, would we be able to statistically detect it? The 295 RA patients cohort that we had collected was sufficiently well–powered to detect main effects similar in size to the *HLA–DRB1* locus (OR$\sim$2.4). However, the lack of replication from many previous non–HLA loci had raised logical concerns about the statistical power to detect true associations in RA. We therefore sought to look for a strategy that would increase the statistical power of our case–control design. The answer came from an epidemiological concept described in 1961 by English physician C Carter and later developed by geneticist B Falconer (Falconer, 1967) called liability. In their search for the heritable component observed in non–Mendelian diseases (i.e. complex diseases), Carter and

Falconer theorized on the existence of a continuous but yet unmeasured variable, directly associated with the cause of the disease. At a certain threshold this variable or liability would give rise to the disease. Consequently, if we are able to select individuals at the two extremes of this liability variable, we will clearly have a greater statistical power to detect true associations. The problem for us was how we could implement this epidemiological feature to increase the power of our analysis. For a long time, physicians have been noticing the presence of an increased aggregation of autoimmune diseases within families. In the clinical experience of our group, we were readily aware that RA patients tend to have relatives with RA but also with other types of autoimmune manifestations. Thus, in order to increase the power of our study, we hypothesized that there should be a shared genetic risk component between RA and other autoimmune diseases. We consequently selected control individuals so that they were as free as possible from this common genetic risk. To fulfil this "hypernormality" condition, the selected control individuals had to be the age of risk for RA (>40 years old) and they should not have any trace of autoimmunity in them or in any of their first order relatives (i.e. parents, siblings and offspring). This was possible thanks to the collaboration with the IRCIS BioBank (Hospital San Joan de Reus, Tarragona) and the access to their randomized control collection, which is one of the best characterized control cohorts in Spain. According to Newton Morton –the inventor of the popular LOD score for family linkage analysis- using this epidemiological strategy we would be entitled to increase our statistical power up to 4 times compared to a study using normal controls (Morton & Collins, 1998).

Available online at www.sciencedirect.com
ScienceDirect
Genomics 90 (2007) 6–13
GENOMICS
www.elsevier.com/locate/ygeno

Identification of a two-loci epistatic interaction associated with susceptibility to rheumatoid arthritis through reverse engineering and multifactor dimensionality reduction

Antonio Julià [a], Jason Moore [b], Laia Miquel [c], Cayetano Alegre [d], Pere Barceló [d], Marylyn Ritchie [e], Sara Marsal [a,*]

[a] *Unitat de Recerca de Reumatologia, Institut de Recerca Hospital Universitari Vall d'Hebron, UAB, 08035 Barcelona, Spain*
[b] *Computational Genetics Laboratory, Dartmouth–Hitchcock Medical Center, Lebanon, NH 03756, USA*
[c] *Departament de Biologia i Bioquímica Molecular, Universitat Autònoma de Barcelona, Barcelona, Spain*
[d] *Unitat de Reumatologia, Hospital General i Universitari Vall d'Hebron, 08035 Barcelona, Spain*
[e] *Department of Molecular Physiology and Biophysics, Vanderbilt University Medical School, Nashville, TN 37240, USA*

Received 20 October 2006; accepted 8 March 2007
Available online 4 May 2007

Abstract

Altered synovial fibroblast (SF) transcriptional activity is a key factor in the disease progression of rheumatoid arthritis (RA). To determine the transcriptional regulatory network associated with SF response to an RA proinflammatory stimulus we applied a CARRIE reverse engineering approach to microarray gene expression data from SFs treated with RA synovial fluid. The association of the inferred gene network with RA susceptibility was further analyzed by a case–control study of promoter single-nucleotide polymorphisms, and the presence of epistatic interactions was determined using the multifactor dimensionality reduction methodology. Our findings suggest that a specific NF-κB transcriptional regulatory network of 13 genes is associated with SF response to RA proinflammatory stimulus and identify a significant epistatic association of two of its genes, *IL6* and *IL4I1*, with RA susceptibility.

Keywords: Rheumatoid arthritis; Synovial membrane; Fibroblast; DNA microarrays; Bioinformatics; Transcription factor; Promoter; SNPs; Genetic epistasis; Genetic susceptibility

Rheumatoid arthritis (RA) is a chronic inflammatory disease with a prevalence of approximately 1% that primarily affects diarthrodial joints, in which synovial inflammation leads to cartilage and bone destruction. The synovial membrane, a rather acellular tissue in normal conditions, becomes hypertrophic and is composed mainly of synovial fibroblasts (SFs) [1]. SFs in RA display an activated phenotype, which significantly contributes to disease initiation and progression [2,3]. Although several transcription factors like AP-1 [4], NF-κB [5], or p53 [6] have been previously associated with SF altered activity, no precise transcriptional regulatory network has been associated with RA pathophysiology. With the advent of microarray technology, global gene expression data can now be used to model transcriptional networks associated with molecular disease mechanisms.

Modeling transcriptional regulatory networks is considered a *reverse engineering* problem. By reverse engineering we understand the process of determining the structure of a system by reasoning backward from observations of its behavior [7]. Different methods have been recently described to determine functional networks from microarray gene expression data. After providing success with lower eukaryotes [8] they are also proving successful in defining regulatory networks in the first studies with human gene expression data [9].

Microarray analysis of cultured SFs treated with a single factor can be useful to study molecular mechanisms relevant to RA [10,11]. However, the synovial environment in RA is extremely complex, with the interplay of cytokines, chemokines, matrix-degrading enzymes, growth factors, and immune

* Corresponding author. Fax: +34 93 489 4015.
E-mail address: smarsal@ir.vhebron.net (S. Marsal).

0888-7543/$ - see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.ygeno.2007.03.011

cell particles [12]. Furthermore, several RA proinflammatory factors like TNF and IL1β can regulate gene transcription via convergent signaling pathways. Synovial fluid is known to contain most of the proinflammatory factors associated with RA pathophysiology. Thus we hypothesize that SF in vitro treatment with a complex proinflammatory stimulus like RA synovial fluid can help to identify the specific SF transcriptional network associated with this disease.

Transcriptional regulatory networks are theoretically prone to the presence of epistatic effects [13]. Epistasis, or more specifically, genetic epistasis, can be defined as the nonindependent effect of genetic polymorphisms over a particular trait in an individual [14]. Until now, the analysis of epistatic effects on human diseases has been limited by the exponential number of combinations to be analyzed in multilocus models. Recently, however, data mining approaches for dimensionality reduction in the analysis of gene × gene and gene × environment interactions have proven useful in the detection and characterization of epistatic effects in human diseases [13,15].

The present study was therefore designed to determine, first, whether a particular transcriptional regulatory network is involved in SF response to RA synovial fluid stimulation and, second, whether promoter polymorphisms in the genes of this network are associated with susceptibility to RA via epistatic interactions. To answer these questions we studied the differential gene expression profiles from cultured synovial fibroblasts with and without RA synovial fluid stimulation. We applied CARRIE, a new method of transcriptional network ascertainment that couples gene expression analysis with promoter sequence information to infer regulatory relationships [16], to the results. After defining the associated transcriptional network, we analyzed the presence of epistatic interactions associated with RA susceptibility between promoter single-nucleotide polymorphisms (SNPs) from the coregulated genes by using the multifactor dimensionality reduction (MDR) method [17].

Results

Differentially expressed genes and significant Gene Ontology (GO) terms

Using conservative criteria for differential expression we obtained a total of 157 genes differentially expressed between treatment groups (Supplementary Table S1). A partial list (fold change >3) of differentially expressed genes is shown in Table 1.

To evaluate the global gene expression changes on SF in response to an RA synovial fluid stimulus we compared GO terms from differentially expressed genes. Statistically over-represented GO terms ($p < 0.05$) were immune response (GO: 0006955), response to biotic stimulus (GO: 0009607), defense response (GO: 0006952), receptor binding (GO: 0005102), cytokine activity (GO: 0005125), and response to wounding (GO: 0009611). The complete list of genes associated with each differentially expressed GO can be found in Supplementary Table S2.

Table 1
Three-fold differentially expressed genes in cultured synovial fibroblasts after RA synovial fluid treatment

Accession No.	Gene	Description	Fold change [a]
NM_016584	*IL23A*	Interleukin 23, α subunit p19	7.9
NM_000641	*IL11*	Interleukin 11	6.7
NM_006329	*FBLN5*	Fibulin 5	5.7
NM_000963	*PTGS2*	Prostaglandin-endoperoxidase synthase 2 (prostaglandin G/H synthase and cyclooxygenase)	5.4
NM_000759	*CSF3*	Colony stimulating factor 3 (granulocyte)	4.4
NM_000675	*ADORA2A*	Adenosine A2a receptor	4.1
AK058127	—	*Homo sapiens* cDNA FLJ25398	3.8
NM_000758	*CSF2*	Colony stimulating factor 2 (granulocyte–macrophage)	3.7
NM_005623	*CCL8*	Chemokine (C–C motif) ligand 8	3.6
NM_001432	*EREG*	Epiregulin	3.6
NM_006443	*C6orf108*	Chromosome 6 open reading frame 108	3.6
NM_002192	*INHBA*	Inhibin, βA (activin A, activin ABα polypeptide)	3.5
NM_000346	*SOX9*	SRY (sex determining region Y)-box 9 (campomelic dysplasia, autosomal sex-reversal)	3.5
NM_033035	*TSLP*	Thymic stromal lymphopoietin	3.4
NM_031437	*RASSF5*	Ras association (RalGDS/AF-6) domain family 5	3.4
NM_021724	*NR1D1*	Nuclear receptor subfamily 1, group D, member 1	3.3
NM_004049	*BCL2A1*	BCL2-related protein A1	3.2
NM_002089	*CXCL2*	Chemokine (C–X–C motif) ligand 2	3.1
NM_001682	*ATP2B1*	ATPase, Ca^{2+} transporting, plasma membrane 1	3.0
NM_002692	*POLE2*	Polymerase (DNA directed), ε2 (p59 subunit)	–3.1
NM_001684	*ATP2B4*	ATPase, Ca^{2+} transporting, plasma membrane 4	–3.1
NM_139314	*ANGPTL4*	Angiopoietin-like 4	–3.2
NM_000331	*SAA1*	Serum amyloid A1	–3.2
NM_006283	*TACC1*	Transforming, acidic coiled-coil-containing protein 1	–3.3
NM_002084	*GPX3*	Glutathione peroxidase 3 (plasma)	–3.4
NM_012242	*DKK1*	Dickkopf homolog 1 (*Xenopus laevis*)	–3.8
NM_006006	*ZBTB16*	Zinc finger and BTB domain containing 16	–4.5

[a] $p < 0.00001$.

Analysis of transcriptional regulatory networks

Determination of significant transcription factor matrices

We determined those transcription factors that most likely control the response of SFs to RA synovial fluid using CARRIE. We found that, from all significant matrices (Fig. 1), the NF-κB distribution matrix stands out as the most clearly associated. NF-κB has a *p* value four orders of magnitude more significant than the immediate associated transcription factor (TF).

Determination of NF-κB regulatory network

We inferred the transcriptional regulatory network of NF-κB involved in the SF response to RA synovial fluid using CARRIE (Fig. 2). Although no significant expression change was observed for NF-κB itself, a significant relationship with 13

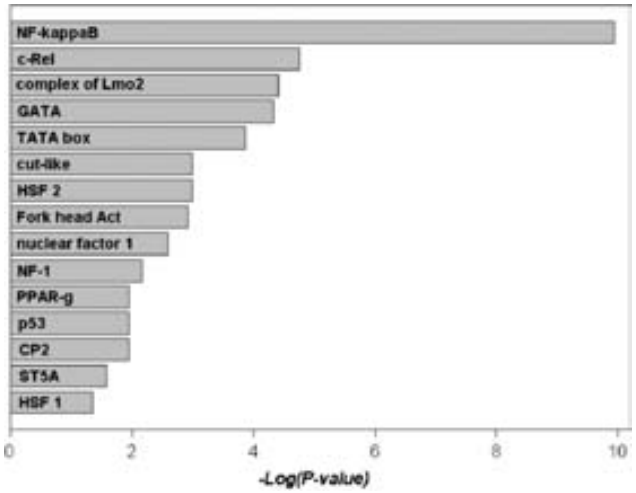


Fig. 1. Detection of the TF responsible for gene expression changes in SF treated with RA synovial fluid. Each horizontal bar represents the log10 p value for overabundance for a specific TF. Only the 15 most significant TFs are plotted.

coregulated genes was assessed. Of these, 11 were up-regulated by NF-κB action, whereas 2 were down-regulated.

Population-based association analysis of NF-κB-coregulated genes

To evaluate the association of the regulatory regions of the 13 genes included in the NF-κB network we genotyped SNPs from or near to the proximal promoter. We finally analyzed a total of 22 SNPs, listed in Table 2. We found all polymorphisms to be in Hardy–Weinberg equilibrium ($p > 0.001$, data not shown). From all polymorphisms tested rs1290754 (*IL4I1*), rs2633958 (*COL7A1*), rs4694636 (*IL8*), and rs344589 (*CD70*) were associated at $p < 0.05$ although significance was not maintained after Bonferroni correction. We found strong pairwise linkage disequilibrium (LD) between markers from the same genetic region ($D' > 0.98$), except for *CD70* ($D' = 0.78$). None of the estimated multimarker haplotypes was significantly associated with RA (Table 2).

Genetic epistasis analysis of NF-κB-coregulated genes

We analyzed the presence of epistatic interactions evaluating all possible two- to seven-way SNP combinations. From these, the two-SNP combination of rs1290754 (*IL4I1*) with rs1800797 (*IL6*) was the best model for RA risk prediction (Fig. 3, left). The testing accuracy of the selected model was 0.599 ($p < 0.02$ based on a 100-fold permutation test). The odds ratio for this model was 2.23 (95% CI 1.51, 3.28). Notably, when these two SNPs were merged as a single variable and the data reanalyzed it clearly came out as the best model, with a testing accuracy of 0.6 and a cross-validation count of 10 of 10 (Fig. 3, right). The entropy analysis [18] of the interaction between SNPs (Fig. 4) clearly shows the strong synergistic interaction between these two SNPs.

Discussion

This study shows that a particular transcriptional regulatory network is involved in SF response to RA synovial fluid stimulation. We found that from all transcriptional networks analyzed, the NF-κB regulatory network is markedly significant. In contrast, other known networks, like AP-1 or p-53, seem to have a secondary role in the response to this complex proinflammatory stimulus.

This study also shows that polymorphisms in genes of this transcriptional network are associated with susceptibility to RA via epistatic interactions. We found that high-order interactions between SNPs from *IL4I1* (rs1290754) and *IL6* (rs1800797) are significantly associated with risk to develop RA.

NF-κB transcriptional network

NF-κB is one of the TFs most strongly associated with RA pathogenesis, regulating the activities of many different genes in many cell types [19,20]. We determined a highly statistical overrepresentation of NF-κB binding sites in the promoter regions of the coregulated genes ($p < 1 \times 10^{-9}$, Fig. 1). In this model, NF-κB is not differentially expressed, which is in

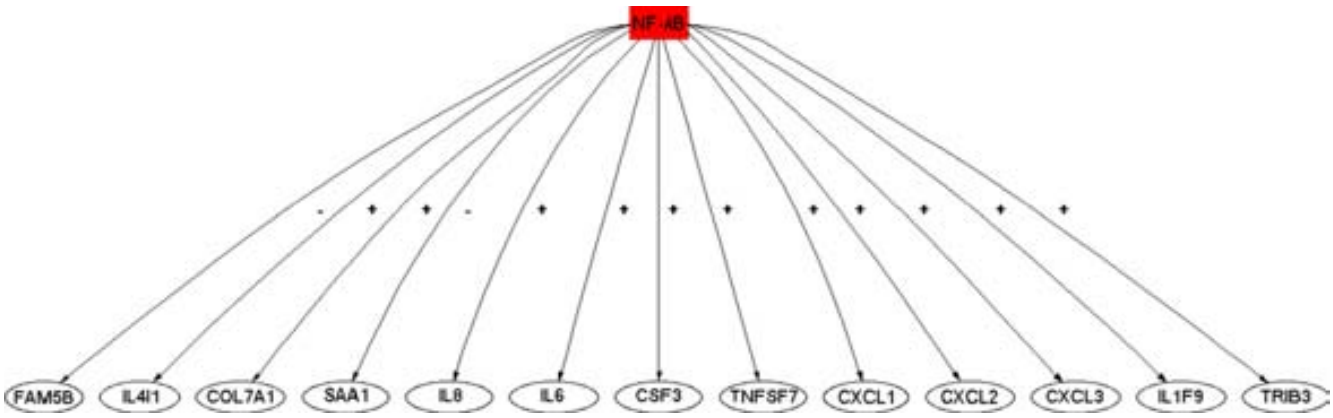


Fig. 2. Inferred transcriptional regulatory network mediating SF response to RA proinflammatory stimulus. Solid arrows between NF-κB (i.e., in red, not differentially expressed) and its target genes represent direct regulation predicted by ROVER. The "+" symbols represent stimulatory relationships and the "−" symbols denote inhibitory relationships.

Table 2
Association analysis of promoter SNPs of NF-κB transcription regulatory network genes

Gene	Accession No.	SNP	MAF-Hap	MAF	*p*-Genot[a]	*p*-Allele[a]	*p*-Haplo[a]
FAM5B	NM_021165	rs2049162	0.06	0.04	0.58	0.37	0.95
FAM5B	NM_021165	rs6665358	0.04	0.04	0.68	0.46	—
FAM5B	NM_021165	rs725416	0.18	0.17	0.42	0.57	—
IL4I1	NM_152899	rs1290751	0.42	0.47	0.65	0.52	0.7
IL4I1	NM_152899	rs1290754	0.46	0.41	**0.028**	0.96	—
COL7A1	NM_000094	rs1264194	0.36	0.31	0.5	0.94	0.08
COL7A1	NM_000094	rs2633958	0.05	0.06	**0.026**	**0.016**	—
SAA1	NM_000331	rs1533723	0.22	0.35	0.78	0.58	0.68
SAA1	NM_000331	rs874957	0	0.35	0.67	0.44	—
IL8	NM_000584	rs2227306	0.34	0.37	0.08	0.23	0.26
IL8	NM_000584	rs4694636	0.42	0.4	0.06	**0.04**	—
IL6	NM_000600	rs1800797	0.48	0.38	0.28	0.94	—
CSF3	NM_000759	rs2227321	0.34	0.36	0.42	0.48	—
CD70	NM_001252	rs344589	0.21	0.14	**0.034**	**0.028**	0.06
CD70	NM_001252	rs168259	0.19	0.12	0.41	0.46	—
CXCL1	NM_001511	rs3117600	0.29	0.31	0.11	0.75	—
CXCL2	NM_002089	rs3806792	0.34	0.41	0.48	0.32	—
CXCL3	NM_002090	rs370655	0.36	0.41	0.91	0.72	—
IL1F9	NM_019618	rs13014143	0.48	0.44	0.92	0.98	0.98
IL1F9	NM_019618	rs13392494	0.15	0.16	0.57	0.8	—
TRIB3	NM 021158	rs6139007	0.18	0.22	0.45	0.3	0.66
TRIB3	NM 021158	rs6084242	0.33	0.36	0.1	0.46	—

MAF-Hap, minor allele frequency in HapMap Caucasian European samples; MAF, minor allele frequency in Spanish hypernormal controls; *p*-Genot, *p* value for genotypic test; *p*-Allele, *p* value for allelic test; *p*-Haplo, *p* value for haplotypic test.
[a] Nominal *p* values, not corrected for multiple tests (in bold $p < 0.05$). All SNPs demonstrated Hardy–Weinberg *p* values > 0.001.

accordance with the fact that its activation takes place mainly at the posttranscriptional level through IκB phosphorylation and degradation [21]. Strikingly, however, we found a clear cutoff between NF-κB and the rest of the TFs (i.e., four orders of magnitude with the second most associated TF). Since synovial fluid is a complex mixture of various stimulating proinflammatory factors we would rather expect a less pronounced main TF prediction. It is even more striking if we compare the similarity between this result and the results obtained using the same inference methodology in yeast perturbation experiments [16]. In these studies different single-factor stimulation experiments, from which the main TF was already known, were used to validate CARRIE methodology.

In the present study we determined a specific transcriptional regulatory network of 13 genes for which there is evidence supporting their roles in SF NF-κB-mediated response in RA. CXC motif chemokines—CXCL1, CXCL2, CXCL3, and IL8—are regulated by NF-κB [22,23], are overexpressed by TNF and IL1β stimulation [24], and have been demonstrated to be overexpressed in RA fibroblasts [22]. Similarly *CSF3*, *IL1F9*, *CD70*, and *COL7A1* can be activated in RA fibroblasts or other cell types after TNF and IL1β stimulation [24–27] or in

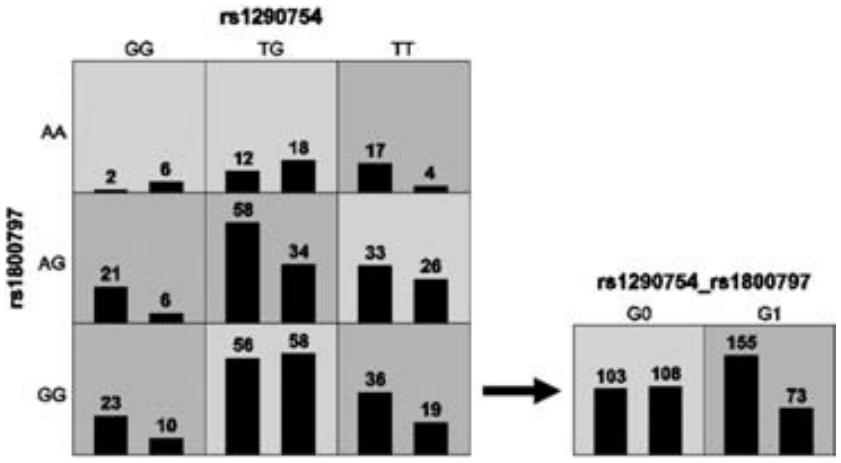


Fig. 3. Summary for *IL4I1* (rs1290754) and *IL6* (rs1800797) promoter SNP combinations associated with risk to RA as (left) independent variables and as (right) a single merged variable. High-risk combinations appear as dark gray and low-risk combinations as light gray: the left bar in the boxes represents the frequency in cases and the right bar represents the frequency in controls.

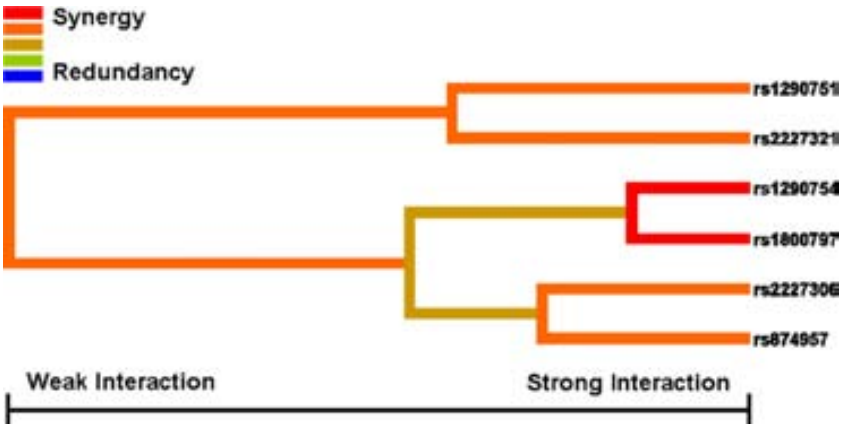


Fig. 4. Dendrogram of interactions between NF-κB transcription regulatory network polymorphisms. The colors used depict the degree of synergy, ranging from red (highest information gain) to blue (highest information redundancy). Note that for the interaction between *IL4I1* (rs1290754) and *IL6* (rs1800797) promoter SNPs, the degree of synergy (gain of information) between them is highest.

RA animal models [28]. IL6, one of the crucial proinflammatory cytokines in RA and which is clearly associated with RA fibroblast altered expression, has a demonstrated NF-κB-induced upregulation [29], together with IL4I1 [30], TRIB3 [31], and SAA1 [32]. The latter, however is significantly down-regulated in our study, an effect for which we have no explanation. Finally, *FAM5B* (i.e., BMP/retinoic acid-inducible neural-specific protein) is a recently characterized gene with a perforin-like conserved domain structure [33] and, to our knowledge, this study is the first evidence for NF-κB transcriptional control.

IL4I1 × *IL6 epistatic model for RA risk*

Several lines of evidence point to the relevance of epistatic effects in RA etiology. The most compelling data come from quantitative trait analysis on mouse [34] and rat [35,36] RA models. Although epidemiological modeling studies of the disease have provided the theoretical framework for epistasis in RA [37], very few studies have analyzed the presence of such interactions in humans [38,39]. One of the reasons for the lack of these studies is the statistical and computational challenge that is associated with the analysis of genetic polymorphism combinations and disease susceptibility [18]. The multifactor dimensionality reduction approach used in this study attempts to address this limitation and is able to determine the multilocus combinations associated with high risk to develop disease [13].

Analysis of main effects in the *IL4I1* promoter SNPs showed a modest association at the genotypic level for rs1290754 ($p < 0.05$, Table 2) although not significant after Bonferroni correction (data not shown). To our knowledge, this is the first study to analyze the association between IL4I1 gene polymorphisms with RA. On the other hand, *IL6* polymorphisms association with RA have already been studied in the Spanish population [40] and, in agreement with this previous study, we did not detect a significant main effect after allelic and genotypic analysis (Table 2). However, by analyzing the high-order interactions between all candidate polymorphisms and RA using MDR we identified a significant interaction between *IL6* promoter SNP rs1800797 and *IL4I1* promoter SNP rs1290754 (OR 2.2; 95% CI 1.5109, 3.2806; $p < 0.02$). Even after merging both SNPs as a unique variable, it is still the model (i.e., single variable model in this case) with the highest testing accuracy (Fig. 3).

How does this epistatic effect increase the susceptibility to develop RA? While IL6 is a well-known proinflammatory cytokine clearly associated with RA pathology, the role of IL4I1 in RA still needs to be explored. One intriguing evidence that could support the active role of this protein in RA pathophysiology could be its identification as an autosomal *H* locus [41]. *H* loci are minor histocompatibility antigens that are processed by MHC class I and class II molecules and that can be responsible for allograft rejection in transplant therapies. Studies in mice found that the H46 locus, containing *Il4i1* (i.e., the murine homologue of *IL4I1*), synthesizes a peptide that is presented by MHC class II that has the potential to elicit $CD4^+$ T cell responses in autoimmunity. Autoantigen activation of T cells is one of the main models for the etiology and pathogenesis of RA [12]. The interaction dendrogram method implemented in MDR software (Fig. 4) allowed us to determine the nature of the interaction as a high-degree synergy between *IL6* and *IL4I1* promoter polymorphisms. Thus, it is tempting to speculate that high-risk combinations in this two genes could led to the proactive role of the synovial fibroblast in the development of RA.

Limitations of the approach

In our view, the scope of the study could be limited mainly in two ways. First, is the stimulation of synovial fibroblasts with RA synovial fluid a valid model to study RA pathophysiology? Second, is CARRIE a valid reverse engineering approach to characterize relevant transcriptional regulatory networks in human disease?

RA synovial fluid is formed by the secretions of multiple activated immune and nonimmune cells, which give rise to a complex proinflammatory environment that contributes to the progression of disease. Thus, the transcriptional profiling of SF stimulated with RA synovial fluid should yield more valuable

insights into disease molecular mechanisms than single cytokine stimulation. Accordingly, several significantly differentially expressed genes we have detected, like *INHBA* [42], *IL11* [43], *PTGS2* [44], *IL6* [22], or *IL23A* [45], have been previously associated with RA synovial pathophysiology (Table 1 and Supplementary Table S1). Furthermore, by using this experimental model we have been able to identify a two-loci epistatic interaction that is significantly associated with susceptibility to developing RA.

Reverse engineering is a new kind of analytical approach that has been only recently used with human data [9]. For this study we have made the assumption that the inference method implemented in CARRIE can efficiently model the transcriptional regulatory network associated with SF response to an RA synovial fluid stimulus. One of the limitations of this methodology is that it is subject to the completeness of the TRANSFAC human database. Therefore, statistical error could have been introduced by absent position-specific scoring matrices (PSSMs) or weak promoter binding sites. However, the highly significant statistical association of the NF-κB transcriptional regulatory network and the previous implication of several of its genes in RA pathophysiology give strong support to the effectiveness of this methodology. To our knowledge, this is the first study to apply CARRIE methodology on human data.

Conclusion

In summary, this study shows that a specific NF-κB transcriptional regulatory network is significantly associated with synovial fibroblast response to RA synovial fluid and that an epistatic interaction between two of its genes, *IL6* and *IL4I1*, is significantly associated with the risk of developing RA. Although other relevant SF regulatory networks cannot be excluded in RA pathophysiology, it is the first demonstration of a transcriptional regulatory network associated with this disease. We believe that the definition of relevant transcriptional networks by reverse engineering can greatly accelerate the search for disease susceptibility genes.

Materials and methods

Cell culture and synovial fluid

Osteoarthritis (OA) SFs have been widely used as a reference for the experimental study of synovial membrane. In this study, synovial membrane was obtained from an OA patient undergoing knee joint replacement surgery. The membrane was thoroughly minced to ~1 mm^3 and incubated for 2 h at 37°C with 1 mg/ml collagenase I A (Sigma, Spain) under continuous agitation. Cells were pelleted and cultured in DMEM with L-glutamine (Gibco Life Technologies, Spain) with 10% FCS and penicillin–streptomycin (50 IU/ml) (Gibco Life Technologies, Spain) at 37°C with 5% CO_2 in a humidified atmosphere.

Synovial fluid was obtained from an inflamed knee joint of a 55-year-old RA female patient in a sterile nonheparinized tube. From the time of extraction to centrifugation (1500*g*, 10 min) the sample was continuously kept at 4°C. The acellular supernatant was kept at −20°C until cell treatment.

Synovial fibroblast stimulation and microarray analysis

After the third passage, the synovial cell culture was divided into control and treatment groups. Control cells were cultured with fresh medium (DMEM) without FCS, whereas treated cells were cultured in a fivefold diluted synovial fluid (80% DMEM). After 12 h of treatment, total RNA was extracted from both cell groups using a column affinity purification method (RNeasy; Qiagen, Spain). For each class three RNA samples were obtained and analyzed separately in Human IA CodeLink Expression Bioarrays (General Electric Healthcare, Spain) representing ~20K UniGene entries. Briefly, each RNA was in vitro amplified, hybridized, stained, and scanned using the manufacturer's instructions. Signal and background intensities were also extracted using the manufacturer's recommended software (CodeLink version 2.3.2). Normalization of background-corrected intensities was performed stepwise: intraclass replicates were normalized using cyclic lowess normalization and both classes were finally normalized using quantile normalization. Data transformation and normalization were carried out using the libraries provided as part of the R statistical language package version 2.1.0 (http://cran.r-project.org). Primary data and supplementary tables can also be accessed at http://www.urr.cat.

Differential gene expression and gene ontologies

Differentially expressed genes in fibroblasts stimulated with RA synovial fluid compared to controls were identified using the two-sample Welch *t* statistic implemented in Bioconductor's "multtest" package (http://www.bioconductor.org). To correct for multiple testing, a modified Bonferroni correction procedure was used. Since for each class we have analyzed RNA samples belonging to the same biological source (i.e., same individual) we decided to use a more stringent statistical cutoff as a measure of differential expression (i.e., ≥2-fold change in mRNA abundance, adjusted p value <0.00001).

Functional analysis of differentially expressed genes was performed using the program GOstat (http://gostat.wehi.edu.au/). Briefly, significantly overexpressed genes were selected as a test group and underexpressed genes were selected as the reference group. GO terms from both groups are then compared via Fisher's exact test and approximated p values are computed for each of them. Since the number of GO terms tested is large, we corrected the nominal p values using the Bonferroni correction procedure.

Reverse engineering of the transcriptional regulatory network associated with SF response to synovial fluid using CARRIE

Normalized data and differential expression significance p values were uploaded into the CARRIE server (http://zlab.bu.edu/CarrieServer/html/). CARRIE is a computational method for transcriptional regulatory network inference from microarray analysis using promoter sequence information [16]. Briefly, microarray results are used to discriminate positive (altered expression) from negative (expressed but constant) groups of genes. From these two groups, statistically overrepresented TFs whose binding sites are more abundant in the positive set relative to the negative set are determined using the ROVER algorithm. Moreover, ROVER also identifies overabundant promoters that are likely to be regulated by the predicted TFs. All the information is finally gathered into CARRIE, where a graphical network is computed highlighting regulatory relationships between TFs with regulated genes.

We used the same criteria for significant expression change as for differential expression analysis (i.e., adjusted $p<0.00001$, fold change ≥2). PSSMs for human TF binding sites were obtained from the TRANSFAC Professional version 7.2 database. Promoter sequences for the list genes in CodeLink Human IA Bioarrays were downloaded using the PromoSer server (http://biowulf.bu.edu/zlab/PromoSer/) version 3.0 (based on NCBI Build 34). For each gene 2000 bases upstream of the transcription start site (TSS) and 50 bases downstream of the TSS were retrieved, selecting only the sequence that was closest upstream to the 5′ end, excluding guessed entries and ignoring promoter region overlaps or assembly gaps. The frequency of significant binding sites for a single TF in a random promoter was set to 0.0001 and the cutoff for binding site overabundance was $p=0.001$.

Since the TRANSFAC collection of matrices contains, in some cases, many matrices for a particular TF, we applied redundancy reduction to obtain only the highest scoring PSSM for each group. Significance values were also corrected for multiple testing using the default Benjamini and Yekutieli stepped correction for control of the FDR.

Gene association study: population selection

To increase the power of the sample-based case–control design we used the hypernormal control group strategy described by Morton and Collins [46]. Hypernormal controls are defined as those individuals with the lowest liability to develop the disease. By using extreme discordant phenotypes the efficiency of the association study is usually much greater than with normal controls. The details of the hypernormal control group ($n=181$) and the cohort of RA patients ($n=257$) are given in Ref. [47]. This study was approved by the ethics committee of the Institut de Recerca Hospital Universitari Vall d'Hebron.

Gene association study: SNP selection and genotyping

For all 13 genes coregulated in NF-κB transcriptional network we selected SNPs in or close to their proximal promoter (2000 bases upstream from TSS). We used Entrez dbSNP (Human Build 35.1) and HapMap (HapMap Public Release 18) database information to select the most informative SNPs. We used the following selection criteria: (i) heterozygosity >0.2 in Caucasian European samples and (ii) in the proximal promoter or in the LD block that encompasses this region, built under standard definition [48]. In those cases in which the second criterion did not apply we favored the selection of multiple close SNPs to increase the informativeness of the locus by building multimarker haplotypes.

Genotyping was done on a MALDI-TOF mass spectrometer (MassArray System) using the Spectrodesigner software (Sequenom) for primer selection and multiplexing and the homogeneous mass-extension process for producing primer extension products. We regenotyped >30% of the samples in cases and controls with a genotype concordance of 100% between independent runs. Genotyping was done at the CEGEN, Nodo de Santiago, Santiago de Compostela, Spain. All primer sequences are available on request.

Population-based association analysis

All association analyses were performed using R (version 2.1.0). Hardy–Weinberg equilibrium was tested using the Genetics package. Association analysis of each SNP was performed using Pearson χ^2 test of the null hypothesis, and exact p values were obtained by Monte Carlo simulation (2000 replicates). Haplotype analysis was performed using the Genecounting/Permute program implemented in the Gap package. Briefly, case–control haplotypes are analyzed using model-free analysis and permutation tests of allelic association [49].

Genetic epistasis analysis using multifactor dimensionality reduction

We analyzed high-order interactions between NF-κB transcriptional regulatory network SNPs using the MDR method. MDR is a nonparametric model-free method designed to analyze gene×gene or gene×environment interactions [13]. Traditionally, the analysis of such high-dimensional data has been hampered by the limitations of statistical modeling techniques. MDR attempts to address this limitation using a data reduction approach called constructive induction [18]. Basically, multilocus genotypes (n dimensions) are pooled into a single risk predictor variable with only two dimensions (i.e., high risk or low risk). The predictive performance of the best model is then assessed through k-fold cross-validation and its significance determined through Monte Carlo permutation testing. It is important to note that MDR assigns a data point as "unknown" when there are no data points for that genotype combination in the training set. This becomes of particular relevance for larger n-way combination analyses, since more contingency-table cells can contain no observations.

The present analysis was performed with the MDR software Beta 1.0.0 RC1 (http://www.epistasis.org/open-source-mdr-project.html). Statistical interpretation of the significance of the MDR model is facilitated by the visualization of an interaction dendrogram. This dendrogram is built from a hierarchical cluster analysis of entropy-based measures of interaction information between polymorphisms.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.ygeno.2007.03.011.

References

[1] L.S. Davis, A question of transformation: the synovial fibroblast in rheumatoid arthritis, Am. J. Pathol. 162 (2003) 1399–1402.
[2] L.C. Huber, et al., Synovial fibroblasts: key players in rheumatoid arthritis, Rheumatology (Oxford) 45 (2006) 669–675.
[3] P.V. Kasperkovitz, et al., Fibroblast-like synoviocytes derived from patients with rheumatoid arthritis show the imprint of synovial tissue heterogeneity: evidence of a link between an increased myofibroblast-like phenotype and high-inflammation synovitis, Arthritis Rheum. 52 (2005) 430–441.
[4] H. Asahara, et al., Direct evidence of high DNA binding activity of transcription factor AP-1 in rheumatoid arthritis synovium, Arthritis Rheum. 40 (1997) 912–918.
[5] K. Aupperle, et al., NF-kappa B regulation by I kappa B kinase-2 in rheumatoid arthritis synoviocytes, J. Immunol. 166 (2001) 2705–2711.
[6] Y. Yamanishi, et al., Regional analysis of p53 mutations in rheumatoid arthritis synovium, Proc. Natl. Acad. Sci. USA 99 (2002) 10025–10030.
[7] A.J. Hartemink, Reverse engineering gene regulatory networks, Nat. Biotechnol. 23 (2005) 554–555.
[8] D. Segre, A. Deluna, G.M. Church, R. Kishony, Modular epistasis in yeast metabolism, Nat. Genet. 37 (2005) 77–83.
[9] K. Basso, et al., Reverse engineering of regulatory networks in human B cells, Nat. Genet. 37 (2005) 382–390.
[10] H.G. Zhang, et al., Novel tumor necrosis factor alpha-regulated genes in rheumatoid arthritis: identification of Naf1/ABIN-1 among TNF-alpha-induced expressed genes in human synoviocytes using oligonucleotide microarrays, Arthritis Rheum. 50 (2004) 420–431.
[11] M. Pierer, et al., Chemokine secretion of rheumatoid arthritis synovial fibroblasts stimulated by Toll-like receptor 2 ligands, J. Immunol. 172 (2004) 1256–1265.
[12] G.S. Firestein, Evolving concepts of rheumatoid arthritis, Nature 423 (2003) 356–361.
[13] M.D. Ritchie, et al., Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer, Am. J. Hum. Genet. 69 (2001) 138–147.
[14] J.H. Moore, A global view of epistasis, Nat. Genet. 37 (2005) 13–14.
[15] Y.M. Cho, et al., Multifactor-dimensionality reduction shows a two-locus interaction associated with Type 2 diabetes mellitus, Diabetologia 47 (2004) 549–554.
[16] P.M. Haverty, U. Hansen, Z. Weng, Computational inference of transcriptional regulatory networks from expression profiling and transcription factor binding site identification, Nucleic Acids Res. 32 (2004) 179–188.
[17] L.W. Hahn, M.D. Ritchie, J.H. Moore, Multifactor dimensionality reduction software for detecting gene–gene and gene–environment interactions, Bioinformatics 19 (2003) 376–382.
[18] J.H. Moore, et al., A flexible computational framework for detecting, characterizing, and interpreting statistical patterns of epistasis in genetic studies of human disease susceptibility, J. Theor. Biol. 241 (2006) 252–261.
[19] P.P. Tak, G.S. Firestein, NF-kappaB: a key role in inflammatory diseases, J. Clin. Invest. 107 (2001) 7–11.
[20] X. Li, S. Makarov, An essential role of NF-kappaB in the "tumor-like" phenotype of arthritic synoviocytes, Proc. Natl. Acad. Sci. USA 103 (2006) 17432–17437.
[21] G.S. Firestein, NF-kappaB: Holy Grail for rheumatoid arthritis? Arthritis Rheum. 50 (2004) 2381–2386.
[22] C. Georganas, et al., Regulation of IL-6 and IL-8 expression in rheumatoid arthritis synovial fibroblasts: the dominant role for NF-kappa B but not C/EBP beta or c-Jun, J. Immunol. 165 (2000) 7199–7206.
[23] S. Haskill, et al., Identification of three related human GRO genes encoding cytokine functions, Proc. Natl. Acad. Sci. USA 87 (1990) 7732–7736.
[24] M. Taberner, K.F. Scott, L. Weininger, C.R. Mackay, M.S. Rolph, Overlapping gene expression profiles in rheumatoid fibroblast-like synoviocytes induced by the proinflammatory cytokines interleukin-1 beta and tumor necrosis factor, Inflammation Res. 54 (2005) 10–16.
[25] S.M. Dunn, et al., Requirement for nuclear factor (NF)-kappa B p65 and NF-interleukin-6 binding elements in the tumor necrosis factor response

region of the granulocyte colony-stimulating factor promoter, Blood 83 (1994) 2469–2479.
[26] R. Debets, et al., Two novel IL-1 family members, IL-1 delta and IL-1 epsilon, function as an antagonist and agonist of NF-kappa B activation through the orphan IL-1 receptor-related protein 2, J. Immunol. 167 (2001) 1440–1446.
[27] H. Takeda, et al., Keratinocyte-specific modulation of type VII collagen gene expression by pro-inflammatory cytokines (tumor necrosis factor-alpha and interleukin-1beta), Exp. Dermatol. 14 (2005) 289–294.
[28] D. Magne, et al., The new IL-1 family member IL-1F8 stimulates production of inflammatory mediators by synovial fibroblasts and articular chondrocytes, Arthritis Res. Ther. 8 (2006) R80.
[29] T.A. Libermann, D. Baltimore, Activation of interleukin-6 gene expression through the NF-kappa B transcription factor, Mol. Cell Biol. 10 (1990) 2327–2334.
[30] C. Copie-Bergman, et al., Interleukin 4-induced gene 1 is activated in primary mediastinal large B-cell lymphoma, Blood 101 (2003) 2756–2761.
[31] M. Wu, L.G. Xu, Z. Zhai, H.B. Shu, SINK is a p65-interacting negative regulator of NF-kappaB-dependent transcription, J. Biol. Chem. 278 (2003) 27072–27079.
[32] J.C. Betts, J.K. Cheshire, S. Akira, T. Kishimoto, P. Woo, The role of NF-kappa B and NF-IL6 transactivating factors in the synergistic activation of human serum amyloid A gene expression by interleukin-1 and interleukin-6, J. Biol. Chem. 268 (1993) 25624–25631.
[33] R.L. Strausberg, et al., Generation and initial analysis of more than 15,000 full-length human and mouse cDNA sequences, Proc. Natl. Acad. Sci. USA 99 (2002) 16899–16903.
[34] M. Johannesson, et al., Identification of epistasis through a partial advanced intercross reveals three arthritis loci within the Cia5 QTL in mice, Genes Immun. 6 (2005) 175–185.
[35] P. Olofsson, P. Wernhoff, J. Holmberg, R. Holmdahl, Two-loci interaction confirms arthritis-regulating quantitative trait locus on rat chromosome 6, Genomics 82 (2003) 652–659.
[36] M. Brenner, et al., The non-MHC quantitative trait locus Cia5 contains three major arthritis genes that differentially regulate disease severity, pannus formation, and joint damage in collagen- and pristane-induced arthritis, J. Immunol. 174 (2005) 7894–7903.
[37] A.S. Rigby, L. Voelm, A.J. Silman, Epistatic modeling in rheumatoid arthritis: an application of the Risch theory, Genet. Epidemiol. 10 (1993) 311–320.
[38] J. Newton, et al., The effect of HLA-DR on susceptibility to rheumatoid arthritis is influenced by the associated lymphotoxin alpha-tumor necrosis factor haplotype, Arthritis Rheum. 48 (2003) 90–96.
[39] A. Martinez, et al., Epistatic interaction between FCRL3 and NFKB1 genes in Spanish rheumatoid arthritis patients, Ann. Rheum. Dis. (2006).
[40] M. Pascual, et al., IL-6 promoter polymorphisms in rheumatoid arthritis, Genes Immun. 1 (2000) 338–340.
[41] H. Sahara, N. Shastri, Second class minors: molecular identification of the autosomal H46 histocompatibility locus as a peptide presented by major histocompatibility complex class II molecules, J. Exp. Med. 197 (2003) 375–385.
[42] F. Ota, et al., Activin A induces cell proliferation of fibroblast-like synoviocytes in rheumatoid arthritis, Arthritis Rheum. 48 (2003) 2442–2449.
[43] H. Okamoto, et al., The synovial expression and serum levels of interleukin-6, interleukin-11, leukemia inhibitory factor, and oncostatin M in rheumatoid arthritis, Arthritis Rheum. 40 (1997) 1096–1105.
[44] I. Siegle, et al., Expression of cyclooxygenase 1 and cyclooxygenase 2 in human synovial tissue: differential elevation of cyclooxygenase 2 in inflammatory joint diseases, Arthritis Rheum. 41 (1998) 122–129.
[45] C.A. Murphy, et al., Divergent pro- and antiinflammatory roles for IL-23 and IL-12 in joint autoimmune inflammation, J. Exp. Med. 198 (2003) 1951–1957.
[46] N.E. Morton, A. Collins, Tests and estimates of allelic association in complex inheritance, Proc. Natl. Acad. Sci. USA 95 (1998) 11389–11393.
[47] A. Julia, et al., Lack of association between the corticotropin-releasing hormone locus and rheumatoid arthritis, Arthritis Rheum. 50 (2004) 2706–2708.
[48] S.B. Gabriel, et al., The structure of haplotype blocks in the human genome, Science 296 (2002) 2225–2229.
[49] J.H. Zhao, D. Curtis, P.C. Sham, Model-free analysis and permutation tests for allelic associations, Hum. Hered. 50 (2000) 133–139.

Chapter 5

Genomics of genetic variation and complexity

5.1 Factors preceding GWAS: technological development

At the beginning of the 20th century everything was ready for whole genome association studies. Susceptibility to common diseases was now thought to be due to common genetic variations in populations, the LD–based association approach had been proven to have increased statistical power compared to the previous linkage approach, and millions of SNPs in the human genome had been already identified. However, the definite step for GWAS studies came when the biotechnological industry finally managed to develop a technology that could massively genotype these polymorphisms in a fast and relatively cheap manner. Thus, from a commercial perspective, it could be said that the GWAS "era" started on June 28^{th} 2004, when Affymetrix publicly announced the availability of microarray–based genotyping systems for over 100,000 SNPs. With the same photolithography technology they had been using for the manufacturing of gene–expression microarrays, and using the same DNA hybridization properties Ed Southern had already exploited in the 70s (Wallace *et al.*, 1979), they were able to offer to biomedical scientists a first tool for genome-wide analysis of genetic variation.

5.2 AMD as a model of a GWAS approach

In 2005, Josephine Hoh and collaborators published the first study that identified a susceptibility variant associated to a common disease (Klein *et al.*, 2005). Using Affymetrix 100K platform, they were able to identify a polymorphism in the complement factor H (*CFH*) gene statistically associated with Acute Macular Degeneration (AMD), the most common form of blindness in the elderly. This finding was the first proof of concept that the hypothesis–free, LD–based approach of whole genome studies could be used to identify relevant genetic variation associated to a complex disease. From this moment to date, an exponential number of genome-wide association studies for complex diseases and other relevant human traits have been published.

5.2.1 Coverage of the genome using the indirect association method

The AMD GWAS study is useful to highlight several fundamental aspects that we encounter in the genome-wide analysis approach. Similar to the microsatellite panels available for family linkage studies, the 100,000 SNPs that composed the Affymetrix 100K array had been selected randomly and evenly distributed throughout the genome. In other words, we should expect that we would analyze a SNP every 30,000 pb of the genome. If LD was sufficiently strong between each neighbouring SNP, then almost all genetic information should be covered up by this approach. However, as Lander's team had already demonstrated in 2001 (Daly *et al.*, 2001), LD is not monotonic along the chromosomes and, instead, it behaves like a block–like structure with regions of high and regions of low LD. Therefore, the even–marker spacing approach used by Affymetrix should be deemed to have lower power to detect relevant associations. If one really wanted to cover the (common) genetic variability information of the genome using the least number of markers possible, one should use its LD information: using less markers in high LD regions and using more markers in regions of low LD. Today we know that the 100K microarray used by Hoh only covers less than 1/3 of the total genetic variation that can be obtained by the GWAS approach (Barrett &

Cardon, 2006). Thus, Hoh and collaborators were quite lucky to hit the right SNP.

5.2.2 Genomic studies: the multiplicity problem

Another relevant aspect of the AMD study relates to the sample size that was used. With only 96 cases and 50 controls they were able to conclude that one SNP was "(...) *significantly associated with disease status*". Whilst this might seem a rather normal situation for a single variant analysis, it is extremely surprising if we take into account that more than 100,000 genetic variants were analyzed in parallel. In statistical analysis, it is assumed that, the more hypothesis one tries to test simultaneously, the more likely that any of them will appear to be significant just by random variation. This is called the multiple test problem and it is a fundamental statistical issue to which extensive research has been devoted. One of the first (and still most used) methods to deal with this problem is the Bonferroni correction method (Hochberg, 1988). In this method, the different significance values (i.e. P values) are assumed to be true and to come from a normal distribution function. Then, the Bonferroni adjusted P value becomes the probability of rejecting at least one Hypothesis given that all null Hypothesis are true. Thus, the Bonferroni corrected family–wise error (i.e. the probability of rejecting a true null Hypothesis) is simply calculated by dividing the desired experimental family–wise error (i.e. alpha value) by the number of tests performed. This means that for a SNP to be statistically associated in Hoh's GWAS design it should have a P value of $0.05/100{,}000 = 5e{-}7$ (or, more precisely 4.8e–7 because the Affymetrix GeneChip consisted of some more SNPs). In the AMD GWAS analysis, the CFH SNP showed a P value of 4.15e–8 and thus it was considered significant.

5.2.3 Factors influencing the statistical power of a GWAS

How probable it was for Hoh's study to obtain a significant result with such a small sample? The answer is not straightforward since it depends on many factors. As we have just seen it depends on which markers we choose: if we include the causal SNP or a SNP in high LD with the causal genetic variant we will have high

power to detect the association. We have also seen that it depends on the type I error rate, alpha, and the multiple testing burden. However, there are several other factors that influence our ability to statistically detect one genetic variant associated with the disease. Some of the most relevant factors that influence the power to detect an association are:

i The frequency of the associated allele (p): the closer it is to 0.5, the higher the probability to detect the association.

ii The sample size: like any other statistical analysis, the power is proportional to the sample sizes of cases and controls that we are studying.

iii The disease prevalence in the general population: the higher the prevalence, the higher the probability we will be able detect a true positive result.

iv The genotyping error: like any other measuring technique, genotyping techniques have a level of error which can negatively influence our ability to detect true results or, even, can induce false positive results.

v The genetic model of disease risk: the power depends on the mathematical description on how each of the possible genotypes contributes to the disease (i.e. much less samples are needed to detect a multiplicative model of risk compared to a recessive model).

vi The effect size: it depends on how strong is the association of the genetic variant with the trait.

vii The level of complexity in the disease: which includes trait and/or genotype heterogeneity, and the presence of epistasis or gene–environment interactions.

viii The presence of confounders: the existence of population stratification or admixture.

Thus, if we assume that we are typing the causal SNP (or in very high LD, r^2 >0.99), the corrected significance after 100,000 simultaneous tests, the reported disease allele frequency (p = 0.45), the number of cases and controls (n = 95 and

n = 50), the prevalence of AMD in the US Caucasian population (p $\sim$ 0.015), absence of genotyping error, a multiplicative model of risk, a high effect size (i.e. the reported allelic OR was 4.2), minimal presence of heterogeneity and/or interaction effects and minimal effect of stratification effects (i.e. all individuals were American "white and non–Hispanic"), the final answer is yes, there was enough power. More specifically, taking into account all of the above parameters, there is a probability P >0.8 of rejecting the null hypothesis given that it is false (i.e. the probability that we will not make a type II error). No wonder, however, why it took two more years to publish a second GWAS study on a complex disease.

5.2.4 Pre–GWAS scenario: scepticism vs. enthusiasm

The AMD study proved that the genome-wide approach was a useful strategy for the study of complex diseases, raising much enthusiasm and expectations within the biomedical community. There was also, however, some level of scepticism on the probability of finding such a favourable combination of parameters in other diseases or other complex traits. Who would be right? Like in the previous whole genome family linkage studies there was no other way out: it had to be tried. It is within this high expectations and scepticism moment that our research group planned to conduct a GWAS study in RA.

5.3 Objectives of the study

The present study was designed to try to answer the next questions:

- Are there new genomic regions associated to RA disease risk?
- Which genomic regions show increased signals when integrating the results of our GWAS with previous GWASs in RA?
- Are there any significant epistatic interactions associated with the susceptibility to RA?

5.4 The study design

5.4.1 Selection of the whole genome genotyping platform

The first step was to choose the genotyping technology. At that time, Affymetrix microarrays (wwww.affymetrix.com) were evidently the first option. They had dominated the commercial microarray gene expression scenario, with a high volume of bioinformatic methodology developed around them. Illumina, a San Diego biotech company (www.illumina.com), had been developing a microarray technology based on new composite material. These were called BeadChips because they were based on arraying micrometric beads, each one carrying a sequence complementary to a specific sequence and specific SNP allele. After allele–specific hybridization, an enzymatic–based extension is performed in which labelled nucleotides are incorporated. After extension, these nucleotides will be visualized by a sandwich–based immunohistochemistry assay (Figure 5.1).

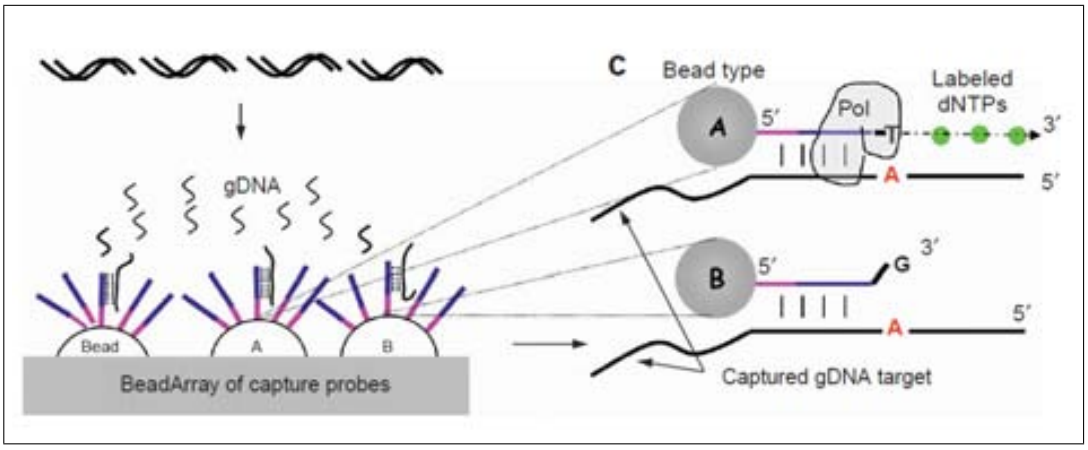


Figure 5.1: Illumina microarray technology for SNP allele identification. (Taken from Gunderson et al. (Gunderson *et al.*, 2005). After the amplification of the genomic DNA, this is hybridized to the 50–mer probes attached to micrometric beads in the array. Finally, an enzymatic allele–specific primer extension is performed. The level of synthesized DNA is quantified using fluorescence and will be used to infer the genotype of the individual.

Similar to the gene expression counterparts, there was a good advantage of the Illumina system over the Affymetrix technology: the probe size. The Illumina

system could fix to their beads 50 nucleotide oligomers whereas Affymetrix could only synthesize probes up to 25 bases long. Thus, although SNP genotyping is not affected by technical noise to the level of mRNA gene expression data, we decided to minimize the risk of false calling due to a lower signal to noise ratio and we chose the Illumina platform. At the time we started planning the study, Illumina had only 1 microarray model available: the Human–1 Beadchip. This model was called "exon–centric" because it was enriched for SNPs within transcribed sequences or in very close proximity to a gene. Although it did not cover untranslated regions like the 100K or the newly introduced (and more expensive) 500K from Affymetrix, the Human 1 seamed the most promising option at the moment. However, the biotechnology industry was already a fast–evolving sector and only some months after the launching of the Human–1 there was a second version capable of genotyping more than 300,000 SNPs. This was called the HumanHap 300 Beadchip and unlike the previous version, it used the LD information generated from the HapMap project (IHC, 2003). The first phase of the HapMap project, which had officially started in 2002 and was completed in 2005, had already characterized the frequencies of several millions of SNPs throughout the genome using three human populations with very different ancestry: European Caucasians, Yoruba Indians from Nigeria and Japanese–Chinese. With this dense genetic information it was now possible to determine the haplotype–block structure of the genome and, from this, select the minimal set of SNPs which would give maximal informativity of the genetic variation (i.e. tagSNPs). Therefore, although it had approximately 180,000 less SNPs than its immediate competitor Affymetrix 500K, its "intelligent" selection of SNPs allowed superior coverage of the whole genome variability (75 vs. 65% coverage, respectively)(Barrett & Cardon, 2006). For all these reasons we finally decided to upgrade to the HumanHap 300 microarray for our GWAS in RA.

5.4.2 Selection of study subjects

The second important aspect in our study design was the sample selection. It was clear that it was going to be a non–randomized retrospective case–control design and it was also clear the number of samples we could afford to genotype.

Although the only precedent was Hoh's successful AMD study with a sample size five times smaller than ours, we chose to implement a study design that could increase the power of our study. Whole genome linkage studies in RA had only successfully replicated the HLA region so it seemed unlikely that there could exist a genetic variant of a similar effect size. Like in the gene expression candidate approach described in Chapter 4, we chose to implement Morton's liability approach to increase the power of our study. This time, however, we did not only select hypernormal control patients but we also extended the liability concept to produce other analysis subgroups. We hypothesized that, on the other extreme of liability, not only patients diagnosed with RA but also patients from other chronic inflammatory disease would be sharing part of their susceptibility background. Also, within RA patients, we considered two levels of heterogeneity: RA patients recently diagnosed as RA following the ACR criteria (broadest heterogeneity) and RA patients diagnosed following the ACR criteria but having a longstanding disease with severe joint damage (highest homogeneity). The collection of such specific sets of patients was possible thanks to the collaboration with Spanish rheumatologists Dr. Javier Ballina (Hospital Universitario de Asturias), Juan de Dios Cañete (Hospital Clínic de Barcelona), Jesús Tornero (Hospital Universitario de Guadalajara) and Alejandro Balsa (Hospital La Paz, Madrid). They are all well–known specialists in the Spanish Rheumatology field, and they are active clinical researchers. Some had their own repository of genomic DNA which helped to speed out the sample collection process.

5.4.3 The importance of QC analysis in GWAS data

An important aspect in any scientific study, quality control analysis is crucial in data generated from high–throughput technologies. Although there was no doubt that the genotyping platform chosen was robust, it was compelling to determine that no technical artefact could be affecting our results. One first quality control measure is the genotyping call rate, that is, the percentage of SNPs in one array that is assigned to a particular genotype with sufficient confidence. As we commented previously, Illumina genotyping technology is based on the complementary hybridization to two sets of beads, each one representing one of the 2

possible alleles (Figure 5.2).

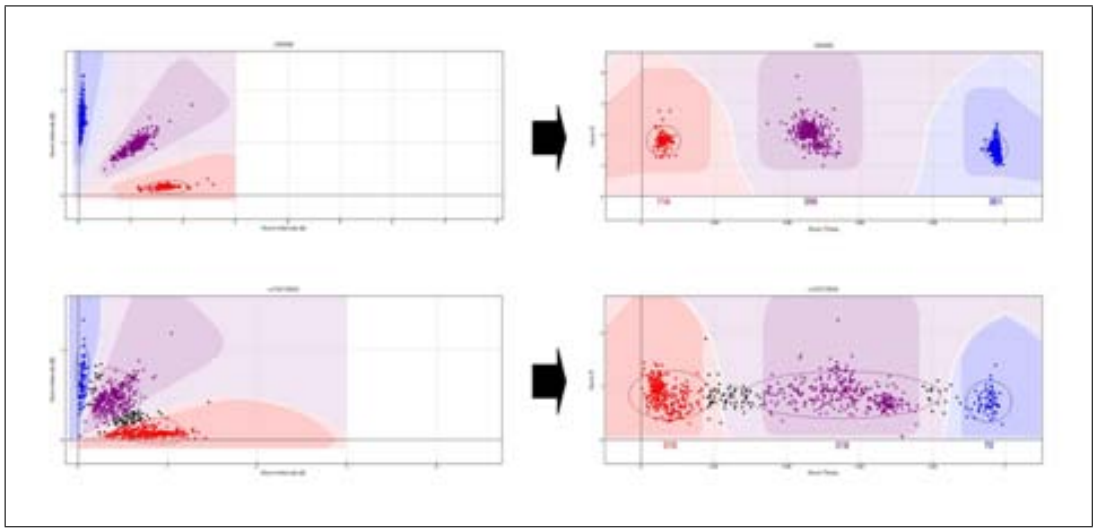


Figure 5.2: Examples of intensity clusters for two different SNPs. The probe intensities for two SNPs are shown in Cartesian (left) and polar (right) coordinates. Each dot represents the genotype of an individual and the three colours, the cluster assignment to each of the 3 possible genotypes. The upper SNP is an example of a neat clustering and genotype call, whilst the lower SNP has a high level of variability which makes it difficult to ascertain certain genotypes with confidence (i.e. missing genotypes, black dots in the figures).

Thus, if one individual is homozygous for a specific SNP, only one of the two sets of bead will hybridize and emit fluorescence. Consequently, in the heterozygous case, a similar amount of hybridization and fluorescence emission will occur in both sets of beads. For each individual and for each SNP we will have three levels of possible intensity per allele bead and two beads per SNPs from we will be able to infer the genotype. Nonetheless, genotype inference (or genotype "calling") is not a white or black process there can be a huge amount of variability depending, principally, on the probe sample labelling, the probe hybridization and the scanning steps. Also, each set of SNP probes has its own particular thermodynamics and, consequently, there is a quite large amount of different signal distributions produced. This means that genotype calling performed in a single individual can be very prone to error. Instead, genotyping algorithms use multi-

ple samples to estimate the 3 possible genotype clusters (i.e. the two homozygous and the heterozygous genotypes) per each SNP. Like for any other unsupervised analysis technique, the more samples we have, the better our cluster estimates will be and the more confident we can be on the estimated genotypes. On average, more than 98% of the SNPs were called with high confidence (Figure 5.3), confirming the quality of the chosen technological platform.

Another SNP–level quality control measure comes from the determination of the deviance from the Hardy Weinberg Equilibrium. In a non–isolated population like the Spanish population, were random mating and lack of selective pressure can be assumed, the genotype distribution for any genetic marker should tend towards Hardy–Weinberg Equilibrium. However, if we observe a high statistical deviance from HWE in the distribution of a SNP in the control group, it is generally assumed to be due to bad quality genotyping and the SNP is consequently discarded from any further analysis. Logically, we cannot assume the same if Hardy Weinberg deviations occur only in the case group: if the SNP is associated with the disease, this will tend to cause altered genotype frequencies. In fact, several authors have proposed to use this property and build association tests based on exclusively genotype frequencies in case individuals (Nielsen *et al.*, 1998).

5.4.4 The fear of population stratification: the TDT test

During the 90's one of the greatest fears of genetic epidemiologists was the possibility of confounding due to population stratification. One most commented work regarding this negative effect is the article from Knowler and colleagues published in 1988 (Knowler *et al.*, 1988). In this study they found a statistically significant association of a particular HLA haplotype with non–insulin–dependent diabetes mellitus (NIDDM) in Pima Indians (i.e. a North American native Indian tribe with a high prevalence of this disease). This haplotype, however, is not present in full–blood native Pima Indians but, instead, it had been introduced by recent admixture with Caucasian European populations. Thus, by comparing two sample groups which differed in the level of admixture in their ancestry, they had created a spurious association between a genetic loci and disease. Genetic

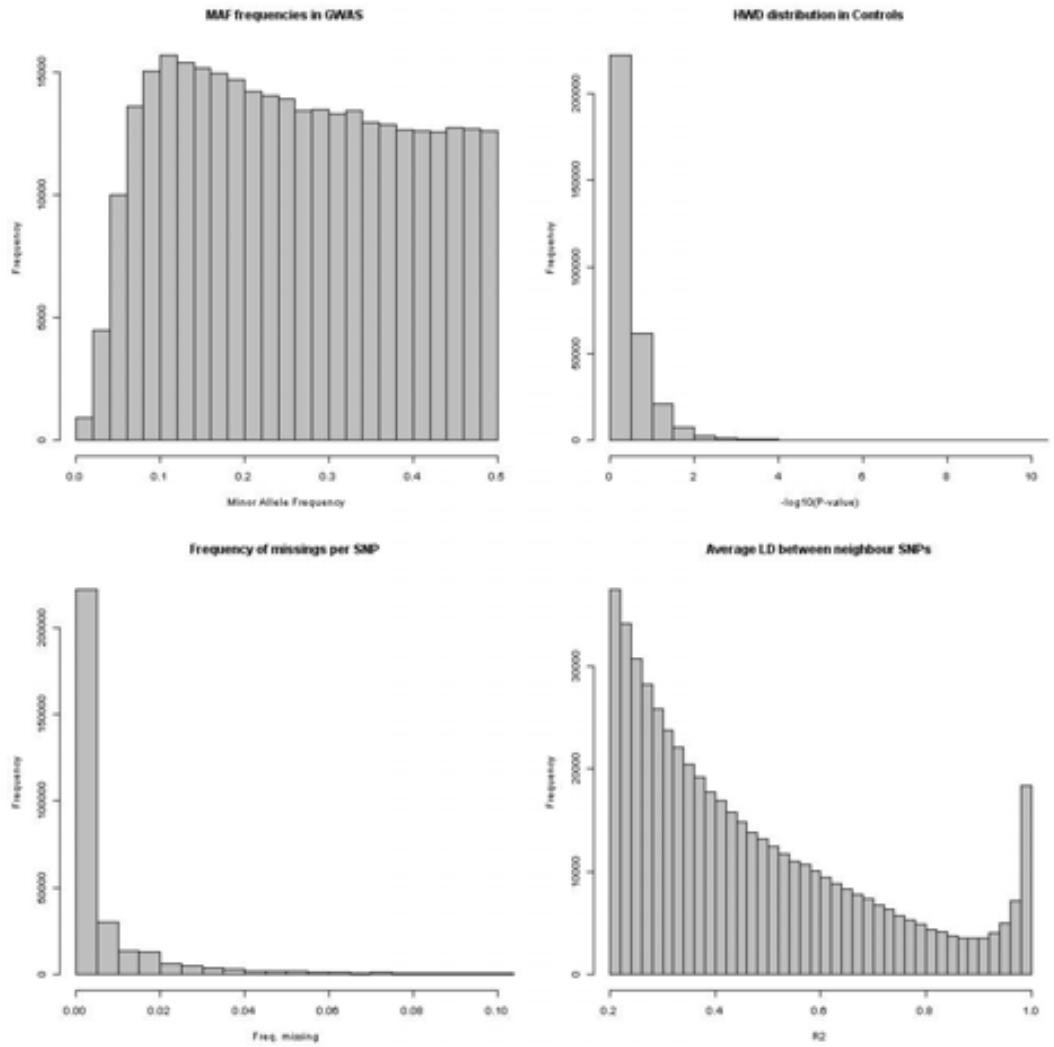


Figure 5.3: Barplots of different GWAS Quality Control measures. As expected, most of the SNPs are common in the general population with minor allele frequencies >5% (upper left plot). Also, deviations from Hardy Weinberg Equilibrium are extremely rare in the control group (upper right plot). The Illumina genotyping platform has a high calling rate, with most of the SNPs having less than 5% missingness (down left plot). The coverage of the array is represented by the LD measures: most SNPs have pairwise LD values r2 >0.2 (down right plot).

epidemiologists feared that this could happen to their studies and so, new analytical approaches were developed to fight against it. From these, the method that became most popular was the Transmission Disequilibrium Test (TDT) by Richard Spielman and Warren Ewens (Spielman *et al.*, 1993). In the first version of the method, family trios (i.e. the proband and its two parents) were analyzed: the parental alleles that were not transmitted to the offspring were then used as a matched pseudo–control, thus avoiding the problem of differential ancestry albeit with a reduction on statistical power. However, after several years of using the so called Family–Based Association Tests (FBATs), the number of loci consistently associated to complex diseases did not increase. It seemed that, if one was careful enough to avoid collecting samples with evident population stratification (i.e. like the Pima Indians), the loss of power of FBATs seemed not to compensate the effort. FBATs were gradually superseded by traditional case–control approaches were the negative effect of stratification was prevented by rational sample ascertainment. Nonetheless, with the progressive introduction of high–throughput genotyping technologies, the search for methods for the identification and control of the population stratification continued. It was evident that, although stratification was not the big problem it had been initially thought, it could be a considerable bias factor when trying to address the association of genetic variations with relatively low penetrance (Clayton *et al.*, 2005).

5.4.5 Bayesian approaches to correct for population stratification

One of the good things of working with a high number of genetic markers is that the problem of population stratification can be more robustly solved. In this case, the basic working assumption is that most genotyped markers will not be associated with disease, which is not a hard assumption when we are genotyping thousands of markers. Bernie Devlin and Katthryng Roeder devised in 1999 a Bayesian approach to deal with stratification called Genomic Control (GC) (Devlin & Roeder, 1999). This approach is based in the assumption that, if there is stratification within the genotyped samples, this should lead to a distribution of association statistics different from what is expected under the null hypothesis.

Thus, under no stratification, the median value of the observed chi–square values should be equal to the median of a 1 degree of freedom chi–square distribution which is 0.456. The ratio of both measures (i.e. observed vs. expected) will inform us about the degree of stratification in our sample or, as they call it, genomic inflation (λ). The more our genomic inflation factor deviates from a ratio of 1, the more we should worry that there is a stratification problem in our sample.

5.4.6 *A priori* identification of population outliers

We do have a method to detect stratification using thousands of markers but, wouldn't it be possible to exclude outlying individuals using a small set of markers and avoid genotyping them in expensive high throughput technologies? At the same time Devlin and Roeder described their GC method, Jonathan Pritchard from Stanford and Noah Rosenberg from Oxford described an alternative method to test for stratification called STRUCTURE (Pritchard & Rosenberg, 1999). In this last method, they used the genetic information from a reduced set of markers to cluster individuals into a predefined number of clusters (i.e. populations). Through simulation studies, they showed that with a reduced number of markers (around 15 microsatellites or approximately 30 SNPs) this objective could be reached. Since at that time there was little notion about the level of population stratification in the Spanish population, we decided to carry out an *a priori* filtering approach to try to discard strong outliers from our case–control cohorts before genotyping them with the Illumina array. In order to look for the best ancestry–informative markers, we used a set of SNPs (n = 34) that had been validated as informative for forensic–based purposes by an international consortium (Phillips *et al.*, 2007). In order to have a positive control for stratification, we included two individuals that had a clear different ancestry from the general Spanish population: a Moroccan and a Peruvian Andean women. The analysis was performed using STRUCTURE by our group and by an independent genetics research group who was blind to the origin status of the individuals. Neither they nor we did find any trace of population stratification but, unfortunately, neither they nor we managed to identify any of the two outliers as such (Figure

5.4). Thus, the validity of the approach was clearly at stake: either the method was inappropriate or the chosen SNPs were not that informative. Regarding the last possibility, using this same set of markers a recent study was able to identify the origin from several of the unmatched DNA samples of the 11^{th} March 2004 terrorist bomb attack (Phillips *et al.*, 2009). However, one crucial difference with our study is that they precisely targeted only two specific populations (i.e. Moroccan and Spanish) and they were able to first train their predictor model with an independent dataset. Still, they were only able to assess with confidence the origin of 4 of the 7 unmatched samples.

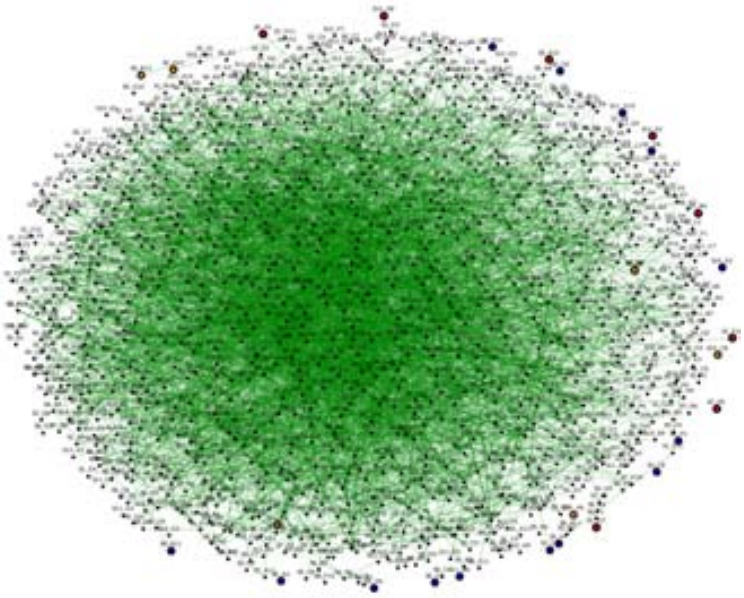


Figure 5.4: Mulitidimensional Scaling Analysis of ancestry–informative markers. Each point in the graphic corresponds to an individual in the study; the distance between individuals is proportional to their genetic differences in their genetic markers.

5.4.7 Ascertainment of geographic origin

With the previous approach we were unable to discriminate with certainty neither the presence of stratification nor individual outliers. Nonetheless, what we could do is to reduce the probability of this happening by an appropriate epidemiological control. As much as possible, we obtained the information from the

geographical origin from the grandparents of the probands. The perfect proband was one from which we could ascertain the Spanish origin from all of the four grandparents. If there was any evidence of non–Spanish origin in either of the ancestors, the individual was discarded and not genotyped. However, obtaining the records of the geographical origin of the grandparents is a costly process, especially, for old aged probands like many longstanding RA patients were it was difficult to obtain such information with reliability. Overall, we were able to obtain the complete information of all 4 grandparents from 30% of the all recruited individuals.

Genetic outliers identified using Principal Component Analysis

Discarding individuals by their place of birth or the place of birth of their ancestors is a good first line of defence against stratification issues. However, there still may be some level of geographic variation which escapes from this epidemiological measure. The history in the Spanish territory is quite rich in episodes of population admixture (Bosch *et al.*, 2001) and, therefore, there is a chance that some of these events might be still hidden in our genomes. The Genomic Control is a good measure to quantify the amount of ancestry heterogeneity present in our samples and it can even be used to adjust the association statistics. However, it tends to be an overconservative measure since it treats all SNPs exactly the same way. STRUCTURE is also not particularly well suited for whole genome data since it requires many computational resources and it also requires that one specifies the number of expected underlying clusters which, in many cases, it is an unknown parameter. Thus, we need a method that is capable to measure the presence of this unwanted genetic variability, that weights each individual according to how much it is influenced by this variability and, finally, that is computationally practical. Almost 30 years ago, Italian population geneticist Luigi Cavalli–Sforza, proposed the use of Principal Component Analysis technique (PCA) to extract the main components of variation (Menozzi *et al.*, 1978) from a group of markers. The PCA technique was first devised by statistician Karl Pearson in 1901 when he searched for a mathematical method to find the plane that would best fit a system of points of multiple variables. With this

method, the original data set is transformed into a new set of variables (the principal components or eigenvectors) which are uncorrelated and which are ordered so that the first few retain most of the variation present in all of the original variables. With this technique, Cavalli–Sforza was able to demonstrate that a high proportion of geographic variation of allele frequencies within the European continent can be explained by only the first principal component of variation. Extending Cavalli–Sforza's implementation, in 2006 Nick Patterson, Alkes Price and David Reich from Harvard University, devised a PCA–based method for GWAS data (Price *et al.*, 2006). Their method is called Eigenstrat and they were able to demonstrate in real GWAS data that it was an effective means on capturing the principal axis of variation and, optionally, correct the association statistics using this information. In the present study, we used the Eigenstrat method to infer the Principal Components of variation in our Spanish cohort.

Exclusion of ascertainment bias

The genomic inflation factor in our sample was very close to the null ($\lambda = 1.01$, $\lambda_{null} = 1$). This was the first evidence that our sample ascertainment criteria had been efficient in minimizing population stratification. Secondly, we determined the principal components of variation in our genomic data. Much like the WTCCC study, we found that the two first principal components captured most of the variation within the genome. However, whilst the WTCCC found a NorthWest to SouthEast trend in the genetic variation in the UK population, we found that the predominant trend of genetic variation in the Spanish population was principally from West to East. Interestingly, three months after the publication of our work, Novembre and collaborators published a comprehensive analysis of the genomic variation present in European populations (Novembre *et al.*, 2008). In this study they used Affymetrix 500K genotyping arrays and the same PC analysis. On one hand, they demonstrated that genomic variation in European subpopulations is sufficiently informative to predict with relatively high accuracy the place of birth of an individual (i.e. 90% individuals predicted within 700 km of their place of birth, 50% within 310 km). On the other hand, it corroborated our finding of the predominance of the West to East trend in the

sample of Spanish origin (n = 131 in their study) compared to the predominant North to South variation in other European populations (Figure 5.5).

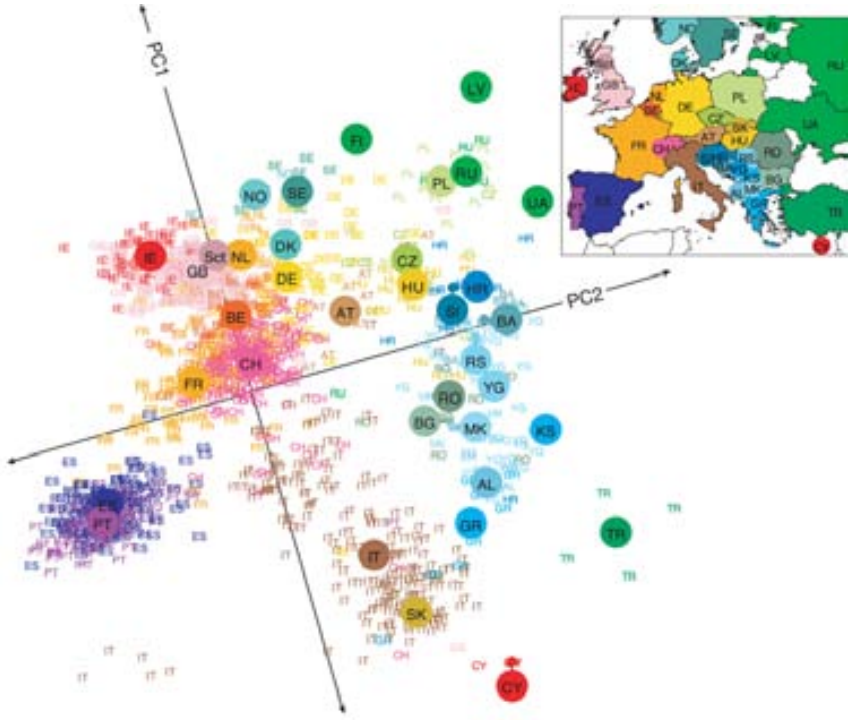


Figure 5.5: Genetic variation in the European population. (Taken from Novembre et al. (Novembre, Johnson et al. 2008)). Using the two principal components of variation inferred from GWAS data, the genetic variation in European Caucasian populations is shown to mirror the geographical variation.

The Eigenstrat method performs an outlier detection and exclusion method based on the relative distance of each individual towards the mean of each PC. If any individual is more than 6 standard deviations away, it is excluded from the sample and the eigenvectors are re–calculated. This process is repeated up to a user defined number of iterations or until no further genetic outlier is found in the sample. In our study we did found a similar number of outliers in the case and control cohorts (χ^2–test, $P = 0.34$), so we could conclude that there was no ascertainment bias.

5.4.8 Replication of GWAS candidate loci

The repeatability of an initial finding is a fundamental aspect of the scientific progress. In the study of the genetics of complex diseases, where the effects of genetic variation upon phenotype are generally small, this is a compelling issue. As we have previously seen, there are many potential factors that affect our study and lead to spurious assumptions. Although one of the most cited GWAS, the WTCCC 2007 study (WTCC, 2007), did not include a replication set, it is now indispensable to perform such validation studies. Even for Hoh's precocious 2005 GWAS, in less than three months a replication study (conducted by Hoh herself and collaborators), was immediately published confirming in an independent dataset the original association of the CFH gene with AMD. Thus, the natural question arises: which SNPs should be chosen to conduct replication? At the time we planned our replication strategy, this was already a highly debated issue. From the most conservative point of view, the answer is straightforward: to replicate only those SNPs that withstand Bonferroni's multiple test correction. However, the experience from other previous GWAS showed that even Bonferroni correction was not a safeguard against false positives. Furthermore, studies including SNPs that were only nominally significant (i.e. only significant without applying the multiple correction procedure) could replicate them in an independent dataset. One striking example is the GWAS study in RA conducted by Robert Plenge and collaborators were they included all SNPs under a nominal P value <0.001, when the Bonferroni–corrected P value is 1,000 times lower ($\sim$ 5e–7). From all 90 tested SNPs, only the SNP ranking in the 77^{th} position in the GWAS study was reproducibly replicated in an independent dataset (Plenge, Cotsapas et al. 2007). In our study, we chose two alternative strategies to select the candidate SNPs for the replication phase. The first strategy was based on a conditional probability approach: we selected those SNPs with nominal association in the extreme liability comparison (i.e. longstanding RA vs. hypernormal controls, P <0.001) that still had a nominal association in the lower liability comparison (i.e. heterogeneous RA vs. hypernormal controls, P <0.001). In the second strategy we used a statistical learning technique called bootstrapping to obtain an alternative measure of significance. Bootstrapping is a resampling–based method

that can give a better statistical inference when the parameters of the underlying distribution are unknown or in doubt (Efron, Halloran et al. 1996). Using both strategies, we finally selected 34 SNPs for replication in the independent set. As a positive control, we also included SNPs from genomic loci previously associated to RA: Protein Tyrosine Phosphatase 22 (*PTPN22*) (Begovich *et al.*, 2004), *PADI4* (Suzuki *et al.*, 2003) and *CTLA4* (Rodriguez *et al.*, 2002) genes.

After performing the replication analysis in an independent cohort of cases and controls (n = 410 and n = 394, respectively) we found that only three SNPs showed nominal significance values (P <0.05). Importantly, from the previously associated genes, only *PTPN22* was positively replicated in this sample (P = 0.022), although *CTLA4* showed a trend towards association (P = 0.06). The other nominally associated SNPs were rs7313861 (in the 3rd intron of *SV2 Related Protein* or *SVOP*, P = 0.043) and rs1324913 (in the 1st intron of *Kruppel–like factor 12* or *KLF12*, P = 0.013). Was validation at the nominal level enough to support the association of these two new candidates to RA susceptibility? We sought to look for further evidence that could support this genetic association. At the time we completed the replication analyses, the first GWAS studies in RA had been recently published in three Caucasian populations: the WTCCC study in UK population (WTCC, 2007), the Brigham and Women's Rheumatoid Arthritis Sequential Study (BRASS) in North–American population (Plenge *et al.*, 2007a), and the North American Rheumatoid Arthritis Consortium (NARAC) and the Swedish Epidemiological Investigation of Rheumatoid Arthritis (EIRA) with patients from North America and Sweden (Plenge *et al.*, 2007b). The integration of the genome-wide association signals from our study together with the signals from these previous studies could give additional information on relevant genetic loci associated to RA. Nonetheless, we had to limit the meta-analysis to those association signals that were made available by the authors of these studies (i.e. complete for WTCCC, P <0.001 for BRASS and P <0.0001 for NARAC–EIRA GWASs). Although not an orthodox meta-analysis in the traditional statistic sense, our bioinformatic approach was able to identify several genomic regions showing high clustering of association signals, like the recently associated *TNFAIP3* locus (Plenge *et al.*, 2007a) between our study and the BRASS study and, importantly, the *KLF12* locus between our study and the WTCCC study.

5.4.9 Genome-wide Scan for Epistasis

Most common diseases like RA have been categorized as complex diseases. The complexity attribute recognizes both the multifactorial origin of the disease as well as the potential existence of interactions between several of these factors. Interactions occur when the effect of one factor upon phenotype (i.e. disease), is modulated by other factors. Thus, unless we don't consider the joint distribution of these interacting factors, we will not be able to find this type of genotype–phenotype effects. Interactions in genetic epidemiology come in two forms: Gene x Environment interactions (GxE) and Gene x Gene interactions (GxG or also called epistasis).

Limitations of interaction analysis

The study of GxE interactions is very appealing for epidemiologists since it opens the possibility to perform preventive strategies: if the environmental input is avoided, the disease could also be avoided. The study of GxE interactions affecting disease risk is however limited by the lack of reliable and extensive environmental data. At present, obtaining reliable and complete measures of environmental exposure of humans is of paramount difficulty, if not impossible. It is envisaged, however, that future initiatives will be carried out to improve the collection of this type of data using large population cohorts. In the study of epistasis there is no such limit since high throughput genotyping and ultimately sequencing technologies have allowed us to capture almost all the genetic information of an individual. However, other types of limits appear that hamper the genome-wide analysis of epistasis: the methodological limit and the computational limit. In other words, do we have the appropriate tools to detect epistasis and, even if we have them, can we use them in a genome-wide scale? But before digging into these two fundamental aspects of genome-wide epistasis analysis, we will overview the evidence and theory supporting this complex genetic mechanism.

Epistasis is pervasive in model organisms

One strong evidence in favour of the existence of a particular biological mechanism in humans is the demonstration of its existence in model organisms. Re-

garding epistasis, there is an increasing evidence that it is a fundamental feature of life ranging from the simple bacteria (Maisnier– Patin *et al.*, 2005) and yeast (Segre *et al.*, 2005) to more complex organisms like birds (Carlborg *et al.*, 2003), mammals (Kim *et al.*, 2001), Drosophila (Sugiyama *et al.*, 2001) and plants (Eshed & Zamir, 1996). But, why should ever exist such a mechanism in nature? Why should evolution permit it? One strong argument in favour of the existence of epistasis comes from the idea of canalization (Moore, 2003). In the life of an organism, there are many threats to is survival either in the form of environmental inputs or in the form of inherited or de novo mutations. Canalization theory says that organisms have evolved towards a system that is resistant to these perturbing phenomena. The existence of interconnected gene networks would be the basis of this compensating mechanism. Thus, whenever one perturbing agent (environmental or genetic) affects one gene of this network, the other non–affected genes will act as compensators and effectively dissipate the negative impact to the organism. The existence of this buffering mechanism would explain, for example, why genetic variants in common diseases explain very little risk: only when multiple elements of this robust gene network are affected, does the system fail and lead to disease.

Historical interpretations of epistasis

William Bateson was the first to use the term "epistasis" (from Greek *"to stand upon"*) to describe the masking effect of one locus upon another locus (Bateson, 1909). He was trying to describe the mechanism by which the offspring of certain dihybrid crossings deviated from the expected Mendelian ratios. Some years later, Fisher used the term "epistacy" to describe those statistical models in which the joint contribution of two factors towards a phenotype deviates from the additive model. Traditionally, the first definition has been adopted by biologists, since it matches the physical conception of molecular interactions (i.e. DNA, RNA or protein) studied in the laboratory. The second definition is purely mathematical and, therefore, it has been more commonly used by statisticians when implementing Fisher's linear modelling framework. How these two different definitions (biological *vs.* mathematical) relate is still a challenge for modern genetics. In the case of the genetic architecture of complex diseases, however, until

we do not find true genetic interactions that are associated with these traits, this will stay as a mere philosophical question rather than a real scientific problem.

The computational limit: using supercomputation to exhaustively search the genome

Genomic technologies are generating vast amounts of biological data. One non–trivial issue is now the necessity to have sufficient computational power to analyze this information. In the case of the genome-wide analysis of epistasis, the problem grows exponential with the number of marker combinations that one wants to analyze (also known as "the curse of dimensionality" as Richard Bellman described it(Bellman, 1957)). In our study design if we want to analyze all possible 2–way combinations of the SNPs in the HumanHap 300 array, we should need to execute 45,000 e6 epistasis tests. If we wanted to explore the 3–SNP dimension, it would take 100,000 times more (45,000 e11 tests). If our analysis algorithm calculates each test in only 0.001 seconds (a fairly fast implementation) it would take 520 days to compute all 2–way interactions and 142,692 years to compute all 3–way interactions. At first sight either option would seam infeasible for obvious reasons. However, in the former case there is one possibility: the use of parallel computation. Supercomputers are big computational infrastructures that can harbour thousands of interconnected computer processors that can be used to perform extremely demanding computational tasks. Fortunately, at the time we were starting the GWAS study, a powerful supercomputation resource had been recently built in Barcelona: the *Mare Nostrum* supercomputer. With more than 4,000 processors it was the most powerful supercomputer in Europe and the fifth in the World. But, compared to other supercomputing centres, *Mare Nostrum*, was built to host both public and private research initiatives, including biomedical research projects like ours. We therefore submitted our project proposal to the Barcelona Supercomputing Centre (www.bsc.es) and we were granted the computational power to perform the first genome-wide epistasis analysis in RA.

The methodological limit: evaluating alternative algorithms

One crucial aspect for the genome-wide analysis of epistasis was the choice of the analysis algorithm. Following our previous candidate–gene approach (Julià *et al.*, 2007), our first option was to evaluate the use of Multifactor Dimensionality Reduction (MDR) (Ritchie *et al.*, 2001). In collaboration with Prof. J Moore and Prof. Josep Lluis Gelpí from BSC Life Science team, we implemented MDR in *Mare Nostrum's* parallel architecture. Briefly, the original method had been implemented in *Java* programming language and had to be translated into a parallel computing amenable language. One of the most commonly used programming languages for this purpose is the *C* language. However, even working with the most powerful supercomputer in Europe there were several restrictions that we had to apply to MDR's approach. Perhaps, the most influential was the reduction of the n–fold cross validation scheme to a simple one–fold validation. Cross–validation is a useful machine learning technique to evaluate the predictability of a model; however it has the important drawback that it is computationally intensive. Thus, in an exponential calculation like the present, it was not possible to fully take advantage of this method. A second problem was the negative influence of main effects. MDR is an epistasis analysis algorithm that can be very powerful "in the absence of main effects" (Pattin *et al.*, 2009). In our MDR analysis, we learnt that this method's limitation was a real burden for the interpretation of the genomewide epistasis results. In an ordinary GWAS scan there are thousands of markers that will show nominal association (P <0.05) just by chance. With the MDR analysis of our GWAS data we found that the most significant SNP–pairs were plagued with SNPs having also moderate to strong main effect. Therefore, we concluded that MDR would not help us sort out the presence of epistasis in RA.

Simple but powerful algorithm: the OR test

Given that MDR method could not be fully exploited and we could not associate SNP pairs to RA susceptibility with confidence, we looked for alternative methods. As we commented previously, the archetypal statistical analysis of interactions is the linear model framework devised by R Fisher. In particular, the logistic regression implementation is more commonly used in the analysis of SNP

to phenotype associations (Henshall & Goddard, 1999). In this method, SNP genotypes are codified as numbers (i.e. 0, 1 and 2 according to the number of minor alleles in the genotype) and the binary response variable (the case–control status) is transformed using the logistic function. The logistic model, in this case including an SNP–SNP interaction coefficient, is then fitted via the least squares or the maximum likelihood estimation methods. Although logistic regression is not suited for high–level interaction analyses due to the sparsity of the factor combinations, it can be well implemented in a 2–way level as the one we require for our study. However, the calculation of the linear regression coefficients is a computationally costly technique and would be highly impractical even with a supercomputer like *Mare Nostrum*. At that time, Harvard statistician Shaun Purcell implemented an open–source software for the analysis of GWAS data called *PLINK* (Purcell *et al.*, 2007). Within the analysis methods implemented, there was a simple algorithm for the analysis of epistasis. This method uses the inter-locus allelic association (i.e. OR) in cases and in controls to compute a z-score that is a measure of the deviance from additivity. This method has two important advantages: first, it is computationally very fast compared to the logistic regression algorithm, and second, we could ascertain that it has a statistical power very close to the logistic regression approach (Figure 5.6).

Fortunately, as well, the method was already implemented in the *C++* language which enormously facilitated its adaptation to *Mare Nostrum* supercomputer. Briefly, one of the fundamental tasks was to include the call to the *Message Passing Interface* protocol which allows the communication between the supercomputer nodes and effectively distributes the millions of epistasis tests between them. Finally, the genome-wide analysis of epistasis in RA took 48 hours using 256 CPUs in parallel.

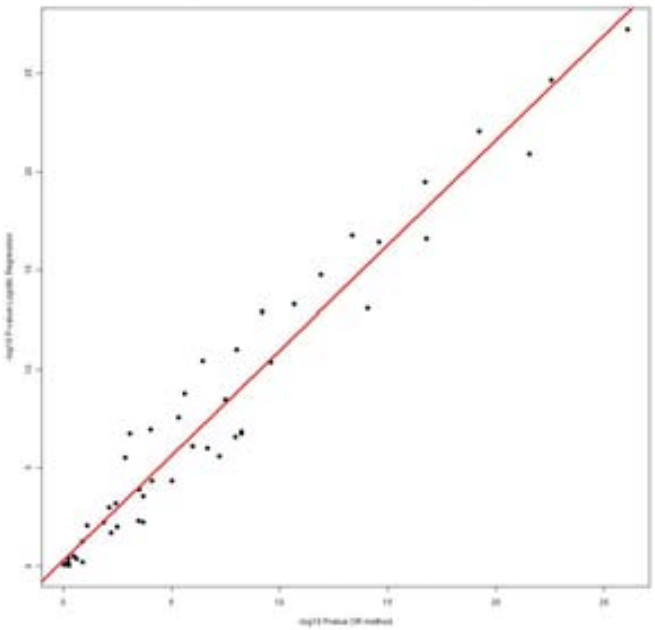


Figure 5.6: Statistical power comparison between logistic regression (Y–axis) and epistasis OR test (X–axis) evaluated of 50 different epistasis models using simulation. The correlation between both analysis methods is shown to be very high (r^2 >0.95).

ARTHRITIS & RHEUMATISM
Vol. 58, No. 8, August 2008, pp 2275–2286
DOI 10.1002/art.23623

Genome-Wide Association Study of Rheumatoid Arthritis in the Spanish Population

KLF12 as a Risk Locus for Rheumatoid Arthritis Susceptibility

Antonio Julià,[1] Javier Ballina,[2] Juan D. Cañete,[3] Alejandro Balsa,[4] Jesus Tornero-Molina,[5] Antonio Naranjo,[6] Mercedes Alperi-López,[2] Alba Erra,[2] Dora Pascual-Salcedo,[4] Pere Barceló,[7] Jordi Camps,[8] and Sara Marsal[1]

Objective. **To identify new genes associated with susceptibility to rheumatoid arthritis (RA), using a 2-stage genome-wide association study.**

Methods. **Following a liability-based study design, we analyzed 317,503 single-nucleotide polymorphisms (SNPs) in 400 patients with RA and 400 control subjects. We selected a group of candidate SNPs for replication in an independent group of 410 patients with RA and 394 control subjects. Using data from the 3 previous genome-wide association studies in RA, we also looked for genomic regions showing evidence of common association signals. Finally, we analyzed the presence of genome-wide epistasis using the binary test implemented in the PLINK program.**

Results. **We identified several genomic regions showing evidence of genome-wide association ($P < 1 \times 10^{-5}$). In the replication analysis, we identified *KLF12* SNP rs1324913 as the most strongly associated SNP ($P = 0.01$). In our study, we observed that this SNP showed higher significance than *PTPN22* SNP rs2476601, in both the genome-wide association studies and the replication analyses. Furthermore, the integration of our data with those from previous genome-wide association studies showed that *KLF12* and *PTPRT* are the unique loci that are commonly associated in 3 different studies ($P = 0.004$ and $P = 0.002$ for *KLF12* in the Wellcome Trust Case Control Consortium study and the Brigham and Women's Rheumatoid Arthritis Sequential Study genome-wide association study, respectively). The genome-wide epistasis analysis identified several SNP pairs close to significance after multiple test correction.**

Conclusion. **The present genome-wide association study identified *KLF12* as a new susceptibility gene for RA. The joint analysis of our results and those from previous genome-wide association studies showed genomic regions with a higher probability of being genuine susceptibility loci for RA.**

Rheumatoid arthritis (RA) is one of the most prevalent autoimmune diseases in the world (1). In RA, chronic inflammation of the synovial joints leads to progressive articular damage, which can result in major functional disability (2). The etiology of RA is unknown, but several family aggregation and twin studies (3,4) clearly demonstrate a heritable component of the disease. Part of this genetic component of susceptibility has been consistently associated with the HLA class II locus variation. The remaining 50–75% of the genetic component includes several other genomic regions that are more difficult to identify due to their lower penetrance or more complex models of action (5,6).

Linkage scans and candidate gene studies have

Supported by the Spanish Ministry of Education and Science (Proyectos Singulares y Estratégicos, PSE-010000-2006-6) and by Schering-Plough SA.

[1]Antonio Julià, Alba Erra, MD, Sara Marsal, MD, PhD: Institut de Recerca, Hospital Universitari Vall d'Hebron, Barcelona, Spain; [2]Javier Ballina, MD, PhD, Mercedes Alperi-López, MD: Hospital Universitario Central de Asturias, Oviedo, Asturias, Spain; [3]Juan D. Cañete, MD, PhD: Hospital Clinic i Provincial de Barcelona, Barcelona, Spain; [4]Alejandro Balsa, MD, PhD, Dora Pascual-Salcedo, PhD: Hospital Universitario La Paz, Madrid, Spain; [5]Jesus Tornero-Molina, MD, PhD: Hospital Universitario de Guadalajara, Castilla-La Mancha, Spain; [6]Antonio Naranjo, MD, PhD: Hospital Universitario de Gran Canaria Dr. Negrin, Las Palmas de Gran Canaria, Spain; [7]Pere Barceló, MD: Hospital Universitari Vall d'Hebron, Barcelona, Spain; [8]Jordi Camps: Barcelona Supercomputing Centre, Barcelona, Spain.

Dr. Tornero-Molina has received consulting fees, speaking fees, and/or honoraria from Wyeth and Bristol-Myers Squibb (more than $10,000 each).

Address correspondence and reprint requests to Sara Marsal, MD, PhD, Unitat de Recerca de Reumatologia, Hospital Universitari Vall d'Hebron, Passeig Vall d'Hebron 119-129, 08035 Barcelona, Spain. E-mail: smarsal@ir.vhebron.net.

Submitted for publication December 5, 2007; accepted in revised form April 14, 2008.

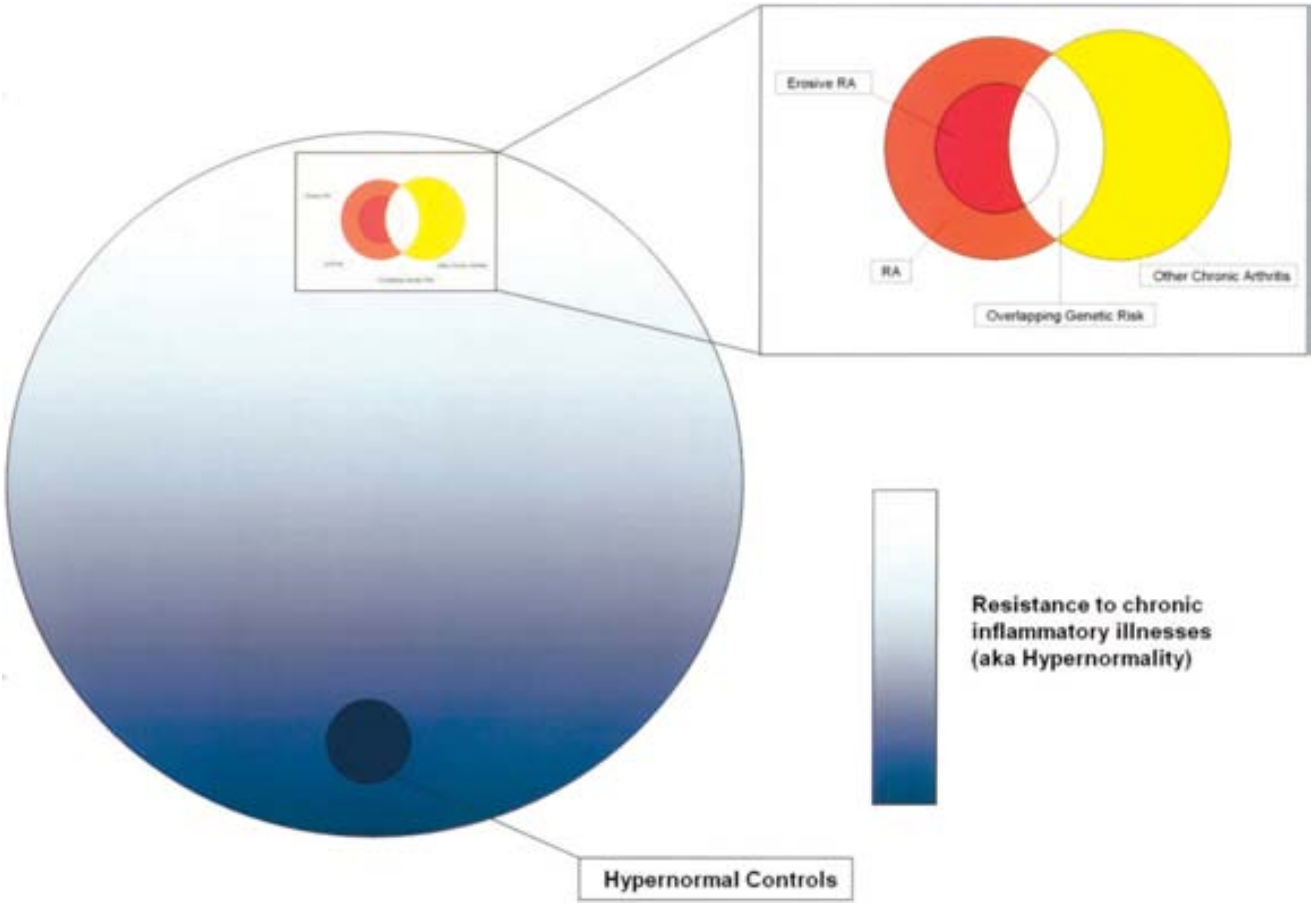


Figure 1. Liability-based model used in the present genome-wide association study. In this model, a continuous latent risk variable is truncated to specify affection, which in this case is a chronic inflammatory disease. Those individuals with the highest liability have a specific group of genes that condition to a particular outcome (i.e., disease diagnostic). Those individuals with the lowest liability are resistant to inflammatory illnesses; hence, they are "hypernormal" for these conditions. RA = rheumatoid arthritis.

successfully identified a small number of candidate genes for RA susceptibility, ranging from the robust association of *PTPN22* (7) to the more modest association of *PADI4* (8) and *CTLA4* (9). Together with these important steps in the characterization of RA genetic architecture, the candidate strategy has also produced a large number of genes that have failed to show convincing association (10). Linkage scans, an extremely powerful methodology for identifying genes with simple genetic models of inheritance, have several limitations for common diseases such as RA (11,12). Recently, genome-wide association studies have enabled the combination of 2 fundamental advantages of the previous approaches: the unbiased analysis of whole genome linkage scans and the power and resolution of case–control studies (13).

To date, 3 genome-wide association studies in RA have been performed, providing important advances in the characterization of genetic susceptibility in RA. The Wellcome Trust Case Control Consortium (WTCCC) performed an unprecedented genome-wide analysis of 7 common diseases in the UK population (14). This approach enabled not only the identification of strong candidate regions for each disease but also the identification of common susceptibility regions between different diseases. More recently, 2 genome-wide studies using North American and Swedish cohorts identified and replicated *TRAF-C5* (15) and *TNFAIP3* (16) as new genetic loci strongly associated with positive anti–cyclic citrullinated peptide antibodies in RA subtype susceptibility. These important findings demonstrate the effectiveness of the genome-wide association study approach and represent important steps toward the identification of RA genetic architecture.

Here, we report the results of a 2-stage genome-

wide association study performed in the Spanish population. In contrast to the 3 previous genome-wide studies, we used a design based on disease liability to both RA and chronic inflammatory diseases. We also performed a replication analysis of a selected group of new candidate single-nucleotide polymorphisms (SNPs) in an independent sample. In order to look for common associated genomic regions, we contrasted our results with those of the 3 previous genome-wide association studies. Finally, we also analyzed more complex genetic models through a genome-wide analysis of gene–gene interactions (i.e., epistasis) associated with RA susceptibility.

PATIENTS AND METHODS

Study design. We performed a 2-stage genome-wide association study in RA. In the first stage, 400 patients with RA and 400 control subjects were analyzed for 317,503 genomic SNPs. From these results, a selection of new candidate SNPs was further genotyped in an independent group of 410 patients with RA and 394 control subjects. In the genome-wide analysis, both case and control groups were formed by 2 subgroups (n = 200 each) based on the liability model shown in Figure 1. This model assumes that there is a continuous latent risk of chronic inflammatory diseases. Those individuals with the lowest risk of developing any type of chronic inflammatory diseases are defined as "hypernormal" (17). In the high-risk zone, the continuous variable is truncated to specify a chronic inflammatory disease. Therefore, this model integrates the increasing evidence of shared genetic risk for common inflammatory diseases (14,18,19) and the specific genetic variants that determine each particular condition.

Informed consent was obtained from all individuals, according to the Declaration of Helsinki. The study was approved by the Institut de Recerca de l'Hospital Universitari Vall d'Hebron ethics committee.

Whole-genome association study subjects. Patients with RA were recruited from 5 Spanish hospitals: Hospital Universitario Central de Asturias, Hospital Universitario de Guadalajara, Hospital Clínic i Provicial de Barcelona, Hospital Universitario de La Paz (Madrid), and Hospital Universitari Vall d'Hebron (Barcelona). All patients fulfilled the revised American College of Rheumatology (ACR; formerly, the American Rheumatism Association) 1987 criteria for the classification of RA (20). Two hundred patients were selected for having a longstanding disease with severe radiologic and functional disability (longstanding RA). The remaining cases were selected from among a group of patients with RA who were attending early arthritis clinics and had been followed up for a minimum of 2 years (early RA).

Control subjects were selected according to the liability model described previously. In order to capture the genetic component that is specific for RA and different from other chronic arthritides, we selected a group of 200 patients with non-RA inflammatory arthritis (non-RA). This group of patients and those with early RA were selected from the same early arthritis clinics, and the non-RA group comprised spondylarthritis (34%), undetermined arthritis (26%), psoriatic arthritis (20%), connective tissue disorders (15%), and other less common inflammatory arthropathies (5%). To increase the efficiency of our study, we selected a group of 200 individuals with the lowest liability for RA or any other chronic inflammatory disease, whom we here describe as hypernormal control subjects. Using the randomized control collection of IRCIS BioBank (Hospital San Joan de Reus, Tarragona, Spain), we selected only those individuals whose age placed them at risk of RA (>40 years old), were Caucasian, and had a 3-generation Spanish origin. We reduced the genetic liability in this group by excluding those individuals with ≥1 first-degree relative with a chronic inflammatory disease (including autoimmune diseases). All 4 subgroups had the female-to-male sex distribution (3:1 ratio) that is characteristic of RA (21).

Replication study subjects. We collected an independent group of 410 patients with RA (347 women and 63 men) from the same 5 hospitals. All patients fulfilled the ACR 1987 revised criteria for the classification of RA and were Caucasian and of Spanish origin. A control group of similar size (n = 394 [284 women and 110 men]) was obtained from the Spanish National DNA Bank repository (Banco Nacional de ADN, Salamanca, Spain). All control subjects were Caucasian and of Spanish origin, were older than age 30 years, and did not have an autoimmune disease.

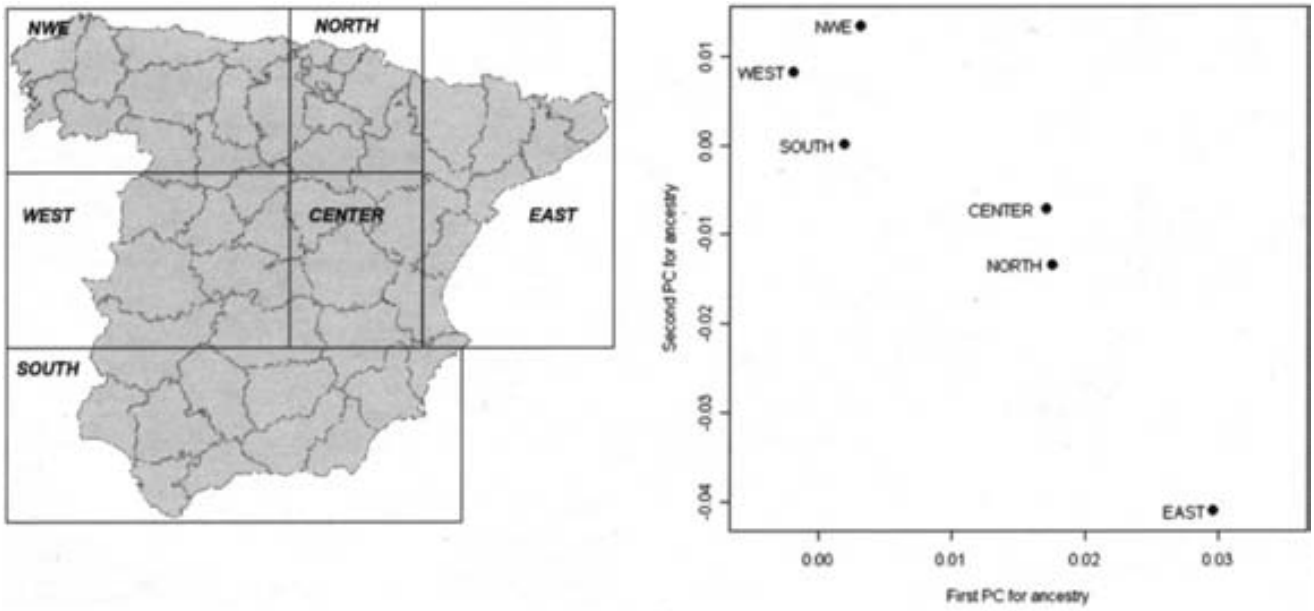
Sample preparation for whole-genome and replication genotyping. Although most samples analyzed were from local DNA collections, ~20% of them were extracted from whole blood using the Flexigene purification system (Qiagen, Chatsworth, CA).

More than 317,000 SNPs were genotyped in each of the 800 individuals in the genome-wide association study, using the HumanHap300 BeadArray system (Illumina, San Diego, CA). The selection of highly informative markers (tagSNPs) included in this system provides strong coverage of the whole genome (22). Samples were amplified, labeled, and hybridized according to the Illumina Infinium II assay. After scanning in an Illumina BeadArray reader, fluorescence intensities were automatically converted to genotypes using Illumina BeadStudio software version 2.0.

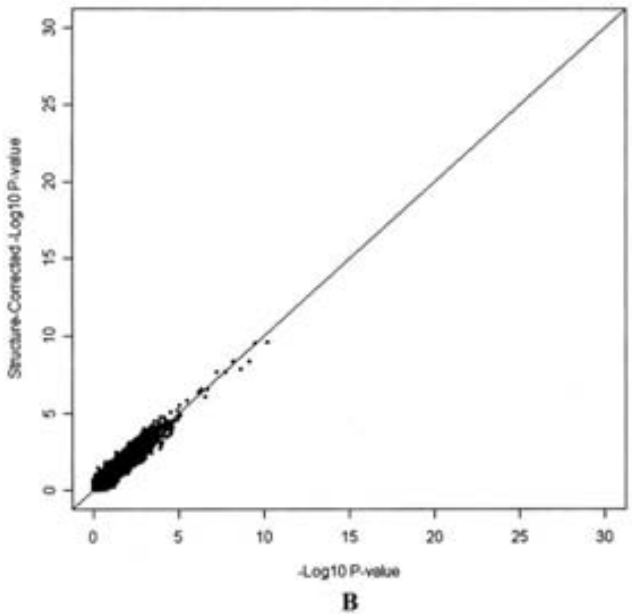
Replication genotyping and cross-platform quality control were performed using the MassARRAY SNP genotyping system (Sequenom, San Diego, CA) (23). Genotype calling was performed using the automatic call system implemented in Sequenom Typer software. All genotyping assays were performed at the Centro Nacional de Genotipado (Barcelona, Spain).

Whole-genome association study quality control. For the 800 individuals analyzed, the average genotyping rate was 98.8%. To this data we subsequently applied several quality control filters, as follows: 1) exclusion of those SNPs with more than 10% of values missing (2.2%), 2) exclusion of individuals with more than 10% of values missing (none), 3) uninformative SNPs (minor allele frequency < 0.01) (0.1%), and 4) SNPs under Hardy-Weinberg disequilibrium ($P < 0.0001$) (0.3%). Although 6,590 SNPs (2%) were in very high pairwise linkage disequilibrium ($r^2 > 0.99$), we did not exclude them from further analyses.

Population structure analysis. The presence of structure in a population can be an important confounder in genetic association studies. In order to detect strong variability components in our genome-wide association study samples, we performed the principal components analysis (PCA) implemented in Eigenstrat software (24). The PCA technique



A



B

Figure 2. A, Principal components (PCs) informative for ancestry. For 246 individuals in the genome-wide association study, complete information regarding the province of birth of all 4 grandparents was available. Based on this information, individuals were divided into 6 geographic regions in Spain. The graph shows the 2 principal components informative for ancestry, demonstrating a west-to-east trend. **B,** Scatter plot of *P* values before (x-axis) and after (y-axis) correction for structure. Correcting for geographic structure using the ancestry-informative PCs as covariates does not show a trend with signals above and below the diagonal line.

effectively decomposes the variability present in high-dimensional data sets into lower dimensions. The top axes of variation (i.e., the principal components) should reflect the geographic trends (if such trends exist) in our sample. As a more indirect measure of population structure, we also calculated the genomic inflation factor ($\lambda_{observed}$) (25), a measure of

the "overdispersion" of the association statistic (i.e., allelic chi-square). The closer this value is to the null value ($\lambda_{null} = 1$), the lower the probability of the presence of population structure in the sample.

Genome-wide association analysis, SNP selection, and replication analysis. We performed chi-square allelic tests for the 299,918 SNPs that remained after quality control filtering using PLINK software (26). Following the liability model described previously, we performed 3 different analyses: a global comparison (all patients with RA versus all control subjects), an extreme liability comparison (patients with long-standing RA versus hypernormal control subjects), and a chronic arthritis liability comparison (all patients with RA and those with non-RA inflammatory arthritis versus hypernormal control subjects). Using Benjamini and Hochberg correction for multiple testing, only HLA class II SNPs and a single marker in chromosome 3 (rs11129989) were significant. Tables showing the complete results are available online at http://www.urr.cat.

Several criteria have been proposed for selecting SNPs for replication that do not withstand conservative multiple test correction methods. Some studies have used the significance rank to select a relative arbitrary number of SNPs (27), while others use biologic information to favor a group of candidate SNPs (28). In our study, we observed an increased number of non-HLA SNPs showing strong signals ($P < 1 \times 10^{-5}$) in the extreme group comparison (7 SNPs) compared with the global comparison (1 SNP). Therefore, we decided to use the information from this comparison in 2 different selection strategies.

In one strategy, we began by genotyping a group of highly significant HLA SNPs (n = 13) (data not shown) in the replication group. All of these SNPs were positively replicated. Next, we calculated a bootstrapped *P* value for the genome-wide association study extreme comparison (n = 1,000 resamplings). From this resample-based rank, we selected all non-HLA SNPs that had higher significance values than any of the positively replicated HLA class II SNPs (7 SNPs). In the second strategy, we selected those SNPs with significance of $P < 1 \times 10^{-3}$ in the extreme group comparison (326 SNPs) and that also had significance of $P < 1 \times 10^{-3}$ when tested in the early RA versus hypernormal control subject data sets (27 SNPs). Both methods yielded a total number of 34 SNPs that were genotyped and analyzed in the replication group. In order to provide a measure of contrast of our results in the replication group, we included SNPs from known candidates for RA. This included *PTPN22* (rs2476601), *CTLA4* (rs231804), and *PADI4* (rs2240340). Like the whole-genome association study analysis, the replication association analysis was performed using the allelic chi-square test ($P < 0.05$).

Genome-wide scan for epistasis. We performed the binary test of epistasis (SNP × SNP method) implemented in PLINK. Performing the $\sim 45 \times 10^{-9}$ pairwise analyses would take several weeks in a typical workstation. In order to make it a feasible analysis, we modified the PLINK software so that it could be run in MareNostrum, a supercomputer with 10.240 64-bit Myrinet-connected processors with a final calculation capacity of 94.21 Teraflops (Barcelona Supercomputing Centre, Barcelona, Spain). Chromosome X SNPs were excluded from the analyses. Tables showing the extended results are available online at http://www.urr.cat.

Comparison with previous genome-wide association studies. Using the available data from each of the 3 previous genome-wide association studies, we looked for genomic regions that share indicative association signals with our study. For the WTCCC study, results for all 500,000 SNPs are available (14), while in the Brigham and Women's Rheumatoid Arthritis Sequential Study (BRASS) (16) and North American Rheumatoid Arthritis Consortium (NARAC) and the Swedish Epidemiological Investigation of Rheumatoid Arthritis (EIRA) (15) studies, only those SNPs with *P* values of $<1 \times 10^{-3}$ and $<1 \times 10^{-4}$, respectively, are directly accessible. In order to perform a more informative analysis, we selected the most significant SNPs in one study ($P < 1 \times 10^{-4}$) (study 1) and searched for those neighboring SNPs in the other study showing an indicative significance ($P < 0.005$) (study 2). We considered 2 SNPs from different studies to be suggestive of a common association if their genomic distance was <200 kb. The results for all analyses are available online at http://www.urr.cat.

RESULTS

Using a liability-based design, we genotyped 317,503 SNPs in 400 patients with RA and 400 control subjects. After applying several filtering criteria, 299,918 high-quality SNPs were finally selected for subsequent analyses.

Population structure. A dense set of SNPs covering the genome enables the robust identification of population outliers, using multidimensional analysis techniques (24,26). In our study, using the PCA technique, we identified and removed 41 outlier individuals (17 patients with RA and 24 control subjects). Analyzing the top principal components, we found that 2 of them captured a west-to-east trend, although they were less efficient in reflecting the north-to-south geographic variation (Figure 2A). Adding these 2 principal components as covariates in the genome-wide association analysis did not show a strong trend in the data (Figure 2B). This is in agreement with the low genomic inflation factor detected ($\lambda_{observed} = 1.01$, $\lambda_{null} = 1$). Therefore, the results reported do not correct for structure.

Genome-wide association study findings. We performed allelic association analyses to identify those loci associated with RA susceptibility and with general susceptibility to chronic inflammatory arthritis. Results for the strongest signals ($P < 1 \times 10^{-5}$) outside the HLA region are shown in Table 1. Except for rs2225966, rs2002842, and rs1328132, the other 19 SNPs are intronic or located within 100 kb from the closest gene. SNP rs11129989 in the extreme liability

Table 1. SNPs showing the strongest evidence of association in the GWAS analysis*

GWAS analysis	SNP	Chr	Gene	MA	MAF	OR	*P*
Global	rs2002842	18	*SALL3*	A	0.49	1.61	5.52×10^{-6}
Extreme liabilities	rs11129989	3	*ZNF662*	G	0.08	0.32	2.47×10^{-7}
Extreme liabilities	rs1328132	6	*OFCC1*	T	0.20	0.46	3.52×10^{-6}
Extreme liabilities	rs11086843	20	*PTPRT*	C	0.54	1.96	3.79×10^{-6}
Extreme liabilities	rs9878975	3	*AGO61*	C	0.11	0.41	6.71×10^{-6}
Extreme liabilities	rs2060396	2	*CTNNA2*	A	0.20	0.47	7.24×10^{-6}
Extreme liabilities	rs2225966	1	*LPHN2*	C	0.26	0.50	8.99×10^{-6}
Extreme liabilities	rs7968375	12	*MANSC1*	A	0.36	0.52	9.06×10^{-6}
Chronic arthritis	rs6739713	2	*R3HDM1*	G	0.36	0.56	1.47×10^{-6}
Chronic arthritis	rs946908	14	*DAAM1*	C	0.16	0.51	1.60×10^{-6}
Chronic arthritis	rs2822383	21	*C21orf81*	T	0.27	2.11	3.21×10^{-6}
Chronic arthritis	rs309137	2	*DARS*	C	0.43	0.58	3.22×10^{-6}
Chronic arthritis	rs309160	2	*DARS*	A	0.43	0.58	3.29×10^{-6}
Chronic arthritis	rs1108929	1	*LOC127540*	A	0.20	0.55	3.66×10^{-6}
Chronic arthritis	rs11129989	3	*ZNF662*	G	0.11	0.49	3.82×10^{-6}
Chronic arthritis	rs4314247	4	*KIAA0992*	G	0.45	0.58	4.44×10^{-6}
Chronic arthritis	rs10915577	1	*AJAP1*	A	0.50	1.73	5.38×10^{-6}
Chronic arthritis	rs4624474	21	*BRWD1*	T	0.48	1.74	6.28×10^{-6}
Chronic arthritis	rs309143	2	*DARS*	G	0.23	0.56	6.47×10^{-6}
Chronic arthritis	rs1324913	13	*KLF12*	A	0.28	0.58	6.53×10^{-6}
Chronic arthritis	rs6986405	8	*SGCZ*	A	0.35	1.83	8.86×10^{-6}
Chronic arthritis	rs2823580	21	*C21orf34*	C	0.25	2.04	9.51×10^{-6}
Chronic arthritis	rs2807873	1	*HLX1*	T	0.24	0.57	9.73×10^{-6}

* Single-nucleotide polymorphisms (SNPs) from the whole-genome association study (GWAS) analyses showing the strongest significance values ($P < 1 \times 10^{-5}$). Only SNPs outside the HLA region (25–35 Mb from chromosome 6) are shown. Chr = chromosome; MA = minor allele; MAF = minor allele frequency; OR = odds ratio.

analysis is the unique non-HLA SNP that was still significant after correction for multiple testing (corrected $P = 0.013$). This SNP was selected for replication in the independent sample.

As expected, several HLA class II–region SNPs showed a strong association in the global and extreme liability analyses ($P < 1 \times 10^{-9}$) but were also the strongest markers in the chronic arthritis analysis (data not shown). In particular, SNPs rs6457617 and rs9275390 were statistically significant in all 3 analyses after correction for multiple testing. Both of these SNPs are between *HLA-DQA1* and *HLA-DQA2*, 5 kb apart from each other.

Replication study findings. Using 2 different approaches, we selected a total group of 34 candidate SNPs for replication in an independent cohort of 410 patients with RA and 394 control subjects. The estimated genotyping error rate was extremely low (0.3%), indicating strong reproducibility of the results.

The results for the final 38 SNPs are shown in Table 2. Among all markers tested, only 5 SNPs showed a nominal association ($P < 0.05$). Two of them (rs10864382 and rs7006821) showed an effect opposite to that detected in the genome-wide association study analysis. The other 3 SNPs were the coding SNP from *PTPN22* (rs2476601; $P = 0.022$) and 2 intronic SNPs, one from the third intron of *SVOP* (rs7313861; $P = 0.043$) and the other from the first intron of *KLF12* (rs1324913; $P = 0.013$). All 3 SNPs showed a good correlation with the size of the genetic effect detected in the genome-wide association study analysis (for the genome-wide association studies and the replication studies, respectively, the odds ratios [ORs] were 1.47 and 1.49 for rs2476601, 1.33 and 1.23 for rs731861, and 0.73 and 0.77 for rs1324913).

Genome-wide epistasis. We performed a genome-wide analysis of all SNP × SNP combinations and their association with susceptibility to RA and chronic arthritis. Although correction for the $>45 \times 10^{-9}$ tests performed determined a very high significance threshold ($P = 1.2 \times 10^{-12}$), we observed several SNP pairs that were very close to this value (Table 3).

Genomic regions common with those in previous genome-wide association studies. The integration of our results with those from the previous genome-wide association studies identified several genomic regions showing common association signals. The closest SNPs (<50 kb) are shown in Table 4. Within this group, 2 SNPs from *CSMD2* showed the most significant common association ($P = 2.99 \times 10^{-5}$ for rs10914783 in WTCCC,

Table 2. Results for selected SNPs in the replication study*

SNP	Chr	Gene	Selection criteria	MA	MAF	*P*, global	OR, global	OR	*P*
rs10889271	1	*INADL*	Bootstrap	T	0.4	0.0016	0.72	1.00	0.977
rs10864382	1	*SLC2A5*	SC	C	0.37	0.0058	1.36	0.79	0.028
rs2807873	1	*HLX1*	Bootstrap	T	0.23	0.00028	0.65	1.17	0.162
rs524331	1	*TRIM67*	SC	T	0.42	5.50×10^{-5}	1.54	0.83	0.069
rs2240340	1	*PADI4*	Known	A	NA	NA	NA	1.09	0.411
rs2476601	1	*PTPN22*	Known	A	0.12	0.029	1.47	1.49	0.022
rs10490105	2	*FANCL*	SC	A	0.21	2.99×10^{-5}	0.61	1.22	0.084
rs2060396	2	*CTNNA2*	Bootstrap	A	0.23	0.0016	0.69	1.01	0.958
rs6739713	2	*R3HDM1*	SC	G	0.38	0.1188	0.85	1.11	0.311
rs231804	2	*CTLA4*	Known	C	0.45	0.02059	0.79	0.83	0.060
rs7609518	2	*GPC1*	Bootstrap	C	0.31	0.0051	1.39	0.95	0.625
rs11129989	3	*ZNF662*	MTS	G	0.1	7.62×10^{-5}	0.54	1.30	0.092
rs4677179	3	*RYBP*	SC	A	0.22	0.0032	1.48	1.11	0.436
rs6802500	3	*PDZRN3*	SC	T	0.4	2.18×10^{-5}	1.59	1.03	0.777
rs1022079	4	*LOC132321*	SC	A	0.29	0.00097	1.48	1.12	0.366
rs306364	4	*LOC132321*	Bootstrap	A	0.49	0.086	1.2	0.95	0.585
rs4314247	4	*KIAA0992*	SC	G	0.45	0.0023	0.73	0.87	0.154
rs289079	4	*PCDH7*	SC	T	0.49	0.00085	1.43	0.93	0.477
rs7725585	5	*DAB2*	SC	A	0.45	0.0001	1.53	1.07	0.494
rs713584	5	*SPOCK*	SC	A	0.35	0.0021	0.72	1.01	0.937
rs3130299	6	*NOTCH4*	SC	G	0.3	0.00048	0.68	0.82	0.069
rs682946	6	*COL9A1*	SC	C	0.33	4.00×10^{-5}	1.61	1.06	0.624
rs1565441	6	*FRMD1*	SC	T	0.49	8.86×10^{-5}	1.51	0.91	0.358
rs3823833	7	*ICA1*	SC	C	0.41	0.00035	0.69	0.93	0.503
rs9656200	7	*GPR85*	SC	A	0.12	0.0014	0.62	0.93	0.596
rs7793728	7	*Sep-07*	SC	G	0.22	4.05×10^{-5}	0.62	0.95	0.672
rs7006821	8	*EYA1*	SC	C	0.09	0.00018	2.29	0.68	0.030
rs1241799	11	*B3GAT1*	SC	A	0.1	0.011	1.62	0.93	0.691
rs1468796	12	*TMPO*	SC	T	0.42	3.68×10^{-5}	1.56	0.84	0.091
rs7313861	12	*SVOP*	SC	T	0.42	0.0079	1.33	1.23	0.043
rs1324913	13	*KLF12*	SC	A	0.28	0.0047	0.73	0.77	0.013
rs1886925	13	*SLC10A2*	SC	A	0.42	0.0022	0.73	1.03	0.768
rs769426	17	*OR1A1*	SC	G	0.3	0.597	0.94	1.08	0.489
rs2002842	18	*SALL3*	SC	A	0.49	5.52×10^{-6}	1.61	0.90	0.287
rs1329820	20	*C20orf23*	Bootstrap	A	0.23	0.0012	1.52	1.04	0.735
rs6030267	20	*PTPRT*	Bootstrap	A	0.21	0.0021	1.52	0.94	0.646
rs2823580	21	*C21orf34*	SC	C	0.24	0.028	1.32	0.92	0.490
rs2836982	21	*BRWD1*	SC	C	0.5	0.0022	1.37	1.12	0.272

* Among all 38 single-nucleotide polymorphisms (SNPs) tested, 5 showed nominal significance. The 2 most associated SNPs are *KLF12* rs1324913 followed by *PTPN22* rs2476601. Two SNPs (*SLC2A5* rs10864382 and *EYA1* rs7006821) show an effect opposite to the estimated effect in the genome-wide association study analysis. Chr = chromosome; MA = minor allele; MAF = minor allele frequency; OR = odds ratio; SC = subgroup comparison; NA = not applicable; MTS = multiple testing significance.

Table 3. Top pairwise SNP × SNP interactions identified in all 3 genome-wide comparisons*

GWAS	SNP 1	Chr	Gene	SNP 2	Chr	Gene	*P*
Global	rs9752494	2	*PPM1B*	rs1569020	12	*GPR133*	1.22×10^{-12}
Global	rs10465885	1	*GJA5*	rs2302502	18	*PTPRM*	3.62×10^{-11}
Global	rs950675	2	*TPO*	rs1569020	12	*GPR133*	4.84×10^{-11}
Global	rs12755965	1	*GJA5*	rs6776932	3	*ACPP*	5.41×10^{-11}
Extreme liability	rs259401	6	*RAB32*	rs2322140	17	*DNAH9*	7.73×10^{-11}
Extreme liability	rs2244817	8	*SULF1*	rs3826296	17	*AKAP1*	8.56×10^{-11}
Extreme liability	rs2244817	8	*SULF1*	rs998113	17	*AKAP1*	9.52×10^{-11}
Chronic arthritis	rs10171653	2	*RTN4*	rs7033413	9	*GLIS3*	5.69×10^{-12}
Chronic arthritis	rs2580768	2	*RTN4*	rs7033413	9	*GLIS3*	2.63×10^{-11}
Chronic arthritis	rs4849025	2	*CNTNAP5*	rs2392829	8	*PXDNL*	9.07×10^{-11}

* The binary test implemented in PLINK was used to identify several single-nucleotide polymorphism (SNP) pairs close to the threshold for correction for multiple testing ($P = 1.11 \times 10^{-12}$). Interaction association can be detected by neighboring SNPs, as can be seen for the *SULF1–AKAP1* interaction in the extreme liability analysis and the *RTN4–GLIS3* interaction in the chronic arthritis analysis. GWAS = genome-wide association study; Chr = chromosome.

Table 4. Genomic loci showing common signals between the present genome-wide association study and the 3 previous genome-wide association studies*

Region	Gene in region	Study 1	Top SNP	*P*	Study 2	Top SNP	*P*
13q22	*KLF12*	URR	rs1887346	6.03×10^{-5}	BRASS	rs9318225	2.00×10^{-3}
		URR	rs9318228	3.66×10^{-5}			
1p35.1–p34.3	*CSMD2*	URR	rs1108929	2.56×10^{-5}	WTCCC	rs10914783	2.99×10^{-5}
5p15	*TAS2R1*	URR	rs13159275	8.92×10^{-5}	WTCCC	rs10513046	0.0021
11p15.1	*NAV2*	URR	rs10833197	4.86×10^{-5}	WTCCC	rs2568127	0.00051
4p14-p12	*ATP8A1*	BRASS	rs10517039	2.00×10^{-6}	URR	rs4370169	0.0021
					URR	rs6447164	0.0044
					URR	rs10517035	0.0009
					URR	rs10517038	0.0048
					URR	rs3811768	0.00042
10q22–q23	*NRG3*	BRASS	rs10509440	6.00×10^{-5}	URR	rs12358407	0.0043
10q23.1	*KIAA1128*	BRASS	rs10491033	1.00×10^{-7}	URR	rs1572430	0.00057
12q24.1	*TBX5*	BRASS	rs10507251	4.00×10^{-5}	URR	rs11830449	0.00029
1p35.1–p34.3	*CSMD2*	WTCCC	rs10914783	2.99×10^{-5}	URR	rs1108929	2.56×10^{-5}
					URR	rs10799004	0.00047
					URR	rs10799006	0.0022
1p31.1	*IFI44*	WTCCC	rs11162922	1.80×10^{-6}	URR	rs7416587	0.0038
					URR	rs4384179	0.0035
5q14.1	*CMYA5*	WTCCC	rs7343	8.28×10^{-5}	URR	rs1129770	0.0049
6q23	*EYA4*	WTCCC	rs2677821	2.48×10^{-13}	URR	rs2327358	0.00073
8q13.3	*EYA1*	WTCCC	rs4133002	6.17×10^{-5}	URR	rs13274769	0.0028
8q23	*OXR1*	WTCCC	rs16874205	9.43×10^{-10}	URR	rs13364828	0.0025
10p12	*PTER*	WTCCC	rs12269329	5.80×10^{-6}	URR	rs11253931	0.0012
15q21.3	*WDR72*	WTCCC	rs1711029	3.61×10^{-12}	URR	rs1021744	0.0047
18q23	*SALL3*	WTCCC	rs2941794	1.27×10^{-10}	URR	rs2002842	0.00092
					URR	rs2941811	0.0047
22q13.1	*C1QTNF6*	WTCCC	rs743777	7.92×10^{-6}	URR	rs229527	0.0038

* The top single-nucleotide polymorphisms (SNPs) from the first study (study 1; $P < 1 \times 10^{-3}$) were examined in the second study (study 2; $P < 0.005$) for signals within a 50-kb flanking region. No matches with the North American Rheumatoid Arthritis Consortium and the Swedish Epidemiological Investigation of Rheumatoid Arthritis studies were found at this genomic distance. URR = Unitat de Recerca de Rheumatologia (present study); BRASS = Brigham and Women's Rheumatoid Arthritis Sequential Study; WTCCC = Wellcome Trust Case Control Consortium.

and $P = 2.56 \times 10^{-5}$ for rs1108929 in our study). This association is the strongest detected, even after extending the analysis to a distance of 200 kb.

We found 2 genomic regions to be common in both the WTCCC study and BRASS. *KLF12* SNPs rs1887346 and rs9318228 in our study were associated with BRASS SNP rs9318225 ($P = 0.002$) (Table 4) and with WTCCC SNP rs1887346 ($P = 0.0049$). *PTPRT* intronic SNPs rs6030267 ($P = 4.08 \times 10^{-5}$) and rs11086843 ($P = 2.07 \times 10^{-6}$) were commonly associated with BRASS SNP rs10485690 ($P = 5 \times 10^{-4}$) and WTCCC SNP rs2223542 ($P = 0.0015$).

DISCUSSION

We performed a 2-stage genome-wide association study for RA in the Spanish population, using 400 patients with RA and 400 control subjects. From these results, we selected a group of candidate SNPs and performed a replication study in an independent group of 410 patients with RA and 394 control subjects. In our study, we found *KLF12* to have stronger significance than previously associated non-HLA SNPs. We also integrated our association results with those of the 3 previous genome-wide association studies in RA. *KLF12* and *PTPRT* are the 2 unique genes that are in common regions in our study and both the WTCCC study and BRASS. Finally, we performed a genome-wide analysis for epistasis and found several SNP pairs with statistical values close to significance even after correction for multiple testing.

In the present study, we followed a liability model that could underlie susceptibility to chronic inflammatory diseases and, thus, susceptibility to RA. Results of several recent studies support this model. *NALP1* has been recently associated with vitiligo and several other autoimmune diseases, including RA (18). Fc receptor–

like protein (29) and STAT-4 (19) have been associated with susceptibility to both RA and systemic lupus erythematosus. *PTPN22* itself was studied and associated with RA after demonstrating its association with susceptibility to type 1 diabetes mellitus (7). The WTCCC genome-wide scan also provides several genomic regions linking chronic inflammatory diseases such as RA, type 1 diabetes mellitus, and Crohn's disease (14). Following this model, we selected individuals in whom the risk of developing chronic inflammatory disease was lowest (hypernormal controls) and also individuals in whom a different chronic inflammatory arthritis (non-RA) was diagnosed. In order to increase the contrast, we also included individuals in the RA group who had a highly erosive phenotype. This strategy always adds substantial power to the traditional case–control design (17), although the difficulties associated with obtaining such individuals generally limit its extended use.

Several lines of evidence support the association of *KLF12* with RA susceptibility. First, in our population, SNP rs1324913 showed a stronger association with RA compared with *PTPN22* SNP rs2476601. This was observed in the genome-wide and the replication analyses. Although this does not imply a stronger genetic effect (for *PTPN22* and *KLF12*, the estimated ORs were 1.49 and 1.3, respectively), it provides further evidence of association. The allelic association analyses performed in the present study assume a multiplicative genetic model (30) (i.e., the risk of developing the disease multiplied by a factor for each susceptibility allele carried). However, exploration of alternative genetic models (i.e., dominant, recessive, and genotypic) can give additional information on relevant genetic associations. In our replication analysis, the dominant model of rs1324913 had a significance of $P = 0.005$; no other replicated SNP showed such an increase in significance (data not shown). In order to check the consistency of this observation, we performed the same model analysis in our genome-wide association study data. We observed that the dominant model also had the strongest association in the extreme liability analysis ($P_{Dominant} = 1 \times 10^{-5}$ versus $P_{Multiplicative} = 1 \times 10^{-4}$) and the global analysis ($P_{Dominant} = 6 \times 10^{-4}$ versus $P_{Multiplicative} = 5 \times 10^{-3}$).

Other important evidence supporting *KLF12* association is that, together with *PTPRT*, they are the only 2 genomic regions commonly found when comparing our study with the BRASS and WTCCC studies. In the extreme liability analysis, 2 SNPs in the *KLF12* transcribed region had a significance of $P < 1 \times 10^{-4}$ (rs1887346 and rs9318228 with $P = 6.02 \times 10^{-5}$ and $P = 3.66 \times 10^{-5}$, respectively). Both SNPs are only 8.2 kb apart from each other, and are only 1.7 kb and 9.9 kb, respectively, from the SNP rs9318225 in BRASS ($P = 2 \times 10^{-3}$) (Table 1). When comparing our study with the WTCCC, we found rs1184596 (same intron, 170 kb upstream) to have an indicative association ($P = 4.9 \times 10^{-3}$). Although it is more distant to SNPs rs1887346 and rs9318228, it is closer to the replicated *KLF12* SNP rs1324913 (38 kb). This finding supports the replicability of this association in different populations.

KLF12 (activator protein 2α [AP-2α] repressor) is a member of the family of Kruppel-like transcriptional regulatory factors (31), which play fundamental roles in differentiation and development. *KLF12* is known to repress the transcription of *AP-2*α transcription factor after binding to the general correpressor protein C-terminal binding protein 1 (32). The expression patterns of AP-2α–regulated genes, including the gene for tumor necrosis factor α (TNFα), have been implicated in malignant transformation and stress responses (33,34). Thus, genetic variations could increase susceptibility to RA through various mechanisms: either by facilitating the transformation of local connective cells (32) or by promoting lymphocyte survival (35). More intriguingly, the recent characterization of AP-2–mediated TNFα gene expression in B19 parvoviral infection could add an alternative mechanism. This type of infection can produce a chronic inflammatory arthritis that can fulfill the diagnostic criteria for RA. For many years, B19 has been studied as a possible trigger for RA, although with controversial results (36,37). The observed genetic association of *KLF12* with RA and its direct implication in TNFα regulation could represent a new perspective on genetic and environmental interactions associated with RA susceptibility.

The integration of our results with those of previous genome-wide association studies identified several relevant genomic regions. *PTPRT* and *KLF12* are the only loci associated in both the WTCCC study and BRASS. Protein tyrosine phosphatase (PTP) receptor T, the most frequently mutated PTP in human cancers, has been recently characterized as a key inhibitor of STAT-3 (38). STAT-3, in turn, mediates transcriptional activation in response to several cytokines, including RA-associated interleukin-6. The common signal at *CSMD2* in our study and the WTCCC study is the strongest association detected. *CSMD2* is a recently cloned gene (39) whose functionality still needs to be identified. Recently, *CSMD1*, a gene with high structural similarity to *CSMD2*, has been implicated in the inhibition of complement activation (40), which could suggest a com-

mon functionality. In addition, the results from our analysis along with previous genome-wide association studies also identified several SNPs from the newly associated *TNFAIP3* locus (rs643571 at $P = 6 \times 10^{-3}$ and rs6915853 at $P = 2 \times 10^{-3}$ in our study). This finding confirms the power of our analysis method to identify genomic regions that are relevant to RA susceptibility.

Although the same technologic platform was used, comparison of our study with the NARAC and EIRA studies did not reveal evidence of association for any of the *TRAF1-C5*–region SNPs. This lack of association, which was also observed in the WTCCC study (15), could be largely attributable to the restricted analysis of anti-CCP antibody–positive RA patients in the NARAC and EIRA cohorts. Therefore, an exact comparison would have required specifically addressing the association in anti-CCP antibody–positive patients only. However, a more detailed analysis of the association in this region showed additional results. We found several indicative signals (4 SNPs at a significance level of $P < 0.01$) in the 5′– and 3′–untranslated regions of *FBXW2*. This gene is 109 kb from *TRAF1* and encodes for a protein that participates in the ubiquitin/proteasome degradation system (41). Therefore, our analysis suggests that, in this genomic region, *FBXW2* seems to be associated with RA susceptibility.

There is increasing evidence that gene–gene interactions (epistasis) could be of major relevance in susceptibility to complex diseases (42,43). A recent study demonstrated that complete genome-wide analysis has more power to detect relevant SNP pairs than methods that use filtering strategies, even after correction for multiple testing (43). In the present study, we performed an exhaustive analysis of all autosomal SNP pairs and found several of them with significance close to the threshold of significance for multiple testing. An important observation is that no known main-effect SNP was observed in this group of top SNPs. In fact, HLA class II SNPs appear only at the level of $P = 1 \times 10^{-8}$ to $P = 1 \times 10^{-7}$. This could probably indicate that, although the possibility of epistasis with this region cannot be discarded, other regions with marginal main effects seem to show stronger interactions associated with disease risk. This also confirms the need to perform exhaustive analyses in the search for epistasis. To our knowledge, none of the top SNP pairs ($P < 1 \times 10^{-10}$) belong to genes from a known common biologic pathway. Protein phosphatase 1B, a regulator of NF-κB transcription factor, has a strong interaction in the global analysis with protein G–coupled receptor 133, which, to date, has no associated biologic function. In the extreme liability analysis, the human sulfatase 1 gene (*SULF1*), a heparan sulfatase involved in tumor progression and inflammation (44), is also interacting with 2 SNPs from the A kinase anchor protein 1 gene (*AKAP1*). The latter has been associated with cAMP-mediated signal transduction and messenger RNA trafficking (45).

The present study is one of the first genome-wide association analyses in RA. Using a liability-based design, we found several new candidate SNPs for RA and chronic inflammatory arthritis. We performed a replication analysis in an independent subset of SNPs, from which *KLF12* emerged as a new candidate susceptibility gene for RA. A comparison of our results with those from the 3 previous genome-wide association studies confirmed the relevance of the *KLF12* locus and also identified several other regions of interest for subsequent studies. In order to search for more complex genetic models involved in RA susceptibility, we performed a genome-wide analysis for epistasis. The results presented here add important aspects to the continuing definition of RA genetic architecture. Consistent replication of these results in different populations will confirm the association of the genomic regions to RA susceptibility.

ACKNOWLEDGMENTS

We thank Professor David Clayton for his critical review of the manuscript. We are indebted to the Barcelona Supercomputing Centre support team. We also are grateful to Genoma España for their support to this project, and especially to Professor J. Bertranpetit CeGen (National Genotyping Center), Professor J. Benitez (CeGen-Nodo de Madrid), Professor A. Carracedo (CeGen-Nodo de Santiago de Compostela), Dr. A. Garcia, and Professor A. Orfao (Banco Nacional de ADN).

AUTHOR CONTRIBUTIONS

Dr. Marsal had full access to all of the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

Study design. Julià, Ballina, Tornero-Molina, Alperi-López, Erra, Barceló, Marsal.

Acquisition of data. Julià, Ballina, Cañete, Balsa, Tornero-Molina, Naranjo, Alperi-López, Erra, Pascual-Salcedo, Marsal.

Analysis and interpretation of data. Julià, Ballina, Cañete, Tornero-Molina, Alperi-López, Camps, Marsal.

Manuscript preparation. Julià, Marsal.

Statistical analysis. Julià, Marsal.

REFERENCES

1. Marrack P, Kappler J, Kotzin BL. Autoimmune disease: why and where it occurs [review]. Nat Med 2001;7:899–905.

2. Firestein GS. Evolving concepts of rheumatoid arthritis [review]. Nature 2003;423:356–61.
3. MacGregor AJ, Snieder H, Rigby AS, Koskenvuo M, Kaprio J, Aho K, et al. Characterizing the quantitative genetic contribution to rheumatoid arthritis using data from twins. Arthritis Rheum 2000;43:30–7.
4. Cornelis F, Faure S, Martinez M, Prud'homme JF, Fritz P, Dib C, et al, for the European Consortium on Rheumatoid Arthritis Families. New susceptibility locus for rheumatoid arthritis suggested by a genome-wide linkage study. Proc Natl Acad Sci U S A 1998;95:10746–50.
5. Jawaheer D, Li W, Graham RR, Chen W, Damle A, Xiao X, et al. Dissecting the genetic complexity of the association between human leukocyte antigens and rheumatoid arthritis. Am J Hum Genet 2002;71:585–94.
6. Chapman J, Clayton D. Detecting association using epistatic information. Genet Epidemiol 2007;31:894–909.
7. Begovich AB, Carlton VE, Honigberg LA, Schrodi SJ, Chokkalingam AP, Alexander HC, et al. A missense single-nucleotide polymorphism in a gene encoding a protein tyrosine phosphatase (PTPN22) is associated with rheumatoid arthritis. Am J Hum Genet 2004;75:330–7.
8. Suzuki A, Yamada R, Chang X, Tokuhiro S, Sawada T, Suzuki M, et al. Functional haplotypes of PADI4, encoding citrullinating enzyme peptidylarginine deiminase 4, are associated with rheumatoid arthritis. Nat Genet 2003;34:395–402.
9. Rodriguez MR, Nunez-Roldan A, Aguilar F, Valenzuela A, Garcia A, Gonzalez-Escribano MF. Association of the CTLA4 3′ untranslated region polymorphism with the susceptibility to rheumatoid arthritis. Hum Immunol 2002;63:76–81.
10. Hirschhorn JN, Lohmueller K, Byrne E, Hirschhorn K. A comprehensive review of genetic association studies [review]. Genet Med 2002;4:45–61.
11. Risch N, Merikangas K. The future of genetic studies of complex human diseases. Science 1996;273:1516–7.
12. Reich DE, Lander ES. On the allelic spectrum of human disease. Trends Genet 2001;17:502–10.
13. Hirschhorn JN, Daly MJ. Genome-wide association studies for common diseases and complex traits [review]. Nat Rev Genet 2005;6:95–108.
14. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 2007;447:661–78.
15. Plenge RM, Seielstad M, Padyukov L, Lee AT, Remmers EF, Ding B, et al. TRAF1-C5 as a risk locus for rheumatoid arthritis: a genomewide study. N Engl J Med 2007;357:1199–209.
16. Plenge RM, Cotsapas C, Davies L, Price AL, de Bakker PI, Maller J, et al. Two independent alleles at 6q23 associated with risk of rheumatoid arthritis. Nat Genet 2007;39:1477–82.
17. Morton NE, Collins A. Tests and estimates of allelic association in complex inheritance. Proc Natl Acad Sci U S A 1998;95:11389–93.
18. Jin Y, Mailloux CM, Gowan K, Riccardi SL, LaBerge G, Bennett DC, et al. NALP1 in vitiligo-associated multiple autoimmune disease. N Engl J Med 2007;356:1216–25.
19. Remmers EF, Plenge RM, Lee AT, Graham RR, Hom G, Behrens TW, et al. STAT4 and the risk of rheumatoid arthritis and systemic lupus erythematosus. N Engl J Med 2007;357:977–86.
20. Arnett FC, Edworthy SM, Bloch DA, McShane DJ, Fries JF, Cooper NS, et al. The American Rheumatism Association 1987 revised criteria for the classification of rheumatoid arthritis. Arthritis Rheum 1988;31:315–24.
21. Doran MF, Pond GR, Crowson CS, O'Fallon WM, Gabriel SE. Trends in incidence and mortality in rheumatoid arthritis in Rochester, Minnesota, over a forty-year period. Arthritis Rheum 2002;46:625–31.
22. Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, et al, for the International HapMap Consortium. A second generation human haplotype map of over 3.1 million SNPs. Nature 2007;449:851–61.
23. Buetow KH, Edmonson M, MacDonald R, Clifford R, Yip P, Kelley J, et al. High-throughput development and characterization of a genomewide collection of gene-based single nucleotide polymorphism markers by chip-based matrix-assisted laser desorption/ionization time-of-flight mass spectrometry. Proc Natl Acad Sci U S A 2001;98:581–4.
24. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 2006;38:904–9.
25. Devlin B, Roeder K, Wasserman L. Genomic control, a new approach to genetic-based association studies [review]. Theor Popul Biol 2001;60:155–66.
26. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 2007;81:559–75.
27. Winkelmann J, Schormair B, Lichtner P, Ripke S, Xiong L, Jalilzadeh S, et al. Genome-wide association study of restless legs syndrome identifies common variants in three genomic regions. Nat Genet 2007;39:1000–6.
28. Wang K, Li M, Bucan M. Pathway-based approaches for analysis of genomewide association studies. Am J Hum Genet 2007. E-pub ahead of print.
29. Kochi Y, Yamada R, Suzuki A, Harley JB, Shirasawa S, Sawada T, et al. A functional variant in FCRL3, encoding Fc receptor-like 3, is associated with rheumatoid arthritis and several autoimmunities. Nat Genet 2005;37:478–85.
30. Balding DJ. Handbook of statistical genetics. West Sussex (UK): John Wiley & Sons; 2001.
31. Bieker JJ. Kruppel-like factors: three fingers in many pies. J Biol Chem 2001;276:34355–8.
32. Schuierer M, Hilger-Eversheim K, Dobner T, Bosserhoff AK, Moser M, Turner J, et al. Induction of AP-2α expression by adenoviral infection involves inactivation of the AP-2rep transcriptional corepressor CtBP1. J Biol Chem 2001;276:27944–9.
33. Kannan P, Buettner R, Chiao PJ, Yim SO, Sarkiss M, Tainsky MA. N-ras oncogene causes AP-2 transcriptional self-interference, which leads to transformation. Genes Dev 1994;8:1258–69.
34. Wajapeyee N, Somasundaram K. Cell cycle arrest and apoptosis induction by activator protein 2α (AP-2α) and the role of p53 and p21WAF1/CIP1 in AP-2α-mediated growth inhibition. J Biol Chem 2003;278:52093–101.
35. Zhou M, McPherson L, Feng D, Song A, Dong C, Lyu SC, et al. Kruppel-like transcription factor 13 regulates T lymphocyte survival in vivo. J Immunol 2007;178:5496–504.
36. Simpson RW, McGinty L, Simon L, Smith CA, Godzeski CW, Boyd RJ. Association of parvoviruses with rheumatoid arthritis of humans. Science 1984;223:1425–8.
37. Takahashi Y, Murai C, Shibata S, Munakata Y, Ishii T, Ishii K, et al. Human parvovirus B19 as a causative agent for rheumatoid arthritis. Proc Natl Acad Sci U S A 1998;95:8227–32.
38. Zhang X, Guo A, Yu J, Possemato A, Chen Y, Zheng W, et al. Identification of STAT3 as a substrate of receptor protein tyrosine phosphatase T. Proc Natl Acad Sci U S A 2007;104:4060–4.
39. Lau WL, Scholnick SB. Identification of two new members of the CSMD gene family small star, filled. Genomics 2003;82:412–5.
40. Kraus DM, Elliott GS, Chute H, Horan T, Pfenninger KH, Sanford SD, et al. CSMD1 is a novel multiple domain complement-regulatory protein highly expressed in the central nervous system and epithelial tissues. J Immunol 2006;176:4419–30.
41. Yang CS, Yu C, Chuang HC, Chang CW, Chang GD, Yao TP, et al. FBW2 targets GCMa to the ubiquitin-proteasome degradation system. J Biol Chem 2005;280:10083–90.

Chapter 6

Discussion

We hope that our work has demonstrated the contributions of two genomic approaches to the characterization of the genetic risk of RA. Using a simple but powerful *in vitro* model of synovial fibroblast activation and a reverse engineering approach, we have identified the basic transcriptional regulatory network associated with the response to the complex proinflammatory environment in RA. We have used this regulatory network as a powerful candidate gene set for RA susceptibility, from which we have identified a significant two–locus epistatic effect. Using a liability–based epidemiological design, we have performed the first genome-wide association scan for RA in the Spanish population. We have provided strong evidence of a new candidate gene for RA, *KLF12*, and we have identified a group of signals which are likely to yield additional susceptibility loci. In this study, we have also performed the first genome-wide exploration for epistasis performed in a complex disease which provides new insights into RA genetic architecture.

In Chapter 1, we have listed the main contributions of the present thesis work, so we will conclude by elaborating a subjective view of the results. We will now discuss potential implications of these results for the study of RA genetic basis. Finally, we will briefly comment on the current and future work based on these results.

There is now little doubt about the usefulness of genomic approaches for the identification of genetic variants associated with complex traits. Gene expression microarrays were the first of such high throughput technologies to be

massively utilized in biomedical research. Actually, almost all common diseases have been explored using microarrays and, with the gradual reduction of manufacturing costs, the number of biomedical features explored using this technology will most likely be exponential. With this increasing amount of biological information stored in public databases it will be able to discover fundamental features of human pathologies. One such approach has been recently conducted (Dudley *et al.*, 2009) with over 8,340 microarrays including data on more than 230 diseases and performed in 122 different tissues. In this meta-analysis they did not only find a high degree of correlation between studies analyzing the same disease but, importantly, they found that this correlation was higher between different tissues of the same disease compared to the same tissue examined under different conditions. Thus, in the next years, systematic approaches like reverse engineering that exploit this connectivity will be essential to discover the fundamental features of common diseases.

The regulation of the gene expression within a cell is a highly complex mechanism. Information is gathered from various sources, processed in multiple intricate ways and finally used, rejected or stored for later use. Transcription factors are the key mediators of this information flow; however, the knowledge of their activity in multiple biological conditions is only starting to be characterized with detail. *Feedforward loops*, *single–input modules* or *dense overlapping regulons* are only some of the complex potential mechanisms by which transcription factors can regulate gene expression (Alon, 2007). However, even before trying to characterize these functional features in human diseases it will be essential to find the appropriate experimental model from which we can generalize. In this sense, we advocate the use of SF stimulation with synovial fluid as a simple but powerful model to find relevant mechanisms in RA pathophysiology. To the best of our knowledge, this is the first study that has performed a genomic analysis using this approach.

How can we say that our SF stimulation model is useful for RA molecular characterization? We have previously demonstrated the robustness of the cytokine profile in RA synovial fluid. We could, nonetheless, criticize that using a unique synovial fibroblast line should have insufficient generalizability on a disease condition. However, the fact that many of the differentially expressed genes

have been associated to RA pathophysiology gives strong support for this simple model. Of interest is *IL23A* gene, which appears to be the most overexpressed gene (∼8–fold) in the synovial fluid stimulated SF and has also been recently found to be highly expressed in RA synovial samples (Brentano *et al.*, 2009) and not in control (i.e. osteoarthritis) samples. The IL23 pathway itself, has been strongly associated through GWAS association to the genetic susceptibility to other chronic inflammatory diseases like Inflammatory Bowel Disease (Duerr *et al.*, 2006), Psoriasis (Nair *et al.*, 2009) and Psoriatic Arthritis (Liu *et al.*, 2008). Although there is no evidence for genetic variation in the *IL23* gene in the genetic susceptibility of RA, the *in vitro* and *in vivo* evidence clearly point out for a role of this gene or its pathway in RA pathophysiology. Of interest, other non–previously associated genes are also associated with RA SF response to synovial fluid; these genes clearly deserve future study to understand their implication in the disease.

It is becoming increasingly clear that the immunological response is a coordinated task that is not only circumscribed to blood–borne cells but also to other cell types. In the case of RA it is clear that SFs are not innocent bystanders but are also active mediators of the immune response. We have shown that SFs express multiple cytokine and chemokine signals under the RA proinflammatory environment, and that the main driver of this activity is NF–kB transcription factor. This proactive role in RA pathology of SF has been very recently strengthened by the discovery that SFs are the responsible for the spreading of RA from affected joints to unaffected joints (Lefevre *et al.*, 2009). This finding has been experimentally identified in human RA SFs and cartilage xenografts in the SCID mice. If this transmigration mechanism, reminiscent of cancer metastasis, is finally confirmed in RA patients, it will be more than ever necessary to fully characterize the intricate mechanisms of SF activity. The inclusion of other systematic approaches like high throughput proteomic and metabolomic technologies will be also essential for this objective.

Comprehensive studies on model organisms are identifying epistasis as a very pervasive mechanism. In Drosophila, for example, it is clearly becoming a fundamental mechanism for complex trait variation (Yamamoto *et al.*, 2008). Basic

mechanisms like Hedgehog or Notch signalling and many others were first discovered in flies and they are now known to exist in humans. Thus, it would be rather unexpected not to find any trace of evolutionary conservation regarding epistatic mechanisms. So, what could be the explanation for the lack of strong and reproducible epistatic interactions associated with disease risk in humans? Although epistasis has been known since the beginning of the twentieth century, only recently there has been the ideal scenario to study it. Genotyping has become increasingly cheap and fast and, above all, the search for main effects in common diseases seems to have finally arrived to its limits, at least, under the common–disease common variant assumptions. In the present study we have proposed two alternative strategies to identify epistatic interactions: a "pathway–based" approach that exploits the identified connectivity between genes, and an "agnostic" approach were all possible combinations are evaluated. Clearly, both strategies have advantages and disadvantages but, *a priori*, it is impossible to predict which one will be more successful. It is our view that only after comprehensive evaluation of all possible approaches and the development of more powerful analytical methods we will be able to identify the epistatic component of complex traits like RA.

In only three years, GWAS studies have provided more than 400 loci robustly associated to common complex diseases. In the case of RA architecture, this has meant passing from an "adobe–like" structure of 2 loci to a sophisticated "building" of more than 20 susceptibility loci. These findings are having a tremendous impact on how we now look at this heterogeneous disease. For example, the everlasting association of the HLA locus with RA seems now to be confined exclusively to those patients having positive autoantibody status (Raychaudhuri *et al.*, 2009b). Nonetheless, anti–CCP and RF negative patients do not show a disease phenotype that is distinguishable from their positive counterparts. True differential genetic origin? Phenocopies? One of the most striking lessons from the study of common chronic diseases using the GWAS approach has been the confirmation that many chronic inflammatory diseases do share a common genetic background. At first glance, this could seem to contradict the observation of genetic differences within RA; however, it is our view that both aspects need not be exclusive and that it will rather depend on the level of characterization

of the phenotype. In our GWAS study we specifically selected a subgroup of RA patients showing advanced erosions in hands irrespective of their autoantibody status. Thus, it could be possible that *KLF12* variation is associated with this specific aspect of RA heterogeneity; if this parameter is not accounted for in future association studies, the replication could potentially be missed.

The characterization of the genetic risk basis for many common complex diseases is actually in a crossroad. How can one explain that after identifying dozens of loci there is a substantial part of heritability that remains still unaccounted for? One possibility is that we have been targeting the wrong genetic marker. In this sense Copy Number Variants (CNVs) have received much interest recently. CNVs are segments of DNA that can range from 500 bases to several megabases in length, which can be present in different numbers in the genome (i.e. as deletions or as amplifications). Although this type of variations had been already known for many years, especially for their association in cancer phenotypes and neurological syndromes, they have been only recently implicated in common disease suceptibility. In particular common CNVs, also called Copy Number Polymorphisms, could well be alternative markers, targeting other genomic loci where SNPs can hardly be found or genotyped. Several recently associated CNP loci like the *β–Defensin* (Hollox *et al.*, 2008) and *LCE3* (de Cid *et al.*, 2009) CNPs in Psoriasis or the *IRGM* locus in Crohn's disease (Hollox *et al.*, 2008) have supported to this possibility. Nevertheless, a recently published whole genome study using the study groups of the WTCCC consortium seems to have sensibly lowered the prospects for his kind of marker (Conrad *et al.*, 2009); in this study they demonstrate that, for most CNVs, there will be a neighbouring SNP in high LD that can capture most of the association to disease, thus giving additional support to the previous SNP–based genomic approach.

Another possibility for the missed genetic heritability is the possibility of a Rare Variant Common Disease scenario. That is, instead of genetic variants having a minor allele frequency of >1% in the general population, highly infrequent but highly penetrant polymorphisms culd be associated with disease. With whole genome resequencing technologies having become today a technical reality, it is possible that these variants, if they exist, will be characterized. However, some caution is in order. Before devoting millions of research funding in resequencing

it should be made clear what kind of realistic expectations can be made from this expensive technology. Pharmaceutical companies, for example, made heavy investments in the characterization of SNPs, from which they expected that they would have gross benefits in return. So far, little transference onto the medical aspects has been obtained from all this heavy investment in genotyping. Should we expect a different scenario after resequencing? Whole genome resequencing must be done but, in our view, until analytical designs that confront biological complexity are not developed and this becomes a mature field, we will repeatedly find ourselves in front of the same obstacle.

The work we have presented here is a small part of a continuing line of research of our group on genomic approaches to the study of Rheumatoid Arthritis and other chronic inflammatory diseases. Gene expression microarrays, for example, have been used for the characterization of multigenic predictors of the response to biological therapies like infliximab (Julià *et al.*, 2009b) or rituximab (Julià *et al.*, 2009a). Also, the Grup de Recerca de Reumatologia is actually coordinating one of the most comprehensive genomic projects in several Immune–Mediated Inflammatory Diseases (IMIDs) including RA. With the collaboration of more than 50 clinical departments from around Spain we are working together to translate the analytical power of the new technologies into meaningful tools for clinicians so that we can contribute, as much as we can, to the improvement of the lives of patients with chronic inflammatory diseases.

Appendix A

Bioinformatic tools used

A.1 General programming languages

- Statically typed languages: *C*, *C++*.
- Dynamic typed languages: *Python*, *Perl*.
- Database managing languages: *MySQL*.
- Text processing languages: *Latex* (http://www.latex-project.org/).

A.2 Statistical software

- Open-source: *R statistical language* (http://cran.r-project.org/).
- Private: *SPSS*.

A.3 General bioinformatics tools

- *Bioconductor* (R extension for Bioinformatic analysis, http://www.bioconductor.org/).
- *Biopython* (Python extension for Bioinformatic analysis, http://biopython.org/).
- *Bioperl* (Perl extension for Bioinformatic analysis, http://biopython.org/).

A.4 Genetic analysis software

- Population genetics: *Arlequin* (http://cmpg.unibe.ch/software/arlequin3/).
- Family-based association testing: *Transmit* (http://www-gene.cimr.cam.ac.uk/clayton/software/).

- Microsatellite association analysis: *CLUMP* (http://www.smd.qmul.ac.uk/statgen/dcurtis/software.html).
- LD, haplotype analysis: *FastEHplus* (http://linkage.rockefeller.edu/soft/), *Haploview* (http://www.broadinstitute.org/haploview/), *Gap* (*R* package).
- GWAS analysis tool: *PLINK* (http://pngu.mgh.harvard.edu/∼purcell/plink/).
- Principal Component Analysis: *EIGENSTRAT* (http://genepath.med.harvard.edu/∼reich/Software.htm).
- Epistasis analysis: *Multifactor Dimensionality Reduction* (http://www.epistasis.org/software.html), *OR test* (*PLINK* function).

A.5 Webserver bioinformatic programs

- Genotyping data management: *SNPator* (http://www.snpator.org/)
- Gene Ontology analysis: *Gostat* (http://gostat.wehi.edu.au/)
- Reverse engineering of transcriptional regulatory networks: *CARRIE* (http://zlab.bu.edu/CarrieServer/html/)

References

Aho, K., Koskenvuo, M., Tuominen, J. & Kaprio, J. (1986). Occurrence of rheumatoid arthritis in a nationwide series of twins. *J Rheumatol*, **13**, 899–902.

Alon, U. (2007). Network motifs: theory and experimental approaches. *Nat Rev Genet*, **8**, 450–61.

Altshuler, D., Brooks, L.D., Chakravarti, A., Collins, F.S., Daly, M.J. & Donnelly, P. (2005). A haplotype map of the human genome. *Nature*, **437**, 1299–320.

Arnett, F., Edworthy, S., Bloch, D., McShane, D., Fries, J., Cooper, N., Healey, L., Kaplan, S., Liang, M. & Luthra, H. (1988). The american rheumatism association 1987 revised criteria for the classification of rheumatoid arthritis. *Arthritis Rheum*, **31**, 315–24.

Avery, O.T., Macleod, C.M. & McCarty, M. (1944). Studies on the chemical nature of the substance inducing transformation of pneumococcal types : Induction of transformation by a desoxyribonucleic acid fraction isolated from pneumococcus type III. *J Exp Med*, **79**, 137–158.

Balding, D.J. (2001). *Handbook of statistical genetics*. John Wiley & Sons, Ltd, West Sussex.

Barrett, J.C. & Cardon, L.R. (2006). Evaluating coverage of genome–wide association studies. *Nat Genet*, **38**, 659–62.

Barton, A., Thomson, W., Ke, X., Eyre, S., Hinks, A., Bowes, J., Plant, D., Gibbons, L.J., Wilson, A.G., Bax, D.E., Morgan, A.W., Emery, P., Steer, S., Hocking, L., Reid, D.M., Wordsworth, P., Harrison, P. & Worthington, J. (2008). Rheumatoid arthritis susceptibility loci at chromosomes 10p15, 12q13 and 22q13. *Nat Genet*, **40**, 1156–9.

Basso, K., Margolin, A.A., Stolovitzky, G., Klein, U., Dalla– Favera, R. & Califano, A. (2005). Reverse engineering of regulatory networks in human B cells. *Nat Genet*, **37**, 382–90.

Bateson, W. (1909). *Mendel's Principles of Heredity*. Cambridge University Press, Cambrige.

Begovich, A.B., Carlton, V.E., Honigberg, L.A., Schrodi, S.J., Chokkalingam, A.P., Alexander, H.C., Ardlie, K.G., Huang, Q., Smith, A.M., Spoerke, J.M., Conn, M.T., Chang, M., Chang, S.Y., Saiki, R.K., Catanese, J.J., Leong, D.U., Garcia, V.E., McAllister, L.B., Jeffery, D.A., Lee, A.T., Batliwalla, F., Remmers, E., Criswell, L.A., Seldin, M.F., Kastner, D.L., Amos, C.I., Sninsky, J.J. & Gregersen, P.K. (2004). A missense single–nucleotide polymorphism in a gene encoding a protein tyrosine phosphatase (PTPN22) is associated with rheumatoid arthritis. *Am J Hum Genet*, **75**, 330–7.

Beissbarth, T. & Speed, T.P. (2004). Gostat: find statistically overrepresented gene ontologies within a group of genes. *Bioinformatics*, **20**, 1464–5.

Bellman, R. (1957). *Dynammic programming.*. Princeton University Press, Princeton, NJ.

Bluestone, J.A., Mackay, C.R., O'Shea, J.J. & Stockinger, B. (2009). The functional plasticity of T cell subsets. *Nat Rev Immunol*, **9**, 811–6.

Bosch, E., Calafell, F., Comas, D., Oefner, P.J., Underhill, P.A. & Bertranpetit, J. (2001). High-resolution analysis of human y-chromosome variation shows a sharp discontinuity and limited gene flow between northwestern africa and the iberian peninsula. *Am J Hum Genet*, **68**, 1019–29.

Botstein, D. & Risch, N. (2003). Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nat Genet*, **33 Suppl**, 228–37.

Brentano, F., Ospelt, C., Stanczyk, J., Gay, R.E., Gay, S. & Kyburz, D. (2009). Abundant expression of the interleukin (IL)23 subunit p19, but low levels of bioactive IL23 in the rheumatoid synovium: differential expression and Toll–like receptor–(TLR) dependent regulation of the IL23 subunits, p19 and p40, in rheumatoid arthritis. *Ann Rheum Dis*, **68**, 143–50.

Bueno, E.M. & Glowacki, J. (2009). Cell–free and cell–based approaches for bone regeneration. *Nat Rev Rheumatol*, **5**, 685–97.

Campbell, P.N., Doniach, D., Hudson, R.V. & Roitt, I.M. (1956). Auto–antibodies in Hashimoto's disease (lymphadenoid goitre). *Lancet*, **271**, 820–1.

Cardon, L.R. & Bell, J.I. (2001). Association study designs for complex diseases. *Nat Rev Genet*, **2**, 91–9.

Carlborg, O., Kerje, S., Schutz, K., Jacobsson, L., Jensen, P. & Andersson, L. (2003). A global search reveals epistatic interaction between QTL for early growth in the chicken. *Genome Res*, **13**, 413–21.

Carmona, L., Villaverde, V., Hernandez– Garcia, C., Ballina, J., Gabriel, R. & Laffon, A. (2002). The prevalence of rheumatoid arthritis in the general population of Spain. *Rheumatology*, **41**, 88–95.

Chabaud, M., Durand, J.M., Buchs, N., Fossiez, F., Page, G., Frappart, L. & Miossec, P. (1999). Human interleukin–17: A T cell–derived proinflammatory cytokine produced by the rheumatoid synovium. *Arthritis Rheum*, **42**, 963–70.

Chang, H.Y., Chi, J.T., Dudoit, S., Bondre, C., van de Rijn, M., Botstein, D. & Brown, P.O. (2002). Diversity, topographic differentiation, and positional memory in human fibroblasts. *Proc Natl Acad Sci U S A*, **99**, 12877–82.

Cherwinski, H.M., Schumacher, J.H., Brown, K.D. & Mosmann, T.R. (1987). Two types of mouse helper T cell clone. III. further differences in lymphokine synthesis between Th1 and Th2 clones revealed by RNA hybridization, functionally monospecific bioassays, and monoclonal antibodies. *J Exp Med*, **166**, 1229–44.

Chikanza, I., Petrou, P., Kingsley, G., Chrousos, G. & Panayi, G. (1992). Defective hypothalamic response to immune and inflammatory stimuli in patients with rheumatoid arthritis. *Arthritis Rheum*, **35**, 1281–8.

Clayton, D.G., Walker, N.M., Smyth, D.J., Pask, R., Cooper, J.D., Maier, L.M., Smink, L.J., Lam, A.C., Ovington, N.R., Stevens, H.E., Nutland, S., Howson, J.M., Faham, M., Moorhead, M., Jones, H.B., Falkowski, M., Hardenbol, P., Willis, T.D. & Todd, J.A. (2005). Population structure, differential bias and genomic control in a large–scale, case–control association study. *Nat Genet*, **37**, 1243–6.

Cobb, M. (2006). Heredity before genetics: a history. *Nat Rev Genet*, **7**, 953–8.

Coffman, R.L. (2006). Origins of the T(H)1–T(H)2 model: a personal perspective. *Nat Immunol*, **7**, 539–41.

Conrad, D.F., Pinto, D., Redon, R., Feuk, L., Gokcumen, O., Zhang, Y., Aerts, J., Andrews, T.D., Barnes, C., Campbell, P., Fitzgerald, T., Hu, M., Ihm, C.H., Kristiansson, K., Macarthur, D.G., Macdonald, J.R., Onyiah, I., Pang, A.W., Robson, S., Stirrups, K., Valsesia, A., Walter, K., Wei, J., Tyler– Smith, C., Carter, N.P., Lee, C., Scherer, S.W. & Hurles, M.E. (2009). Origins and functional impact of copy number variation in the human genome. *Nature*, 1–9.

Consortium, G. (2001). Creating the gene ontology resource: design and implementation. *Genome Res*, **11**, 1425–33.

Cordell, H.J. (2002). Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. *Hum Mol Genet*, **11**, 2463–8.

Cordell, H.J. & Clayton, D.G. (2005). Genetic association studies. *Lancet*, **366**, 1121–31.

Cornelis, F., Faure, S., Martinez, M., Prud'homme, J.F., Fritz, P., Dib, C., Alves, H., Barrera, P., de Vries, N., Balsa, A., Pascual– Salcedo, D., Maenaut, K., Westhovens, R., Migliorini, P., Tran, T.H., Delaye, A., Prince, N., Lefevre, C., Thomas, G., Poirier, M., Soubigou, S., Alibert, O., Lasbleiz, S., Fouix, S., Weissenbach, J. & et al. (1998). New susceptibility locus for rheumatoid arthritis suggested by a genome–wide linkage study. *Proc Natl Acad Sci U S A*, **95**, 10746–50.

Courtenay, J.S., Dallman, M.J., Dayan, A.D., Martin, A. & Mosedale, B. (1980). Immunisation against heterologous type II collagen induces arthritis in mice. *Nature*, **283**, 666–8.

Crick, F. (1970). Central dogma of molecular biology. *Nature*, **227**, 561–3.

Daly, M.J., Rioux, J.D., Schaffner, S.F., Hudson, T.J. & Lander, E.S. (2001). High–resolution haplotype structure in the human genome. *Nat Genet*, **29**, 229–32.

Davis, L.S. (2003). A question of transformation: the synovial fibroblast in rheumatoid arthritis. *Am J Pathol*, **162**, 1399–402.

de Cid, R., Riveira– Munoz, E., Zeeuwen, P.L., Robarge, J., Liao, W., Dannhauser, E.N., Giardina, E., Stuart, P.E., Nair, R., Helms, C., Escaramis, G., Ballana, E., Martin– Ezquerra, G., den Heijer, M., Kamsteeg, M., Joosten, I., Eichler, E.E., Lazaro, C., Pujol, R.M., Armengol, L., Abecasis, G., Elder, J.T., Novelli, G., Armour, J.A., Kwok, P.Y., Bowcock, A., Schalkwijk, J. & Estivill, X. (2009). Deletion of the late cornified envelope LCE3B and LCE3C genes as a susceptibility factor for psoriasis. *Nat Genet*, **41**, 211–5.

Deighton, C., Walker, D., Griffiths, I. & Roberts, D. (1989). The contribution of HLA to rheumatoid arthritis. *Clin Genet*, **36**, 178–82.

Del Puente, A., Knowler, W.C., Pettitt, D.J. & Bennett, P.H. (1989). High incidence and prevalence of rheumatoid arthritis in Pima Indians. *Am J Epidemiol*, **129**, 1170–8.

DeLisi, C. (2008). Meetings that changed the world: Santa Fe 1986: Human genome baby–steps. *Nature*, **455**, 876–7.

Devlin, B. & Roeder, K. (1999). Genomic control for association studies. *Biometrics*, **55**, 997–1004.

Distler, J.H., Jungel, A., Huber, L.C., Seemayer, C.A., Reich, r., C. F., Gay, R.E., Michel, B.A., Fontana, A., Gay, S., Pisetsky, D.S. & Distler, O. (2005). The induction of matrix metalloproteinase and cytokine expression in synovial fibroblasts stimulated with immune cell microparticles. *Proc Natl Acad Sci U S A*, **102**, 2892–7.

Donis-Keller, H., Green, P., Helms, C., Cartinhour, S., Weiffenbach, B., Stephens, K., Keith, T.P., Bowden, D.W., Smith, D.R., Lander, E.S. & et al. (1987). A genetic linkage map of the human genome. *Cell*, **51**, 319–37.

Dudley, J.T., Tibshirani, R., Deshpande, T. & Butte, A.J. (2009). Disease signatures are robust across tissues and experiments. *Mol Syst Biol*, **5**, 307.

Duerr, R.H., Taylor, K.D., Brant, S.R., Rioux, J.D., Silverberg, M.S., Daly, M.J., Steinhart, A.H., Abraham, C., Regueiro, M., Griffiths, A., Dassopoulos, T., Bitton, A., Yang, H., Targan, S., Datta, L.W., Kistner, E.O., Schumm, L.P., Lee, A.T., Gregersen, P.K., Barmada, M.M., Rotter, J.I., Nicolae, D.L. & Cho, J.H. (2006). A genome–wide association study identifies IL23R as an inflammatory bowel disease gene. *Science*, **314**, 1461–3.

Edwards, J.C., Szczepanski, L., Szechinski, J., Filipowicz– Sosnowska, A., Emery, P., Close, D.R., Stevens, R.M. & Shaw, T. (2004). Efficacy of B–cell–targeted therapy with rituximab in patients with rheumatoid arthritis. *N Engl J Med*, **350**, 2572–81.

Ehrenstein, M.R., Evans, J.G., Singh, A., Moore, S., Warnes, G., Isenberg, D.A. & Mauri, C. (2004). Compromised function of regulatory T cells in rheumatoid arthritis and reversal by anti–TNFalpha therapy. *J Exp Med*, **200**, 277–85.

Elenbaas, B. & Weinberg, R.A. (2001). Heterotypic signaling between epithelial tumor cells and fibroblasts in carcinoma formation. *Exp Cell Res*, **264**, 169–84.

Eshed, Y. & Zamir, D. (1996). Less–than–additive epistatic interactions of quantitative trait loci in tomato. *Genetics*, **143**, 1807–17.

Falconer, D.S. (1967). The inheritance of liability to diseases with variable age of onset, with particular reference to diabetes mellitus. *Ann Hum Genet*, **31**, 1–20.

Fife, M., Fisher, S., John, S., Worthington, J., Shah, C., Ollier, W., Panayi, G., Lewis, C. & Lanchbury, J. (2000). Multipoint linkage analysis of a candidate gene locus in rheumatoid arthritis demonstrates significant evidence of linkage and association with the corticotropin–releasing hormone genomic region. *Arthritis Rheum*, **43**, 1673–8.

Firestein, G.S. (1996). Invasive fibroblast–like synoviocytes in rheumatoid arthritis. passive responders or transformed aggressors? *Arthritis Rheum*, **39**, 1781–90.

Firestein, G.S. (1999). Starving the synovium: angiogenesis and inflammation in rheumatoid arthritis. *J Clin Invest*, **103**, 3–4.

Firestein, G.S. (2003). Evolving concepts of rheumatoid arthritis. *Nature*, **423**, 356–61.

Francis– West, P.H., Parish, J., Lee, K. & Archer, C.W. (1999). BMP/GDF–signalling interactions during synovial joint development. *Cell Tissue Res*, **296**, 111–9.

Gentleman, R.C., Carey, V.J., Bates, D.M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., Hornik, K., Hothorn, T., Huber, W., Iacus, S., Irizarry, R., Leisch, F., Li, C., Maechler, M., Rossini, A.J., Sawitzki, G., Smith, C., Smyth, G., Tierney, L., Yang, J.Y. & Zhang, J. (2004). Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol*, **5**, R80.

Gregersen, P.K. (1999). Genetics of rheumatoid arthritis: confronting complexity. *Arthritis Res*, **1**, 37–44.

Gregersen, P.K., Silver, J. & Winchester, R.J. (1987). The shared epitope hypothesis. an approach to understanding the molecular genetics of susceptibility to rheumatoid arthritis. *Arthritis Rheum*, **30**, 1205–13.

Gunderson, K.L., Steemers, F.J., Lee, G., Mendoza, L.G. & Chee, M.S. (2005). A genome–wide scalable SNP genotyping assay using microarray technology. *Nat Genet*, **37**, 549–54.

Gusella, J.F., Wexler, N.S., Conneally, P.M., Naylor, S.L., Anderson, M.A., Tanzi, R.E., Watkins, P.C., Ottina, K., Wallace, M.R., Sakaguchi, A.Y. & et al. (1983). A polymorphic DNA marker genetically linked to Huntington's disease. *Nature*, **306**, 234–8.

Harrington, L.E., Hatton, R.D., Mangan, P.R., Turner, H., Murphy, T.L., Murphy, K.M. & Weaver, C.T. (2005). Interleukin 17–producing CD4+ effector T cells develop via a lineage distinct from the T helper type 1 and 2 lineages. *Nat Immunol*, **6**, 1123–32.

Harvey, J., Lotze, M., Stevens, M.B., Lambert, G. & Jacobson, D. (1981). Rheumatoid arthritis in a Chippewa Band. I. pilot screening study of disease prevalence. *Arthritis Rheum*, **24**, 717–21.

Haverty, P.M., Hansen, U. & Weng, Z. (2004). Computational inference of transcriptional regulatory networks from expression profiling and transcription factor binding site identification. *Nucleic Acids Res*, **32**, 179–88.

Haworth, O., Hardie, D.L., Burman, A., Rainger, G.E., Eksteen, B., Adams, D.H., Salmon, M., Nash, G.B. & Buckley, C.D. (2008). A role for the integrin alpha6beta1 in the differential distribution of CD4 and CD8 T–cell subsets within the rheumatoid synovium. *Rheumatology (Oxford)*, **47**, 1329–34.

Henshall, J.M. & Goddard, M.E. (1999). Multiple–trait mapping of quantitative trait loci after selective genotyping using logistic regression. *Genetics*, **151**, 885–94.

Hirschhorn, J.N. & Daly, M.J. (2005). Genome–wide association studies for common diseases and complex traits. *Nat Rev Genet*, **6**, 95–108.

Hirschhorn, J.N., Lohmueller, K., Byrne, E. & Hirschhorn, K. (2002). A comprehensive review of genetic association studies. *Genet Med*, **4**, 45–61.

Hochberg, Y. (1988). A sharper bonferroni procedure for multiple tests of significance. *Biometrika.*, 800–802.

Hollox, E.J., Huffmeier, U., Zeeuwen, P.L., Palla, R., Lascorz, J., Rodijk– Olthuis, D., van de Kerkhof, P.C., Traupe, H., de Jongh, G., den Heijer, M., Reis, A., Armour, J.A. & Schalkwijk, J. (2008). Psoriasis is associated with increased beta–defensin genomic copy number. *Nat Genet*, **40**, 23–5.

Horton, R., Wilming, L., Rand, V., Lovering, R.C., Bruford, E.A., Khodiyar, V.K., Lush, M.J., Povey, S., Talbot, J., C. C., Wright, M.W., Wain, H.M., Trowsdale, J., Ziegler, A. & Beck, S. (2004). Gene map of the extended human MHC. *Nat Rev Genet*, **5**, 889–99.

Huber, L.C., Distler, O., Tarner, I., Gay, R.E., Gay, S. & Pap, T. (2006). Synovial fibroblasts: key players in rheumatoid arthritis. *Rheumatology (Oxford)*, **45**, 669–75.

Ihaka, R. & Gentleman, R. (1996). R: a language for data analysis and graphics. *J Comput Graphical Statist*, 299–314.

IHC (2003). The International HapMap Project. *Nature*, **426**, 789–96.

Jawaheer, D. & Gregersen, P.K. (2002). The search for rheumatoid arthritis susceptibility genes: a call for global collaboration. *Arthritis Rheum*, **46**, 582–4.

Jawaheer, D., Seldin, M., Amos, C., Chen, W., Shigeta, R., Monteiro, J., Kern, M., Criswell, L., Albani, S., Nelson, J., Clegg, D., Pope, R., Schroeder, H.J., Bridges, S.J., Pisetsky, D., Ward, R., Kastner, D., Wilder, R., Pincus, T., Callahan, L., Flemming, D., Wener, M. & Gregersen, P. (2001). A genomewide screen in multiplex rheumatoid arthritis families suggests genetic overlap with other autoimmune diseases. *Am J Hum Genet*, **68**, 927–36.

Julià, A. & Marsal, S. (2003). Complex diseases: rheumatoid arthritis as a model of study. *Med Clin (Barc)*, **121**, 616–8.

Julià, A., Gallardo, D., Vidal, F., De, J.J., Barcelo, P., Vilardell, M. & Marsal, S. (2003). Association study between corticotrophin–releasing hormone genomic region (8q13) and rheumatoid arthritis in the Spanish population. *Rheumatology*, **42**, 1–5.

Julià, A., Gallardo, D., Vidal, F., Tomas, C., Barcelo, P., Vilardell, M. & Marsal, S. (2004). Lack of association between the corticotropin–releasing hormone locus and rheumatoid arthritis. *Arthritis Rheum*, **50**, 2706–8.

Julià, A., Moore, J., Miquel, L., Alegre, C., Barcelo, P., Ritchie, M. & Marsal, S. (2007). Identification of a two–loci epistatic interaction associated with susceptibility to rheumatoid arthritis through reverse engineering and multifactor dimensionality reduction. *Genomics*, **90**, 6–13.

Julià, A., Ballina, J., Cañete, J., Balsa, A., Tornero– Molina, J., Naranjo, A., Alperi– López, M., Erra, A., Pascual– Salcedo, D., Barceló, P., Camps, J. & Marsal, S. (2008). Genome–wide association study of rheumatoid arthritis in the Spanish population: KLF12 as a risk locus for rheumatoid arthritis susceptibility. *Arthritis Rheum*, **58**, 2275–2286.

Julià, A., Barcelo, M., Erra, A., Palacio, C. & Marsal, S. (2009a). Identification of candidate genes for rituximab response in rheumatoid arthritis patients by microarray expression profiling in blood cells. *Pharmacogenomics*, **10**, 1697–708.

Julià, A., Erra, A., Palacio, C., Tomas, C., Sans, X., Barcelo, P. & Marsal, S. (2009b). An eight–gene blood expression profile predicts the response to infliximab in rheumatoid arthritis. *PLoS One*, **4**, e7556.

Kantha, S.S. (1992). The legacy of von Behring and Kitasato. *Immunol Today*, **13**, 374.

Khurana, J.S. (2009). *Bone pathology*. SpringerLink, Berlin.

Kim, J.H., Sen, S., Avery, C.S., Simpson, E., Chandler, P., Nishina, P.M., Churchill, G.A. & Naggert, J.K. (2001). Genetic analysis of a new mouse model for non–insulin–dependent diabetes. *Genomics*, **74**, 273–86.

Kirkham, B.W., Lassere, M.N., Edmonds, J.P., Juhasz, K.M., Bird, P.A., Lee, C.S., Shnier, R. & Portek, I.J. (2006). Synovial membrane cytokine expression is predictive of joint damage progression in rheumatoid arthritis: a two–year prospective study (the DAMAGE study cohort). *Arthritis Rheum*, **54**, 1122–31.

Klein, R.J., Zeiss, C., Chew, E.Y., Tsai, J.Y., Sackler, R.S., Haynes, C., Henning, A.K., SanGiovanni, J.P., Mane, S.M., Mayne, S.T., Bracken, M.B., Ferris, F.L., Ott, J., Barnstable, C. & Hoh, J. (2005). Complement factor H polymorphism in age–related macular degeneration. *Science*, **308**, 385–9.

Knowler, W.C., Williams, R.C., Pettitt, D.J. & Steinberg, A.G. (1988). Gm3;5,13,14 and type 2 diabetes mellitus: an association in American Indians with genetic admixture. *Am J Hum Genet*, **43**, 520–6.

Kong, Y.Y., Feige, U., Sarosi, I., Bolon, B., Tafuri, A., Morony, S., Capparelli, C., Li, J., Elliott, R., McCabe, S., Wong, T., Campagnuolo, G., Moran, E., Bogoch, E.R., Van, G., Nguyen, L.T.,

Ohashi, P.S., Lacey, D.L., Fish, E., Boyle, W.J. & Penninger, J.M. (1999). Activated T cells regulate bone loss and joint destruction in adjuvant arthritis through osteoprotegerin ligand. *Nature*, **402**, 304–9.

Kotake, S., Udagawa, N., Takahashi, N., Matsuzaki, K., Itoh, K., Ishiyama, S., Saito, S., Inoue, K., Kamatani, N., Gillespie, M.T., Martin, T.J. & Suda, T. (1999). IL–17 in synovial fluids from patients with rheumatoid arthritis is a potent stimulator of osteoclastogenesis. *J Clin Invest*, **103**, 1345–52.

Kremer, J.M., Westhovens, R., Leon, M., Di Giorgio, E., Alten, R., Steinfeld, S., Russell, A., Dougados, M., Emery, P., Nuamah, I.F., Williams, G.R., Becker, J.C., Hagerty, D.T. & Moreland, L.W. (2003). Treatment of rheumatoid arthritis by selective inhibition of t–cell activation with fusion protein CTLA4Ig. *N Engl J Med*, **349**, 1907–15.

Lander, E.S. (1996). The new genomics: global views of biology. *Science*, **274**, 536–9.

Lander, E.S. & Botstein, D. (1986). Strategies for studying heterogeneous genetic traits in humans by using a linkage map of restriction fragment length polymorphisms. *Proc Natl Acad Sci U S A*, **83**, 7353–7.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J.P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange– Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer,

S., Milne, S., Mullikin, J.C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R.H., Wilson, R.K., Hillier, L.W., McPherson, J.D., Marra, M.A., Mardis, E.R., Fulton, L.A., Chinwalla, A.T., Pepin, K.H., Gish, W.R., Chissoe, S.L., Wendl, M.C., Delehaunty, K.D., Miner, T.L., Delehaunty, A., Kramer, J.B., Cook, L.L., Fulton, R.S., Johnson, D.L., Minx, P.J., Clifton, S.W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J.F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M. *et al.* (2001). Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.

Lawrence, J.S. (1970). Heberden Oration, 1969. Rheumatoid arthritis—nature or nurture? *Ann Rheum Dis*, **29**, 357–79.

Lee, H.Y., Jeon, H.S., Song, E.K., Han, M.K., Park, S.I., Lee, S.I., Yun, H.J., Kim, J.R., Kim, J.S., Lee, Y.C., Kim, S.I., Kim, H.R., Choi, J.Y., Kang, I., Kim, H.Y. & Yoo, W.H. (2006). CD40 ligation of rheumatoid synovial fibroblasts regulates rankl–mediated osteoclastogenesis: evidence of NF–kappaB–dependent, CD40–mediated bone destruction in rheumatoid arthritis. *Arthritis Rheum*, **54**, 1747–58.

Lefevre, S., Knedla, A., Tennie, C., Kampmann, A., Wunrau, C., Dinser, R., Korb, A., Schnaker, E.M., Tarner, I.H., Robbins, P.D., Evans, C.H., Sturz, H., Steinmeyer, J., Gay, S., Scholmerich, J., Pap, T., Muller– Ladner, U. & Neumann, E. (2009). Synovial fibroblasts spread rheumatoid arthritis to unaffected joints. *Nat Med*, 1414–1420.

Lewontin, R.C. & Kojima, K. (1960). The evolutionary dynamics of complex polymorphisms. *Evolution*, **14**, 458–472.

Lipshutz, R.J., Fodor, S.P., Gingeras, T.R. & Lockhart, D.J. (1999). High density synthetic oligonucleotide arrays. *Nat Genet*, **21**, 20–4.

Lipsky, P.E., van der Heijde, D.M., St Clair, E.W., Furst, D.E., Breedveld, F.C., Kalden, J.R., Smolen, J.S., Weisman, M., Emery, P., Feldmann, M., Harriman, G.R. & Maini, R.N. (2000). Infliximab

and methotrexate in the treatment of rheumatoid arthritis. Anti–Tumor Necrosis Factor Trial in Rheumatoid Arthritis with Concomitant Therapy Study Group. *N Engl J Med*, **343**, 1594–602.

Liu, Y., Helms, C., Liao, W., Zaba, L.C., Duan, S., Gardner, J., Wise, C., Miner, A., Malloy, M.J., Pullinger, C.R., Kane, J.P., Saccone, S., Worthington, J., Bruce, I., Kwok, P.Y., Menter, A., Krueger, J., Barton, A., Saccone, N.L. & Bowcock, A.M. (2008). A genome–wide association study of psoriasis and psoriatic arthritis identifies new disease loci. *PLoS Genet*, **4**, e1000041.

Lubberts, E., Koenders, M.I., Oppers– Walgreen, B., van den Bersselaar, L., Coenen– de Roo, C.J., Joosten, L.A. & van den Berg, W.B. (2004). Treatment with a neutralizing anti–murine interleukin–17 antibody after the onset of collagen–induced arthritis reduces joint inflammation, cartilage destruction, and bone erosion. *Arthritis Rheum*, **50**, 650–9.

MacGregor, A., Snieder, H., Rigby, A., Koskenvuo, M., Kaprio, J., Aho, K. & Silman, A. (2000). Characterizing the quantitative genetic contribution to rheumatoid arthritis using data from twins. *Arthritis Rheum*, **43**, 30–7.

MacKay, K., Eyre, S., Myerscough, A., Milicic, A., Barton, A., Laval, S., Barrett, J., Lee, D., White, S., John, S., Brown, M., Bell, J., Silman, A., Ollier, W., Wordsworth, P. & Worthington, J. (2002). Whole–genome linkage analysis of rheumatoid arthritis susceptibility loci in 252 affected sibling pairs in the United Kingdom. *Arthritis Rheum*, **46**, 632–9.

Maisnier– Patin, S., Roth, J.R., Fredriksson, A., Nystrom, T., Berg, O.G. & Andersson, D.I. (2005). Genomic buffering mitigates the effects of deleterious mutations in bacteria. *Nat Genet*, **37**, 1376–9.

Marsal, S., Hall, M.A., Panayi, G.S. & Lanchbury, J.S. (1994). Association of TAP2 polymorphism with rheumatoid arthritis is secondary to allelic association with HLA-DRB1. *Arthritis Rheum*, **37**, 504–13.

Masood, E. (1999). As consortium plans free SNP map of human genome. *Nature*, **398**, 545–6.

Maxam, A.M. & Gilbert, W. (1977). A new method for sequencing dna. *Proc Natl Acad Sci U S A*, **74**, 560–4.

McInnes, I.B. & Schett, G. (2007). Cytokines in the pathogenesis of rheumatoid arthritis. *Nat Rev Immunol*, **7**, 429–42.

Menozzi, P., Piazza, A. & Cavalli– Sforza, L. (1978). Synthetic maps of human gene frequencies in europeans. *Science*, **201**, 786–92.

Meyer, L.H., Franssen, L. & Pap, T. (2006). The role of mesenchymal cells in the pathophysiology of inflammatory arthritis. *Best Pract Res Clin Rheumatol*, **20**, 969–81.

Misra, N., Bayry, J., Lacroix– Desmazes, S., Kazatchkine, M.D. & Kaveri, S.V. (2004). Cutting edge: human CD4+CD25+ T cells restrain the maturation and antigen–presenting function of dendritic cells. *J Immunol*, **172**, 4676–80.

Mohr, J. (1951). Estimation of linkage between the Lutheran and the Lewis blood groups. *Acta Pathol Microbiol Scand*, **29**, 339–44.

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

Moore, J.H. & Williams, S.M. (2009). Epistasis and its implications for personal genetics. *Am J Hum Genet*, **85**, 309–20.

Morton, N.E. (1956). The detection and estimation of linkage between the genes for elliptocytosis and the Rh blood type. *Am J Hum Genet*, **8**, 80–96.

Morton, N.E. & Chung, S. (1978). *Genetic Epidemiology*. Elsevier Science & Technology Books, Elsevier.

Morton, N.E. & Collins, A. (1998). Tests and estimates of allelic association in complex inheritance. *Proc Natl Acad Sci U S A*, **95**, 11389–93.

Muller– Ladner, U., Kriegsmann, J., Franklin, B.N., Matsumoto, S., Geiler, T., Gay, R.E. & Gay, S. (1996). Synovial fibroblasts of patients with rheumatoid arthritis attach to and invade normal human cartilage when engrafted into SCID mice. *Am J Pathol*, **149**, 1607–15.

Nair, R.P., Duffin, K.C., Helms, C., Ding, J., Stuart, P.E., Goldgar, D., Gudjonsson, J.E., Li, Y., Tejasvi, T., Feng, B.J., Ruether, A., Schreiber, S., Weichenthal, M., Gladman, D., Rahman, P., Schrodi, S.J., Prahalad, S., Guthery, S.L., Fischer, J., Liao, W., Kwok, P.Y., Menter, A., Lathrop, G.M., Wise, C.A., Begovich, A.B., Voorhees, J.J., Elder, J.T., Krueger, G.G., Bowcock, A.M. & Abecasis, G.R. (2009). Genome–wide scan reveals association of psoriasis with IL–23 and NF–kappaB pathways. *Nat Genet*, **41**, 199–204.

Nakae, S., Nambu, A., Sudo, K. & Iwakura, Y. (2003). Suppression of immune induction of collagen–induced arthritis in IL–17–deficient mice. *J Immunol*, **171**, 6173–7.

Nelson, M.R., Kardia, S.L., Ferrell, R.E. & Sing, C.F. (2001). A combinatorial partitioning method to identify multilocus genotypic partitions that predict quantitative trait variation. *Genome Res*, **11**, 458–70.

Nielsen, D.M., Ehm, M.G. & Weir, B.S. (1998). Detecting marker–disease association by testing for hardy–weinberg disequilibrium at a marker locus. *Am J Hum Genet*, **63**, 1531–40.

Novembre, J., Johnson, T., Bryc, K., Kutalik, Z., Boyko, A.R., Auton, A., Indap, A., King, K.S., Bergmann, S., Nelson, M.R., Stephens, M. & Bustamante, C.D. (2008). Genes mirror geography within europe. *Nature*, **456**, 98–101.

Page-McCaw, A., Ewald, A.J. & Werb, Z. (2007). Matrix metalloproteinases and the regulation of tissue remodelling. *Nat Rev Mol Cell Biol*, **8**, 221–33.

Pap, T., Muller– Ladner, U., Gay, R.E. & Gay, S. (2000). Fibroblast biology. role of synovial fibroblasts in the pathogenesis of rheumatoid arthritis. *Arthritis Res*, **2**, 361–7.

Pattin, K.A., White, B.C., Barney, N., Gui, J., Nelson, H.H., Kelsey, K.T., Andrew, A.S., Karagas, M.R. & Moore, J.H. (2009). A computationally efficient hypothesis testing method for epistasis analysis using multifactor dimensionality reduction. *Genet Epidemiol*, **33**, 87–94.

Peltonen, L. & McKusick, V.A. (2001). Genomics and medicine. dissecting human disease in the postgenomic era. *Science*, **291**, 1224–9.

Perricone, R., Perricone, C., De Carolis, C. & Shoenfeld, Y. (2008). NK cells in autoimmunity: a two–edg'd weapon of the immune system. *Autoimmun Rev*, **7**, 384–90.

Phillips, C., Salas, A., Sanchez, J.J., Fondevila, M., Gomez– Tato, A., Alvarez– Dios, J., Calaza, M., de Cal, M.C., Ballard, D., Lareu, M.V. & Carracedo, A. (2007). Inferring ancestral origin using a single multiplex assay of ancestry–informative marker SNPs. *Forensic Sci Int Genet*, **1**, 273–80.

Phillips, C., Prieto, L., Fondevila, M., Salas, A., Gomez– Tato, A., Alvarez– Dios, J., Alonso, A., Blanco– Verea, A., Brion, M., Montesino, M., Carracedo, A. & Lareu, M.V. (2009). Ancestry analysis in the 11–M Madrid bomb attack investigation. *PLoS One*, **4**, e6583.

Pierer, M., Rethage, J., Seibl, R., Lauener, R., Brentano, F., Wagner, U., Hantzschel, H., Michel, B.A., Gay, R.E., Gay, S. & Kyburz, D. (2004). Chemokine secretion of rheumatoid arthritis synovial fibroblasts stimulated by Toll–like receptor 2 ligands. *J Immunol*, **172**, 1256–65.

Plenge, R.M., Padyukov, L., Remmers, E.F., Purcell, S., Lee, A.T., Karlson, E.W., Wolfe, F., Kastner, D.L., Alfredsson, L., Altshuler, D., Gregersen, P.K., Klareskog, L. & Rioux, J.D. (2005). Replication of putative candidate–gene associations with rheumatoid arthritis

in >4,000 samples from North America and Sweden: association of susceptibility with PTPN22, CTLA4, and PADI4. *Am J Hum Genet*, **77**, 1044–60.

Plenge, R.M., Cotsapas, C., Davies, L., Price, A.L., de Bakker, P.I., Maller, J., Pe'er, I., Burtt, N.P., Blumenstiel, B., Defelice, M., Parkin, M., Barry, R., Winslow, W., Healy, C., Graham, R.R., Neale, B.M., Izmailova, E., Roubenoff, R., Parker, A.N., Glass, R., Karlson, E.W., Maher, N., Hafler, D.A., Lee, D.M., Seldin, M.F., Remmers, E.F., Lee, A.T., Padyukov, L., Alfredsson, L., Coblyn, J., Weinblatt, M.E., Gabriel, S.B., Purcell, S., Klareskog, L., Gregersen, P.K., Shadick, N.A., Daly, M.J. & Altshuler, D. (2007a). Two independent alleles at 6q23 associated with risk of rheumatoid arthritis. *Nat Genet*, **4**, 1477–82.

Plenge, R.M., Seielstad, M., Padyukov, L., Lee, A.T., Remmers, E.F., Ding, B., Liew, A., Khalili, H., Chandrasekaran, A., Davies, L.R., Li, W., Tan, A.K., Bonnard, C., Ong, R.T., Thalamuthu, A., Pettersson, S., Liu, C., Tian, C., Chen, W.V., Carulli, J.P., Beckman, E.M., Altshuler, D., Alfredsson, L., Criswell, L.A., Amos, C.I., Seldin, M.F., Kastner, D.L., Klareskog, L. & Gregersen, P.K. (2007b). TRAF1–C5 as a risk locus for rheumatoid arthritis—a genomewide study. *N Engl J Med*, **357**, 1199–209.

Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A. & Reich, D. (2006). Principal components analysis corrects for stratification in genome–wide association studies. *Nat Genet*, **38**, 904–9.

Pritchard, J.K. & Rosenberg, N.A. (1999). Use of unlinked genetic markers to detect population stratification in association studies. *Am J Hum Genet*, **65**, 220–8.

Purcell, S., Neale, B., Todd– Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., de Bakker, P.I., Daly, M.J. & Sham, P.C. (2007). PLINK: a tool set for whole–genome association and population–based linkage analyses. *Am J Hum Genet*, **81**, 559–75.

Raghavan, S., Cao, D., Widhe, M., Roth, K., Herrath, J., Engstrom, M., Roncador, G., Banham, A.H., Trollmo, C., Catrina, A.I. & Malmstrom, V. (2009). FOXP3 expression in blood, synovial fluid and synovial tissue during inflammatory arthritis and intra–articular corticosteroid treatment. *Ann Rheum Dis*, **68**, 1908–15.

Rajewsky, K. (1996). Clonal selection and learning in the antibody system. *Nature*, **381**, 751–8.

Raychaudhuri, S., Remmers, E.F., Lee, A.T., Hackett, R., Guiducci, C., Burtt, N.P., Gianniny, L., Korman, B.D., Padyukov, L., Kurreeman, F.A., Chang, M., Catanese, J.J., Ding, B., Wong, S., van der Helm– van Mil, A.H., Neale, B.M., Coblyn, J., Cui, J., Tak, P.P., Wolbink, G.J., Crusius, J.B., van der Horst– Bruinsma, I.E., Criswell, L.A., Amos, C.I., Seldin, M.F., Kastner, D.L., Ardlie, K.G., Alfredsson, L., Costenbader, K.H., Altshuler, D., Huizinga, T.W., Shadick, N.A., Weinblatt, M.E., de Vries, N., Worthington, J., Seielstad, M., Toes, R.E., Karlson, E.W., Begovich, A.B., Klareskog, L., Gregersen, P.K., Daly, M.J. & Plenge, R.M. (2008). Common variants at CD40 and other loci confer risk of rheumatoid arthritis. *Nat Genet*, **40**, 1216–23.

Raychaudhuri, S., Plenge, R.M., Rossin, E.J., Ng, A.C., Purcell, S.M., Sklar, P., Scolnick, E.M., Xavier, R.J., Altshuler, D. & Daly, M.J. (2009a). Identifying relationships among genomic disease regions: predicting genes at pathogenic SNP associations and rare deletions. *PLoS Genet*, **5**, e1000534.

Raychaudhuri, S., Thomson, B.P., Remmers, E.F., Eyre, S., Hinks, A., Guiducci, C., Catanese, J.J., Xie, G., Stahl, E.A., Chen, R., Alfredsson, L., Amos, C.I., Ardlie, K.G., Barton, A., Bowes, J., Burtt, N.P., Chang, M., Coblyn, J., Costenbader, K.H., Criswell, L.A., Crusius, J.B., Cui, J., De Jager, P.L., Ding, B., Emery, P., Flynn, E., Harrison, P., Hocking, L.J., Huizinga, T.W., Kastner, D.L., Ke, X., Kurreeman, F.A., Lee, A.T., Liu, X., Li, Y.,

Martin, P., Morgan, A.W., Padyukov, L., Reid, D.M., Seielstad, M., Seldin, M.F., Shadick, N.A., Steer, S., Tak, P.P., Thomson, W., van der Helm– van Mil, A.H., van der Horst– Bruinsma, I.E., Weinblatt, M.E., Wilson, A.G., Wolbink, G.J., Wordsworth, P., Altshuler, D., Karlson, E.W., Toes, R.E., de Vries, N., Begovich, A.B., Siminovitch, K.A., Worthington, J., Klareskog, L., Gregersen, P.K., Daly, M.J. & Plenge, R.M. (2009b). Genetic variants at CD28, PRDM1 and CD2/CD58 are associated with rheumatoid arthritis risk. *Nat Genet*, **41**, 1313–8.

Remmers, E.F., Plenge, R.M., Lee, A.T., Graham, R.R., Hom, G., Behrens, T.W., de Bakker, P.I., Le, J.M., Lee, H.S., Batliwalla, F., Li, W., Masters, S.L., Booty, M.G., Carulli, J.P., Padyukov, L., Alfredsson, L., Klareskog, L., Chen, W.V., Amos, C.I., Criswell, L.A., Seldin, M.F., Kastner, D.L. & Gregersen, P.K. (2007). STAT4 and the risk of rheumatoid arthritis and systemic lupus erythematosus. *N Engl J Med*, **357**, 977–86.

Rigby, A.S., Voelm, L. & Silman, A.J. (1993). Epistatic modeling in rheumatoid arthritis: an application of the risch theory. *Genet Epidemiol*, **10**, 311–20.

Risch, N. (1987). Assessing the role of HLA–linked and unlinked determinants of disease. *Am J Hum Genet*, **40**, 1–14.

Risch, N. (1990). Linkage strategies for genetically complex traits. i. multilocus models. *Am J Hum Genet*, **46**, 222–8.

Risch, N. & Merikangas, K. (1996). The future of genetic studies of complex human diseases. *Science*, **273**, 1516–7.

Ritchie, M.D., Hahn, L.W., Roodi, N., Bailey, L.R., Dupont, W.D., Parl, F.F. & Moore, J.H. (2001). Multifactor–dimensionality reduction reveals high–order interactions among estrogen–metabolism genes in sporadic breast cancer. *Am J Hum Genet*, **69**, 138–47.

Ritchlin, C. (2000). Fibroblast biology. effector signals released by the synovial fibroblast in arthritis. *Arthritis Res*, **2**, 356–60.

Rodriguez, M.R., Nunez– Roldan, A., Aguilar, F., Valenzuela, A., Garcia, A. & Gonzalez– Escribano, M.F. (2002). Association of the CTLA4 3' untranslated region polymorphism with the susceptibility to rheumatoid arthritis. *Hum Immunol*, **63**, 76–81.

Ropes, M.W., Bennett, G.A., Cobb, S., Jacox, R. & Jessar, R.A. (1957). Proposed diagnostic criteria for rheumatoid arthritis. *Ann Rheum Dis*, **16**, 118–25.

Rose, H.M., Ragan, C. & et al. (1948). Differential agglutination of normal and sensitized sheep erythrocytes by sera of patients with rheumatoid arthritis. *Proc Soc Exp Biol Med*, **68**, 1–6.

Rose, N.R. & Bona, C. (1993). Defining criteria for autoimmune diseases (Witebsky's postulates revisited). *Immunol Today*, **14**, 426–30.

Rose, N.R., Kite, J., J. H., Doebbler, T.K., Spier, R., Skelton, F.R. & Witebsky, E. (1965). Studies on experimental thyroiditis. *Ann N Y Acad Sci*, **124**, 201–30.

Saiki, R.K., Scharf, S., Faloona, F., Mullis, K.B., Horn, G.T., Erlich, H.A. & Arnheim, N. (1985). Enzymatic amplification of beta–globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. *Science*, **230**, 1350–4.

Sakaguchi, S., Sakaguchi, N., Asano, M., Itoh, M. & Toda, M. (1995). Immunologic self–tolerance maintained by activated t cells expressing IL–2 receptor alpha–chains (CD25). breakdown of a single mechanism of self–tolerance causes various autoimmune diseases. *J Immunol*, **155**, 1151–64.

Sanger, F., Nicklen, S. & Coulson, A.R. (1977). DNA sequencing with chain–terminating inhibitors. *Proc Natl Acad Sci U S A*, **74**, 5463–7.

Sarkar, S. & Fox, D.A. (2008). Regulatory T cells in rheumatoid arthritis. *Curr Rheumatol Rep*, **10**, 405–12.

Schena, M., Shalon, D., Davis, R.W. & Brown, P.O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, **270**, 467–70.

Schena, M., Shalon, D., Heller, R., Chai, A., Brown, P.O. & Davis, R.W. (1996). Parallel human genome analysis: microarray–based expression monitoring of 1000 genes. *Proc Natl Acad Sci U S A*, **93**, 10614–9.

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

Shiozawa, S., Hayashi, S., Tsukamoto, Y., Goko, H., Kawasaki, H., Wada, T., Shimizu, K., Yasuda, N., Kamatani, N., Takasugi, K., Tanaka, Y., Shiozawa, K. & Imura, S. (1998). Identification of the gene loci that predispose to rheumatoid arthritis. *Int Immunol*, **10**, 1891–5.

Silman, A., MacGregor, A., Thomson, W., Holligan, S., Carthy, D., Farhan, A. & Ollier, W. (1993a). Twin concordance rates for rheumatoid arthritis: results from a nationwide study. *Br J Rheumatol*, **32**, 903–7.

Silman, A.J., Ollier, W., Holligan, S., Birrell, F., Adebajo, A., Asuzu, M.C., Thomson, W. & Pepper, L. (1993b). Absence of rheumatoid arthritis in a rural Nigerian population. *J Rheumatol*, **20**, 618–22.

Silman, A.J., Newman, J. & MacGregor, A.J. (1996). Cigarette smoking increases the risk of rheumatoid arthritis. results from a nationwide study of disease–discordant twins. *Arthritis Rheum*, **39**, 732–5.

Sladek, R., Rocheleau, G., Rung, J., Dina, C., Shen, L., Serre, D., Boutin, P., Vincent, D., Belisle, A., Hadjadj, S., Balkau, B., Heude, B., Charpentier, G., Hudson, T.J., Montpetit, A., Pshezhetsky, A.V., Prentki, M., Posner, B.I., Balding, D.J., Meyre, D., Polychronakos, C. & Froguel, P. (2007). A genome–wide

association study identifies novel risk loci for type 2 diabetes. *Nature*, **445**, 881–5.

Smolen, J.S., Aletaha, D., Koeller, M., Weisman, M.H. & Emery, P. (2007). New therapies for treatment of rheumatoid arthritis. *Lancet*, **370**, 1861–74.

Solomon, E. & Bodmer, W.F. (1979). Evolution of sickle variant gene. *Lancet*, **1**, 923.

Southern, E.M. (1975). Detection of specific sequences among DNA fragments separated by gel electrophoresis. *J Mol Biol*, **98**, 503–17.

Spielman, R., McGinnis, R. & Ewens, W. (1993). Transmission test for linkage disequilibrium: the insulin gene region and insulin–dependent diabetes mellitus (IDDM). *Am J Hum Genet*, **52**, 506–16.

Spitz, F. & Duboule, D. (2001). Development. The art of making a joint. *Science*, **291**, 1713–4.

Stastny, P. (1976). Mixed lymphocyte cultures in rheumatoid arthritis. *J Clin Invest*, **57**, 1148–57.

Sternberg, E., Hill, J., Chrousos, G., Kamilaris, T., Listwak, S., Gold, P. & Wilder, R. (1989a). Inflammatory mediator–induced hypothalamic–pituitary–adrenal axis activation is defective in streptococcal cell wall arthritis–susceptible Lewis rats. *Proc Natl Acad Sci U S A*, **86**, 2374–8.

Sternberg, E., Young, W., Bernardini, R., Calogero, A., Chrousos, G., Gold, P. & Wilder, R. (1989b). A central nervous system defect in biosynthesis of corticotropin–releasing hormone is associated with susceptibility to streptococcal cell wall–induced arthritis in Lewis rats. *Proc Natl Acad Sci U S A*, **86**, 4771–5.

Storey, G.D. (2001). Alfred Baring Garrod (1819–1907). *Rheumatology (Oxford)*, **40**, 1189–90.

Sturtevant, A.H. (1913). The linear arrangement of six sex–linked factors in drosophila, as shown by their mode of association. *J Exp Zool*, **14**, 43–59.

Sugiyama, F., Churchill, G.A., Higgins, D.C., Johns, C., Makaritsis, K.P., Gavras, H. & Paigen, B. (2001). Concordance of murine quantitative trait loci for salt–induced hypertension with rat and human loci. *Genomics*, **71**, 70–7.

Suzuki, A., Yamada, R., Chang, X., Tokuhiro, S., Sawada, T., Suzuki, M., Nagasaki, M., Nakayama– Hamada, M., Kawaida, R., Ono, M., Ohtsuki, M., Furukawa, H., Yoshino, S., Yukioka, M., Tohma, S., Matsubara, T., Wakitani, S., Teshima, R., Nishioka, Y., Sekine, A., Iida, A., Takahashi, A., Tsunoda, T., Nakamura, Y. & Yamamoto, K. (2003). Functional haplotypes of PADI4, encoding citrullinating enzyme peptidylarginine deiminase 4, are associated with rheumatoid arthritis. *Nat Genet*, **34**, 395–402.

Thomson, W., Barton, A., Ke, X., Eyre, S., Hinks, A., Bowes, J., Donn, R., Symmons, D., Hider, S., Bruce, I.N., Wilson, A.G., Marinou, I., Morgan, A., Emery, P., Carter, A., Steer, S., Hocking, L., Reid, D.M., Wordsworth, P., Harrison, P., Strachan, D. & Worthington, J. (2007). Rheumatoid arthritis association at 6q23. *Nat Genet*, **39**, 1431–3.

Thornton-Wells, T.A., Moore, J.H. & Haines, J.L. (2004). Genetics, statistics and human disease: analytical retooling for complexity. *Trends Genet*, **20**, 640–7.

van den Berg, W.B. & Miossec, P. (2009). IL–17 as a future therapeutic target for rheumatoid arthritis. *Nat Rev Rheumatol*, **5**, 549–53.

Vaquerizas, J.M., Kummerfeld, S.K., Teichmann, S.A. & Luscombe, N.M. (2009). A census of human transcription factors: function, expression and evolution. *Nat Rev Genet*, **10**, 252–63.

Verheijden, G.F., Rijnders, A.W., Bos, E., Coenen– de Roo, C.J., van Staveren, C.J., Miltenburg, A.M., Meijerink, J.H., Elewaut, D., de Keyser, F., Veys, E. & Boots, A.M. (1997). Human cartilage glycoprotein–39 as a candidate autoantigen in rheumatoid arthritis. *Arthritis Rheum*, **40**, 1115–25.

Vischer, E. & Chargaff, E. (1948). Studies on the composition of nucleic acids. *Fed Proc*, **7**, 197.

Visscher, P.M., Hill, W.G. & Wray, N.R. (2008). Heritability in the genomics era—concepts and misconceptions. *Nat Rev Genet*, **9**, 255–66.

Waaler, E. (1940). On the occurrence of a factor in human serum activating the specific agglutination of sheep blood corpuscles. *Acta Pathol Microbiol Scand*, **17**, 172–88.

Wallace, R.B., Shaffer, J., Murphy, R.F., Bonner, J., Hirose, T. & Itakura, K. (1979). Hybridization of synthetic oligodeoxyribonucleotides to phi chi 174 DNA: the effect of single base pair mismatch. *Nucleic Acids Res*, **6**, 3543–57.

Wang, D.G., Fan, J.B., Siao, C.J., Berno, A., Young, P., Sapolsky, R., Ghandour, G., Perkins, N., Winchester, E., Spencer, J., Kruglyak, L., Stein, L., Hsie, L., Topaloglou, T., Hubbell, E., Robinson, E., Mittmann, M., Morris, M.S., Shen, N., Kilburn, D., Rioux, J., Nusbaum, C., Rozen, S., Hudson, T.J., Lipshutz, R., Chee, M. & Lander, E.S. (1998). Large–scale identification, mapping, and genotyping of single–nucleotide polymorphisms in the human genome. *Science*, **280**, 1077–82.

Watson, J.D. & Crick, F.H. (1953). Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature*, **171**, 737–8.

Weissenbach, J., Gyapay, G., Dib, C., Vignal, A., Morissette, J., Millasseau, P., Vaysseix, G. & Lathrop, M. (1992). A second–generation linkage map of the human genome. *Nature*, **359**, 794–801.

Whitacre, C.C. (2001). Sex differences in autoimmune disease. *Nat Immunol*, **2**, 777–80.

Williams, R.O., Feldmann, M. & Maini, R.N. (1992). Anti–tumor necrosis factor ameliorates joint disease in murine collagen–induced arthritis. *Proc Natl Acad Sci U S A*, **89**, 9784–8.

Wilson, A. & Trumpp, A. (2006). Bone–marrow haematopoietic–stem–cell niches. *Nat Rev Immunol*, **6**, 93–106.

Woolley, D.E. (2003). The mast cell in inflammatory arthritis. *N Engl J Med*, **348**, 1709–11.

WTCC (2007). Genome–wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, **447**, 661–78.

Yamamoto, A., Zwarts, L., Callaerts, P., Norga, K., Mackay, T.F. & Anholt, R.R. (2008). Neurogenetic networks for startle–induced locomotion in drosophila melanogaster. *Proc Natl Acad Sci U S A*, **105**, 12393–8.

Yao, Z., Painter, S.L., Fanslow, W.C., Ulrich, D., Macduff, B.M., Spriggs, M.K. & Armitage, R.J. (1995). Human IL–17: a novel cytokine derived from T cells. *J Immunol*, **155**, 5483–6.

Zhang, H.G., Hyde, K., Page, G.P., Brand, J.P., Zhou, J., Yu, S., Allison, D.B., Hsu, H.C., Mountz, J.D., Gallagher, J., Howlin, J., McCarthy, C., Murphy, E.P., Bresnihan, B., FitzGerald, O., Godson, C., Brady, H.R. & Martin, F. (2004). Novel tumor necrosis factor alpha–regulated genes in rheumatoid arthritis: Identification of Naf1/ABIN–1 among TNF–alpha–induced expressed genes in human synoviocytes using oligonucleotide microarrays. *Arthritis Rheum*, **50**, 420–31.

Zinkernagel, R.M. & Doherty, P.C. (1974). Restriction of in vitro T cell–mediated cytotoxicity in lymphocytic choriomeningitis within a syngeneic or semiallogeneic system. *Nature*, **248**, 701–2.