Development and Evaluation of Web Applications for Investigating Candidate Genes in Rat Models of Complex Diseases

Lars Andersson

Department of Cell and Molecular Biology – Genetics
Lundberg Institute, Faculty of Science

2008

UNIVERSITY OF GOTHENBURG

ABSTRACT

Many human diseases, such as rheumatoid arthritis and type 2 diabetes mellitus, have a very complex development, depending on both environmental and multiple genetic factors. By crossing inbred rat strains susceptible to a genetic disorder with strains resistant to the same disorder, genomic regions associated with the disease can be identified, so called quantitative trait loci (QTLs). A QTL region is often rather large, sometimes covering hundreds of genes. To help selecting the most likely causative candidate genes from such QTLs in rat, we have created a publicly available application, called candidate gene capture (CGC).

The CGC application was primarily applied on experimentally induced arthritis QTLs in rat. CGC uses an array of keywords compared to the reference term "arthritis". For each keyword, this results in a keyword score that reflects the percentage of PubMed abstracts containing the keyword that also contain the reference term. OMIM records for human genes localized to human regions homologous to rat QTL regions, are scanned for all keywords. The sum of all matching keyword scores is used to rank candidate genes within each QTL. When evaluated, the CGC application is able to rank candidate genes for arthritis-associated QTLs in a manner very similar to what is done manually.

In a second application, CGC was applied on non-insulin dependant diabetes mellitus QTLs in rat. Here, the number of included keywords was dramatically increased. In the CGC-Diabetes application the user can choose from 25 different reference terms, to which the keywords are compared. The reference terms are selected to represent sub-phenotypes of diabetes so that the user can choose which distinct characteristics to analyze. A "phylogenetic tree" was created to give an overview of how much the gene rankings would differ when different reference terms are used. Just like the CGC-Arthritis application, the CGC-Diabetes application proves to be successful in ranking candidate genes in a manner very similar to what is done manually.

In an extended version of the CGC-Arthritis application, CGC-RefLink, candidate genes identified for a QTL using CGC can be functionally connected to candidate genes in other QTLs via hyperlinks in the respective OMIM records. In a comparative study, CGC-RefLink was applied on arthritis QTLs from two distinct rat crosses. In this way, we were able to find functional connections between genes in QTLs from the two crosses that could contribute to a similar arthritis phenotype.

Finally, using the CGC-Arthritis and the CGC-RefLink applications, we analyzed the localization of candidate genes in the rat genome. We concluded that i) certain QTLs from two different rat crosses harbor a number of genes involved in similar functions, which could be associated to arthritis and ii) candidate genes are randomly distributed between QTL and non-QTL regions.

Keywords: rat, complex disease, rheumatoid arthritis, type 2 diabetes mellitus, QTL, candidate genes, web application, text mining.

ISBN 978-91-628-7647-0

LIST OF PAPERS

This thesis is based on the following papers (I-IV), which will be referred to in the text by their roman numerals:

I. **Andersson L,** Petersen G, Johnson P, Ståhl F: A web tool for finding gene candidates associated with experimentally induced arthritis in the rat. *Arthritis Res Ther* 2005, **7**(3):R485-492.

II. **Andersson L,** Petersen G, Ståhl F: Ranking candidate genes in rat models of type 2 diabetes. *Submitted 2008*

III. **Andersson L,** Ståhl F: CGC-RefLink, a tool for finding functional gene pairs related to experimentally induced arthritis in rats. *Submitted 2008*

IV. **Andersson L,** Ståhl F: Distribution of candidate genes for experimentally induced arthritis in rats. *Submitted 2008*

TABLE OF CONTENTS

ABBREVIATIONS

Aia	Adjuvant-induced arthritis
BB/Dr	Bio-breeding diabetes-resistant (inbred rat strain)
BN	Brown Norway (inbred rat strain)
CGC	Candidate gene capture
Cia	Collagen-induced arthritis
DA	Dark agouti (inbred rat strain)
EIA	Experimentally induced arthritis
F1	First generation of a cross
F2	Second generation of a cross (F1 x F1)
GO	Gene ontology
HSA	Human chromosome (Homo Sapiens)
Mb	Mega base pairs
MeSH	Medical Subject Headings
Niddm	Non-insulin dependent diabetes mellitus
Oia	Oil-induced arthritis
OMIM	Online Mendelian inheritance in man
Pia	Pristane-induced arthritis
QTL	Quantitative trait locus
RA	Rheumatoid arthritis
RNO	Rat chromosome (Rattus Norvegicus)
Scwia	Streptococcal cell wall-induced arthritis
T2D	Type 2 diabetes mellitus

INTRODUCTION

Complex diseases

Genetic diseases that occur due to a single gene mutation and that interfere with a specific function can often be identified by the pattern of inheritance. These diseases are defined as autosomal, x-linked, recessive or dominant and follow a strict Mendelian inheritance pattern.

However, many chronic human diseases do not follow a Mendelian pattern and the onset of the disease is usually influenced by a number of genes as well as environmental factors [1]. Human conditions that exhibit such complexity include rheumatoid arthritis, diabetes mellitus, various cancers, hypertension and many more. The complexity of these conditions makes them much more difficult to study than monogenic diseases. The more genes that are involved, the smaller the contribution of each gene to the disease and the harder they become to detect [2]. In addition, a specific phenotype can develop through the action of different susceptibility genes in different individuals (genetic heterogeneity). To find common genetic aberrations, much effort needs to be made in screening samples from large numbers of patients [3].

Animal models/ Rattus Norvegicus

Animal models are often used to study complex diseases. One advantage in using animal models is that the genetic and environmental heterogeneity, which will always be a complication when studying a human population, can be greatly reduced by controlled breeding and identical habitats. The brown rat (Rattus Norvegicus) is one of the most commonly used animal models for complex human diseases. The brown rat has been bred as a laboratory animal since the 19th century and has been used in a wide spectrum of physiological, psychological and genetic studies [4]. More than 500 inbred rat strains have been developed for a wide range of phenotypes and as models for human diseases [5]. 90% of the genes identified in rat have homologous counterparts in both mouse and human, making it a good model for human genetics [6]. Other advantages in using rat as a model for human diseases include large litter size, short gestation period and a well-studied biology [7]. In addition, compared to smaller rodents, the physiology of rats is more similar to human. The almost complete sequence of the rat genome was released in April 2004, making the rat even more useful as a model of human genetic diseases. Consequently the identification of genes and mutations associated with disease has accelerated in the recent years [8].

QTL analysis

Inbred rat strains, that are susceptible to a condition mimicking a human disease, can be used to map genetic regions to a given phenotype. By crossing rats susceptible to a disease (P1) to rats resistant to the same disease (P2) and then making intercrosses of the offspring (F1xF1), alternatively backcrosses to one of the parents (F1xP), a generation of rats whose genome will be a mix of the two P-animals will be generated. Using genetic markers, genetic regions that are inherited from the two inbred strains can be identified. By statistically linking regions inherited from the susceptible strains to animals displaying characteristics of the disease under current investigation, genomic regions harboring disease-enhancing alleles can be discovered. Such regions are called Quantitative Trait Loci, or QTLs [9].

A common problem with QTL analysis however, is that the genomic regions found to be associated with a disease are rather large, covering hundreds of genes. In addition, a QTL can be identified because it harbors a single gene with a relatively strong effect on a phenotype, or it can be identified due to several genes, each adding a smaller contribution to the studied phenotype. Thus, in most cases it is unknown how many disease-enhancing alleles a given QTL actually contains. For example, the collagen-induced arthritis QTL Cia5 has been shown to comprise at least three loci regulating arthritis severity [10]. To find which gene or genes within a QTL that actually contributes to the phenotype under study is the major challenge. A first step in the search for such causative genes most often includes a time consuming work to find candidate genes.

Sources of gene information

One source of information that can be very useful when deciding whether a gene is a probable candidate gene for a certain disease is the OMIM database (Online Mendelian Inheritance in Man) [11]. OMIM contains textual information for over 12,000 human genes as well as for all known Mendelian disorders. The information is in the form of summaries of scientific papers with a focus on the relationship between genotype and phenotype and with references to the original papers included. Hence, OMIM gives a comprehensive, up to date overview of what is known about the functions of specific genes and associated phenotypes.

Another source of functional information, which is often used in bioinformatics applications, is gene ontology (GO) [12]. GO is a collaborative effort of several databases associated with the GO consortium. Genes in GO are annotated with terms from a controlled vocabulary divided into three hierarchical categories: cellular component, biological process and molecular function. Although not as detailed as OMIM, GO can give a comprehension of the functions of a gene. In addition, the form of a controlled hierarchical vocabulary makes GO a valuable bioinformatics resource.

Available tools for selecting candidate genes

There are a number of different bioinformatics tools available, which are all built to aid the prioritization of candidate genes. A majority of these were developed in the same time period as the candidate gene capture (CGC) application (Paper I). A possible explanation may be that the need for these kinds of tools and the possibility to create them coincided at this time. Some representative examples are listed below.

GeneSeeker searches several databases simultaneously and filters positional candidate genes based on expression and phenotypic data from both human and mouse [13]. The application is best suited for disorders where the expression pattern is expected to be aberrant in the affected tissues. The application presents a list of candidate genes based on the cytogenetic region and expression locations entered by the user. No prioritization is made between the presented suggestions.

POCUS (prioritization of candidate genes using statistics) is a downloadable software that ranks candidate genes within a susceptibility region based on similarity to genes within other susceptibility regions for the same disease [14]. This similarity is measured using GO terms, InterPro domain IDs and expression pattern terms. Genes within two susceptibility regions sharing similarities, where the probability of finding this by chance is less than 5%, are considered good candidates. This method is not that useful if the susceptibility regions analyzed contain many genes however. In addition, if two genes contributing to the same disease are not annotated with the same functional data, they will not be found.

SNPs3D is an application available online that uses keywords weighted against a selected disease to search genome wide for candidate genes using PubMed abstract as the information source [15]. The scoring of keywords in SNPs3D, which was released about a year after CGC, is very similar to ours and is based on the number of abstract associated with a disease that contain the keyword divided by the total number of abstracts containing the keyword. The information used for text mining is PubMed abstracts associated with genes. SNPs3D can also give information on relationships between candidate genes in a graphical interface, based on keywords that are associated with different genes. However, this tool is not designed for searches within restricted chromosomal regions, such as QTLs.

G2D (candidate genes to inherited diseases) prioritizes candidate genes within a genomic region (maximum limit 50 Mb) from one of three starting points: a disease phenotype described as an OMIM identifier, genes associated with the disease or another genomic region associated with the disease [16, 17]. If an OMIM ID is given, the MeSH terms of all articles cited in the OMIM record are used to score similar GO terms. The sequences of the genes within the selected region are then compared to a set of genes that are scored based on the GO terms. A gene with high sequence homology to a gene with a high GO score is considered to be a good candidate gene. If known disease-associated genes are given, the candidate genes are prioritized based on their similarity of GO terms and sequence similarity. If another region associated with the disease is given, all genes with protein-protein interactions with genes within the second region, are suggested as candidate genes. The interactions are collected from the STRING database, which contains known and predicted protein interactions

[18]. The candidate genes are ranked based on the STRING score, which indicates how likely it is to be a true connection.

GFSST (gene function similarity search tool) is a web application that finds genes from searches with a protein identifier [19]. GFFST extracts GO terms associated with the given protein and ranks genes based on similar GO profiles. Alternatively, a set of GO terms can be entered as the search criteria. The search can be performed towards the human or mouse proteome, but searches limited to genomic regions are not implemented.

Endeavour is a downloadable application that prioritizes candidate genes based on similarities to genes known to be associated to a specific phenotype [20]. The similarities are based on multiple information sources, such as PubMed abstracts, GO terms, InterPro protein domains, KEGG pathways, expression data, sequence similarity (BLAST) and more.

Suspects is another web application that ranks candidate genes based on similarity to genes known to be associated with a complex trait [21]. Within a given genomic region, Suspects first collects genes that are likely to be involved in disease based on sequence features. The GO terms, InterPro domains and expression profiles for the genes are then compared to the set of genes that are known to be associated with the investigated disease.

TOM (Transcriptomics of OMIM) is another example of a web application that within a genomic locus prioritizes among candidate genes based on similarities to genes known to be involved in a disease [22]. The similarities are scored based on GO terms and expression profiles. Alternatively, a second locus associated with the disease can be entered, and similarities between genes from the two loci are presented.

BioMercator is a statistical tool that based on QTL maps from different experiments generates consensus loci and hence minimize the number of candidate genes [23]. BioMercator is available on request.

QTL Mixer searches QTLs associated with similar traits from three different species (rat, mouse and human). Homologous genes localized within QTLs in all three species are considered to be the best candidates [24]. At present, the only diseases available for analysis are multiple sclerosis (experimental allergic encephalomyelitis in mouse and rat) and rheumatoid arthritis (collagen-induced arthritis in mouse and rat).

From this overview it can be concluded that a majority of the applications available for candidate gene prioritization are focused on finding candidate genes that share certain characteristics with genes that are already known to be associated with a disease (POCUS, G2D, GFSST, Endeavour and TOM). A complex disorder can however be caused by susceptibility genes involved in very different biochemical pathways, cellular functions and cellular localizations and with totally different DNA sequences. Hence, candidate genes that differ too much in functional annotation, sequence features and expression profile will probably not be selected using a similarity-based prioritization.

A second strategy applied by BioMercator and QTL Mixer, involve methods based on genomic localization. They may both decrease the number of candidate genes, but will not tell much about the remaining genes.

Finally, SNPs3D and our CGC application use a rather similar text mining approach with weighted keywords (Paper I). However, our method is focused on rat QTLs, while SNPs3D makes genome wide searches in human. The textual information used in SNPs3D is PubMed abstracts, while we use PubMed abstracts to score keywords and OMIM as the textual information source. Once again, we like to emphasize that SNPs3D was published about one year after the first publication of CGC (Paper I).

Our choice of using a text mining approach in CGC is based on the belief that written text might be more informative when prioritizing among candidate genes, since a step of human consideration is added to the data. In addition, information extraction from text requires no similarity to other disease-associated genes (connections to well-established disease-associated genes can of course be valuable information in the form of text as well). More specifically, we think that OMIM records is a good choice of information source, since they contain an up to date and comprehensive resume of the knowledge of specific genotype – phenotype interactions.

Rheumatoid arthritis

Rheumatoid arthritis (RA) is one of the most common autoimmune diseases. RA is characterized by chronic inflammation of the joints accompanied by variable degrees of erosive bone loss and cartilage destruction. The prevalence of RA worldwide is about 1% and women are affected about 2.5 times as often as men. Based on twin studies, the genetic contribution to RA is calculated to be around 60%, with the major histocompatibility complex as the predominant contributor [25, 26]. Environmental factors connected to the incidence of RA include smoking, coffee consumption, body mass index and physical activity [27]. To find genes contributing to RA, rat models are often used. Several inbred rat strains have been identified that are susceptible to experimentally induced arthritis, conditions very much resembling human RA. The arthritis phenotype can be induced in these rats with an injection of a number of agents, such as collagen, pristane, oil, streptococcal cell wall or adjuvant [28]. By linkage studies of crosses of such arthritis susceptible strains to arthritis resistant strains, at least 68 experimentally induced arthritis QTLs have been identified in rat [29, 30]. Even though many of these are overlapping and may have been identified due to allelic variations of the same genes, they cover more than 50% of the rat genome [31]. Since each QTL contains at least one disease-enhancing gene, the genetic interactions behind the disease are very complex.

The DA rat

The inbred Dark Agouti (DA) rat strain was established in 1965 [32]. It is susceptible to an array of autoimmune disorders including experimentally induced arthritis [33]. Of the inbred rat strains, DA is considered to be the most arthritis-prone, responding with severe arthritis to a number of inducing agents. DA is also the only known rat strain to respond with arthritis to injection of non-immunogenic oil [34]. In crosses with different arthritis resistant rat strains, over 50 QTLs associated with experimentally induced arthritis have been identified using the DA strain [33, 35-48].

The BB/Dr rat

The Bio-Breeding Diabetes-resistant rat (BB/Dr) is susceptible to several autoimmune diseases, such as experimentally induced insulin dependent diabetes mellitus, autoimmune thyroid disease and collagen-induced arthritis [49-52]. In crosses with the BN strain, at least 14 QTLs associated with experimentally induced arthritis have been identified using this strain [53].

AIMS OF THE STUDY

The overall aim of this PhD project was to develop web applications for investigating candidate genes in rat models of complex diseases. More specifically, we sought to:

- Develop general methods for ranking candidate genes in QTLs for complex diseases in different rat models and implement these methods as publicly available web applications.

- Based on the methods developed, analyze the genetic interactions behind complex diseases in specific rat models.

- Based on the methods developed, investigate the association of candidate genes, gene functions and QTL regions.

MATERIALS AND METHODS

The CGC application is at the center of this work. It was developed in three steps: i) a basic application applied on RA models in rat, ii) an extended application including the possibility to change so called reference terms, applied on a diabetes model in rat, and iii) a complementary application for connecting candidate genes between QTLs, applied on two crosses of collagen-induced arthritis in rat. Finally, the CGC applications were used as a basis for exploring the distribution of candidate genes and connections between specific loci in rat arthritis models. More detailed information on the methods used in this study is available in the corresponding papers.

CGC - Basic application (Paper I)

In a first step the CGC application was designed to rank candidate genes for experimentally induced arthritis QTLs in rat. 37 QTLs for pristane-induced arthritis (Pia), collagen-induced arthritis (Cia), oil-induced arthritis (Oia), adjuvant-induced arthritis (Aia) and streptococcal cell wall-induced arthritis (Scwia) were collected from RatMap [54]. For each QTL, all genes within the homologous human region were selected and included in the web application. The OMIM records for the homologous human genes within each region were made searchable for 49 preset keywords, each with a keyword score reflecting the connection to arthritis. The keyword scores were calculated from the number of PubMed abstracts containing the keyword together with the reference term ”arthritis”, divided by the number of PubMed abstracts containing the keyword alone. This value was then multiplied with 100 to get the final keyword score. The user has the possibility to add up to ten keywords of his own choice, and the keyword score can automatically be calculated. When a candidate gene search is performed for a QTL of interest, the keyword scores for the keywords found within each OMIM record are added, and the candidate genes are ranked based on the sum of the matching keyword scores (the CGC score).

Database construction

All tables included in the application were stored in a MySQL database. The tables were made searchable from the web page through PHP-scripts.

Information on the 37 experimentally induced arthritis QTLs was collected from RatMap [54]. The data included flanking markers defining the region of the QTL, locus information and a short description of each QTL. Human gene information was collected from National Centre for Biotechnology Information (NCBI) [55]. Human gene position data was collected from University of California Santa Cruz (UCSC) [56]. By combining the information from NCBI and UCSC, a MySQL table containing human gene information ordered by codon start position was generated. Rat gene symbols and positions were collected from RatMap and integrated into the human gene table based on gene homologies.

To find human regions homologous to the rat QTLs, an integrated linkage map containing polymorphic markers and rat genes was used. For each QTL, the first rat genes that were found outside the flanking markers and that had homologous human genes with known positions were selected. The region defined by these two flanking human genes were considered to be the homologous region. In many cases however, due to recombinations throughout the evolution, the human homologous regions are comprised of several smaller regions on different chromosomes. To overcome this problem, all rat genes with known human counterparts within each QTL were used to define the final human homologous gene list.

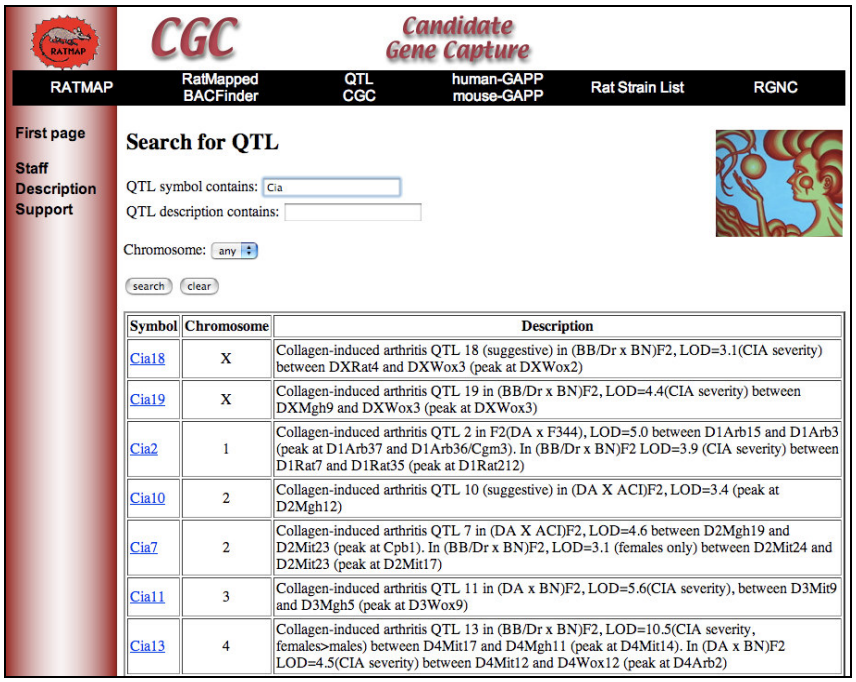
Functional gene information to be used for keyword querying was downloaded from OMIM and stored in a local MySQL database. The 49 keywords were selected from literature describing rheumatoid arthritis and from Medical Subject Headings (MeSH) terms from the PubMed MeSH term database under the headings “autoimmune diseases” and “rheumatoid arthritis”. The keyword scores were calculated as described above and stored in a MySQL table.

Running the application

The application can be found at www.ratmap.org/cgc/arthritis.php. At the first page of the CGC application the user has the option to search for a QTL based on QTL symbol, words in the QTL description or chromosome number. A list of QTLs matching the search criteria is collected from the MySQL database and presented on the same page. The user then selects the QTL of interest from the presented list (Figure 1).

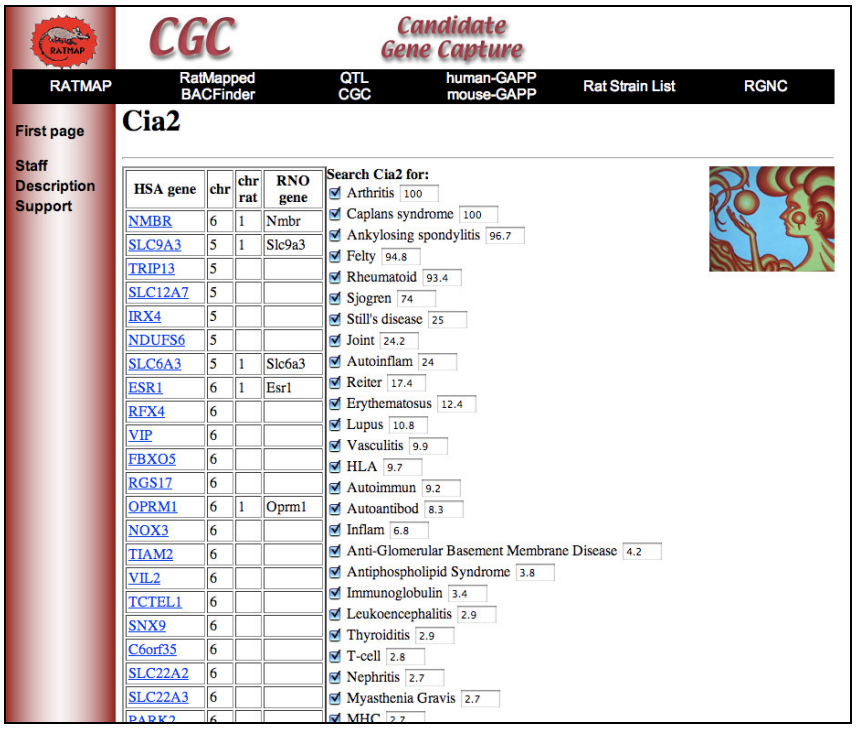
At the resulting (second) page of the CGC application, known rat genes within the selected QTL are presented as well as human genes within the homologous human region (Figure 2). A list of 49 preset keywords are also collected from the database and presented together with their individual keyword scores. The keyword scores can be overwritten and keywords can be deselected. In addition, the user can add up to ten keywords and the keyword scores can automatically be calculated. When the search button is pressed, OMIM records for all human genes presented are scanned for the selected keywords.

The CGC search results in a page where all genes within the selected QTL with at least one matching keyword are presented together with the sum of the keyword scores (the CGC score). Candidate genes with a CGC score of 100 or above are marked in red color and candidate genes with a CGC score between 50 and 100 are marked in yellow color (Figure 3).



Symbol	Chromosome	Description
Cia18	X	Collagen-induced arthritis QTL 18 (suggestive) in (BB/Dr x BN)F2, LOD=3.1(CIA severity) between DXRat4 and DXWox3 (peak at DXWox2)
Cia19	X	Collagen-induced arthritis QTL 19 in (BB/Dr x BN)F2, LOD=4.4(CIA severity) between DXMgh9 and DXWox3 (peak at DXWox3)
Cia2	1	Collagen-induced arthritis QTL 2 in F2(DA x F344), LOD=5.0 between D1Arb15 and D1Arb3 (peak at D1Arb37 and D1Arb36/Cgm3). In (BB/Dr x BN)F2 LOD=3.9 (CIA severity) between D1Rat7 and D1Rat35 (peak at D1Rat212)
Cia10	2	Collagen-induced arthritis QTL 10 (suggestive) in (DA X ACI)F2, LOD=3.4 (peak at D2Mgh12)
Cia7	2	Collagen-induced arthritis QTL 7 in (DA X ACI)F2, LOD=4.6 between D2Mgh19 and D2Mit23 (peak at Cpb1). In (BB/Dr x BN)F2, LOD=3.1 (females only) between D2Mit24 and D2Mit23 (peak at D2Mit17)
Cia11	3	Collagen-induced arthritis QTL 11 in (DA x BN)F2, LOD=5.6(CIA severity), between D3Mit9 and D3Mgh5 (peak at D3Wox9)
Cia13	4	Collagen-induced arthritis QTL 13 in (BB/Dr x BN)F2, LOD=10.5(CIA severity, females>males) between D4Mit17 and D4Mgh11 (peak at D4Mit14). In (DA x BN)F2 LOD=4.5(CIA severity) between D4Mit12 and D4Wox12 (peak at D4Arb2)

Figure 1. Selection of a QTL of interest on the first page of the CGC application.



HSA gene	chr	chr rat	RNO gene
NMBR	6	1	Nmbr
SLC9A3	5	1	Slc9a3
TRIP13	5		
SLC12A7	5		
IRX4	5		
NDUFS6	5		
SLC6A3	5	1	Slc6a3
ESR1	6	1	Esr1
RFX4	6		
VIP	6		
FBXO5	6		
RGS17	6		
OPRM1	6	1	Oprm1
NOX3	6		
TIAM2	6		
VIL2	6		
TCTEL1	6		
SNX9	6		
C6orf35	6		
SLC22A2	6		
SLC22A3	6		

Figure 2. The second page of the CGC application. Rat genes and human genes within the homologous region are presented and keywords are selected.

				search
USF2	19			
FXYD5	19			244.9 AXL rheumatoid, autoimmun, inflam, Tcell, pemphigus, lymphocyte, antigen, cytokine, erythematosus, arthritis, lupus,
FXYD7	19			202.2 CHRNA7 rheumatoid, inflam, cytokine, arthritis,
FXYD3	19			121.3 ESR1 Tcell, nephritis, glomerulonephritis, antigen, arthritis, lupus,
FXYD1	19			308.5 GPI rheumatoid, sjogren, inflam, immunoglobulin, joint, Tcell, hemolytic, arthritis, anemia,
HPN	19	1	Hpn	128.2 TGFB1 autoimmun, inflam, Tcell, lymphocyte, antigen, cytokine, arthritis, infecti, diabetes,
SCN1B	19			210.7 ZFP36 rheumatoid, autoimmun, inflam, arthritis, dermatitis,
ZNF302	19			233.6 VIP rheumatoid, autoimmun, inflam, joint, arthritis,
ZNF345	19			131.8 FCGRT autoimmun, inflam, immunoglobulin, Tcell, mhc, lymphocyte, antigen, cytokine, arthritis,
SPINT2	19			204.6 SIGLEC5 rheumatoid, immunoglobulin, Tcell, lymphocyte, arthritis, infecti, diabetes,
KCNK6	19			128.3 HAMP HLA, inflam, Tcell, mhc, cytokine, arthritis, anemia, infecti, diabetes,
PSMD8	19			54.1 APOE joint, Tcell, antigen, erythematosus, lupus, diabetes, lambert,
LGALS4	19	0	Lgals4	35.8 RYR1 inflam, joint, Tcell, cytokine,
LGALS7	19			24.2 HAS1 joint,
ACTN4	19	1	Actn4	36.6 PEPD immunoglobulin, Tcell, lymphocyte, erythematosus, lupus, anemia, infecti, dermatitis,
MAP4K1	19			25.6 SNRPD2 antigen, erythematosus, lupus,
RYR1	19			10.8 NUCB1 lupus,
ECH1	19	1	Ech1	15.5 LU HLA, immunoglobulin, antigen,
HNRPL	19			11.5 FTL HLA, anemia,
SIRT2	19			10.6 SOD2 HLA, diabetes,
NFKBIB	19			22.3 PLG HLA, inflam, antigen, anemia, infecti,
MRPS12	19			9.7 TCP1 HLA,
PAK4	19			17.1 ACTN4 HLA, lymphocyte, antigen, infecti, diabetes,
MAP3K10	19			18.3 DMPK HLA, immunoglobulin, Tcell, antigen,
AKT2	19	1	Akt2	17.5 CD22 autoimmun, immunoglobulin, lymphocyte, antigen,
PSMC4	19			9.2 FBL autoimmun,
FCGBP	19			11.6 SNRP70 autoimmun, antigen,
FBL	19			21.5 BAX autoimmun, inflam, Tcell, nephritis,
DYRK1B	19			8.4 FPR1 inflam, infecti,
CLC	19			11.6 FPRL1 inflam, Tcell, cytokine,
				9.6 LHB inflam, Tcell,
				9.2 PLAUR inflam, antigen,
				6.8 PTGIR inflam,
				9.6 MAP3K4 inflam, Tcell,
				21.7 TYROBP inflam, immunoglobulin, Tcell, mhc, antigen, cytokine, infecti,

Figure 3. The result of a CGC search. Gene rankings and CGC scores are presented.

CGC – Multiple reference terms (Paper II)

In order to make CGC into a more versatile application that could take more than one single phenotype into consideration, a rebuilt version was made. This version of the application was designed to rank candidate genes for different aspects of non-insulin dependent diabetes mellitus (Niddm) QTLs in rat. Information on 55 Niddm QTLs was collected from RatMap and RGD [30, 54]. A total of 789 keywords related to diabetes were selected. These keywords were chosen from terms frequently found in literature describing different aspects of diabetes as well as from MeSH terms. The keyword value for each keyword was calculated in a similar manner as in the arthritis application described above, but in this application each keyword was compared to 25 different ”reference terms”. The reference terms were selected to reflect different aspects of diabetic syndromes as suggested by experts in the field. The reference terms include words such as ”insulin secretion”, ”pancreas development”, ”insulin resistance” and ”hyperinsulinemia”. More general terms such as ”diabetes” were also included. This gives the user the possibility to use the reference term that best reflects the diabetes model under study. The list of keywords and keyword values are generated according to the selected reference term. Only keywords with a keyword value of 0.1 or greater are included in the list.

A CGC-Diabetes search with several hundreds of keywords might take several minutes to perform however. We therefore created a quick search version containing 28 keywords. These keywords were selected because they were all frequently found in literature describing diabetes. In addition, these terms all received significant keyword scores when compared to the reference term “diabetes”. The 28 keywords were compared to the 25 reference terms in the same way, and the quick version is available as a complement to the CGC-Diabetes application.

Database construction

The CGC-Diabetes application and the accompanying database were constructed in the same way as the CGC-Arthritis application, with some differences. In the CGC-Diabetes application, the human-rat homologous regions are not manually created, but instead generated when a QTL is selected. A PHP script selects the position of the flanking markers from a MySQL table. All rat genes known to be situated between these markers are selected from a database. Human genes homologous to these rat genes are selected, as well as human genes situated within the regions defined by the selected human homologues. These human genes are included in the CGC search.

Running the application

The CGC-Diabetes application can be found at www.ratmap.org/cgc/diabetes.php. The application works in the same way as CGC-Arthritis. On the first page, the user can search for and select a Niddm QTL of interest. In contrast to the CGC-Arthritis application, at the second page the user can select from a list of 25 reference terms (Figure 4). The preset reference term is "diabetes", which is a rather general term, and the user may want to select a more specific term. The keyword scores of 789 keywords are changed to represent co-occurrences with the selected reference term. Only keywords with keyword scores above 0.1 are included in the following CGC search. The resulting list of candidate genes is ordered based on the genes individual CGC scores. The user may select another reference term and then rerun the search with the new keyword scores to retrieve a different gene ranking.

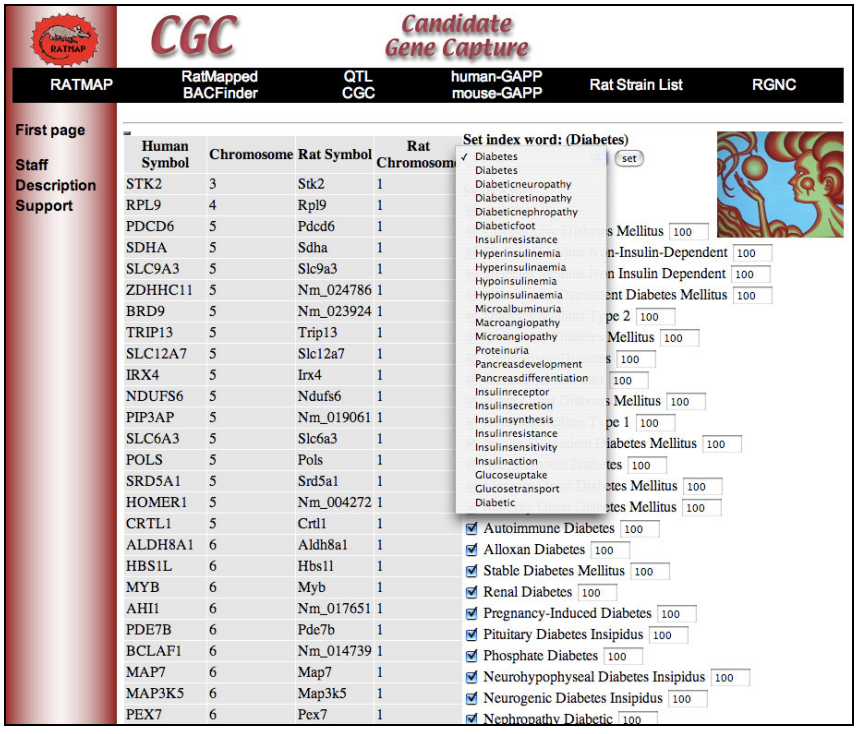


Figure 4. The second page of the CGC–Diabetes application involves the selection of reference terms to which the keyword are to be compared.

Tree phylogeny of reference terms

To evaluate how much the resulting candidate gene rankings from CGC will differ if different reference term are used we compared gene rankings for a single QTL (Niddm46) using the quick CGC version (presented above). Gene rankings were generated for the QTL using each of the 25 different reference terms. The differences in ranking position of ten top candidate genes between searches with each reference term were then compared. (For example, if gene A is ranked 1 when the reference term ”diabetes” is used and ranked 4 when the reference term ”hypoinsulinemia” is used, the difference in candidate gene position for that gene is 3. The sum of the differences of the top ten genes between two reference terms was used as a measure of difference in ranking.)

To give an overview of how much the gene rankings will differ if two different reference terms are used, a graphical tree was created using the FITCH software from Phylip (Phylogeny interface package version 3.66) [57]. The software was developed to create phylogenic trees based on distances calculated from molecular sequences, restriction sites, restriction fragments or gene frequencies. FITCH uses a distance-based optimization created by Fitch & Margoliash [58]. The FITCH software creates a tree with the smallest squared distance between the computed distances and the predictions in the tree. The phylogenies are estimated from a distance matrix under an additive tree model, in which the distances are expected to equal the sums of branch length between the species compared. In our evaluation, we used the differences in gene rankings when using different reference terms in CGC as the distance matrix.

Linking candidate genes between QTLs (Paper III)

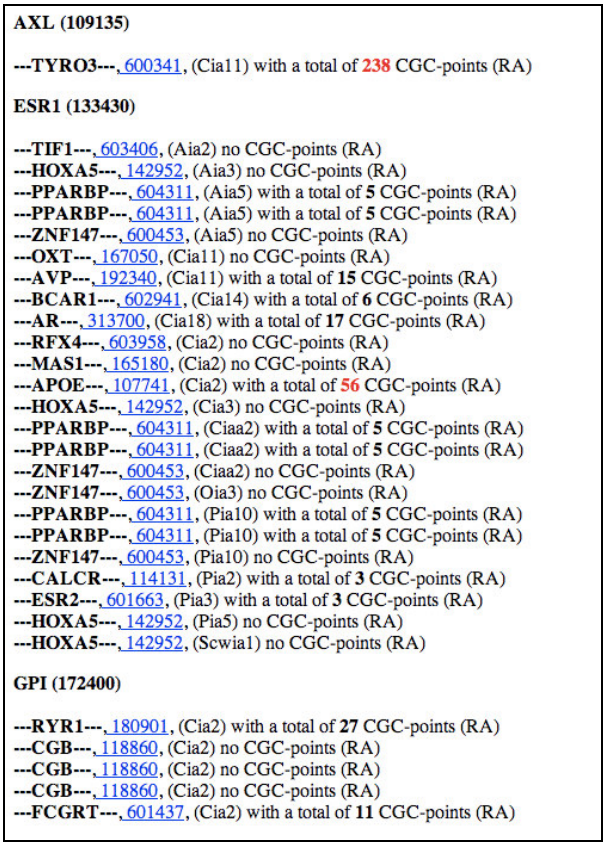
To be able to find possible candidate genes originally ranked low by the CGC application and to find common pathways leading to the same phenotype, but with different genes disrupted in different rat strains, an additional function was added to the arthritis CGC application, called CGC-RefLink. The CGC-RefLink application uses the links within individual OMIM records to other OMIM records to connect different candidate genes. Thus, when a candidate gene ranking has been performed by CGC for one QTL, all genes within all the arthritis QTLs included in the application that are linked to the ranked genes, can be found. The genes that are found to be linked to the selected candidate genes are presented together with individual CGC scores.

Database construction

The CGC-RefLink application is based on a local OMIM database stored in MySQL. The OMIM database was made searchable for hyperlinks containing the six digit OMIM IDs of selected genes via PHP. By running basic CGC searches for all included QTLs, CGC scores were retrieved for all genes, and the results were stored in a MySQL database. These CGC scores are retrieved each time a CGC-RefLink connection is found.

Running the application

The CGC-RefLink application is available at www.ratmap.org/cgc/ra_to_omim.php. To use the CGC-RefLink application, the user first performs a CGC search in the same way as for CGC-Arthritis. When the ranking of candidate genes has been retrieved, the user can choose which genes to search for connections to. The checkboxes for interesting genes are marked and the "compare" button is pressed. The OMIM records for all human genes situated within a region homologous to an arthritis QTL are searched for hyperlinks to the selected candidate genes. The results are presented on a new page (exemplified in Figure 5).



AXL (109135)

---TYRO3---, 600341, (Cia11) with a total of **238** CGC-points (RA)

ESR1 (133430)

---TIF1---, 603406, (Aia2) no CGC-points (RA)
---HOXA5---, 142952, (Aia3) no CGC-points (RA)
---PPARBP---, 604311, (Aia5) with a total of **5** CGC-points (RA)
---PPARBP---, 604311, (Aia5) with a total of **5** CGC-points (RA)
---ZNF147---, 600453, (Aia5) no CGC-points (RA)
---OXT---, 167050, (Cia11) no CGC-points (RA)
---AVP---, 192340, (Cia11) with a total of **15** CGC-points (RA)
---BCAR1---, 602941, (Cia14) with a total of **6** CGC-points (RA)
---AR---, 313700, (Cia18) with a total of **17** CGC-points (RA)
---RFX4---, 603958, (Cia2) no CGC-points (RA)
---MAS1---, 165180, (Cia2) no CGC-points (RA)
---APOE---, 107741, (Cia2) with a total of **56** CGC-points (RA)
---HOXA5---, 142952, (Cia3) no CGC-points (RA)
---PPARBP---, 604311, (Ciaa2) with a total of **5** CGC-points (RA)
---PPARBP---, 604311, (Ciaa2) with a total of **5** CGC-points (RA)
---ZNF147---, 600453, (Ciaa2) no CGC-points (RA)
---ZNF147---, 600453, (Oia3) no CGC-points (RA)
---PPARBP---, 604311, (Pia10) with a total of **5** CGC-points (RA)
---PPARBP---, 604311, (Pia10) with a total of **5** CGC-points (RA)
---ZNF147---, 600453, (Pia10) no CGC-points (RA)
---CALCR---, 114131, (Pia2) with a total of **3** CGC-points (RA)
---ESR2---, 601663, (Pia3) with a total of **3** CGC-points (RA)
---HOXA5---, 142952, (Pia5) no CGC-points (RA)
---HOXA5---, 142952, (Scwia1) no CGC-points (RA)

GPI (172400)

---RYR1---, 180901, (Cia2) with a total of **27** CGC-points (RA)
---CGB---, 118860, (Cia2) no CGC-points (RA)
---CGB---, 118860, (Cia2) no CGC-points (RA)
---CGB---, 118860, (Cia2) no CGC-points (RA)
---FCGRT---, 601437, (Cia2) with a total of **11** CGC-points (RA)

Figure 5. Example of a result from a CGC-RefLink search. *AXL*, *ESR1* and *GPI* were selected from the result page in a CGC search for Cia2 (Figure 3). All genes within any experimentally induced arthritis QTL with an OMIM-hyperlink to the three selected genes are presented together with their CGC scores. These CGC scores were obtained from basic CGC searches for all arthritis QTLs.

Investigating connections between candidate genes in QTLs from two rat crosses

We used the CGC-RefLink application to establish OMIM links between genes within collagen induced arthritis QTLs from two different crosses, DA x BN and BB/Dr x BN. We chose to only include connections between genes where at least one of them had a CGC score of 100 or greater. In addition, only gene connections where one of the genes was localized in a QTL from DA x BN and the other gene was localized in a

QTL from BB/Dr x BN were included. The gene connections were divided into three groups based on the CGC scores of the connected genes: a) Two connected genes with CGC scores of 100 or above. b) One gene with a CGC score less than 100 connected to two genes with CGC scores of 100 or above. c) One gene with a CGC score less than 100 connected to a single gene with a CGC score of 100 or above. All connections in group A and B were manually surveyed to evaluate the functional association between the genes and to arthritis. In group C, connections including genes in QTLs where no connections were found in group A or B were manually evaluated.

Functional connections between QTLs (Paper IV)

To investigate if several genes within specific QTL regions share functions with several genes within other QTLs, we analyzed the 13 QTLs identified in the two crosses DA x BN and BB/Dr x BN. We used the CGC-RefLink application to search OMIM records for all hyperlinks connecting genes from a QTL of the DA x BN cross to a gene in a QTL of the BB/Dr x BN cross. Based on the number of genes within each QTL and on the total number of connections between specific QTLs from the two crosses, we then calculated if certain pairs of QTLs had an overrepresentation of shared gene pairs. This was done with a chi-square test with Yates correction.

The pairs of QTLs that were found to be connected by a significantly high number of gene pairs were then dissected into two groups: a) two QTLs connected by a significantly high number of gene pairs, where at least one gene in each pair received a CGC score of 50 or above, b) two QTLs connected by a significantly high number of genes where both genes in each pair received a CGC score lower than 50.

To investigate if the genes connecting two QTLs share similar functions that could be associated with arthritis, we used the DAVID Gene Functional Classification Tool [59, 60]. This tool generates groups of functionally related genes based on co-occurrences of terms collected from 14 annotation sources, such as GO, Swiss-Prot keywords, KEGG-Pathways and OMIM. For each QTL pair, all genes connecting the two QTLs in either the high CGC score group or the low CGC score group were functionally compared. We recorded functional groups of genes where the geometrical mean of the p-values for the shared terms was lower than 0.05 (Geo). However, since this is a mean of the terms found in each cluster, clusters with higher Geo-values might still contain interesting terms with low p values.

Investigating localization of candidate genes (Paper IV)

To investigate if candidate genes for arthritis in rats identified with the CGC application is more likely to be found within QTL regions than in non-QTL regions, we made a genome wide CGC search. 68 QTL regions associated with experimentally induced arthritis in rat were collected from RGD [30]. The CGC-Arthritis application was modified to rank genes in a genome wide manner for 2403 rat genes with known genomic position collected from RGD and Ensembl [30, 61]. The genes were subsequently divided into groups reflecting their CGC scores. For each group, the numbers of genes localized in QTL regions and in non-QTL regions were recorded.

RESULTS AND DISCUSSION

CGC-Arthritis - Basic application (Paper I)

The CGC application uses keywords and keyword values based on co-occurrences of a keyword and a reference term in PubMed abstracts. Several alternative strategies could be used to put different weights on keywords. For example, hierarchical terms such as GO terms or MeSH terms could be applied, where more specific terms render higher keyword values. This kind of controlled vocabulary however, is dependent on the opinion of experts in the fields, who have decided which terms to be selected and how they should be arranged. Naturally, such an approach limits usable keywords to terms included in the selected vocabulary. Our method on the other hand, makes it possible to include an, in principal, unlimited number of keywords. In the CGC-Diabetes application for example, 789 keywords underwent calculation of keyword values (Paper II). Irrelevant or unspecific keywords will generate very low keyword values and will consequently have neglectable impacts on the final candidate gene ranking.

Gene information source

We chose to use OMIM records as the genetic information source in the CGC application, since written text based on human evaluations, as opposed to more unrefined data, often brings an interpretation that gives more meaning to the data. OMIM records summarizes the most relevant data known for a gene and are as such sufficient as a primary data sources. A more obvious text source would of course be PubMed, which is frequently used for text-based searches [62]. OMIM records are in our opinion a better choice since they are always based on a manual consideration in relation to the function of a single gene. Furthermore, our choice of OMIM as the gene information source, as opposed to, for instance OMIA or similar animal gene resources, was based on the fact that most human genes are described in more detail than genes of non-human origin [63]. In order to use OMIM as the primary text resource, for all QTLs we established the homologous regions between rat and human.

Of course, using human gene data instead of rat data, may be questionable considering the specific gene changes in the rat model used. On the other hand, the rat models included in our application are, after all, models for human diseases, designed to be able to study and understand human diseases. Furthermore, most OMIM records include descriptions of animal studies, making them appropriate also from a rat model point of view.

First evaluation of CGC: Keyword values and gene ranking

To evaluate how good the CGC-application is in ranking candidate genes compared to a human evaluation for the experimentally induced arthritis QTLs tested, we randomly picked four QTLs. The genes within the four QTLs were ranked by the CGC application. Independently, we performed manual evaluations of the OMIM records describing the genes within the four QTLs. In the manual evaluation, genes were scored on a scale from one to five, where five meant that we found no connection to arthritis and one meant that the connection to arthritis was obvious.

To compare the gene rankings obtained from CGC with the manual ratings, the CGC results for each QTL were divided in three separate groups:

Group 1 (the high group) contained the top two ranked genes for each QTL.
Group 2 (the middle group) contained the genes ranked in position 3 – 6.
Group 3 (the low group) contained the rest of the genes in each QTL.

We found that out of the eight highest ranked genes in group 1, seven were manually evaluated as 1 or 2 based on the OMIM records alone. The eighth gene, CD74 on Cia17, was manually evaluated to 3 based on the OMIM records alone. When additional literature was surveyed though, CD74 too turned out to be a very likely candidate gene. The similarity between the manually evaluation and the CGC rankings was consistent in the second and third group as well. In group 2 (genes ranked in position 3–6 by CGC) the average manual rating was 2.7 and in group 3 (genes ranked low by CGC) the average was 3.75 (Table 1).

Table 1. CGC rankings and manual evaluation of genes in three groups and four tested QTLs. The genes were manually rated 1 – 5, where 1 is an obvious candidate gene.

	Best two		Middle group		Low group	
QTL	CGC	Manual	CGC	Manual	CGC	Manual
Cia4	152.6	1.5	10.7	3.5	2.3	3.9
Cia10	128.5	1	14.9	2	3.6	3.9
Cia14	20.4	1	9.5	2.7	4.6	3.4
Cia17	26	2.5	14.4	2.5	4.9	3.8

The CGC-Arthritis application proves to be able to rank candidate genes very similar to what would be done manually, hence we believe that the application should be very useful when selecting likely candidate genes for experimentally induced arthritis QTLs.

CGC-Diabetes – Multiple reference terms (Paper II)

Many clinical diagnoses such as type 2 diabetes mellitus and rheumatoid arthritis comprise a multitude of sub-phenotypes, and the disease might differ from patient to patient. Hence, the genetics involved in the development of the disease will differ between individuals. Because of this discrepancy we wanted to offer a tool where the user can choose from an array of sub-phenotypes when ranking genes. This second version of the CGC application was implemented on non-insulin dependent diabetes mellitus QTLs in rat. 25 different reference terms that reflect different aspects of type 2 diabetes and a total of 789 keywords were included. The keyword values of each keyword were individually calculated based on co-occurrence frequencies with each of the reference terms.

Tree phylogeny of reference terms

The respective values of the 789 keywords will differ greatly depending on which reference term that is used. For example 330 terms are found with a score of 0.1 or higher when the reference term ”diabetes” is used, while only 24 are found with the term ”diabetic foot”.

When we compared the outcome of queries using different reference terms we found a subset of keywords that were frequently found in the query results and that made substantial contributions to the final CGC score. Taking advantage of this circumstance, we also made a quick version of the diabetes CGC application, containing 28 keywords that are frequently found in diabetes literature. Keyword values for these 28 keywords were calculated using all 25 reference terms. For most queries, this quick version generates more or less the same rankings as the full CCG application, but much faster. This also indicates that the 789 keywords in the full CGC application should be more than sufficient to cover most aspects of type 2 diabetes mellitus.

A multitude of reference terms can make it hard to select which one or which ones to use. To get an overview of how much the gene rankings will differ if different reference terms are used, we calculated the differences in gene position for the top ten ranked genes in a single QTL. The sum of the differences in gene rankings between every combination of reference terms were used as a distance matrix to construct a phylogenetic tree using the FITCH software from Phylip [57].

In the tree, the lengths of the horizontal branches between two reference terms represent the difference in gene rankings when using the two terms (Figure 5). For example, the reference terms “glucose uptake” and “glucose transport” are situated very closely, indicating that CGC searches with keywords compared against these two terms would generate very similar candidate gene rankings. In the same way there are a cluster of five reference terms, all including “insulin”, that are situated in the vicinity of each other in the tree (“hyperinsulinemia”, “hyperinsulinaemia”, insulin sensitivity”, “insulin resistance” and “insulin action”). Thus, it seems as if CGC searches based on reference terms that are functionally related will generate similar gene rankings. On the other hand, terms that are localized far apart and thus rank genes very differently, also seems to be less functionally related.

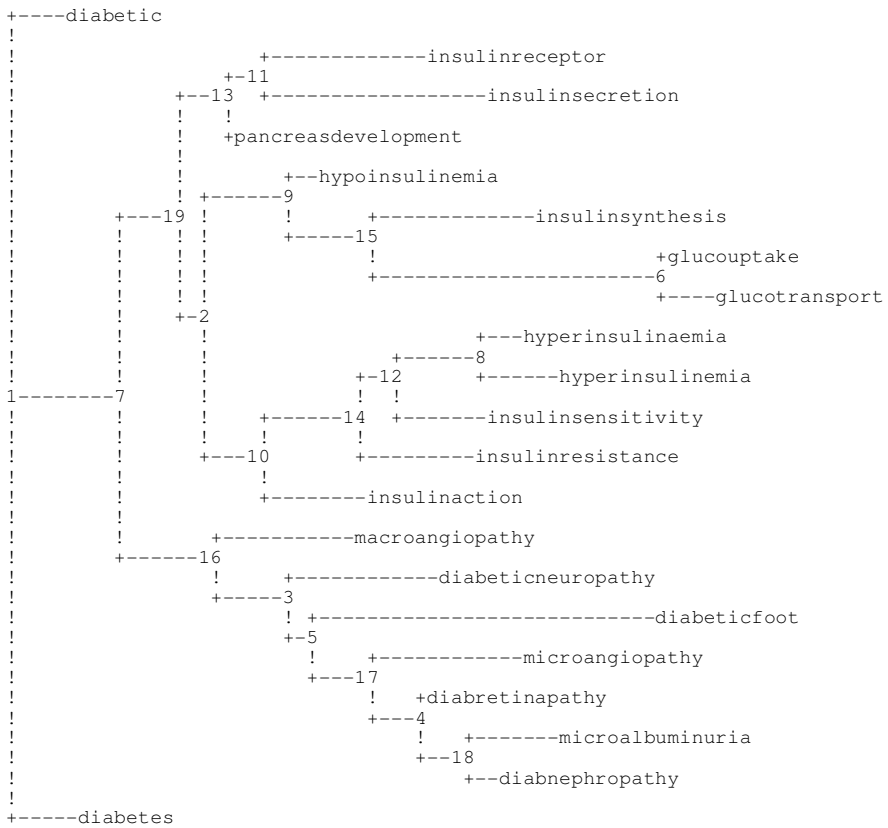


Figure 6. The "phylogenetic tree", based on differences in gene rankings when using different reference terms.

Second Evaluation of CGC: Multiple keywords and keyword selection

To evaluate how successful the diabetes CGC application is in ranking genes for non-insulin dependent diabetes QTLs, we made a manual candidate gene rating for four randomly selected QTLs in a manner similar to what was done for the arthritis CGC application (Paper 1 – see above). Candidate gene rankings was obtained from CGC using the reference term "diabetes". This term generated the longest list of significant keywords and was considered to be the most general reference term.

In summary, when comparing the manual candidate gene ratings and the CGC gene rankings for the four QTLs analyzed, out of the 21 genes with a CGC score of 100 or greater, 52% were manually rated as "obvious" candidates (manual rating 1), 38% were manually rated as "likely" candidates (rating 2) and 10% were manually rated as "possible" candidates (rating 3). No candidate genes with a CGC score of 100 or greater was manually rated as "unlikely" (rating 4) or "irrelevant" (rating 5). Of the genes with a CGC score less than 100, no obvious candidate (rating 1) was manually found. 16% were manually rated as "likely" candidates (rating 2), 14% were rated as "possible" candidates (rating 3), 34% were rated as "unlikely" candidates (rating 4) and 36% were rated as "irrelevant" (rating 5).

Table 2. Genes manually evaluated as likely candidate genes for Niddm on a scale from 1 to 5 as compared with the rankings obtained from CGC-Diabetes (given in percentage values).

	CGC >100	CGC <100
1. Obvious	52	0
2. Likely	38	16
3. Possible	10	14
4. Unlikely	0	34
5. Irrelevant	0	36

To conclude the observations from our evaluation of the CGC-Diabetes application, it obviously ranks candidate genes very similar to what would be done manually. Thus, the application should be of great use when evaluating which genes are the most likely candidate genes within a region associated with diabetes.

CGC-RefLink – Connecting candidate genes (Paper III)

Our evaluation of the original CGC application shows that it succeeds very well in ranking gene candidates. However, in many cases there may be several genes in a single QTL that all have a high CGC score, which can make it hard to select the most likely candidate genes. In order to discriminate between different highly rated candidate genes we reasoned that a second selection criterion was needed. Preferably, this selection criterion should be independent and not associated with our weighted keywords and reference terms. Still using the same information source, we found that one such independent criterion could be the citations between different OMIM records. Each time another gene is mentioned in an OMIM gene record, a hyperlink to the OMIM record of that gene is included. By scanning OMIM records for such links to other OMIM records, genes involved in the same functions, or that affect each other in one way or another can be found. This method also makes it possible to re-evaluate previously low-ranked genes if they are linked to genes that are highly rated as a candidate genes for the investigated phenotype.

To test if this method was able to present a clear view of how gene mutations in QTL regions of a given strain may affect the phenotype, we selected thirteen collagen induced arthritis QTL regions discovered with the two arthritis susceptible rat strains DA and BB/Dr crossed to the arthritis resistant rat strain BN. 9 QTLs were identified in F2 animals of BB/Dr x BN crosses and 5 QTLs were identified in F2 animals of DA x BN. Of these QTLs, only one (Cia13) was found in both variants of crosses. This indicates that the genes contributing to the collagen induced arthritis phenotype in offspring of the two variants of crosses are different. Since the resulting condition is the same in affected F2 animals of the two crosses we hypothesize that genes involved in the same pathways are disrupted in the two variants. Hence, a gene variant contributing to the phenotype within a QTL from one of the rat crosses corresponds to a variant of another gene within a QTL from the other rat cross.

In the study, only QTLs from the two susceptible rat strains which both had been crossed to the same resistant strain were included. Thus, the genetic differences in the models would all come from DA and BB/Dr. In addition, for the QTLs selected, the

arthritis phenotype was induced by collagen in both crosses to limit environmental differences. To find corresponding gene pairs between genes within QTLs from DA x BN and BB/Dr x BN, all 13 QTLs were scanned for candidate genes using the arthritis CGC application described above. 34 genes with a CGC score of 100 or greater were found within the selected QTLs. These highly rated genes were used as the basis for establishing gene pairs. The OMIM records for these genes were scanned for references to OMIM records for any gene situated within a QTL from the opposite cross. In the same way, all OMIM records for genes within any of the included QTLs were scanned for a reference to any of the highly rated genes of the opposite cross. In this way, 15 of the highly rated genes were found to be linked to 38 genes with a CGC score less than 100. In addition, four interconnections between highly rated genes only were found.

The results of this study were divided into three groups:

Group A: Two connected genes, one from each cross, where both has a CGC score of 100 or above.
Group B: One gene with a CGC score less than 100 from one cross connected to two genes with CGC scores of 100 or greater from the other cross.
Group C: One gene with a CGC score less than 100 from one cross connected to one gene with a CGC score of 100 or greater from the other cross.

Four gene pairs were found to match the criteria for group A. Since both genes in each pair are already rated as good candidate genes for arthritis by the CGC application they here become confirmed as top candidates. In addition, since our hypothesis is that more or less the same functions should be disrupted in the two variants of rat crosses, each gene pair should have a function in common and that function should be related to arthritis. When analyzing the gene pairs in group A, it is possible to find at least four such key functions:

CD44 (Cia11) -*TGFB1* (Cia2) - Decreased downregulation of inflammation [64].
CD69 (Cia13) -*TGFB1* (Cia2) - Decreased downregulation of inflammation [65].
CD44 (Cia11) -*TNF* (Cia1) - Increased bone loss [66].
AXL (Cia2) -*TYRO3* (Cia11) - Increased longevity of inflammation [67, 68].
AXL (Cia2) -*TYRO3* (Cia11) - Decreased self tolerance [69].

In group B, we found four groups of genes where one of the genes from one of the crosses had a CGC score less than 100, but had connections to two genes from the other cross with a CGC score of 100 or above. Although the connections to arthritis were not as striking as in group A, all four genes, that were previously rated rather low by the CGC application, were all re-evaluated as very likely candidate genes for collagen induced arthritis when their connections to the highly rated genes were studied.

Some possible dysfunctions contributing to the arthritis phenotypes could be concluded in this group as well:

TAP1 (Cia1) – *B2M* (Cia11) – *FCGRT* (Cia2)
- Impaired MHCI presentation (TAP1-B2M) and misregulation of IgG-levels (*B2M* - *FCGRT*) [70-72].

TNF (Cia1) – *TNFRSF1A* (Cia13) – *LTBR* (Cia13)
- Misregulation of TNF-induced immune regulators (*TNF* - *TNFRSF1A*) [73, 74].

TNF (Cia1) - *DDX11*(Cia13) – *TGFB1* (Cia2)
- Reduced telomeric length (*TNF* - *DDX11* - *TGFB1*) [75, 76].

CCR5 (Cia6) – *CD4* (Cia13) – *CD69* (Cia13)
- Misregulation of T-cell migration (*CCR5* - *CD4*) [77-79].

In group C, 38 gene pairs were found, where one gene from one of the crosses had a CGC score of 100 or greater and the other gene from the other cross had a CGC score less than 100. We chose to analyze gene pairs from this group so that every QTL included in this study would have at least one candidate gene with a link to a candidate gene for a QTL from the opposite cross. In other words, QTLs with no functional gene pairs in group A or B were further analyzed in group C.

For each of the QTLs Cia12, Cia7, Cia6, Cia17 and Cia19 only one gene pair was found to match the criteria for group C. For the QTLs Cia5ab and Cia14, three gene pairs each were found to match group C. In these cases, the gene pairs manually considered to contribute to an arthritic phenotype are presented below. For the two remaining QTLs, Cia16 and Cia18, no gene pairs were found matching any of the groups.

ICAM2 (Cia5ab) – *TNF* (Cia1)	-Misregulated apoptosis of B-cells [80].
ELN (Cia12) – *TNXB* (Cia1)	-Abnormal elastin bodies [81].
HMOX1 (Cia14) – *TNF* (Cia1)	-Decreased downregulation of inflammation [82].
IL21 (Cia7) – *CD44* (Cia11)	-Decreased downregulation of inflammation and increased bone loss [66, 83, 84].
MBD2 (Cia17) – *CHD4* (Cia13)	-Connection to dermatomyositis [85, 86].
PHEX (Cia19) – *FGF23* (Cia13)	-Connection to hypophosphatemia and osteoarthritis [87, 88].

No clear connection to rheumatoid arthritis could be concluded for the gene pairs found for Cia17 and Cia19. It should be noted that several genes with very high CGC scores are present in some of the QTLs above, although no connections to genes within QTLs from the opposite cross were found. For instance, in Cia12 we find the highly rated gene NCF1 (CGC score 240,9) but with no connections to genes from the opposite cross. It has been shown by positional cloning that the Pia4 QTL, which overlaps with Cia12, is a polymorphism in Ncf1 [89]. However, such highly rated genes are, of course of great interest, but in the context of interconnected gene pairs they cannot be considered as candidate genes.

To conclude our findings, by combining the CGC ranking with citations between OMIM records we are able to suggest candidate genes for most of the DA x BN and BB/Dr x BN QTLs where a gene is found within a QTL from the opposite cross with a similar function. In these gene pairs, a mutated form of one of the genes in one of the crosses are hypothesized to correspond to a mutation of the other gene in the other cross, which would make the similar phenotypical outcome.

Functional connections between QTL regions (Paper IV)

Since some of our data indicate that a majority of QTLs contain several genes that contribute to a phenotype under study, we investigated if it is possible to find genes within specific QTL regions that share functions with genes within other specific QTL regions. Once again, we analyzed the 13 experimentally induced arthritis QTLs identified in the DA x BN and BB/Dr x BN crosses. We based this study on the same assumption as earlier, that since animals from the two crosses develop similar phenotypes but with different loci involved, they should have the same biochemical or cellular functions disrupted, but with different genes mutated or disregulated.

We used the CGC-RefLink application to find connections between genes localized in QTLs identified in the two respective rat crosses. Pairs of QTLs that were linked by a significantly high number of gene pairs were considered to be functionally connected.

Four pairs of QTLs were found with significantly high numbers of gene connections. One pair with a p value of 0.054, Cia13 – Cia1, was also included in the further dissection (Table 4). The gene connections for these pairs of QTLs were divided into connections where at least one of the genes had a CGC score > 50 and connections where both genes had a CGC score < 50. All gene pairs connecting the five pairs of QTLs are presented in Table 5. It turned out that four of the connected QTL pairs had significantly high numbers of connections involving high-ranked candidate genes (Cia11 – Cia1, Cia11 – Cia2, Cia13 – Cia1 and Cia5ab – Cia1). Cia11 – Cia2 were found to have significantly high numbers of gene connections in both groups while only one pair of the connected QTLs had significant connections in the low CGC score group (Cia14 – Cia7). Thus, six groups of gene pairs were defined. In all significant connections between the QTLs studied except one, candidate genes with a relatively high CGC score are involved. When investigating Cia14 and Cia7, which were connected by genes with CGC scores lower than 50, there are in fact no genes with higher CGC scores. The genes with the highest CGC scores within these regions (IL15 - 27.3 CGCp, IL21 – 25.7 CGCp and IL2 – 21.4 CGCp) are all involved in the connections between Cia14 and Cia7.

Table 3. Pairs of QTLs connected by a significantly large number of gene pairs.

QTL1	Number of genes in QTL1	QTL2	Number of genes in QTL2	Number of gene pairs	P value (Yates)
Cia11	58	Cia1	136	10	0.0017
Cia11	58	Cia2	260	23	7.00E-06
Cia13	135	Cia1	136	18	0.054
Cia5ab	25	Cia1	136	6	0.018
Cia14	53	Cia7	19	4	0.0024

Table 4. Gene pairs connecting QTLs in a significant manner. Gene pairs where at least one of the genes received a CGC score of 50 or above are shaded in the table.

Cia11			Cia1	
CD44	233.7	-	TNF	383.4
PCNA	2.8	-	CDKN1A	242.4
TP53BP1	5.3	-	CDKN1A	242.4
PLA2R1	8.8	-	TNF	383.4
B2M	23.5	-	TAP1	117.7
B2M	23.5	-	TAPBP	30.4
CDC25B	0	-	MAPK14	24.6
WT1	7.7	-	BCR	2.8
B2M	23.5	-	CSTB	0
GCG	12.5	-	GLP1R	0

Cia11			Cia2	
TYRO3	244.9	-	AXL	244.9
AVP	15.3	-	ESR1	121.3
OXT	0	-	ESR1	121.3
CD44	233.7	-	TGFB1	128.2
ITGB6	6.8	-	TGFB1	128.2
ACVR1	0	-	TGFB1	128.2
GCG	12.5	-	VIP	233.6
B2M	23.5	-	FCGRT	131.8
CD44	233.7	-	HAS1	24.2
CD44	233.7	-	VIL2	9.7
PRNP	25	-	SOD2	10.6
PRNP	25	-	PLG	22.3
BDNF	9.6	-	PLG	22.3
SN	15.8	-	CD22	17.5
PRNP	25	-	BAX	21.5
PRNP	25	-	FPRL1	11.6
CD59	12.5	-	PLAUR	9.2
CKMT1	0	-	CKM	6.2
SN	15.8	-	SIGLEC8	5.8
TBR1	0	-	T	5.5
SN	15.8	-	CD33	2.4
PCNA	2.8	-	POLD1	2.4
BDNF	9.6	-	GRIN2D	0

Table 4. (Continued) Gene pairs connecting QTLs in a significant manner. Gene pairs where at least one of the genes received a CGC score of 50 or above are shaded in the table.

Cia13			**Cia1**	
CDKN1B	10.6	-	CDKN1A	242.4
TNFRSF1A	55.9	-	TNF	383.4
LAG3	13.8	-	TNF	383.4
TNFRSF7	12.7	-	TNF	383.4
SIAT8A	2.8	-	TNF	383.4
OLR1	5.8	-	TNF	383.4
DDX11	2	-	TNF	383.4
VWF	34.8	-	CYP21A2	50
LTBR	107.5	-	LTB	2.5
TNFRSF1A	55.9	-	LTB	2.5
A2M	5.8	-	C4B	43.3
AICDA	9.5	-	PIM1	18.4
PPARG	21.2	-	RXRB	15.2
PPARG	21.2	-	PPARD	12.7
GPR19	0	-	DRD1	2.8
VWF	34.8	-	BCR	2.8
PTPN6	14	-	G6B	0
ETV6	4.6	-	TEL2	0
Cia5ab			**Cia1**	
ACE	3.6	-	TNF	383.4
ICAM2	10.2	-	TNF	383.4
GH1	4.8	-	TNF	383.4
CD79B	8.3	-	PIM1	18.4
MAP2K6	0	-	MAPK14	24.6
FALZ	0	-	HMGIY	6.2
Cia14			**Cia7**	
IL15	27.3	-	IL2	21.4
NFATC3	11.3	-	IL2	21.4
IL15	27.3	-	IL21	25.7
GLG1	2.8	-	FGF2	3.4

When investigating the functional gene clusters obtained from the functional classification application DAVID [60] for the QTL pairs, we find significant clusters containing a majority of the genes for five out of six groups of genes, indicating that there is indeed a functional connection between each identified QTL pair. When inspecting the actual annotation terms associating the genes, functions that may be involved in the development of arthritis can be identified for four out of five QTL pairs.

A pair of QTLs sharing several gene pairs that also have functions related to arthritis, may suggest that these QTLs harbor several genes contributing to the phenotype, as opposed to one gene with a strong phenotypic effect. In addition, the functions shared by a pair of QTLs from the two different rat crosses might be disrupted in animals from both crosses, and hence explain how they can develop similar arthritis phenotypes with different loci involved.

Localization of candidate genes (Paper IV)

Since the arthrits QTLs in rat cover more than 50% of the genome [31], it seems reasonable to believe that a majority of the genes that are associated with arthritis in rat would be localized within these regions. To explore this notion we took the CGC-Arthritis application and modified it to rank candidate genes for arthritis in rat in a genome wide manner.

In total, out of the 2403 genes, 1160 (48%) were found to be localized within experimentally induced arthritis QTLs and 1243 (52%) were localized in non-QTL regions. When analyzing genes with respect to their CGC score, no correlation between highly ranked candidate genes and their position within QTL regions can be made. Thus, to our surprise, the distribution of genes with different CGC scores inside and outside of QTL regions seems to be completely random. The only small deviation is that 57% of the genes with a CGC score of 100 or above are localized in non-QTL regions.

Table 5. Localization of candidate genes within different ranges of CGC scores within QTL regions and non-QTL regions.

CGC score	Genes within QTLs	Genes outside of QTLs
200 -	23	22
100 -200	23	38
50 -100	12	13
30 -50	18	16
20 -30	59	54
10 –20	138	154
0 -10	887	946
Total	1160	1243

These results are contradictory to the findings presented by Xiong and coworkers, who report that 124 of 185 genes identified as RA associated genes are localized within experimentally induced arthritis-QTL regions [31]. Xiong and coworkers used a method to find RA associated genes similar to ours but with some crucial differences. They searched for combinations of gene symbols and the term ”arthritis” within OMIM and PubMed records. Then they turned to a manual evaluation of the references for the primary candidate genes to select RA associated genes. This is in contrast to our method, which is based on automatic selections. We believe that our choice of an automatic procedure is to prefer in this kind of evaluation, since a manual selection of candidate genes may bring a bias towards established candidate genes situated within QTLs.

Our conclusion that approximately 50 percent of candidate genes for arthritis genome wide in rat are localized within QTL regions, which do in fact cover half of the rat genome, indicates that there are still many genes with a capacity to contribute to an arthritis phenotype that have not yet been detected through QTL analyses. This is not too surprising however, since a very limited number of rat strains susceptible to arthritis have been used in these studies. Each inbred rat strain has a unique allelic combination, and the experimentally induced arthritis studied using these models are probably just a small subset of the genetic combinations that could lead to an arthritis phenotype. From this reasoning a hypothetical consequence could be that with enough rat models of arthritis, more or less the whole rat genome would be covered by QTLs for this disease. In some sense, this seems contradictory since the purpose of QTL analysis is to limit the number of possible candidate genes. But this may very well be the case if many different candidate genes in different QTLs interfere with the same cellular functions. In this perspective, a rat model would be nothing but a specific case of gene disruptions in a limited set of biochemical pathways.

CONCLUDING REMARKS

This thesis focuses on the development and evaluation of web based tools for finding candidate genes for QTLs associated with complex diseases in rat. More specifically, the thesis describes:

- The construction of a web application that can make reliable selections of candidate genes in QTLs associated with experimentally induced arthritis in rat (CGC-Arthritis).

- The construction of a web application that can make reliable selections of candidate genes for Niddm QTLs in rat. This application can be adjusted to select candidate genes for a multitude of sub-phenotypes of diabetes (CGC-Diabetes).

- The development of a method for identifying functional connections between candidate genes in different QTLs associated with the same phenotype (CGC-RefLink). Based on this method, we could find functional gene connections between collagen-induced arthritis QTLs that might explain how rats from two different crosses can develop similar arthritis phenotypes, but with different loci involved.

- An investigation of how functional gene pairs identified using CGC-RefLink are localized to specific arthritis QTLs. We found five pairs of QTLs from two rat crosses that share a significantly large number of functional gene connections. The gene pairs for each connected pair of QTLs were found to share functions, which could be related to arthritis.

- An investigation of how candidate genes identified using the CGC-Arthritis application are distributed in the rat genome. Based on this study, we conclude that these genes are not more frequently found in QTL regions than in non-QTL regions.

ACKNOWLEDGEMENTS

I wish to express my sincere gratitude to all people who in one way or another have contributed to this thesis. I especially want to thank:

Fredrik Ståhl, my supervisor, who introduced me to the fields of genetics and bioinformatics. I am grateful for all your encouragement and support. Thank you for all the fruitful discussions and help when preparing the manuscripts. There has also been some great fun during the years.

Barbara Gawronska, my co-supervisor, for good collaboration.

Dan Röhme, for inspirational and interesting discussions, both of scientific and non-scientific nature.

Greta Petersen, for being a great friend, for support and collaboration and for all the fun we have shared.

Sandra Karlsson for the good times shared on our study camps at Od.

Nancy Ilenius, for always being helpful and for proofreading.

Agneta Proft for being a great person and an enthusiastic teacher.

Per Johnson and **Pedro Gomez-Fabre** for getting me started with the database construction and programming.

Past and present people at the Department of Genetics: Profossor **Göran Levan, Karin Klinga-Levan**, **Afrouz Behboudi**, **Emma Samuelson, Carola Nordlander, Tatjana Adamovic, Fredrik Trossö, Emma Lü, Leyla Roshani** and **Ahmad Hamta** for being nice guys and creating a nice atmosphere.

My family and all my friends for all the support and inspiration.

Last but not least, **Camilla**, for all your love, care and support, especially during these last weeks!

REFERENCES

1. Motulsky AG: **Genetics of complex diseases**. *J Zhejiang Univ Sci B* 2006, **7**(2):167-168.
2. Konig IR, Schafer H, Muller HH, Ziegler A: **Optimized group sequential study designs for tests of genetic linkage and association in complex diseases**. *Am J Hum Genet* 2001, **69**(3):590-600.
3. Farrall M, Morris AP: **Gearing up for genome-wide gene-association studies**. *Hum Mol Genet* 2005, **14 Spec No. 2**:R157-162.
4. Jacob HJ: **Functional genomics and rat models**. *Genome Res* 1999, **9**(11):1013-1016.
5. Aitman TJ, Critser JK, Cuppen E, Dominiczak A, Fernandez-Suarez XM, Flint J, Gauguier D, Geurts AM, Gould M, Harris PC *et al*: **Progress and prospects in rat genetics: a community view**. *Nat Genet* 2008, **40**(5):516-522.
6. Lindblad-Toh K: **Genome sequencing: three's company**. *Nature* 2004, **428**(6982):475-476.
7. Gill TJ, 3rd, Smith GJ, Wissler RW, Kunz HW: **The rat as an experimental animal**. *Science* 1989, **245**(4915):269-276.
8. Gibbs RA, Weinstock GM, Metzker ML, Muzny DM, Sodergren EJ, Scherer S, Scott G, Steffen D, Worley KC, Burch PE *et al*: **Genome sequence of the Brown Norway rat yields insights into mammalian evolution**. *Nature* 2004, **428**(6982):493-521.
9. Lander ES, Schork NJ: **Genetic dissection of complex traits**. *Science* 1994, **265**(5181):2037-2048.
10. Brenner M, Meng HC, Yarlett NC, Joe B, Griffiths MM, Remmers EF, Wilder RL, Gulko PS: **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* 2005, **174**(12):7894-7903.
11. Amberger J, Bocchini CA, Scott AF, Hamosh A: **McKusick's Online Mendelian Inheritance in Man (OMIM(R))**. *Nucleic Acids Res* 2008.
12. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT *et al*: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium**. *Nat Genet* 2000, **25**(1):25-29.
13. van Driel MA, Cuelenaere K, Kemmeren PP, Leunissen JA, Brunner HG, Vriend G: **GeneSeeker: extraction and integration of human disease-related information from web-based genetic databases**. *Nucleic Acids Res* 2005, **33**(Web Server issue):W758-761.
14. Turner FS, Clutterbuck DR, Semple CA: **POCUS: mining genomic sequence annotation to predict disease genes**. *Genome Biol* 2003, **4**(11):R75.
15. Yue P, Melamud E, Moult J: **SNPs3D: candidate gene and SNP selection for association studies**. *BMC Bioinformatics* 2006, **7**:166.
16. Perez-Iratxeta C, Bork P, Andrade MA: **Association of genes to genetically inherited diseases using data mining**. *Nat Genet* 2002, **31**(3):316-319.

17. Perez-Iratxeta C, Bork P, Andrade-Navarro MA: **Update of the G2D tool for prioritization of gene candidates to inherited diseases**. *Nucleic Acids Res* 2007, **35**(Web Server issue):W212-216.
18. von Mering C, Jensen LJ, Kuhn M, Chaffron S, Doerks T, Kruger B, Snel B, Bork P: **STRING 7--recent developments in the integration and prediction of protein interactions**. *Nucleic Acids Res* 2007, **35**(Database issue):D358-362.
19. Zhang P, Zhang J, Sheng H, Russo JJ, Osborne B, Buetow K: **Gene functional similarity search tool (GFSST)**. *BMC Bioinformatics* 2006, **7**:135.
20. Aerts S, Lambrechts D, Maity S, Van Loo P, Coessens B, De Smet F, Tranchevent LC, De Moor B, Marynen P, Hassan B *et al*: **Gene prioritization through genomic data fusion**. *Nat Biotechnol* 2006, **24**(5):537-544.
21. Adie EA, Adams RR, Evans KL, Porteous DJ, Pickard BS: **SUSPECTS: enabling fast and effective prioritization of positional candidates**. *Bioinformatics* 2006, **22**(6):773-774.
22. Rossi S, Masotti D, Nardini C, Bonora E, Romeo G, Macii E, Benini L, Volinia S: **TOM: a web-based integrated approach for identification of candidate disease genes**. *Nucleic Acids Res* 2006, **34**(Web Server issue):W285-292.
23. Arcade A, Labourdette A, Falque M, Mangin B, Chardon F, Charcosset A, Joets J: **BioMercator: integrating genetic maps and QTL towards discovery of candidate genes**. *Bioinformatics* 2004, **20**(14):2324-2326.
24. Serrano-Fernandez P, Ibrahim SM, Koczan D, Zettl UK, Moller S: **In silico fine-mapping: narrowing disease-associated loci by intergenomics**. *Bioinformatics* 2005, **21**(8):1737-1738.
25. Deighton CM, Walker DJ, Griffiths ID, Roberts DF: **The contribution of HLA to rheumatoid arthritis**. *Clin Genet* 1989, **36**(3):178-182.
26. MacGregor AJ, Snieder H, Rigby AS, Koskenvuo M, Kaprio J, Aho K, Silman AJ: **Characterizing the quantitative genetic contribution to rheumatoid arthritis using data from twins**. *Arthritis Rheum* 2000, **43**(1):30-37.
27. Pedersen M, Jacobsen S, Klarlund M, Pedersen BV, Wiik A, Wohlfahrt J, Frisch M: **Environmental risk factors differ between rheumatoid arthritis with and without auto-antibodies against cyclic citrullinated peptides**. *Arthritis Res Ther* 2006, **8**(4):R133.
28. Joe B: **Quest for arthritis-causative genetic factors in the rat**. *Physiol Genomics* 2006, **27**(1):1-11.
29. **The Rat Genome Database, RGD** [http://www.rgd.mcw.edu/]
30. Twigger SN, Shimoyama M, Bromberg S, Kwitek AE, Jacob HJ: **The Rat Genome Database, update 2007--easing the path from disease to data and back again**. *Nucleic Acids Res* 2007, **35**(Database issue):D658-662.
31. Xiong Q, Jiao Y, Hasty KA, Stuart JM, Postlethwaite A, Kang AH, Gu W: **Genetic and molecular basis of quantitative trait loci of arthritis in rat: genes and polymorphisms**. *J Immunol* 2008, **181**(2):859-864.
32. Wilson DB: **Quantitative Studies on the Behavior of Sensitized Lymphocytes in Vitro. Ii. Inhibitory Influence of the Immune Suppressor, Imuran, on the Destructive Reaction of Sensitized Lymphoid Cells against Homologous Target Cells**. *J Exp Med* 1965, **122**:167-172.

33. Kawahito Y, Cannon GW, Gulko PS, Remmers EF, Longman RE, Reese VR, Wang J, Griffiths MM, Wilder RL: **Localization of quantitative trait loci regulating adjuvant-induced arthritis in rats: evidence for genetic factors common to multiple autoimmune diseases**. *J Immunol* 1998, **161**(8):4411-4419.
34. Lorentzen JC, Klareskog L: **Susceptibility of DA rats to arthritis induced with adjuvant oil or rat collagen is determined by genes both within and outside the major histocompatibility complex**. *Scand J Immunol* 1996, **44**(6):592-598.
35. Cannon GW, Woods ML, Clayton F, Griffiths MM: **Induction of arthritis in DA rats by incomplete Freund's adjuvant**. *J Rheumatol* 1993, **20**(1):7-11.
36. Dracheva SV, Remmers EF, Gulko PS, Kawahito Y, Longman RE, Reese VR, Cannon GW, Griffiths MM, Wilder RL: **Identification of a new quantitative trait locus on chromosome 7 controlling disease severity of collagen-induced arthritis in rats**. *Immunogenetics* 1999, **49**(9):787-791.
37. Griffiths MM, Remmers EF: **Genetic analysis of collagen-induced arthritis in rats: a polygenic model for rheumatoid arthritis predicts a common framework of cross-species inflammatory/autoimmune disease loci**. *Immunol Rev* 2001, **184**:172-183.
38. Griffiths MM, Wang J, Joe B, Dracheva S, Kawahito Y, Shepard JS, Reese VR, McCall-Vining S, Hashiramoto A, Cannon GW *et al*: **Identification of four new quantitative trait loci regulating arthritis severity and one new quantitative trait locus regulating autoantibody production in rats with collagen-induced arthritis**. *Arthritis Rheum* 2000, **43**(6):1278-1289.
39. Gulko PS, Kawahito Y, Remmers EF, Reese VR, Wang J, Dracheva SV, Ge L, Longman RE, Shepard JS, Cannon GW *et al*: **Identification of a new non-major histocompatibility complex genetic locus on chromosome 2 that controls disease severity in collagen-induced arthritis in rats**. *Arthritis Rheum* 1998, **41**(12):2122-2131.
40. Jansson AM, Jacobsson L, Luthman H, Lorentzen JC: **Susceptibility to oil-induced arthritis is linked to Oia2 on chromosome 4 in a DA(DA x PVG.1AV1) backcross**. *Transplant Proc* 1999, **31**(3):1597-1599.
41. Lorentzen JC, Glaser A, Jacobsson L, Galli J, Fakhrai-rad H, Klareskog L, Luthman H: **Identification of rat susceptibility loci for adjuvant-oil-induced arthritis**. *Proc Natl Acad Sci U S A* 1998, **95**(11):6383-6387.
42. Meng HC, Griffiths MM, Remmers EF, Kawahito Y, Li W, Neisa R, Cannon GW, Wilder RL, Gulko PS: **Identification of two novel female-specific non-major histocompatibility complex loci regulating collagen-induced arthritis severity and chronicity, and evidence of epistasis**. *Arthritis Rheum* 2004, **50**(8):2695-2705.
43. Nordquist N, Olofsson P, Vingsbo-Lundberg C, Petterson U, Holmdahl R: **Complex genetic control in a rat model for rheumatoid arthritis**. *J Autoimmun* 2000, **15**(4):425-432.
44. Olofsson P, Holmberg J, Pettersson U, Holmdahl R: **Identification and isolation of dominant susceptibility loci for pristane-induced arthritis**. *J Immunol* 2003, **171**(1):407-416.
45. Olofsson P, Lu S, Holmberg J, Song T, Wernhoff P, Pettersson U, Holmdahl R: **A comparative genetic analysis between collagen-induced arthritis and pristane-induced arthritis**. *Arthritis Rheum* 2003, **48**(8):2332-2342.

46. Olofsson P, Wernhoff P, Holmberg J, Holmdahl R: **Two-loci interaction confirms arthritis-regulating quantitative trait locus on rat chromosome 6**. *Genomics* 2003, **82**(6):652-659.
47. Remmers EF, Longman RE, Du Y, O'Hare A, Cannon GW, Griffiths MM, Wilder RL: **A genome scan localizes five non-MHC loci controlling collagen-induced arthritis in rats**. *Nat Genet* 1996, **14**(1):82-85.
48. Vingsbo-Lundberg C, Nordquist N, Olofsson P, Sundvall M, Saxne T, Pettersson U, Holmdahl R: **Genetic control of arthritis onset, severity and chronicity in a model for rheumatoid arthritis in rats**. *Nat Genet* 1998, **20**(4):401-404.
49. Kloting I, Vogt L, Stark O, Fischer U: **Genetic heterogeneity in different BB rat subpopulations**. *Diabetes Res* 1987, **6**(3):145-149.
50. Watson WC, Thompson JP, Terato K, Cremer MA, Kang AH: **Human HLA-DR beta gene hypervariable region homology in the biobreeding BB rat: selection of the diabetic-resistant subline as a rheumatoid arthritis research tool to characterize the immunopathologic response to human type II collagen**. *J Exp Med* 1990, **172**(5):1331-1339.
51. Zipris D: **Evidence that Th1 lymphocytes predominate in islet inflammation and thyroiditis in the BioBreeding (BB) rat**. *J Autoimmun* 1996, **9**(3):315-319.
52. Zipris D, Lien E, Xie JX, Greiner DL, Mordes JP, Rossini AA: **TLR activation synergizes with Kilham rat virus infection to induce diabetes in BBDR rats**. *J Immunol* 2005, **174**(1):131-142.
53. Furuya T, Salstrom JL, McCall-Vining S, Cannon GW, Joe B, Remmers EF, Griffiths MM, Wilder RL: **Genetic dissection of a rat model for rheumatoid arthritis: significant gender influences on autosomal modifier loci**. *Hum Mol Genet* 2000, **9**(15):2241-2250.
54. **RatMap** [http://www.ratmap.org]
55. **Human Genome Resources, National Center for Biotechnology Information, National Library of Medicine (Bethesda, MD)** [http://www.ncbi.nlm.nih.gov/genome/guide/human/]
56. **Genome Bioinformatics Group at University of California Santa Cruz (UCSC)** [http://genome.ucsc.edu/]
57. Felsenstein J: **PHYLIP (Phylogeny Inference Package) version 3.6. Distributed by the author.** *Department of Genome Sciences, University of Washington, Seattle* 2005.
58. Fitch WM, Margoliash E: **Construction of phylogenetic trees**. *Science* 1967, **155**(760):279-284.
59. **The Database for Annotation, Visualization and Integrated Discovery (DAVID) 2008** [http://david.abcc.ncifcrf.gov/]
60. Dennis G, Jr., Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, Lempicki RA: **DAVID: Database for Annotation, Visualization, and Integrated Discovery**. *Genome Biol* 2003, **4**(5):P3.
61. Birney E, Andrews TD, Bevan P, Caccamo M, Chen Y, Clarke L, Coates G, Cuff J, Curwen V, Cutts T *et al*: **An overview of Ensembl**. *Genome Res* 2004, **14**(5):925-928.
62. **PubMed** [http://www.pubmed.gov]
63. **Online Mendelian Inheritance in Animals** [http://omia.angis.org.au/]

64. Teder P, Vandivier RW, Jiang D, Liang J, Cohn L, Pure E, Henson PM, Noble PW: **Resolution of lung inflammation by CD44**. *Science* 2002, **296**(5565):155-158.
65. Sancho D, Gomez M, Viedma F, Esplugues E, Gordon-Alonso M, Garcia-Lopez MA, de la Fuente H, Martinez AC, Lauzurica P, Sanchez-Madrid F: **CD69 downregulates autoimmune reactivity through active transforming growth factor-beta production in collagen-induced arthritis**. *J Clin Invest* 2003, **112**(6):872-882.
66. Hayer S, Steiner G, Gortz B, Reiter E, Tohidast-Akrad M, Amling M, Hoffmann O, Redlich K, Zwerina J, Skriner K *et al*: **CD44 is a determinant of inflammatory bone loss**. *J Exp Med* 2005, **201**(6):903-914.
67. O'Neill LA: **TAMpering with toll-like receptor signaling**. *Cell* 2007, **131**(6):1039-1041.
68. Rothlin CV, Ghosh S, Zuniga EI, Oldstone MB, Lemke G: **TAM receptors are pleiotropic inhibitors of the innate immune response**. *Cell* 2007, **131**(6):1124-1136.
69. Seitz HM, Camenisch TD, Lemke G, Earp HS, Matsushima GK: **Macrophages and dendritic cells use different Axl/Mertk/Tyro3 receptors in clearance of apoptotic cells**. *J Immunol* 2007, **178**(9):5635-5642.
70. Abele R, Tampe R: **The ABCs of immunology: structure and function of TAP, the transporter associated with antigen processing**. *Physiology (Bethesda)* 2004, **19**:216-224.
71. D'Urso CM, Wang ZG, Cao Y, Tatake R, Zeff RA, Ferrone S: **Lack of HLA class I antigen expression by cultured melanoma cells FO-1 due to a defect in B2m gene expression**. *J Clin Invest* 1991, **87**(1):284-292.
72. Ghetie V, Ward ES: **Multiple roles for the major histocompatibility complex class I- related receptor FcRn**. *Annu Rev Immunol* 2000, **18**:739-766.
73. Micheau O, Tschopp J: **Induction of TNF receptor I-mediated apoptosis via two sequential signaling complexes**. *Cell* 2003, **114**(2):181-190.
74. Stauber GB, Aiyer RA, Aggarwal BB: **Human tumor necrosis factor-alpha receptor. Purification by immunoaffinity chromatography and initial characterization**. *J Biol Chem* 1988, **263**(35):19098-19104.
75. Frank S, Werner S: **The human homologue of the yeast CHL1 gene is a novel keratinocyte growth factor-regulated gene**. *J Biol Chem* 1996, **271**(40):24337-24340.
76. Vasa-Nicotera M, Brouilette S, Mangino M, Thompson JR, Braund P, Clemitson JR, Mason A, Bodycote CL, Raleigh SM, Louis E *et al*: **Mapping of a major locus that determines telomere length in humans**. *Am J Hum Genet* 2005, **76**(1):147-151.
77. Browning J, Horner JW, Pettoello-Mantovani M, Raker C, Yurasov S, DePinho RA, Goldstein H: **Mice transgenic for human CD4 and CCR5 are susceptible to HIV infection**. *Proc Natl Acad Sci U S A* 1997, **94**(26):14637-14641.
78. Mack M, Bruhl H, Gruber R, Jaeger C, Cihak J, Eiter V, Plachy J, Stangassinger M, Uhlig K, Schattenkirchner M *et al*: **Predominance of mononuclear cells expressing the chemokine receptor CCR5 in synovial effusions of patients with different forms of arthritis**. *Arthritis Rheum* 1999, **42**(5):981-988.

79. Yang YF, Mukai T, Gao P, Yamaguchi N, Ono S, Iwaki H, Obika S, Imanishi T, Tsujimura T, Hamaoka T *et al*: **A non-peptide CCR5 antagonist inhibits collagen-induced arthritis by modulating T cell migration without affecting anti-collagen T cell responses**. *Eur J Immunol* 2002, **32**(8):2124-2132.
80. Perez OD, Kinoshita S, Hitoshi Y, Payan DG, Kitamura T, Nolan GP, Lorens JB: **Activation of the PKB/AKT pathway by ICAM-2**. *Immunity* 2002, **16**(1):51-65.
81. Burch GH, Gong Y, Liu W, Dettman RW, Curry CJ, Smith L, Miller WL, Bristow J: **Tenascin-X deficiency is associated with Ehlers-Danlos syndrome**. *Nat Genet* 1997, **17**(1):104-108.
82. Kirino Y, Takeno M, Murakami S, Kobayashi M, Kobayashi H, Miura K, Ideguchi H, Ohno S, Ueda A, Ishigatsubo Y: **Tumor necrosis factor alpha acceleration of inflammatory responses by down-regulating heme oxygenase 1 in human peripheral monocytes**. *Arthritis Rheum* 2007, **56**(2):464-475.
83. Naor D, Nedvetzki S: **CD44 in rheumatoid arthritis**. *Arthritis Res Ther* 2003, **5**(3):105-115.
84. Zeng R, Spolski R, Finkelstein SE, Oh S, Kovanen PE, Hinrichs CS, Pise-Masison CA, Radonovich MF, Brady JN, Restifo NP *et al*: **Synergy of IL-21 and IL-15 in regulating CD8+ T cell expansion and function**. *J Exp Med* 2005, **201**(1):139-148.
85. Le Guezennec X, Vermeulen M, Brinkman AB, Hoeijmakers WA, Cohen A, Lasonder E, Stunnenberg HG: **MBD2/NuRD and MBD3/NuRD, two distinct complexes with different biochemical and functional properties**. *Mol Cell Biol* 2006, **26**(3):843-851.
86. Seelig HP, Moosbrugger I, Ehrfeld H, Fink T, Renz M, Genth E: **The major dermatomyositis-specific Mi-2 autoantigen is a presumed helicase involved in transcriptional activation**. *Arthritis Rheum* 1995, **38**(10):1389-1399.
87. Bowe AE, Finnegan R, Jan de Beur SM, Cho J, Levine MA, Kumar R, Schiavi SC: **FGF-23 inhibits renal tubular phosphate transport and is a PHEX substrate**. *Biochem Biophys Res Commun* 2001, **284**(4):977-981.
88. Hardy DC, Murphy WA, Siegel BA, Reid IR, Whyte MP: **X-linked hypophosphatemia in adults: prevalence of skeletal radiographic and scintigraphic features**. *Radiology* 1989, **171**(2):403-414.
89. Olofsson P, Holmberg J, Tordsson J, Lu S, Akerstrom B, Holmdahl R: **Positional identification of Ncf1 as a gene that regulates arthritis severity in rats**. *Nat Genet* 2003, **33**(1):25-32.