# Development of an HLA microarray and its application in Inflammatory Bowel Disease

Dissertation zur Erlangung des akademischen Grades des
Doktors der Naturwissenschaften (Dr. rer. nat.)

eingereicht im Fachbereich Biologie, Chemie, Pharmazie
der Freien Universität Berlin

vorgelegt von

Martin Kerick

aus Krefeld

August, 2008

Die vorliegende Arbeit wurde in der Zeit von Februar 2002 bis Dezember 2003 am Max Planck Institut für molekulare Genetik in Berlin und von Januar 2004 bis April 2006 am Institut für klinische Molekularbiologie in Kiel angefertigt.

| | |
|---|---|
| 1. Gutachter | Prof. Dr. Hans Lehrach<br>Max Planck Institut für molekulare Genetik |
| 2. Gutachter | Prof. Dr. Rupert Mutzel<br>Freie Universität Berlin |

| | |
|---|---|
| Disputation | 3. November 2008 |

**to**  
**my beloved parents**  
**Annette and Wolfgang**  
**my siblings Renate and Robert**  
**Anne and my friends**  
**without whom**  
**I would have never ended**  
**this pretty challenging project**

*...*
*And as imagination bodies forth*
*The form of things unkown, the poet's pen*
*Turns them to shapes, and gives to airy nothing*
*A local habitation and a name*

*WilliamShakespeare*

# Contents

# List of Figures

# List of Tables

# Abbreviations

Table 1: Abbreviations

| Abbreviation | Description |
|---|---|
| 5-ASA | 5-Aminosalicylate |
| ANOVA | Analysis of Variance |
| BLAST | Basic Local Alignment Search Tool |
| BLAT | BLAST-Like Alignment Tool |
| CD | Crohn's disease |
| CDi | Crohn's disease, inflamed |
| cDNA | complementary DNA |
| CDni | Crohn's disease, non-inflamed |
| Ct | number of cycles to reach threshold in Real-time PCR |
| Dis | ANOVA: Difference of CD and UC |
| DisInfl | ANOVA: CD or UC specific effect of inflammation |
| DisReg | ANOVA: GI-region specific difference of CD and UC |
| DNA | deoxyribonucleic acid |
| DSC | disease specificity control |
| DSS | dextran sulfate sodium |
| EDTA | ethylenediaminetetraacetic acid |
| F | female |
| FAM | 6-carboxyfluorescein |
| FDR | false discovery rate |
| g | gram |
| GI | Gastro-intestinal |
| GO | Gene Ontology |
| HSPG | heparan sulphate proteoglycans |
| HLA | human leukocyte antigen |
| IBD | Inflammatory Bowel Disease |
| IBDi | CD and UC, inflamed |
| IBDiN | IBDi vs healthy controls |
| IBDni | CD and UC, non-inflamed |
| IBDniN | IBDni vs healthy controls |

Continued on Next Page...

Table 1 – Continued

| Abbreviation | Description |
|---|---|
| Infl | ANOVA: inflamed vs non-inflamed |
| IQR | interquartile range |
| L | liter |
| M | male |
| m | milli; $10^{-3}$ |
| M | molar (mol/L) |
| MHC | major histocompatibility complex |
| min | minute |
| MOPS | 3-(4-morpholino)propane sulfonic acid |
| mRNA | messenger RNA |
| N | normal/healthy control |
| NC | normal/healthy control |
| NSAID | non-steroidal anti-inflammatory drug |
| PCR | polymerase chain reaction |
| RegInfl | ANOVA: GI-region specific effect of inflammation |
| rpm | revolutions per minute |
| s | second |
| Th1 | type 1 helper T cell |
| Th2 | type 2 helper T cell |
| Th3 | suppressor T-cells |
| UC | ulcerative colitis |
| UCi | ulcerative colitis, inflamed |
| UCni | ulcerative colitis, non-inflamed |
| xg | acceleration due to the force of gravity |
| y | year |
| $\mu$ | micro; $10^{-6}$ |

# 1 Introduction

## 1.1 The HLA region on chromosome 6

The human leukocyte antigen (HLA) complex or major histocompatibility complex (MHC) on chromosome 6 is the most important genetic region in the human genome in relation to infection, autoimmunity and transplantation[77]. It encodes cell-surface antigen-presenting proteins and many other genes involved in the immune system. The antigen presenting proteins can be divided into two classes. Class I antigens (HLA A, B & C) present peptides from inside the cell (including viral peptides if present). Class II antigens (HLA DR, DP, & DQ) present phagocytosed antigens from outside of the cell.

Over the past 50 years, the study of the HLA genes and gene products has resulted in important contributions to immunology, population genetics and transplant medicine. But despite best efforts, more than 100 diseases linked to chromosome 6p[208] remain chief contributors to seriously reduced life quality and global mortality. Developing countries e.g. suffer from more than 13 million deaths each year caused by infectious diseases and industrialized countries show rising incidence of auto-immune diseases affecting currently 4% of the population[32]. The etiopathogenesis of these global health threads is often complex (for example, polygenic and non-Mendelian) and involve genetic, epigenetic and environmental factors[217].

The MHC/HLA complex on chromosome 6p contains the most polymorphic genes in the genome[138] but it has been difficult to identify the precise genes associated with infection or autoimmunity, mainly because of the strong linkage disequilibrium and high gene density over the region[210]. Therefore many studies have either failed to pinpoint a single gene or suffer from insufficient evidence.

## 1.2 Inflammatory Bowel Disease

Inflammatory bowel disease (IBD) are relapsing chronic inflammatory disorders of the gastrointestinal (GI) tract. Although the etiology of IBD is unknown it is clear that it is a multifactorial disease. As typical complex disorder the interplay of genetic, environmental and immunological factors is thought to constitute the pathogenesis[49].

Clinical symptoms of IBD include abdominal pain, diarrhea, weight loss, and

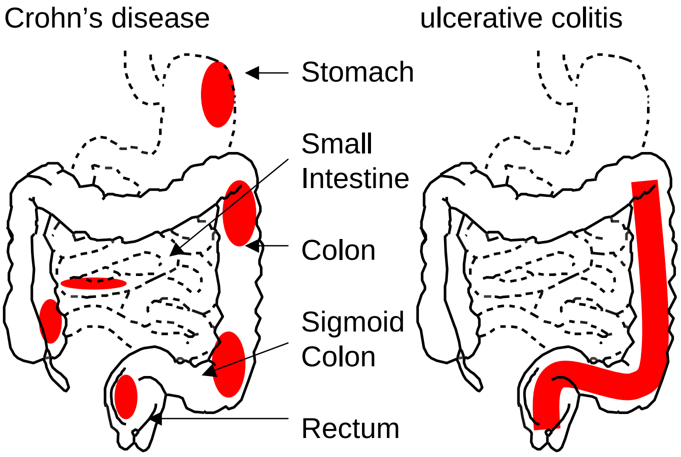


Figure 1: Inflammatory patterns found in the two major forms of IBD: Crohn's disease and ulcerative colitis. Red color exemplarily highlights affected regions. Crohn's disease is characterized by discontinuous inflammation of the whole gastrointestinal tract, while ulcerative colitis is predominantly identified by continuous inflammation of the colon.

fever[228]. The disease incidence peaks in early adulthood and is rising worldwide, with a current lifetime prevalence of 0.1-0.5% in Western countries[157,186].

Inflammatory bowel diseases can be subdivided into two major forms based on clinical, radiologic, endoscopic and pathological criteria[68,155]:

*Crohn's disease* is characterized by discontinuous inflammation of the GI tract with thickened intestinal walls, fissures or fistulae, fibrosis, the presence of granulomas in all tissue layers, and inflammation affecting all layers of the intestinal mucosa. *Ulcerative Colitis* in contrast predominantly affects the colon and is characterized by continuous inflammation mainly found in superficial tissue layers and ulcers penetrating into the submucosa.

Besides disease-subtype specific aberrations of the immune response it is considered that general defects of the immune system underlie the diseases. Therefore the phenotype of patients with IBD often resembles that of patients suffering from other auto-immune diseases. In detail extra intestinal complications such as joint disorders (ankylosing spondylitis), skin disorders (erythema nodosum, epidermolysis bullosa acquisita), hepatobiliary diseases (primary sclerosing cholangitis), and eye diseases (uveitis) often co-occur with IBD. UC is often associated with ankylosing spondylitis, arthritis, primary sclerosing cholangitis, and an increased risk of colon carcinoma[206], while CD is more often associated with psoriasis and thrombotic vascular complications[228].

### 1.2.1 IBD – Genetic factors

There are many studies that support the influence of genetics in IBD, sometimes affecting one subtype more than the other. IBD is a complex polygenic disease with differential concordance rates in twins. In CD, monozygotic twins have a concordance rate of 58%, while dizygotic twins have rates similar to siblings. In UC, the concordance rates are lower than in CD (monozygotic twins, 6-17%; dizygotic twins, 0-5%)[148,207,211]. Further evidence for the genetic component of IBD is reflected in the nine susceptibility loci that have been identified for IBD, as outlined in Table 2.

Most prominently, variations in *CARD15* (IBD1), an innate immunity receptor, have been strongly associated solely with CD pathogenesis. Homozygotes carrying two risk mutations have a 44 times greater probability of contracting CD than non-carriers of the mutations[81].

Chromosome 6 has been implicated in the etiology of IBD, by association and genetic-linkage studies dating as far back as 1972. Previously reported associations include class I[12,61,159,214], class II[51,140,209], and tumor necrosis factor-a (TNF$\alpha$) alleles[17,154]. Stratification analysis recently revealed that the association seems to be gender specific as was pointed out by Fisher et al.[50] (figure2). This sex-specific linkage occurs in both CD and UC families.

Animal models also strongly support the genetic component of inflammatory bowel disease. The mice strains Samp/Yit and C3H/HeJBir are murine models of intestinal inflammation and spontaneous colitis, respectively[98,100]. In particular, the C3H/HeJBir strain was also shown to be sensitive to chemically induced colitis by administration of dextran sulfate sodium in comparison to other strains of mice[124]. Furthermore, though no one single model manifests all the characteristics of human IBD, different genetic defects in transgenic animal models (knock-in or knock-out) can reproduce the same clinical phenotype. For example, knocking-out *IL10*, a cytokine whose normal function is to suppress macrophage function, yields a mouse model with over-production of Th1 cytokines, resulting in intestinal inflammation[125,219]. An *IL2* (T-cell proliferation) knock-out mouse results in increased *growth factor B*, *CD14* and inducible *NOS*, and ultimately, colitis[65,131,176,182]. Fi-

Table 2: Genetic loci associated to Inflammatory Bowel Diseases

| Locus | Chromosomal Location | Candidate genes |
|---|---|---|
| IBD1 | 16p12-q13 | CARD15, CD11, ITGA[M,L,X,D], CD19, IL4R, SPN |
| IBD2 | 12p13.2-q24.1 | IFNG, SLC11A2 |
| IBD3 | 6p | TNF$\alpha$ |
| IBD4 | 14q11-q12 | T-cell receptor [$\alpha$,$\gamma$] complex |
| IBD5 | 5q31 | SLC22A[4,5], IL[3,4,5,13], CSF2, SPINK5 |
| IBD6 | 19p13 | ICAM-1, C3, TBXA2R, CYP4F3, TYK2, JAK3 |
| IBD7 | 1p36 | |
| IBD8 | 16p | |
| IBD9 | 3p26 | |
| unnamed | 10q23 | DLG5 |
| unnamed | 7q21.1 | ABCB1 |
| unnamed | 7q22 | MUC3, HGF, EGFR |
| unnamed | 7p21.3 | AGR2 |
| unnamed | 3p21.2 | GNAI2 |

nally, cross-breeding a DSS-susceptible mouse strain (C3H/HeJ) with a partially resistant strain (C57BL/6) yielded significant genetic linkage to colitis at marker D5Mit216, a locus which is syntenic with the human linkage on chromosome 5[123]. Together, these examples illustrate that genetic changes can cause colonic inflammation.

### 1.2.2 IBD – Environmental factors

Non-genetic factors, such as the environment, have an influence on the incidence of IBD, because genetic data cannot fully explain all variability seen in IBD patients[228]. A wide variety of environmental factors have been suggested as risk factors for the development of inflammatory disease, including smoking, diet, hygiene, and technological advances that cause changes in bacterial flora. Indeed, IBD may generally be described as a side effect of 20th century industrialization, since incidence of IBD has increased within industrialized nations since the 1950's[129].

Of all environmental factors, smoking has the strongest influence on inflammatory bowel disease, with opposite effects in Crohn's disease and ulcerative colitis. Whereas smoking increases the risk of development, relapse, and the need for surgi-

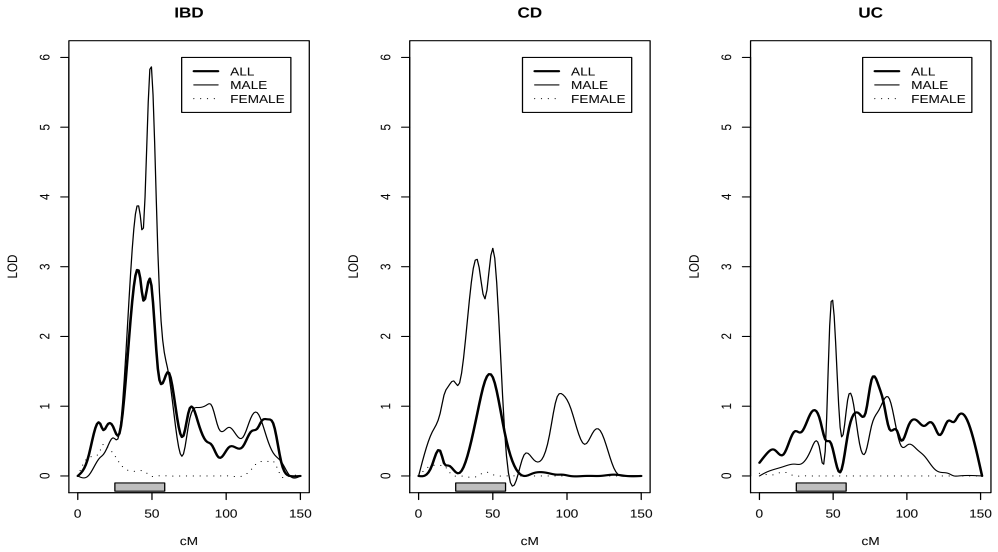


Figure 2: Genetic Association with Inflammatory Bowel Diseases found on chromosome 6 in the extended HLA region. The picture is reproduced from a study by Fisher et al.[50]. Depicted are the LOD scores to measure the amplitude of association found. Stratification analysis revealed that the association is gender specific, with strong association found for male subjects in both CD and UC families. The gray bar, depicts the region covered by the HLA microarray design developed in this study.

cal intervention in Crohn's disease, smoking is viewed as beneficial against UC, due to the fact that many UC patients are non-smokers or former smokers[119,163]. The mechanism by which smoking influences different effects in CD and UC is unknown; however, one possibility is that smoking may impair antimicrobial response in CD, while promoting protective effects of luminal bacteria in UC[14].

With respect to diet, breastfeeding is suggested to be protective against IBD[95], whereas increased intake of simple carbohydrates and fast food are risk factors for CD[153]. High sucrose intake is a risk factor for IBD, while high fat intake is a risk factor for UC[164]. Factors, which affect availability, transport or metabolism of the major energy source of the colonocytes, the short chain fatty acids (SCFA) and in particular, butyrate, may introduce a risk for IBD by causing colonocyte starvation. For example, increased ingestion of food containing sulfur may be associated with UC, since sulfur compounds inhibit metabolism of butyrate[169]. Lack of dietary fiber to breakdown into SCFA by bacterial flora or impaired transport of SCFAs can cause colon pathology[34,180].

Changes in hygienic practices have also been hypothesized to be related to the increasing rate of IBD. It has been suggested that overall improvements in public sanitation may predispose for IBD by removing the stimulus for a mucosal immune response, which would normally be required in an unsanitary environment[172]. The pre-disposing immune response genes, when not sufficiently activated in a *clean* environment, may lead to hyper-stimulation of the immune response if challenged by a microbial antigen later in life. Alternatively, the lack of exposure to microbial antigens may leave the mucosal immunological system in a perpetually primed state which makes the system vulnerable to dysregulation in the form of an autoimmunity reaction[222]. The loss of the intestinal parasite, the helminth, is an example of the hygiene hypothesis. As an integral part of the intestinal flora, helminths affect T regulatory cells to stop over-activation of T cells. The absence of the helminth upsets T cell regulation, a process that is probably common in diseases with an autoimmunity component (IBD, asthma, multiple sclerosis, allergies)[97,218].

Interestingly, the advent of refrigeration has been proposed to cause Crohn's disease in genetically susceptible individuals. The use of refrigeration in commercial and domestic food preparation encourages the growth of psychrotrophic bacteria, such as *Yersinia enterocolitica* and *Listeria monocytogenes*, which are able to grow at cold temperatures (between -1°C and 10°C). The *cold chain hypothesis* proposes that chronic ingestion of psychrotrophic bacteria from refrigerated foods invokes an over-active host reaction in genetically susceptible individuals who have lost the ability to tolerate the bacteria[82].

### 1.2.3 IBD – Cellular Immunity

The gut immune system must maintain a delicate balance between tolerating food antigens and commensal bacteria, while being able to appropriately respond to pathogens. Oral tolerance is thought to occur by three mechanisms: 1) clonal selection takes place to eliminate the effector T cells that recognize commensal bacterial as *self-antigen*; 2) clonal anergy is induced (i.e. lymphocytes are rendered unresponsive to the oral antigen); and 3) suppressor T-cells (Th3 cells) are induced[18]. In inflammatory bowel disease, different defects in the maintenance of oral tolerance have the potential to cause chronic inflammation. Though there are no prospective studies to definitively show that oral tolerance is defective in IBD, there are studies that suggest that there is some truth to the hypothesis. In one example, a non-prospective study has shown that T-cells from IBD patients proliferated and produced cytokines in response to their own microflora, whereas T-cells from control subjects did not[41]. In a more recent example, IBD patients and controls were administered an oral antigen (keyhole limpet hemocyanin) before initial immunization and booster[102]. In contrast to the control group, tolerance was not induced in CD and UC patients, as shown by enhanced levels of T-cell proliferation after immunization, suggesting a defect in suppression response in IBD. In a follow-up study, the same authors repeated the protocol on IBD patients and first-degree relatives and found that intestinal permeability could not be faulted for the lack of tolerance induction and suggested that there is a genetic defect responsible for oral tolerance in IBD[101].

An upset in balance between effector T-cells (Th1 or Th2), which cause inflammation, and regulatory T-cells, which prevent inflammation, can also result in mucosal inflammation. According to mouse models, inflammation may result from either an over-active effector T-cell response or a weakened regulatory T-cell response[18]. Within the effector T-cells, mucosal inflammation can be a consequence of excessive T helper 1 cell (Th1) response, which induces secretion of *IL2*, *IL12*, *IFN*$\gamma$ and/or *TNF*$\alpha$. Alternatively, inflammation may result from an excessive Th2 response, which induces secretion of *IL4*, *IL5*, *IL10* and/or *IL13*. These cytokine profiles are clear distinguishing characteristics of the two subtypes of IBD. The CD cytokine profile looks more like a Th1-mediated inflammation. Many studies have showed that macrophages isolated from CD, but not UC patients, produce increased amounts of *IL12*, which activates Th1 cells[136,151]. In addition, T cells from CD patients show increased levels of *IFN*$\gamma$ and decreased levels of *IL4* in comparison to controls[53,136,151]. A final piece of evidence supporting the role of Th1 response in CD is the result that CD patients treated with an antibody against the *p40* chain of *IL12* show an immediate reduction in inflammation, which is paralleled by decreased *IL12* and *IFN*$\gamma$ levels in mononuclear cells from colonic lamina propria[126]. In contrast to CD, the UC cytokine profile resembles a Th2-mediated inflammation. Although the hallmark cytokine of Th2 response, *IL4*, has not been shown to be increased in UC Th2 cells, there are other studies that suggest that UC is driven by Th2 mediated inflammation[18]. For example, the development of auto-antibodies, such as anti-neutrophil cytoplasmic antibody (pANCA), is more prevalent in UC in contrast to CD[178], while the anti-Saccharomyces cerevisiae antibody (ASCA) is more prevalent in CD than in UC, which provides a possibility to distinguish UC and CD by serological testing[9]. Moreover, Th2-related subclasses of *IgG* (*IgG1* and *IgG4*) are associated with UC, as opposed to Th1-related subclasses (*IgG2*), which are more associated with CD[92]. Together, these studies support distinct T-helper-cell responses in CD and UC.

Abnormalities in any aspect of the innate immune response may ultimately lead to inflammation in IBD. The most prominent case in point is the association of mutations in *CARD15* to CD. The protein encoded by this gene, *NOD2*, is an intracellular receptor for bacterial peptidoglycan, in particular, muramyl dipeptide. *NOD2* is expressed in epithelial cells, dendritic cells, granulocytes and monocytes. There are three hypotheses as to how defects in *NOD2* act to cause CD[18]: 1) the lack of functional *NOD2* leads to impaired macrophage activation and infection; 2) lack of functional *NOD2* in epithelial cells negates activation of an epithelial cell defense response (in the form of chemokine and defensin secretion), allowing bacterial colonization of the crypt and penetration through the epithelial barrier; 3) impaired recognition of bacterial antigen leads to inappropriate activation of antigen presenting cells, resulting in an imbalance between effector and regulatory T-cells.

### 1.2.4 IBD – The epithelial barrier

The mucosal epithelium provides a physical barrier separating an outside environment (food antigens, microflora and pathogens) from the inner workings of the body. This physical barrier consists of a thick glycocalyx mucous layer coating the apical surface of the epithelial cell layer and the epithelial cells themselves, which are joined laterally by tight and adherens junctions and basally to the extracellular matrix of the lamina propria. Breaches of the epithelial barrier can allow components of the outside environment to come into direct contact with the gut immunological machinery, thereby causing an overwhelming immune response. In the case of IBD, is not clear whether impaired barrier function causes a chronic inflammation first or if an impaired immune response leads to degradation of the barrier first. Due to the multifactorial nature of IBD, it is possible that one or both events contribute to the initiation of chronic inflammation, depending on the characteristics of the individual patient. Many animal models support the hypothesis that the loss of barrier function causes inflammation. In one example, severe inflammation is observed in transgenic mice expressing a dominant negative *N-cadherin* transgene that impairs intra-epithelial cell adhesion[69]. Not only does the intestinal epithelial layer provide a physical barrier, it also plays an active role in immune response. Epithelial cells are able to detect microbial antigens and further transmit the appropriate response. For example, Toll-like receptor 5 (*TLR5*) is specifically expressed on the basolateral side of the epithelial cell, where, if it is activated by bacterial flagellin, it initiates an inflammatory response to combat the bacterial breach[60]. And finally, recent in vitro studies show that the epithelial tight junction can alter its structure in response to cytokines such as tumor necrosis factor alpha ($TNF\alpha$)[229] or interferon-gamma ($IFN\gamma$)[212].

## 1.3 Differential splicing and disease

The impact of splicing on physiological events has been investigated in various species and model systems (for a review see Black et al.[15]). Within the recent years several studies have focused on splicing in common human diseases, indicating a key role in pathophysiology. However, the current knowledge of splicing in human diseases is still very limited. Differential splicing of pre-mRNA is a highly regulated process. Splicing factors (SF), consisting of a combination of more than a hundred proteins and small RNAs, are needed to splice pre-mRNAs (for a review see[15]). Besides the splicing factors that excise an intron, a growing number of reports concentrate on splicing factors determining and activating splicing sites. These regulatory splicing factors, specifically known as splicing enhancers and splicing inhibitors, interact to guide splice site selection[15]. It has been proposed that regulatory splicing factors modulate differential splicing by controlling the ratio of splicing enhancers and splicing inhibitors, which in turn, determine splice site selection[55,83,130,135,160,215]. Several recent studies have pointed out that aberrant differential splicing is a potential cause for human disease[24,48,73,146]. For example differential splicing affects apoptotic processes[26,183,195,201] and signal transduction pathways[19,31,227]. Differential splicing exerts its action by changing either the function, location or expression level of a protein. Aberrant intron-retention, for instance, was shown to cause the dysfunction of *ATRX* leading to *acquired alpha thalassemia*[143].

## 1.4 Microarray technology

A DNA microarray is a high-throughput technology used in molecular biology and in medicine. Their use for gene expression profiling was first reported by Schena et al.[179] in 1995, and a complete eukaryotic genome (*Saccharomyces cerevisiae*) on a microarray was published in 1997[108]. Microarray technology evolved from Southern blotting and consists of an arrayed series of thousands of microscopic spots of DNA oligonucleotides, called probes or features, each containing pico-moles of a specific DNA sequence. This can be a short section of a gene or other DNA element that are used as probes to hybridize a cDNA or cRNA sample (called target) under high-stringency conditions. The types of DNA arrays differ by the length of probe DNA arrayed to catch the target.

**Short oligonucleotide arrays** known as *Affymetrix arrays* consist of up to 20 single-stranded oligonucleotides (25mers) per gene. The expression level of a gene is calculated by different algorithms integrating the probe expression levels.

**Long oligonucleotide arrays** *Longmer arrays* measure the expression level of a gene by hybridization of cDNA to only one single stranded oligonucleotide of 50 to 70 bases.

**cDNA arrays** consist of an array of double-stranded DNA of variable length. Probe length usually is in the range of 200-5000 base-pairs.

Probe-target hybridization is usually detected and quantified by fluorescence-based detection of fluorophore-labeled targets to determine relative abundance of nucleic acid sequences in the target. Figure 1.4 depicts microarray preparation and a typical microarray experiment.

DNA microarrays can be used to measure changes in expression levels of genes or to detect single nucleotide polymorphisms (SNPs).

Microarray technology creates several challenges: to name just a few, multiple levels of replication have to be addressed in experimental design, the data has to be standardized prior to analysis, the amount of multiple testing as compared to the small number of biological replicates poses many statistical questions and specificity issues have to be acknowledged when it comes to interpret the results.

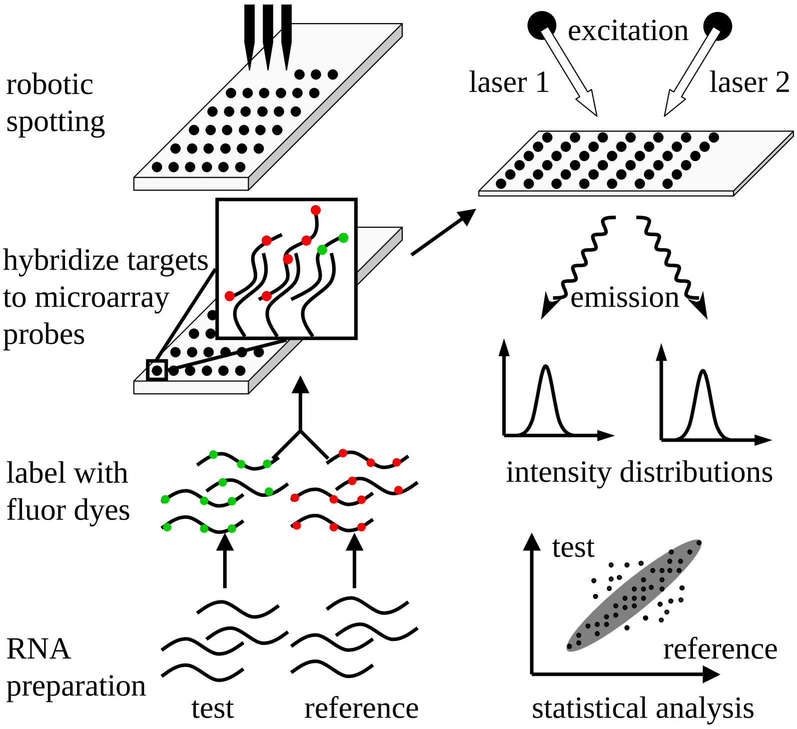


Figure 3: Microarray preparation and experiment. Specific sequences from thousands of genes get spotted onto small, solid supports at fixed locations. Equal amounts of cDNA from two sources to be compared are labeled and hybridized to the microarray. Scanning reveals each fixed location where microarray probe and target sequence match for each fluor. Fluorescence intensity is translated into gene activity for each gene and source. Statistic reveals if gene activity differs for the two sources.

## 1.5 Aims of the study

Genetic studies repeatedly confirmed linkage of the HLA region on chromosome 6p21 to the pathogenesis of Inflammatory bowel disease. Due to the extensive haplotype structures in the HLA region on chromosome 6 it is almost impossible to disentangle the genetic association found for IBD and to single out potentially disease causing genes. The assumption that disease associated genes might exhibit aberrant expression or splicing patterns initiated this study.

The first aim of this study was to construct a tool to investigate the transcriptome of the HLA linkage region on chromosome 6p21.

The second objective of this study was to use the constructed HLA-microarray to investigate differential gene expression in both inflamed and non-inflamed IBD samples and normal controls.

As a follow-up to the microarray results, high-throughput 384-well plate format real-time PCR was used to further verify the gene transcript expression in a larger number of patient samples (195 samples, including normal controls, IBD samples and disease specificity controls).

In summary, the objectives of this study were three-fold: 1) to construct a HLA-microarray; 2) to use microarray expression screening to identify novel genes involved in the pathogenesis of IBD; 3) to focus on processes identified by microarray screening and to analyze transcript expression in a larger number of patients using quantitative real-time PCR.

# 2 Methods

## 2.1 Establishment of a chromosome 6p21 microarray

### 2.1.1 Data sources

**Ensembl core** The Ensembl database[79] version 25.34 based on NCBI assembly 34 was used as reference database. Next to the transcript model database of its own annotation pipeline, the Ensembl database is more of an assembly of databases. Within its database scheme, data from the Vega database as well as data from the ESTgene database can be found. The Ensembl transcript database contained 34111 transcript models (22291 genes) of which 798 transcripts (504 genes) were located in the HLA (20 Mb - 50 Mb) region on chromosome 6.

**Sanger Center Vega** Data from the manually curated Vega database[223] was extracted from the Ensembl core database. The Vega database contained 9557 manually curated transcript models (5501 genes). 705 transcripts (438 genes) were found located in the HLA region.

**Ensembl ESTgene** The Ensembl core database additionally harbored 43710 transcript models (24980 genes) from the ESTgene database. 803 transcripts models (472 genes) were found for the HLA region.

**NCBI RefSeq** The RefSeq release 8 of 2004-10-31 of Homo sapiens was downloaded from NCBI (ftp://ftp.ncbi.nih.gov/RefSeq/) as files in FASTA format. Only reviewed transcripts with name prefix NM have been used in this study. Of 22239 sequences 21073 sequences were already mapped into the Ensembl database scheme. Positional information, if present, was down-loaded from UCSC[58]. Of 22239 sequences 548 (426 genes) mapped to the HLA region on chromosome 6.

### 2.1.2 Merging databases

**Grouping transcripts** The algorithm made heavy use of the *Perl* module *IntSpan* which handles spans of integers. It treated transcripts as sets of numeric intervals. Set operations have been used to compare the transcripts against each other. Grouping transcripts from different databases was a three step process. First; transcripts were assigned to each other if they shared exon sequence of at least two exons. If the total number of intervals of the intersect of both transcripts was greater equal than two the transcripts were assigned to each other. If there was only one interval of at least 100 bases in the intersect and the exon-count of one transcript was smaller than three the transcripts got assigned to each other. If there was only one interval of at least 200 bases in the intersect the transcripts got assigned to each other. In all other cases no assignment was done. If a transcript got assigned to two gene models in one databases scheme manual curation was needed so that a transcript belonged to only one gene model in one database scheme. In step two a transcript model was build as union of all exons of all transcripts which have been assigned to each other so far. The third step searched for transcripts, which had been missed in the first step. Again manual curation was needed if one transcript was assigned to two gene models in one database scheme.

**Down-sizing redundancy** Global redundancy was defined as transcripts models, which have multiple copies of itself distributed in the genome and was assessed by BLAT sequences search *all against all.* In case, two transcripts had sequence identity of 99% and the length difference was smaller than 5% (max. 30 bases) and their gene name / gene description were identical one copy was masked, while keeping the information from both copies. Local redundancy was addressed as follows. If two transcript models were identical (length and sequence) we mask one copy in the database. Two transcript models were assumed identical if the intersect had the same number of exons and the length of each exon of the intersect differed to the individual exons by less than 6 bases. If one transcript model was a subset of another transcript model we determined if the two models shared the structure (exon-intron boundaries). In case of shared structure the algorithm kept the smaller transcripts only of manually curated databases (RefSeq and Vega). In case of different exon-intron structure both transcripts were kept in the database.

### 2.1.3 Limits of microarray technology

**Sequence quality of transcripts** An *all against all* BLAT[89] sequence search was set up to investigate the uniqueness of the sequence for each transcript-model. The Ensembl database was used as source of transcript models. The BLAT sequence search was optimized in that each query sequence was divided in chunks of 28 bases. *tileSize* was set to 14 bases, *stepSize* to 7 bases, *repMatch* to 100000 and minimal score (match minus mismatch) was required to be 20. Each hit for each query sequence was treated as numeric interval and intervals were gathered by union operation. Hits for each query-sequence were aggregated by hit name and hit chunks smaller than 20 bases were removed, since they do not generate significant cross-hybridization according to Kane et al.[87]. To determine transcript-specific sequence all hits were gathered by union operation and subtracted from the query-sequence (relative complement of hits in query-sequence). The remaining query-sequence is unique and was analyzed for windows of pre-defined size. Window size one was used to investigate if microarray probes of length 25 bases could be designed. The assumption here was, that a mismatch of size one abolishes hybridization. Window size 31 and 51 were used to check if microarray probes of length 50 and 70 could be designed. The finding of Kane et al.[87] that 19 bases do not generate significant cross-hybridization explains the difference in size. To determine gene-specific sequence all hits from the same gene were dropped, the remaining hits were gathered by union operation and subtracted from the query-sequence (relative complement of hits in query-sequence). The remaining query-sequence is gene-specific and was analyzed as described above.

**Sequence quality of exons and exon-junctions** Exon boundary information was extracted from Ensembl database. An exon-junction was defined as sequence window placed centered on the exon-junction. The exon-junction sequence window size was 25, 40 and 50 bases to determine if microarray probes of length 25, 50 and 70 bases could be designed on the exon-junction. To determine the sequence properties of exons and exon-junctions of a transcript the set of transcript-specific bases was calculated as described above in section 2.1.3. For each exon or exon-junction the intersect with the set of transcript-specific bases was calculated. The intersect sequence was analyzed for windows of pre-defined size to investigate if microarray probes of different length could be designed as described above. To determine gene-specific exons and exon-junctions the same strategy was followed except that the intersect of exon/exon-junction sequence with the set of gene-specific bases of the query-sequence was calculated.

### 2.1.4 Microarray fabrication

**Coating of slides** Glass slides were coated with *poly-L-lysine* according to protocol developed by Schena et al.[179]. In short; slides were cleaned over night in cleaning solution (2.5 M NaOH, 60%(v/v) EtOH) and then vigorously washed in $ddH_2O$. Slides were transferred to *poly-L-lysine* solution (10%(v/v) *poly-L-lysine*, 10%(v/v) Tissue culture PBS, 80%(v/v) $ddH_2O$) and incubated (shaking) for 45 min. Slides were washed twice in $ddH_2O$ (second wash solution contained 20$\mu$l Tween20 per 250ml) and immediately dried by 10 min. centrifugation at 1000 rpm (178 xg).

**Spotting** Oligonucleotide probes of length 57±7 (fabricated by Metabion, Germany) were dissolved in spotting buffer (3M SSC, 0.5M betaine) at 25 $\mu$M concentration and printed on *poly-L-lysine* coated slides (*reinweiss slides*, Beyer Walter) using an in-house (Max Planck Institute for Molecular Genetics) modified QArray (Genetics LTD, UK) robot with 48 Stealth Micro Spotting Pins (TeleChem International) at a dot pitch of 187.5$\mu$m. Probes of target transcripts were spotted in quadruplicate while control probes (Scorecard Control (Amersham), Arabidopsis transcripts *PR* & *rcab* (Scienion, Germany)) were spotted 8 times and guide probes (*GAPDH*, *ACTB*) were spotted 240 times.

**Post-processing** Microarrays were post-processed according to protocol developed by Diehl et al.[40]. Spotted slides were left at room temperature overnight and then heat-treated on a metal block at 80°C for 5s. Oligonucleotide probes were cross-linked twice at 1200J using a Stratalinker (Stratagene, Netherlands) on auto-cross-link setting. For the blocking process, 1g succinic anhydride (Fluka, Deisendorf, Germany) was freshly dissolved in 200ml anhydrous 1,2-dichloroethane (DCE; Fluka). To this solution, 2.5ml of N-methylimidazol (Fluka) was added and immediately poured into the slide chamber. Incubation was for 1h, placed on an orbital shaker for slight agitation. Subsequently, the slides were briefly washed in 200ml of fresh DCE and incubated in boiling water for 2 min for DNA denaturation. After a brief rinse in 95% ethanol, the slides were dried by centrifugation (5 min 500rpm, 44 xg).

### 2.1.5 Microarray quality control

**Labeled random primer hybridization** Cy3$^{TM}$-labeled random 6mer primer was resuspended in 90 $\mu$l hybridization buffer (Microarray 4x Hyb buffer Amersham

No.rpk0325) and then hybridized in a hybridization chamber (Scienion, Berlin) overnight at 42°C under a coverslip with humidity maintained by hybridization buffer. After hybridization, slides were washed consecutively in 2x SSC/0.05% SDS, 0.2x SSC, 0.1x SSC for 5 minutes at room temperature (22°C) and then dried by centrifugation.

**Dynamic range and ratio control** To determine the dynamic range of the microarray, the data gathered as described in section 3.2 was used. In total the raw data of 60 microarray experiments was analyzed. Each calibration control was spotted eight times resulting in 480 measurements per calibration control and dye. To inspect the expression ratio measured by the microarray, a subset of the data gathered as described in section 3.2 was used. In total the raw and normalized data of 24 microarray experiments was analyzed. Normalization was carried out as described in section 2.2.4. Each ratio control was spotted eight times resulting in 192 measurements per ratio control.

## 2.2 Splicing factors and intron-retention in IBD

### 2.2.1 Patients

**Informed consent and approval by the ethics committee** All patients included in this study consented to additional research biopsies being taken 24 h prior to endoscopy. The procedures in the study protocol were approved by the Ethics Committee of the Medical Faculty of the Christian-Albrechts-University prior to the start of the study.

**Subject group one** A group of 30 male individuals (recruited between 1999 and 2003; mean age 31 years (sd: 5.2 years)) was analyzed by microarray technology. The study group (table 3) comprised six healthy individuals as control subjects, with endoscopic and histological examination yielding no significant pathological findings. To investigate gene expression in IBD, twelve patients with CD and twelve patients with UC were recruited. All endoscopic biopsies were taken from a defined area either of the sigmoid colon (at 20-30 cm measured during withdrawal) or the terminal ileum (5-10 cm into the ileum), and immediately snap-frozen in liquid nitrogen. Inflammation status of the biopsy was determined by the trained endoscopist.

Table 3: Patient Groups analyzed in this study

| | Healthy Control<br>non-inflamed | DSC<br>inflamed | DSC<br>non-inflamed | CD<br>inflamed | CD<br>non-inflamed | UC<br>inflamed | UC<br>non-inflamed |
|---|---|---|---|---|---|---|---|
| | Subject group 1 - analyzed by microarray | | | | | | |
| n | 6 | - | - | 6 | 6 | 6 | 6 |
| Sex (F/M) | -/6 | - | - | -/6 | -/6 | -/6 | -/6 |
| Mean Age±Sd | 35.5±7.1 | - | - | 28.7±4.3 | 29.9±4.7 | 30.0±5.2 | 30.6±4.8 |
| | Subject group 2 - analyzed by Real-Time PCR | | | | | | |
| n | 30 | 15 | 30 | 30 | 30 | 30 | 30 |
| Sex (F/M) | 15/15 | 8/7 | 15/15 | 15/15 | 15/15 | 15/15 | 15/15 |
| Mean Age±Sd | 50.4±13.6 | 46.9±18.3 | 49.0±17.6 | 30.5±8.5 | 39.8±11.3 | 34.7±9.8 | 42.4±12.4 |

**Subject group two** To verify and investigate the molecular epidemiology of selected signals identified by microarray experiments, subject group one was extended from 30 to 195 individuals and analyzed by real-time PCR (table 3). IBD patients, part of group 2 consisted of 60 CD patients and 60 UC patients. The normal control population consisted of 30 healthy individuals (HN), who had no significant pathological findings following endoscopic examination. As disease specificity controls (DSC), 45 patients with colonic disease but not IBD, were included. The basic cause for endoscopy of DSC patients was infectious diarrhea, other forms of gastrointestinal inflammation, or irritable bowel syndrome. All subgroups of this large cohort were balanced concerning inflammation status, age and gender. Inflammation status of the biopsy was determined by the trained endoscopist.

### 2.2.2 Experimental design

To compare experimental design strategies using the *daMA* library in R or ANOVA in general one needs to code the experimental design and the biological questions (contrasts) in matrix notation. To help the inexperienced reader to replicate the setup, the design and contrast matrix used in this study are given below.

**Design matrix** The design matrix given was replicated three times, resulting in 60 rows for 60 microarray experiments.

| Array | G | R | NnS | UnS | UiS | CnS | CiS | NnT | UnT | UiT | CnT | CiT |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | -1 | -1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2 | 1 | -1 | 0 | 0 | -1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| 3 | 1 | -1 | 0 | 0 | 0 | 1 | -1 | 0 | 0 | 0 | 0 | 0 |
| 4 | 1 | -1 | 0 | 1 | 0 | -1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 5 | 1 | -1 | 1 | -1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 6 | 1 | -1 | 0 | 0 | 0 | 0 | 0 | -1 | 0 | 1 | 0 | 0 |
| 7 | 1 | -1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | -1 | 0 | 1 |
| 8 | 1 | -1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | -1 |
| 9 | 1 | -1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | -1 | 0 |
| 10 | 1 | -1 | 0 | 0 | 0 | 0 | 0 | 1 | -1 | 0 | 0 | 0 |
| 11 | 1 | -1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | -1 | 0 | 0 |
| 12 | 1 | -1 | 0 | 1 | 0 | 0 | 0 | -1 | 0 | 0 | 0 | 0 |
| 13 | 1 | -1 | 0 | 0 | 0 | 1 | 0 | 0 | -1 | 0 | 0 | 0 |
| 14 | 1 | -1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | -1 | 0 |
| 15 | 1 | -1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | -1 |
| 16 | 1 | -1 | -1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| 17 | 1 | -1 | 0 | -1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| 18 | 1 | -1 | 0 | 0 | 0 | -1 | 0 | 0 | 0 | 0 | 0 | 1 |
| 19 | 1 | -1 | 0 | 0 | 0 | 0 | -1 | 0 | 0 | 1 | 0 | 0 |
| 20 | 1 | -1 | 0 | 0 | -1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |

**Contrast matrix** In the matrix of biological questions (= contrasts) the columns correspond to factors (factor combinations) and the rows to questions. Using ANOVA a coefficient for each factor (factor combination) is calculated and the contrast matrix represents the mathematical formula to calculate the comparisons of interest.

| | G | R | NnS | UnS | UiS | CnS | CiS | NnT | UnT | UiT | CnT | CiT |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Dis | 0 | 0 | 0 | -0.25 | -0.25 | 0.25 | 0.25 | 0 | -0.25 | -0.25 | 0.25 | 0.25 |
| Infl | 0 | 0 | 0 | -0.25 | 0.25 | -0.25 | 0.25 | 0 | -0.25 | 0.25 | -0.25 | 0.25 |
| DisInfl | 0 | 0 | 0 | -0.25 | 0.25 | 0.25 | -0.25 | 0 | -0.25 | 0.25 | 0.25 | -0.25 |
| IBD vs N | 0 | 0 | -1 | 0.25 | 0.25 | 0.25 | 0.25 | -1 | 0.25 | 0.25 | 0.25 | 0.25 |
| Reg | 0 | 0 | 0 | 0.25 | 0.25 | 0.25 | 0.25 | 0 | -0.25 | -0.25 | -0.25 | -0.25 |
| RegDis | 0 | 0 | 0 | -0.25 | -0.25 | 0.25 | 0.25 | 0 | 0.25 | 0.25 | -0.25 | -0.25 |
| RegInf | 0 | 0 | 0 | -0.25 | 0.25 | -0.25 | 0.25 | 0 | 0.25 | -0.25 | 0.25 | -0.25 |
| IBDn vs N | 0 | 0 | -0.5 | 0.25 | 0 | 0.25 | 0 | -0.5 | 0.25 | 0 | 0.25 | 0 |
| IBDi vs N | 0 | 0 | -0.5 | 0 | 0.25 | 0 | 0.25 | -0.5 | 0 | 0.25 | 0 | 0.25 |

### 2.2.3 Isolation of RNA

All reagents, glassware and laboratory utensils to be used for RNA isolation were specially treated in order to minimize RNA degradation by RNAses. Solutions were prepared with 0.1% DEPC-treated distilled water and sterilized by autoclaving. All glassware, ceramic mortar and pestles, Teflon pestles and metal spatulas were cleaned with common laboratory washing detergent, rinsed thoroughly in distilled water and air-dried before wrapping in aluminum foil and baking at 180°C for 12-16 h before use, in order to inactivate any contaminating RNAses. All plasticware was purchased as UV-sterilized consumables (50 mL conical tubes, pipet tips with aerosol filters) or RNAse-free consumables (microfuge tubes).

**Homogenization of mucosa biopsies** Mucosal biopsies were stored in liquid nitrogen prior to use. To minimize premature thawing of the biopsies during RNA isolation, homogenization equipment (mortar, pestle, polyethylene tubes) was cooled with liquid nitrogen prior to the start of the procedure. A commercial kit was used to isolate total RNA from biopsy material (RNeasy mini-kit, Qiagen) according to the manufacturer's protocol. Briefly, one mucosal biopsy was ground into a fine powder using a cooled Teflon mortar in a 1.5 mL microfuge tube. The powder was then mixed with 650 $\mu$L of RLT lysis solution, incubated at room temperature for 10 min and vortexed for 30 s. The sample was then further treated by centrifuging the lysate through a QIA shredder column for 2 min at 14000 rpm (16000 xg) in a microfuge. The eluate was centrifuged for 3 min at 14000 rpm (16000 xg) and the cleared lysate was transferred into a new microfuge tube. One volume of 70% ethanol was then added to the lysate, mixed well by vortexing and loaded onto an RNA binding column fitted in a 2 mL collection tube. The solution was centrifuged through the binding column for 15 s at 10000 rpm (8000 xg). The mini-column was washed once by pipetting 350 $\mu$L of RW1 wash buffer on to the column, incubating for 10 min at room temperature and then centrifuging 15s at 10000 rpm (8000 xg).

**DNAse treatment of RNA and final isolation of RNA** DNAse treatment of the RNA was carried out while the RNA was still bound on the column. First, DNAse stock solution was prepared by reconstituting lyophilized DNAse enzyme (Qiagen) by the addition of 550 $\mu$L of RNAse-free water. This DNAse stock solution (10

$\mu$L, 2.7 Kunitz units/$\mu$L) was further diluted in 70 $\mu$L RDD buffer (as provided in RNAse-free DNase Kit). The diluted DNAse (80 $\mu$L) was pipetted directly on to the spin column membrane and allowed to incubate at room temperature for 15 min. Upon completion of the digestion time, the mini-column was washed once with 350 $\mu$L buffer RW1 and twice with 500 $\mu$L buffer RPE. Each wash step was completed by centrifuging the tube 15 s at 10000 rpm (8000 xg). To completely rid the column of wash solution, the column was then centrifuged 1 min at 14000 rpm (16000 xg). Total RNA was eluted from the mini-column into an RNAse-free microfuge tube by pipetting 50 $\mu$L of RNAse-free water directly onto the membrane, allowing to completely soak through the membrane at room temperature for 5 min and then centrifuging 1 min at 10000 rpm (8000 xg).

**Quality control of RNA** RNA concentration was determined by measuring the absorbance of a diluted RNA sample (1:40) in a spectrophotometer at 280 nm and 260 nm. The concentration and was calculated using the following formula: [RNA]$\mu$g/$\mu$L = A260 X 40 $\mu$g RNA /mL/A260 X 1000 $\mu$L/mL Typically, the A260/280 nucleic acid to protein ratio was between 1.8-2.2, indicating good quality of the isolated RNA. The integrity of the RNA was assessed by one of two methods: 1) denaturing RNA gel electrophoresis; or 2) electrophoresis on an RNA 6000 Nano LabChip (Agilent). In the first instance, a 1.5% agarose gel was prepared in 1 X MOPS and 0.9M formaldehyde. Approximately 1 $\mu$g of total RNA was heat-denatured in 10 $\mu$L of RNA loading solution at 70°C for 10 min and cooled on ice. RNA samples were checked by electrophoresis on an agarose gel at 4-5 V/cm in running buffer of 0.1 X MOPS. After one hour of migration, the gel was photographed under UV light. In the second instance, RNA quality was measured by the Agilent 2100 Bioanalyzer, which used electrophoresis through channels in a microfluidic chip (Nano LabChip) to separate RNA molecules based on size. In both methods, the strong presence of 18S and 28S RNA molecules and the absence of smears at lower molecular weights indicated intact, high quality RNA suitable for further studies. Genomic DNA contamination of RNA was assessed by amplification of a housekeeper gene (*GAPDH*) using the following primers.

| | forward | reverse |
|---|---|---|
| *GAPDH* | 5'-acccactcctccacctttgac-3' | 5'-ctgttgctgtagccaaattcgt-3' |

Approximately 1 $\mu$g of total RNA was used as a template for a PCR reaction containing 1X PCR Buffer, 1.5 mM MgCl2, 0.2mM dNTPs, 0.4 $\mu$M forward primer, 0.4 $\mu$M reverse primer, and 0.1 $\mu$L Taq polymerase in a total reaction volume of 50 $\mu$L. The cycling program was used as follows: 95°C at 2 min, then 40 cycles of 95°C 30s, 60°C 30 s, 72°C 30 s, and a final extension step of 72°C for 5 min. PCR product was mixed with 10 X DNA loading dye before electrophoresis on a 3% agarose mini-gel in 1X TAE buffer. After 45 min of electrophoresis, the gel was photographed under UV light. The presence of a 101 bp band indicated that a genomic contamination was present in the RNA. If the total RNA failed to meet any of the quality criteria, the total RNA was re-isolated from additional biopsies or RNA solution was subject to another round of binding, washing and DNAse I digest on a fresh mini-column.

### 2.2.4 Microarray analysis

**RNA amplification and labeling** 1 $\mu$g total RNA was amplified using the *aminoallyl MessageAmp kit* (Ambion) according to the manufacturer's protocol. In short, cDNA synthesis was carried out using an T7 oligo(dT) primer with two hours incubation time at 42°C followed by second strand synthesis at 16°C for two hours. cDNA purification was carried out using the provided filter cartridge. aRNA amplification was run overnight (~12hours) with a UTP:aaUTP ratio of 2:3. After purification on a provided filter cartridge, the aRNA concentration was measured, partitioned and dried using a SpeedVac©(Savant). aRNA was then coupled to DMSO dissolved Cy-dyes™(Cy3, Cy5; Amersham Bioscience) and excessive dye were quenched using Hydroxylamine. Labeled aRNA was purified on a provided filter cartridge and immediately used for microarray hybridization.

**Microarray hybridization** 3 $\mu$g of each Cy3/5 (Amersham) labeled amplified RNA was resuspended in 90 $\mu$l hybridization buffer containing 30% formamide (Microarray 4x Hyb buffer Amersham No. rpk0325) and then hybridized in a hybridization chamber (Scienion) overnight (minimum 16 hours) at 42°C under a coverslip with humidity maintained by hybridization buffer. After hybridization, slides were washed consecutively in 2xSSC/0.05%SDS, 0.2xSSC, 0.1xSSC for 5 minutes at room temperature (22°C) and then dried by centrifugation.

**Data acquisition** Scan data was acquired using a GenePix©3000 scanner (Molecular Devices). Photo-multiplier voltage was manually adjusted for each array and dye, so that less than 10 spots per microarray reach the measurement maximum. Image analysis was performed by AIDA software (Raytest, version 4.01). For each microarray a readout grid had to be manually placed to help the spot finding algorithm used by the software.

**Data analysis** All data analysis was done using software libraries of *Bioconductor*[59] in the statistical computing environment *R*[161] and in-house software based on the *limma* library[192]. Intra-array replicates were summarized using *Tukey's biweight* algorithm[71]. Systematic variation in the measured intensity levels was removed using printtiploess normalization[191]. Following normalization only 148 probes detecting splicing factors and 145 probes detecting intron-retention were further analyzed. ANOVA analysis (from the *limma* package) was performed using the factors *inflammation*, *IBD subtype* and *GI-region*. Specifically, each observation was modeled by

$$Y_{osir} = \sum_{sir}(C_{sir} * F_{sir}) + \epsilon \tag{1}$$

In an expression analysis framework, $Y_{osir}$ represents the observed log intensity of oligonucleotide $o$, measured in an experiment with disease subtype $s$, inflammation-status $i$ and colon region $r$. $C_{sir}$ depicts coefficients to be modeled for each different factor combination of disease subtype $s$, inflammation-status $i$ and colon region $r$ specified by $F_{sir}$. $\epsilon$ depicts a normally distributed error term. All questions of interest were being addressed by contrasts of $C_{sir}$ (after fitting the model). We extracted significant changes in expression using the R-package *limma*[192] at a cut-off of $p < 0.01$.

### 2.2.5 RT-PCR analysis

**cDNA synthesis** Total RNA was isolated from patient biopsies from patients as described in section 2.2.3. For each patient sample, cDNA was reverse-transcribed from 1 $\mu$g of total RNA in a total volume of 100 $\mu$L using the MultiScribe™Reverse Transcriptase reagents (Applied Biosystems). Components of the reverse transcriptase reaction were mixed on ice. As a negative control for the reverse transcriptase action, the above amounts were halved and transcriptase was omitted from the reaction. All reactions were prepared on ice and the incubated at 25°C for 10 min, 48°C for 30 min and 95°C for 5 min. The reaction was stopped by addition of 2 $\mu$L of 0.5 M EDTA. The success of the first strand cDNA synthesis was then checked by using 2.5 $\mu$L of the first stand synthesis as a template for PCR reaction with *GAPDH* as previously described in section 2.2.3. The amplification of the 101 bp *GAPDH* PCR product in the positive reverse transcriptase reaction, but not in the negative reverse transcriptase reaction, indicated successful first stand synthesis without contamination from genomic DNA. The final theoretical concentration of cDNA, assuming 100% efficiency of reverse transcription, was 10 ng cDNA/ $\mu$L.

**384 well plate production** cDNA prepared in section 2.2.5 was diluted 1:5 to a theoretical concentration of 2 ng/$\mu$L before being added to 384-deep well plates. For relative quantitation, standard curve cDNA was prepared from reverse-transcription of a mixture of total RNA obtained from inflamed IBD and normal colon mucosa to provide the greatest range of gene transcript expression. The standard curve was serially diluted in ten dilutions from 1:1 to 1:200 or 1:1000, in a total volume of 250 $\mu$L per deep well. For the real-time PCR reaction, 5 $\mu$L of the cDNA solution was added to the 384 well plate in quadruplicate: two wells for *ACTB* quantitation and two wells for target gene quantitation.

**Gene expression assays on ABI prism 7900HT** Real time PCR assays were carried out using, either commercial primers and fluorescent probes (Applied Biosystems: Assay-onDemand$^{TM}$) or self-designed primers and an intercalating fluorescent agent (Applied Biosystems: SYBR$^{TM}$-green). To render even spurious DNA contamination impossible all total RNA samples were digested by DNAse twice and checked by PCR for genomic contamination as described in section 2.2.3. Sufficient volumes of reaction mix were prepared in bulk and then dispensed by multipipet to 384-well PCR plates containing 5 $\mu$L of cDNA (2 ng/$\mu$L). All reagents were kept cooled on ice during the pipetting procedure. Plates were sealed with PCR film and stored at 4 °C or -20°C if they were not immediately run. Plates were briefly centrifuged and then run on an ABI 7900HT Sequence Detection System using the following thermo-cycling profile for Assay-on-Demand$^{TM}$Assays (CYBR$^{TM}$-green assays): 95°C for 10 min; and 40 (45) cycles of: 95°C for 15 s; 60°C (62°C) for 1 min. Assay-On-Demand$^{TM}$products used:

Table 4: Components for Real-Time$^{TM}$PCR

| Assay-on-Demand$^{TM}$ | | |
|---|---|---|
| | Reaction volume($\mu$L) | Final concentration |
| cDNA template (2ng/$\mu$L) | 5 | 1 ng/$\mu$L |
| 20x Assay-on-Demand$^{TM}$ | 0.5 | 1x |
| 2x TagMan© Universal PCR Master Mix | 4.5 | 0.9x |
| Total volume | 10 | |
| self-designed primers | | |
| | Reaction volume ($\mu$L) | Final concentration |
| cDNA template (2ng/$\mu$L) | 5 | 1 ng/$\mu$L |
| ? nM forward primer | 0.15 | x nM |
| ? nM reverse primer | 0.15 | y nM |
| 2x CYBR$^{TM}$-green PCR Master Mix | 4.5 | 0.9x |
| Total volume | 10 | |
| Concentration for self-designed primers | | |
| | forward primer (nM) | reverse primer (nM) |
| IER3 intron-excised form | 250 | 250 |
| IER3 intron-retained form | 125 | 125 |
| PARC intron-excised form | 50 | 50 |
| PARC intron-retained form | 125 | 50 |
| FGD2 intron-excised form | 250 | 250 |
| FGD2 intron-retained form | 125 | 125 |

| | | | | | |
|---|---|---|---|---|---|
| DUSP11 | Hs00186058_m1 | HNRPAB | Hs00258679_m1 | HNRPH3 | Hs00247221_m1 |
| SF3B14 | HS00255423_m1 | SFPQ | Hs00192574_m1 | SFR2IP | Hs00190882_m1 |
| SLU7 | Hs00197528_m1 | REG1A | Hs00602710_g1 | IL8 | Hs00174103_m1 |

**Detection of intron-retention** Primers for *SYBR-green* real time PCR were designed using *Primer3* [173] and sequence information from the Ensembl[79] database. To exclusively amplify either the intron-excised or the intron-retained form, we chose to design the reverse primer either across the exon junction (to catch the intron-excised form) or within the intron (to catch the intron-retained form) and the forward primer within the exon next to the intron of interest.

| Transcript | forward primer | reverse primer |
|---|---|---|
| PARC intron-retained form | 5’-caaaccctgctactcctgtgc-3’ | 5’-caacgtgtctctcagctcttgg-3’ |
| PARC intron-excised form | 5’-atactgaggggtgctcttctgc-3’ | 5’-gcgtgtggtaggcataggttc-3’ |
| FGD2 intron-retained form | 5’-caggttgccttgagtgattcc-3’ | 5’-gggagcaacacggagagg-3’ |
| FGD2 intron-excised form | 5’-cgcatccagagcagcg-3’ | 5’-ggatgtactccttgagcagcag-3’ |
| IER3 intron-retained form | 5’-gtgagtatcgccgaagtgg-3’ | 5’-gacaaaacaggagacaggtcagg-3’ |
| IER3 intron-excised form | 5’-ctcgagtggtccggcg-3’ | 5’-agggatgcggcgttagg-3’ |

**Data analysis** Upon completion of the run, the data was analyzed by the Sequence Detection System (SDS) 2.0 software using the following parameters:

1. threshold line set as low as possible such that the line intersects the amplification curves in the exponential phase yet any amplification in the non-template controls is reduced to negligible levels;

2. baseline set separately for housekeeper and target such that the baseline starts at three cycles and ends just before the first amplification curves rise above threshold;

3. extreme outliers in the standard curve eliminated from the calculation of the standard curve.

Upon adjusting all of these parameters, the SDS software automatically calculated the relative quantities, which were then exported as a tab-delimited text file for further data processing. Statistical analysis on comparisons of interest was carried out using the Wilcoxon rank sum test. P-values lower than 0.05 were considered significant.

### 2.2.6 PCR analysis

As template we pooled the total RNA from three randomly chosen individuals for IBD, UC, CD and healthy controls (see table 3). Primers for PCR were designed using *Primer3*[173] and sequence information from the Ensembl[79] database. To amplify the intron-retained form, we chose to design the reverse primer within the intron and the forward primer within the exon next to the intron of interest.

**Primers** Primers for IER3, PARC and FGD2 are described in section 2.2.5.

| Transcript | forward primer | reverse primer |
|---|---|---|
| YIPF3 | 5'-cttgtacgccaacatcgacatc-3' | 5'-tggtgtgctattattacttgtttcag-3' |
| GRM4 | 5'-ggcccaaagcaagtgtgg-3' | 5'-ctcgtacggacatcctcttgg-3' |
| TMEM63B | 5'-tgggtcacaacacacagatacc-3' | 5'-agctagcaagccatgaggtagg-3' |

**Setup** To reduce unspecific amplification the PCR was run using a touchdown thermo-cycling profile:
initiation:95°C for 10 min
touchdown: 20 cycles of: 95°C for 15 s; 72°C for 1 min; each cycle 0.5°C less down to 62°C
productive: 30 cycles of: 95°C for 15 s; 62°C for 1 min

# 3 Results

## 3.1 Establishment of an HLA microarray

To investigate the transcriptome of the HLA region on chromosome 6, the microarray was designed to query as many sequences as possible in order to provide comprehensive coverage of the HLA region.

### 3.1.1 Merging databases to enrich the HLA transcriptome

None of the publicly available transcriptome databases contained all transcript models since the process of genome annotation was far from being finished. Therefore, we maximized the sequence pool by merging different data sources. Four different databases (see figure 4) have been chosen. Most of the data was extracted from the Ensembl database assembly which harbored, at least in part, all four databases.

**Ensembl core database** was chosen to provide the framework because its features were easily accessible by computer scripts and it used a consistent coordinate system to merge transcript models of different databases.

**NCBI's Reference Sequences** were integrated because of their manually curated high quality transcript models.

**Sanger's Vega database** was incorporated because it extended the Ensembl core database with manual curation of selected chromosomal regions and

**Ensembl ESTgene database** was used because it provided a rich source of transcript models for unknown genes.

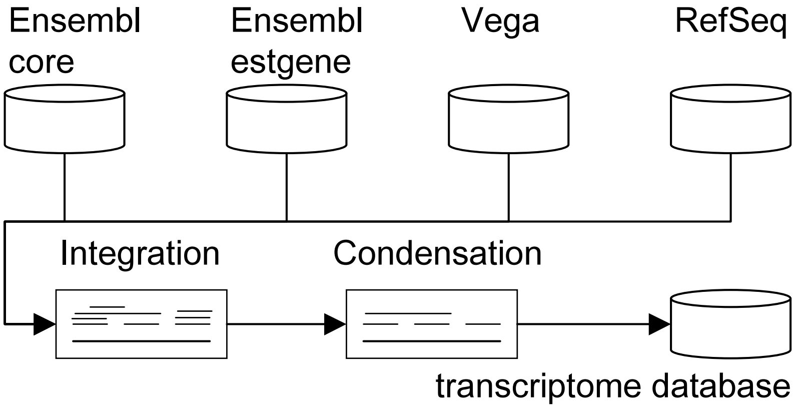


Figure 4: Work-flow to construct an extended transcriptome database. The core database from Ensembl was chosen to provide the framework because of its consistent coordinate and naming system.

Each database had its own transcript identifiers and efforts to cross reference the corresponding identifiers were still incomplete. This problem was facilitated by the fact that all databases were already mapped to one consensus version of the human genome called the *golden path*. Each transcript was then addressed by its coordinates on the *golden path*. We developed a three-step algorithm to merge the databases using the *golden path* coordinates of their transcript models.

The algorithm treated transcripts as sets of intervals and then applied set operations to find overlapping sets. As an empirical guideline, two transcripts formed a group of related transcripts (a gene) if they shared the sequence and structural information (exon-intron boundaries) of at least two exons. Figure 5 schematically depicts the merging algorithm. The first step assigned transcripts to the same model, based on the fact that they shared at least two common exons. The second step constructs a *meta* transcript model as union of all transcript models found related by step one. The *meta* transcript model is used in step three to gather transcripts missed in step one. In the rare case that a single transcript mapped to two different genes within one database scheme, the transcripts were manually resolved. Detailed description of the numerical cutoffs can be found in section 2.1.2.

Pooling the data sources introduced transcript redundancy, but also provided a depth of coverage of the transcriptome, which was not met by any of the source databases alone.

### 3.1.2 Down-sizing redundancy of the enriched transcriptome database

The redundancy created by pooling the data sources was down-sized by a second algorithm. We made a distinction between global and local redundancy and treated them separately.

Global redundancy is found with transcripts models, which have multiple copies of itself distributed in the genome. Local redundancy is based on (near-) identical transcript models from different databases. The algorithm to down-size redundancy is schematically depicted in figure 6. Global redundancy was only addressed, if the transcripts were nearly identical copies of each other. Our threshold used demanded sequence homology of 99% or greater and coverage of at least 95%. Given this case, one copy was masked.

Local redundancy was addressed in two ways. In the first instance we masked one

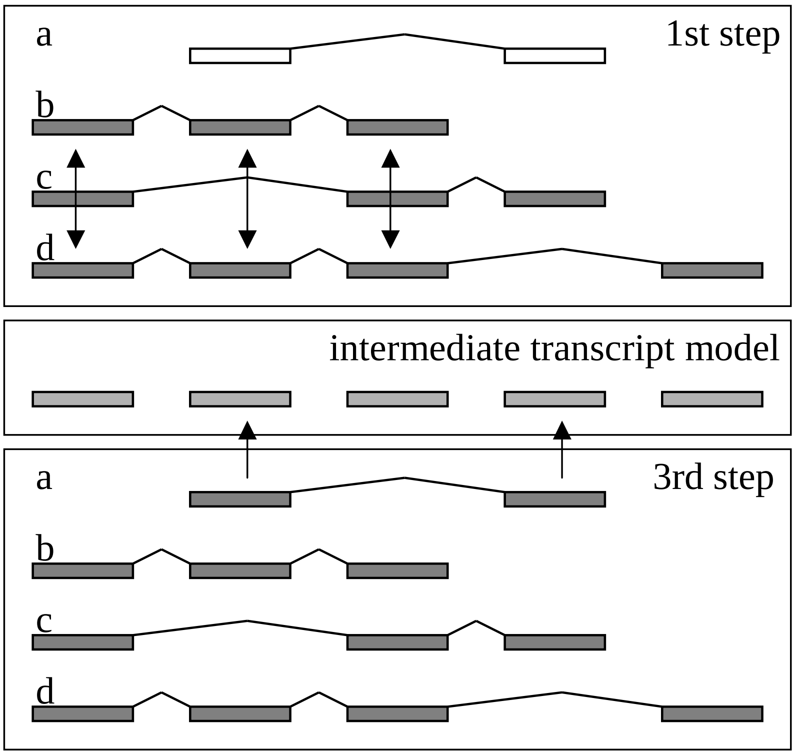


Figure 5: Algorithm to group transcripts from different databases to *genes* of the enriched transcriptome database in a three-step process. First; transcripts were assigned to each other if they shared exon sequence of at least two exons. If a transcript was assigned to two gene models within one databases scheme manual curation was needed so that the transcript belonged to only one gene model within one database scheme. In step two a transcript model was built as union of all exons of all transcripts which have been assigned to each other so far. The third step searched for transcripts, which had been missed in the first step. Again manual curation was needed if one transcript was assigned to two gene models within one database scheme. In the example given, the first step groups transcripts b to d. The second step assigns transcript a to transcripts b to d using the intermediate gene model.

copy in the database, if the two transcript models were identical. Two transcript models were assumed identical if they had the same number of exons and the length difference of each related exon pair was smaller than 6 bases. This *fuzzy* matching algorithm was applied to integrate the Ensembl ESTgene database, which was reported to have exon-boundary specificity of 30% only[47]. In the second instance, one transcript model was a subset of another transcript model. In this case, we had

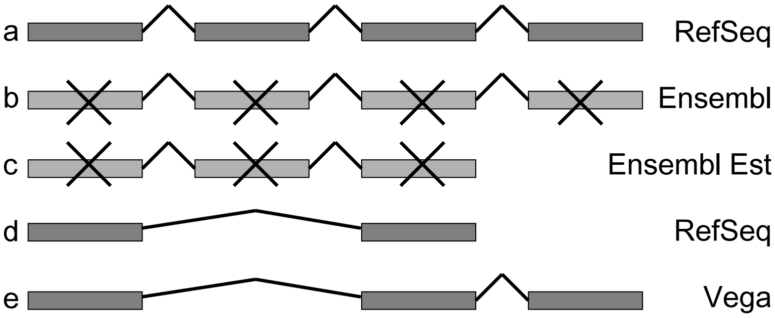


Figure 6: Algorithm to reduce local redundancy in the extended transcriptome database. Transcript models which were masked by the algorithm are highlighted by a cross. If two transcript models were identical (transcripts A and B) the algorithm masked one copy in the database. If one transcript model was a subset of another transcript model (C is a subset of B and D is a subset of E) and shared the exon-intron structure the algorithm masked the smaller transcript if it did not originate from a manually curated database (RefSeq, Vega). A transcript model with different exon-intron structure was kept in the database, even if its sequence was a subset of another transcript (D, E are subsets of A).

to distinguish wether the two models shared the exon-intron structure or had each a unique exon-intron structure. If a smaller transcript model shared the structure and sequence of a longer transcript model, it could either indicate different splicing or an incomplete transcript model. Our algorithm kept in those cases only transcripts of manually curated databases (RefSeq and Vega). A different exon-intron structures of two transcripts is characteristic of differential splicing and therefore both transcripts were kept in the database.

Both forms of redundancy were difficult to address, since there was a smooth transition from 100% identity of two transcripts/genes to transcripts/genes one would assume as different. To summarize the achievement of the *database merging* strategy used table 5 gives an overview of the gene and transcript counts in the extended HLA region and its coverage by the different databases and commercial microarray products.

### 3.1.3 The limits of microarray technology

The common feature of all types of microarrays is the measurement of mRNA concentrations by a specific hybridization of dye-coupled-cDNA (the target) to DNA arrayed on hydrophobic surfaces (the probe).

Table 5: Coverage of genes and transcripts in the extended HLA region

| Chromosome 6 extended HLA region (Mb20-50) | | |
|---|---|---|
| Database | transcript models | gene models |
| Our Assembly | 1653 | 1081 |
| Ensembl core 25.34 | 798 | 504 |
| Ensembl ESTgene (GP34) | 803 | 472 |
| Vega (GP34) | 705 | 438 |
| RefSeq 8 | 548 | 432 |
| Ensembl core 37.35 | 947 | 676 |
| Ensembl ESTgene (GP35) | 1131 | 551 |
| Vega 37.35 | 1571 | 779 |
| RefSeq 15 | 560 | 443 |
| Affymetrix HU133a2 | 631 | 402 |
| Clontech Atlas15K | 326 | 317 |

All types of microarrays have different properties concerning sensitivity and specificity. Kane et al.[87] investigated the limits of specific hybridization by evaluating the minimum homology needed for unspecific binding of target to probe. Their results suggest that unspecific binding is observed if two sequences exhibit more than 70% identity or have a common sequence interval of more than 20 base-pairs of 100% identity. Although short nucleotide probes should provide the greatest discrimination between related sequences[16], several studies reported that they had poorer hybridization yields[103,113,121,165]. Hughes et al.[80] studied the influence of various experimental parameters on the hybridization sensitivity and specificity. By testing oligonucleotides of several lengths, they showed that 60-mers represented the best compromise between sensitivity and specificity. Given these findings we chose to design a microarray based on oligonucleotide probes and to investigate the limits of specificity of oligonucleotide microarray technology.

An *all against all* sequence search was set up, using BLAT[89] software to investigate the specificity of the sequence of each transcript-model. The Ensembl database was used as source of transcript models because of its coherent nomenclature and ease of use.

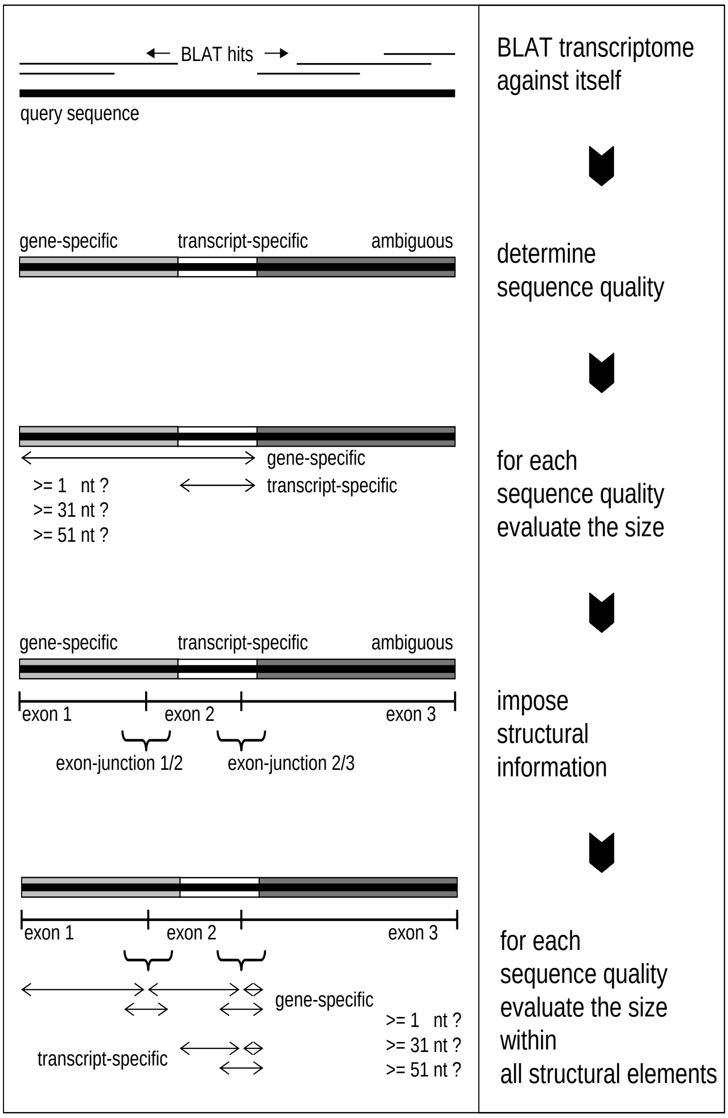


Figure 7: Algorithm to test the limits of oligonucleotide microarrays. A BLAT sequence search *all against all* was setup using Ensembl as data source. The algorithm discriminated three levels of specificity: transcript-specific, gene-specific and ambiguous. For each transcript the size of transcript-specific and gene-specific sequence was evaluated if it was sufficient to be used for probe design of length 25nt, 50nt or 70nt. Given the findings of Kane et al.[87] this translates to sequence windows needed of size 1, 31 and 51 nucleotides. To determine if probes of both specificity levels can be designed for each exon or exon-junction structural information about each transcript was added. For each exon and exon junction the size of transcript-specific and gene-specific sequence was evaluated if it was sufficient to be used for probe design of length 25nt, 50nt or 70nt.

Figure 7 schematically depicts the algorithm to determine the limits of microarray technology. Parsing the BLAT output, the algorithm developed discerned three levels of specificity: transcript-specific, gene-specific and ambiguous. A sequence was called transcript-specific if it was found only within one transcript-model of the source database. Within this line a sequence was called gene-specific if it was only found within transcript-models of one gene of the source database. If a sequence was found within multiple unrelated transcript-models it was masked as ambiguous. After resolving the specificity of the sequence of each transcript-model, we searched for transcript- or gene-specific sequence windows of three different sizes;

one, thirty-one and fifty-one bases. The window of size one was chosen to represent 25mer oligonucleotides, as used by Affymetrix$^{\text{TM}}$microarrays. Here, we assume that a single unique base was sufficient to discern two sequences which were highly similar. The other window sizes (31 and 51) were used to explore the limits of long oligonucleotide arrays with probe lengths of fifty and seventy nucleotides, respectively. The algorithm assumed that an unspecific stretch of sequence smaller than twenty bases could be neglected, in line with the findings of Kane et al.[87]. Figure 8 summarizes our findings on the sequence properties of transcript models. About 83% of all transcript models had at least one transcript-specific base, while only 51% (49%) had a transcript-specific stretch of at least 31 (51) bases. Lowering the specificity constraints by including sequences that were only gene-specific, increased the numbers of representable transcript models to 95% (79/78%) for sequence windows of size one (31/51).

To deepen this analysis, we imposed structural information on the transcript-models and checked each exon and each exon-junction separately for specific sequence windows of size one, 51 and 71. As is shown in figure 8 only 21% of all exons have had at least one transcript-specific base, while only 13% (12%) of all exons included a transcript-specific stretch of 31 (51) bases.

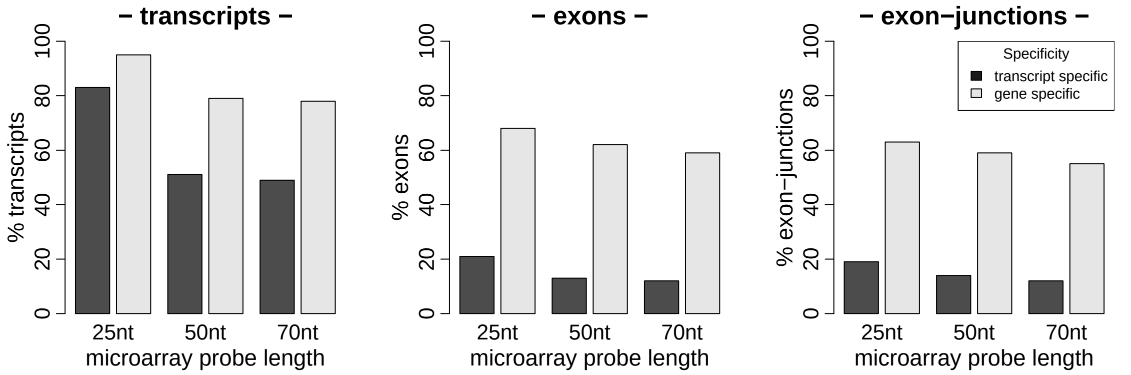


Figure 8: Limits of oligonucleotide microarrays in addressing transcripts, exons and exon-junctions. Depicted is the fraction of all transcripts, exons and exon-junctions where oligonucleotide probes of selected specificity and length could be designed.

The results of our analysis on exon-junctions were very similar to our findings on exon sequence. Gene-specific sequences were found for 68(62/59)% of exons. Again the numbers for gene-specific exon-junctions were very similar (63/59/55%). Therefore, our results confirmed the finding that the shorter the probe-length the more transcripts can be investigated. Another conclusion was that the fraction of transcripts that can be investigated by a technology using hybridization never reaches 100%. This constraint is even stronger if one wanted to discern the expression levels of exons and/or exon-junctions.

Our findings proposed that it is reasonable to address gene expression at transcript level but not at the level of exons or exon-junctions.

### 3.1.4 Probe placement strategies

The key to surveying the structural fluidity of the transcriptome by differential splicing is to address as many transcript variants of a gene as possible using clever probe placement strategies. It is likely that the number of detected exon-intron boundaries increases by time, so the only strategy to maximize the number of addressable transcript variants is to use tiling arrays, where probes are designed every $n$ bases along a chromosome. However, the limits of specificity will render many probes useless because probes may match repetitive sequence for instance and the costs of this strategy would be extremely high.

We chose to only compare probe placement strategies, which depend on *a priori* knowledge about exon-intron boundaries present. Different probe placement strategies can be envisioned, as depicted schematically by figure 9.

The strategies differ by the required number of probes, the depth of coverage, and total cost per gene. The first strategy placed a probe on each exon. Strategy number two addressed every exon-junction known so far. The third strategy investigated only the exon-junctions needed to distinguish between known transcript variants. Strategy number four placed one probe which addressed the maximum number of known transcripts and one probe for each transcript with transcript-specific sequence. The technical limits of oligonucleotide microarrays (shown above) constrained our aims to reasonable goals.

To compare the different probe design strategies, we chose to calculate four parameters as measurements of placement strategy fitness. First, the total number of addressed transcripts was calculated to investigate if the transcription level of the whole gene could be integrated from all measurements. Second, the number of transcripts addressed by transcript-specific probes was counted. Third, the number

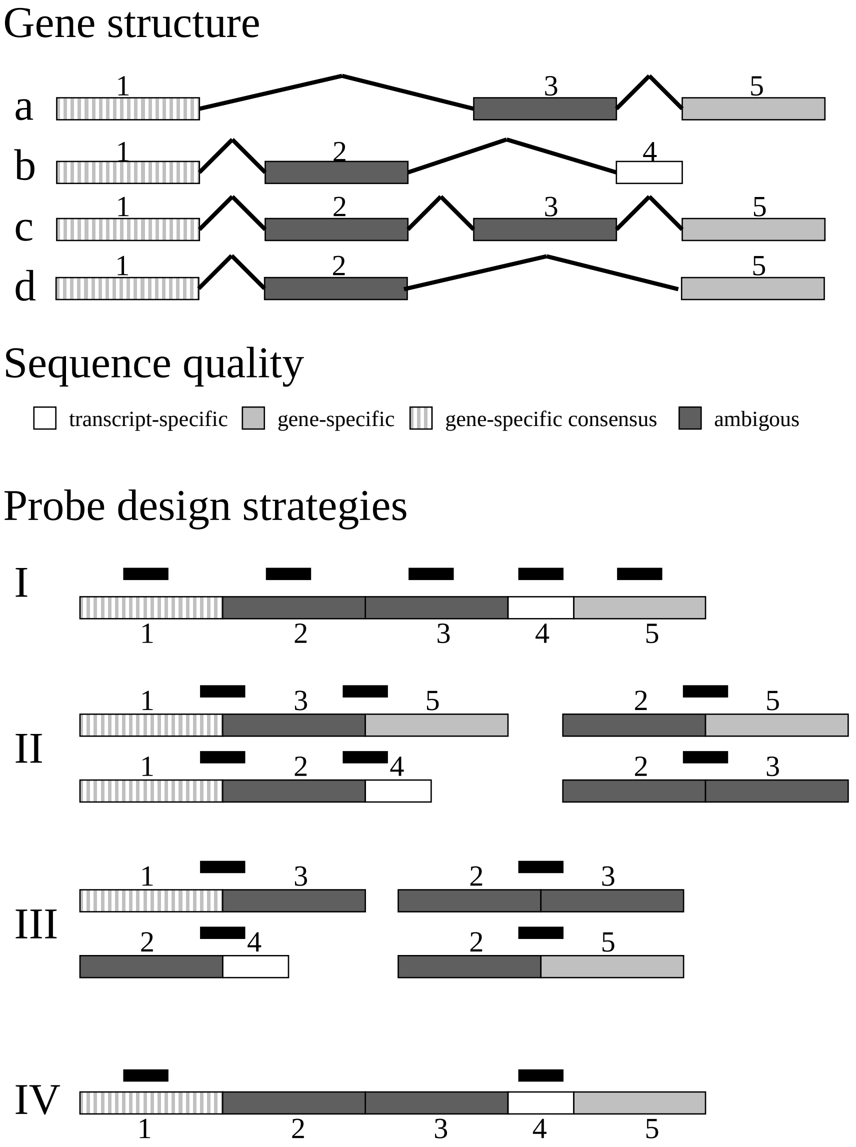


Figure 9: Design strategies for microarray probes. The exon-intron structure of four (a-d) different transcripts of a prospective gene is depicted with exons shown as boxes and introns as edges. Specificity of the sequence is shown by color coding. The percentage of each specificity represent our results from section 3.1.3. The lower part depicts oligonucleotide probe design strategies to distinguish between different transcripts. Exons are shown with numbering and color coding according to the upper part. Suggested locations of probes are shown as black bars. The first strategy designs a probe for each exon. Strategy number two addresses every exon-junction known so far. The third strategy investigates only the exon-junctions needed to distinguish between known transcript variants. Strategy number four designs one probe which addresses the maximum number of known transcripts and one probe for each transcript with transcript-specific sequence.

of exons addressed, being at least gene-specific, was metered. And fourth, the count of probes needed for each strategy was assessed to represent the costs. To take the robustness of the design strategies into account, parameters one to three were calculated as an average over all *leave one out* possible scenarios. For instance, if in strategy IV the probe on exon one fails parameter one and three decline to one, while parameter two stays at value one. If the probe on exon four fails in strategy IV, parameter two and three decline to zero and one respectively, while parameter one stays at value one. Calculating the average over these *leave one out* scenarios, quantity one to three resulted in values 2, 0.5 and 1 respectively.

To decide which strategy was the most efficient the four quality parameters were integrated. The sum of (robust) values for parameters one to three was calculated and then divided by the costs (parameter four). As is summarized by table 6 strategy number IV was the most efficient according to the selected criteria.

Table 6: Discerning the efficiency of all probe design strategies. Three coverage parameters as measurements of design strategy fitness and one cost parameter were calculated. 1) coverage of all transcripts. 2) coverage of transcripts with transcript-specific sequence. 3) coverage of gene specific exons. 4) costs as number of probes needed. Robustness of the design strategies was integrated by calculating parameters one to three as average over all *leave one out* scenarios possible. Integration of design strategy fitness values was done taking the sum of parameters one to three and dividing by the costs (four).

| Strategy: | I | II | III | IV |
|---:|---:|---:|---:|---:|
| Transcripts addressed | 3.20 | 4.00 | 3.00 | 2.00 |
| Transcripts addressed specifically | 0.80 | 0.83 | 0.75 | 0.50 |
| Gene-specific Exons addressed | 2.40 | 2.83 | 2.25 | 1.00 |
| Costs (probes needed) | 5.00 | 6.00 | 4.00 | 2.00 |
| Efficiency | 1.28 | 1.28 | 1.50 | 1.75 |

### 3.1.5 Probe design and gene selection

Taking the above considerations into account, we decided work with long oligonucleotide probes applying probe placement strategy IV (see figure 9). Since uniform length of oligonucleotide probes lead to high discrepancies between oligonucleotide melting temperatures, and subsequently to high differences in hybridization behaviors[7,194], we chose to use software developed by Haas et al.[63], which designs oligonucleotide probes of variable length by taking their melting temperature into account. Our target melting temperature of 62°C lead to an average probe length of 57nt±7nt.

In total, 1902 microarray probes were designed as part of this thesis, including 1175 probes covering 30 Megabases on chromosome 6 from 20Mb to 50Mb. The specificity of probes selected to cover transcripts and / or genes within the HLA region on chromosome six was assessed using the enriched transcriptome database. Our design yielded 824 transcript-specific probes and 351 gene-specific probes. The probes designed were able to cover transcripts from 752 genes, 69% of all genes of the HLA region in the enriched database. We were able to design probes for the following 24 HLA genes :

Table 7: HLA molecules covered by the chromosome 6 microarray. For genes displayed with an asterix, no class could be assigned.

| HLA class I | | HLA class II | |
|---|---|---|---|
| Gene | probe specificity | Gene | probe specificity |
| HLA-A | gene-specific | CD74 | gene-specific |
| HLA-B | transcript-specific | HLA-DMA | transcript-specific |
| HLA-C | transcript-specific | HLA-DOA | gene-specific |
| HLA-E | transcript-specific | HLA-DOB | gene-specific |
| HLA-F | transcript-specific | HLA-DPA1 | gene-specific |
| HLA-G | transcript-specific | HLA-DPA2 | transcript-specific |
| HCP5 | transcript-specific | HLA-DPB1 | gene-specific |
| HCG27* | transcript-specific | HLA-DQA1 | transcript-specific |
| HCG9* | transcript-specific | HLA-DQB2 | transcript-specific |
| HCG18* | transcript-specific | HLA-DRA | transcript-specific |
| HCG22* | transcript-specific | HLA-DRB1 | transcript-specific |
| HCG25* | transcript-specific | HLA-DRB5 | transcript-specific |

The HLA chip containd 6011 different probes measuring transcripts or genes and 30 probes which serve as internal controls. 4110 probes were designed in cooperation with Scienion AG (Berlin). Genes outside the HLA region on chromosome 6 were selected for their implication in various functional settings such as the immune system, signal transduction, splicing or ubiquitination. The genes were included to supplement the functional context of the measurements on chromosome 6 genes.

### 3.1.6 The HLA chip – quality control

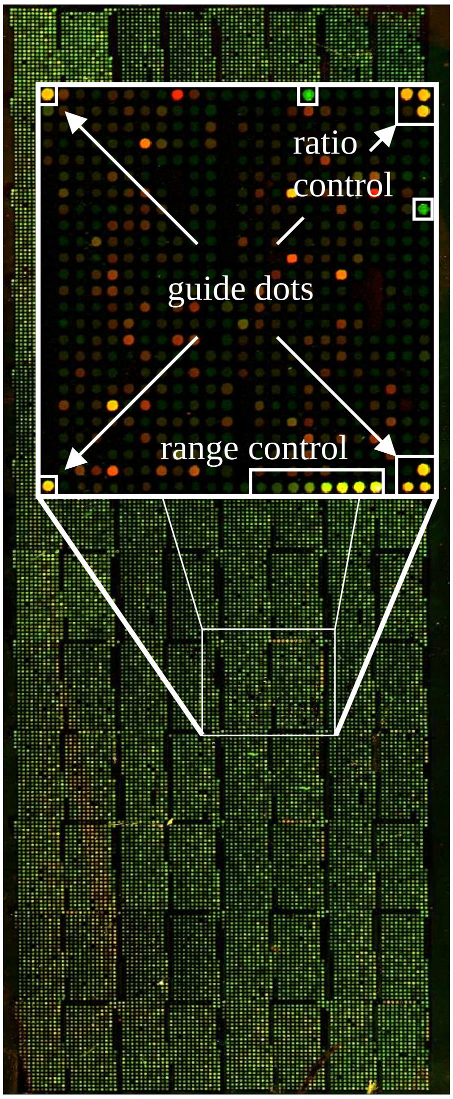


Figure 10: HLA chip random primer hybridization to control drop-out-rate. Blow up shows guide dots, range and ratio controls.

In addition to 6011 different probes measuring transcripts or genes the HLA chip contained 30 probes which serve as internal controls. All transcript/gene probes were spotted four times which enabled the calculation of array and spot effects. To address spotting uniformity, labeled random primer hybridizations were carried out, as exemplarily visualized in figure 10. Typical drop-out rate was found below 5 per 1000 spots (data not shown). With the help of Claus Hultschig, the spotting grid design contained 240 *guide dots* (made of probes for housekeeping genes *ACTB* and *GAPDH*), which proved to be essential for proper function of the spot quantitation software. Figure 10 gives a visual impression of the guide dots at the corners of the blow-up square.

The internal Scorecard (Amersham) controls were used to determine the dynamic range and to control the ratio measured by the HLA-chip (figure 10). As data source an experiment with 60 HLA microarrays conducted as described in section 3.2 was used. *Spike In* Controls (Amersham: Scorecard$^{TM}$) were used to represent different copy numbers. While the dynamic range shown in figure 11 was judged fairly well (3 - 4 orders of magnitude) the ratio controls shown in figure 11 displayed clear signs of quenching of ratios. Although all artificial ratios were observed to always show the right tendency after normalization, the size of the observed log ratios were only 20% the size of the expected log ratios, which translates into a quenching factor of ∼5. The quenching factor for log ratios was found nearly independent of *spike in* copy number/abundance with a factor of 5 for low abundance spikes and a factor of 4.2 for high abundance spikes.

We used microarray data and expression data generated by real-time PCR (experiment: section 3.2) of nine genes to check if the same quenching of ratios is reproducible with *real* genes. As can be calculated from table 8 mean quenching factor found was 4.1 (SD: 3.4), only marginally lower than the quenching factor of the ratio controls.

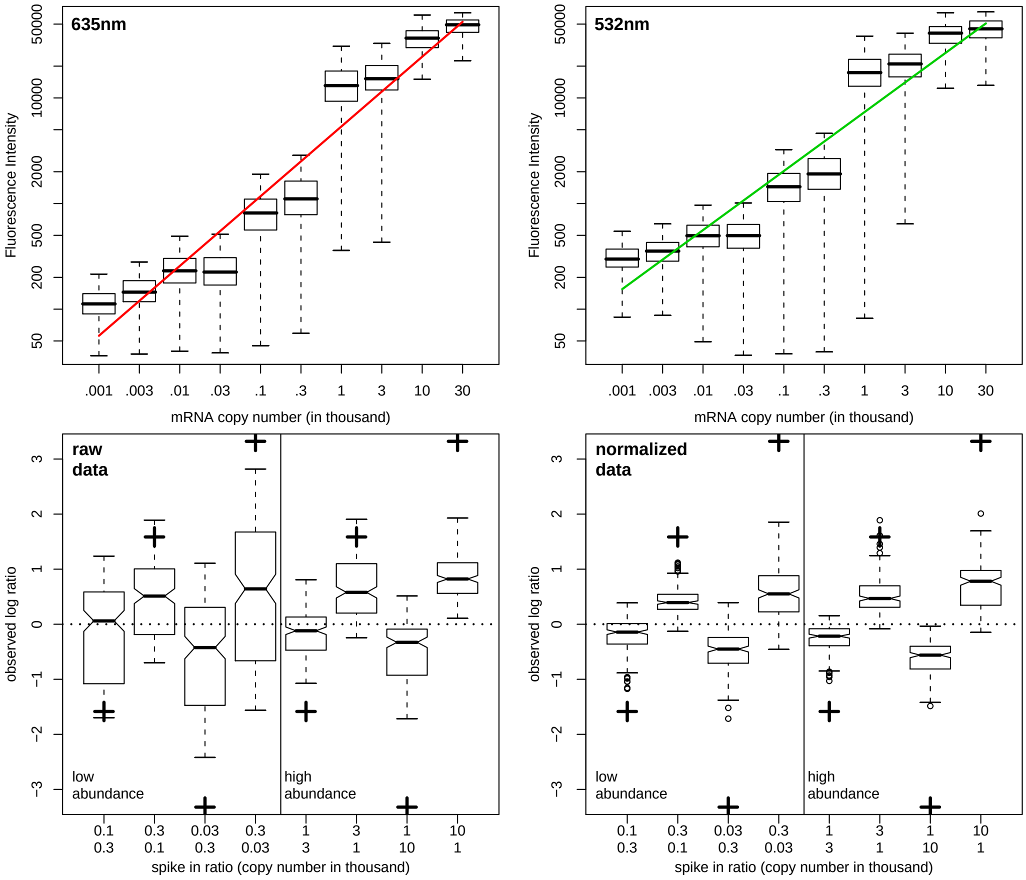


Figure 11: Dynamic range and ratio controls of the HLA chip. Dynamic range is displayed on a logarithmic scale for signal intensities at two wavelengths (532nm and 635nm). *Spike In* Ratio Controls showed quenched ratios in raw and normalized data. Artificial ratios were introduced at high and low concentrations. The expected log ratio is schematically depicted by a plus sign for each box-plot.

Although the underestimation of ratios did not affect the statistical analysis it was necessary to consider this in the interpretation of the results. The average Pearson correlation between technical replicates on each array was calculated on raw and normalized data (section3.2) and was found to be fairly good (median pearson = 0.93 (raw) and median pearson = 0.96 (normalized), n = 60 arrays).

Table 8: Ratio comparison of microarray data to real-time PCR data. Log$_2$ratios are taken from section 3.2

| | DUSP11 | HNRPAB | HNRPH3 | SLU7 | SF3B14 | SFR2IP | SFPQ | IL8 | REG1A |
|---|---|---|---|---|---|---|---|---|---|
| log$_2$-ratio Array | -0.36 | -0.43 | -0.28 | -0.26 | -0.27 | -0.22 | -0.33 | 0.83 | 1.30 |
| log$_2$-ratio RTPCR | -0.67 | -0.49 | -0.70 | -0.77 | -0.57 | -0.98 | -0.67 | 7.21 | 13.9 |
| quenching factor | 1.85 | 1.15 | 2.49 | 2.95 | 2.10 | 4.43 | 2.03 | 8.79 | 10.7 |

### 3.1.7 Detection of intron-retention

Transcript models change or vanish in the process of dynamic genome annotation. In particular, transcript models derived solely from EST sequences were subject to change. In this process, some oligonucleotide probes designed for EST-derived transcript models lost *their* transcripts. These probes can nevertheless be useful to measure intron-retention, a splicing process of growing interest. A signal from the probes measuring intronic sequence can be compromised three fold. Genomic contamination, cross-hybridization and unknown exons might lead to signals misleadingly classified as intron-retention. While cross-hybridization was ruled out by including only probes that match a single locus in the genome and genomic contamination was controlled by stringent quality controls on total RNA, the possibility for unknown exons still remained. We qualitatively graded the supporting evidence for present intron-retention using the genome browser at UCSC[88,90]. Evidence for intron-retention was rated strong if the supporting sequences bridged the whole intron. An inter-

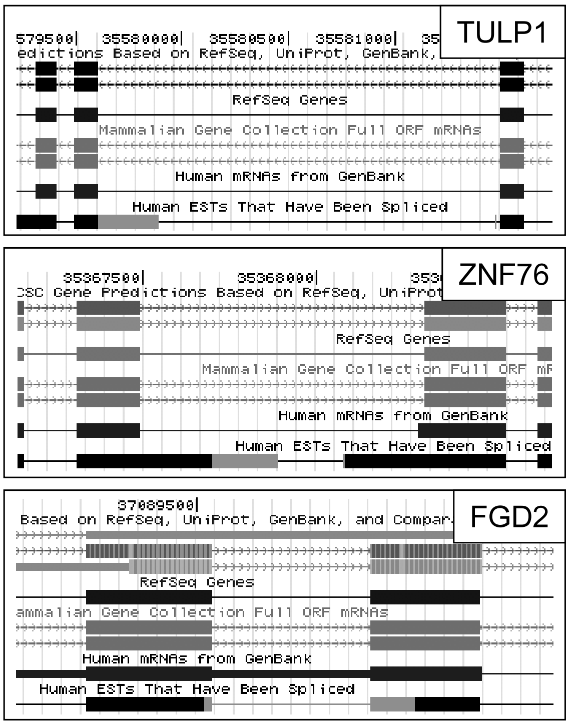

TULP1
579500| 35580000| 35580500| 35581000| 35
edictions Based on RefSeq, UniProt, GenBank,
RefSeq Genes
Mammalian Gene Collection Full ORF mRNAs
Human mRNAs from GenBank
Human ESTs That Have Been Spliced
ZNF76
35367500| 35368000| 353
CSC Gene Predictions Based on RefSeq, UniProt
RefSeq Genes
Mammalian Gene Collection Full ORF mR
Human mRNAs from GenBank
Human ESTs That Have Been Spliced
FGD2
37089500|
Based on RefSeq, UniProt, GenBank, and Compar
RefSeq Genes
ammalian Gene Collection Full ORF mRNAs
Human mRNAs from GenBank
Human ESTs That Have Been Spliced

Figure 12: Supporting evidence for intron-retention in the UCSC databases.

mediate rating was assigned, if intron-retention was only partially supported by EST/RNA sequences. If the supporting evidence consisted of only one EST, a weak rating was assigned. Figure 12 gives an example of each level of supporting evidence for intron-retention found in the UCSC database.

Of 145 events, seventy-four (51%) showed strong, forty-eight (33%) intermediate and twenty-three (16%) weak evidence for intron-retention in the UCSC database. We experimentally verified present intron-retention in six genes using PCR and three randomly chosen samples from each subgroup of patients (table 3) as a template. The possibility of a hidden exon was ruled out in three out of six cases with only weak evidence for an unknown exon in two cases, however, intron-retention was detected in all six genes.

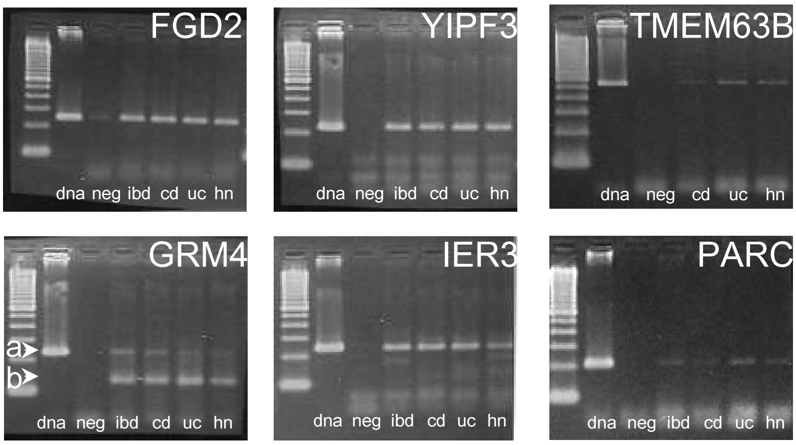


Figure 13: Supporting evidence for intron-retention in FGD2, YIPF3, TMEM63B, GRM4, IER3 and PARC as found by PCR. In case of an unknown exon a PCR product smaller than the positive control (template = DNA) was expected. For instance the gene GMR4 most likely contained an unknown exon (size b) and intron-retention (size a) since two PCR products were observed. For all genes, the data supports present intron-retention.

## 3.2 Splicing Factors and Intron-Retention in Inflammatory Bowel Disease

This study presents a systematic analysis of differential expression of 149 splicing factors and 145 intron sequences in IBD. We use the term *splicing factor* for genes and transcripts that are known to interact and modulate splicing events. Table 15 groups the splicing factors investigated based on classification by Black et al. [15,233].

### 3.2.1 Experimental setup and design

To study potential changes in gene expression in IBD we analyzed mucosal biopsies taken from the gastrointestinal (GI) tract. Gene expression in mucosal epithelial cells is influenced by a number of factors. To exclude gender effects we chose to analyze male patients only, because the genetic association of the HLA region to IBD is more pronounced in males as can be seen in figure 2. Our main focus of analysis was to discern Crohn's disease from ulcerative colitis concerning inflammatory patterns of gene expression. We therefore studied the transcriptome from mucosal biopsies taken from inflamed and non-inflamed regions of the GI-tract of both patient groups.

Table 9: Biopsy stratification by biological factor combinations. Each tissue sample is described by a three letter code. The first letter (N, U, C) depicts disease status: Normal/Healthy, CD, UC; the second letter (n, i) depicts inflammation status: inflamed, non-inflamed and the third letter (S, T) depicts GI-region (Sigma, T.ileum).

| | **disease** | | | **inflammation** | | **GI-region** | |
|---|---|---|---|---|---|---|---|
| Code | healthy | CD | UC | inflamed | non-inflamed | sigmoid colon | terminal ileum |
| NnS | x | | | | x | x | |
| NnT | x | | | | x | | x |
| CnS | | x | | | x | x | |
| CiS | | x | | x | | x | |
| CnT | | x | | | x | | x |
| CiT | | x | | x | | | x |
| UnS | | | x | x | | x | |
| UiS | | | x | | x | x | |
| UnT | | | x | x | | | x |
| UiT | | | x | | x | | x |

Table 10: Biological questions to evaluate experimental design strategies.

| Code | Question |
|---|---|
| **Dis** | Difference of disease subgroups (CD vs UC) |
| **Infl** | Effect of inflammation (infl. vs non-infl.) |
| **DisInfl** | UC or CD specific effects of inflammation |
| **Reg** | Difference of GI-regions |
| **RegDis** | GI-region specific difference of CD vs UC |
| **RegInfl** | GI-region specific effect of inflammation |
| **IBDniN** | Comparison of non-inflamed IBD samples to healthy controls |
| **IBDiN** | Comparison of inflamed IBD samples to healthy controls |

To investigate disease effects, which do not manifest in inflammation, we also included biopsies from a group of healthy individuals. The region of sampling also influences gene expression levels, as many genes have been shown to exhibit tissue specific transcription levels. The clinical observation that ileal inflammation is found in ulcerative colitis patients basically as secondary inflammation called backwash ileitis (Schreiber, personal communication) adds to the importance of GI-region specific measurements. In studying biopsies from the sigmoid colon and the terminal ileum our experimental setup not only accounted for regional differences in gene expression, but would also allow to discern backwash ileitis from ulcerative colitis.

Taken together, the selection of patients excluded the study of gender effects but included effects of disease group (CD or UC), inflammation (inflamed, non-inflamed) and GI-region (sigmoid colon, terminal ileum). Since healthy individuals do not exhibit mucosal inflammation, patients for ten different factor combinations had to be found. Table 9 depicts the factor combinations addressed in this study. Patient samples, which fulfilled the stringent selection criteria, were scarce and the amount of total RNA harvested from each biopsy allowed for analysis on four microarrays only. Additionally, only cyclic designs are able to compare many individuals, while controlling for dye bias.

Experimental design strategies, which fit these constraints had to be found and compared concerning their efficiency to answer a number of biological questions depicted in table 10. In an ANOVA setup the partitioning of variance and degrees of freedom per biological question and experimental design can be used to compare dif-

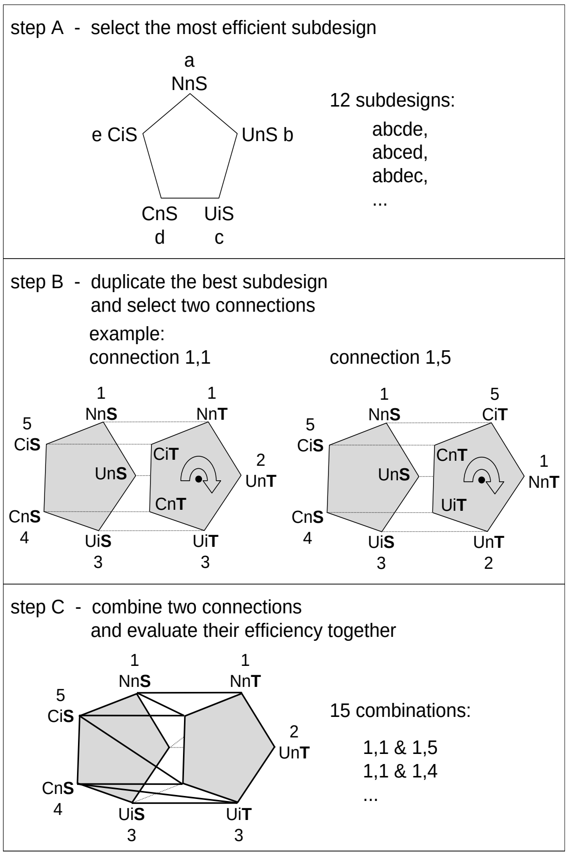


Figure 14: Strategy to find an efficient experimental design. **A** the design-space for only five factor combinations (table 9: sigmoid colon) of one GI-region was explored, using biological questions (table 10) without GI-region effects. Letters a to e depict the order of factor combinations within the cyclic sub-designs. Each edge depicts a microarray experiment comparing two factor combinations. **B** the best sub-design was used twice, once for each GI-region. Connections between the two GI-region specific sub-designs had to be found. Here factor combinations are coded as numbers to describe the sub-design connections. **C** Each factor combination had to be connected twice to its own GI-region and twice to the other GI-region. Hence, sub-design connections were evaluated in groups of two.

ferent experimental design strategies[20,106]. As pointed out by Kerr and Churchill[91] in the context of fixed models, the indirect comparison design (e.g. common reference) leads to a higher variance of the statistic as the direct comparison designs, i.e., it is less efficient than the direct comparison designs. We used the R package *daMA*[20,106] to evaluate the efficiency for each design strategy.

The overwhelming number ($\frac{n!}{2*n}$ =181440) of distinct cyclic undirected graphs form the design-space of ten factor combinations (table 9) and motivated us to divide the problem into three steps. First we compared the efficiency of 12 subdesigns of the experimental-design-space for the five factor combinations of only one GI-region, using our biological questions from table 10 that do not include GI-region effects. In step two the best sub-design found was applied to the factor combinations for each GI-region and hence duplicated. In step three connections between the duplicated sub-design had to be found, which connect both GI-regions. We searched for connection combinations so that each factor combination connects to two factor combinations of the other GI-region. Figure 14 schematically depicts our strategy used to find the most efficient experimental design. All efficiencies were calculated

Table 11: Relative efficiency of 12 sub-designs concerning 5 biological questions (Table 10). Letters a to e depict the order of factor combinations within the cyclic sub-designs (see figure 14).

| Sub-design | biological questions | | | | |
|---|---|---|---|---|---|
| Code | Dis | Infl | DisInfl | IBDniN | IBDiN |
| abecd | 1.00 | 0.60 | 0.43 | 1.00 | 0.50 |
| acbde | 0.43 | 0.60 | 1.00 | 0.50 | 1.00 |
| acdbe | 1.00 | 0.60 | 0.43 | 0.50 | 1.00 |
| abced | 0.43 | 0.60 | 1.00 | 1.00 | 0.50 |
| acebd | 1.00 | 0.43 | 0.60 | 0.63 | 0.63 |
| abcde | 0.43 | 1.00 | 0.60 | 0.71 | 0.71 |
| abdce | 1.00 | 0.43 | 0.60 | 0.63 | 0.63 |
| acbed | 0.43 | 1.00 | 0.60 | 0.71 | 0.71 |
| abedc | 0.60 | 1.00 | 0.43 | 0.71 | 0.71 |
| adbce | 0.60 | 0.43 | 1.00 | 0.63 | 0.63 |
| acedb | 0.60 | 0.43 | 1.00 | 0.63 | 0.63 |
| adcbe | 0.60 | 1.00 | 0.43 | 0.71 | 0.71 |
| CR | 0.30 | 0.30 | 0.30 | 0.33 | 0.33 |

Table 12: Relative efficiency of 15 experimental sub-design combinations concerning 8 biological questions. Coding of sub-design connections is described in figure 14

| Connection | Biological questions | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Code | Dis | Infl | DisInfl | Reg | RegDis | RegInfl | IBDn | IBDi |
| 1: 1,1&1,1 | 0.40 | 0.29 | 0.50 | 1.00 | 1.00 | 1.00 | 0.38 | 0.38 |
| 2: 2,2&2,2 | 0.59 | 0.39 | 0.90 | 0.88 | 0.50 | 0.50 | 0.54 | 0.54 |
| 3: 3,3&3,3 | 0.86 | 0.85 | 0.66 | 0.81 | 0.37 | 0.28 | 0.86 | 0.86 |
| 4: 4,4&4,4 | 0.86 | 0.85 | 0.66 | 0.81 | 0.37 | 0.28 | 0.86 | 0.86 |
| 5: 5,5&5,5 | 0.59 | 0.39 | 0.90 | 0.88 | 0.50 | 0.50 | 0.54 | 0.54 |
| 6: 1,1&1,2 | 0.56 | 0.39 | 0.73 | 0.96 | 0.78 | 0.79 | 0.52 | 0.52 |
| 7: 1,1&1,3 | 0.71 | 0.63 | 0.62 | 0.95 | 0.72 | 0.65 | 0.70 | 0.70 |
| 8: 1,1&1,4 | 0.71 | 0.63 | 0.62 | 0.95 | 0.72 | 0.65 | 0.70 | 0.70 |
| 9: 1,1&1,5 | 0.56 | 0.39 | 0.73 | 0.96 | 0.78 | 0.79 | 0.52 | 0.52 |
| 10: 1,2&1,3 | 0.81 | 0.63 | 0.87 | 0.88 | 0.50 | 0.43 | 0.78 | 0.78 |
| 11: 1,2&1,4 | 0.91 | 0.78 | 0.83 | 0.90 | 0.56 | 0.53 | 0.89 | 0.89 |
| 12: 1,2&1,5 | 0.80 | 0.57 | 1.00 | 0.93 | 0.63 | 0.73 | 0.75 | 0.75 |
| 13: 1,3&1,4 | 1.00 | 1.00 | 0.75 | 0.85 | 0.44 | 0.34 | 1.00 | 1.00 |
| 14: 1,3&1,5 | 0.91 | 0.78 | 0.83 | 0.90 | 0.56 | 0.53 | 0.89 | 0.89 |
| 15: 1,4&1,5 | 0.81 | 0.63 | 0.87 | 0.88 | 0.50 | 0.43 | 0.78 | 0.78 |

as relative efficiencies. Relative efficiency values are obtained by dividing the absolute efficiency values by the maximal value for a given question. Table 11 and table 12 contain the results of step one and three of the experimental design finding process explained in figure 14.

To decide, which experimental design strategy to follow we had to prioritize our biological questions, since each experimental design strategy favored certain biological questions. In the context of IBD our question of outstanding interest was to discover disease subtype specific gene expression changes due to inflammation. As a consequence, we chose the design with maximum efficiency regarding this question and thus the design was constructed from sub-design 11 by connection combination 1,2&1,5. Figure 15 schematically depicts the experimental design used in this study.

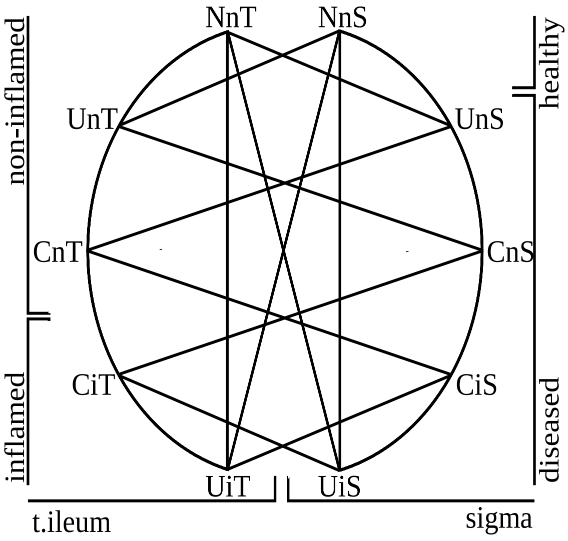


Figure 15: Experimental design used to extract the effetcs of Inflammation, Disease and GI-region on gene expression. Each biopsy sample is described by a three letter code. The first letter (N, U, C) depicts disease status: Normal/Healthy, CD, UC; the second letter (n, i) depicts inflammation status: inflamed, non-inflamed and the third letter (S, T) depicts GI-region (Sigma, T.ileum). Each edge depicts a microarray experiment that compares two nodes (e.g. UnS [UC non-inflamed Sigma] against CnS [CD non-inflamed Sigma]). This design was repeated three times.

### 3.2.2 Quality Control of biopsy cell type and inflammation status

The mucosa is a heterogeneous cell population composed of immune and non-immune cells. Marker genes for epithelial cells and leukocytes indicate that the gene expression profiles of the analyzed biopsy specimens mainly represent the gene expression of epithelial cells and that immune cells do not significantly contribute to the observed transcript patterns (table 13).

The presence of inflammation in the second patient group (table 3) samples, which were used for real-time PCR quantification was confirmed by quantifying the inflammation markers IL8 and REG1A in these samples (Figure 16). Apparently *IL8* ex-

Table 13: Control genes of cell types represented in biopsy material. The control expression values are given for the terminal ileum and the colon. These numbers represent average fluorescence values ($\log_2$ scale) across 60 arrays. Genes with an average fluorescence intensity of 150 were considered as significantly expressed.

| Cell type/gene | Average expression sigmoid colon | Average expression terminal ileum |
|---|---|---|
| Epithelial markers[21] | | |
| E-cadherin | 322 | 254 |
| Villin 1 | 476 | 567 |
| Villin 2 | 354 | 348 |
| Laminin $\beta$1 | 1284 | 1147 |
| Mucin 13 | 1054 | 976 |
| EPLIN | 1248 | 825 |
| Leucocyte markers[54] | | |
| CD16 | 150 | 161 |
| CD18 | 180 | 185 |
| CD80 | 104 | 125 |
| CD86 | 201 | 179 |
| Mesenchymal markers[21] | | |
| CD31 | 179 | 205 |
| CD11c | 107 | 114 |

pression showed strong concordance (85%) with the endoscopic observation of acute inflammation, while the expression pattern of *REG1A* was more complex. Elevated levels of *REG1A* expression was also found in non-inflamed tissue, which might be explained by a third tissue state of regenerating but non-inflamed tissue.

### 3.2.3 Altered expression of splicing factors in inflamed mucosa of IBD patients

Using the HLA chip, we found thirty-three splicing factors differentially regulated in IBD patients comparing inflamed tissue to non-inflamed tissue. The Sm and Sm-like (LSm) protein transcripts showed a strong association to inflammation ($p = 0.002$) with four out of five investigated transcripts being up-regulated in inflamed IBD tissue as compared to non-inflamed IBD tissue. Comparing the expression fingerprint of healthy control subject biopsies to non-inflamed biopsies and inflamed biopsies

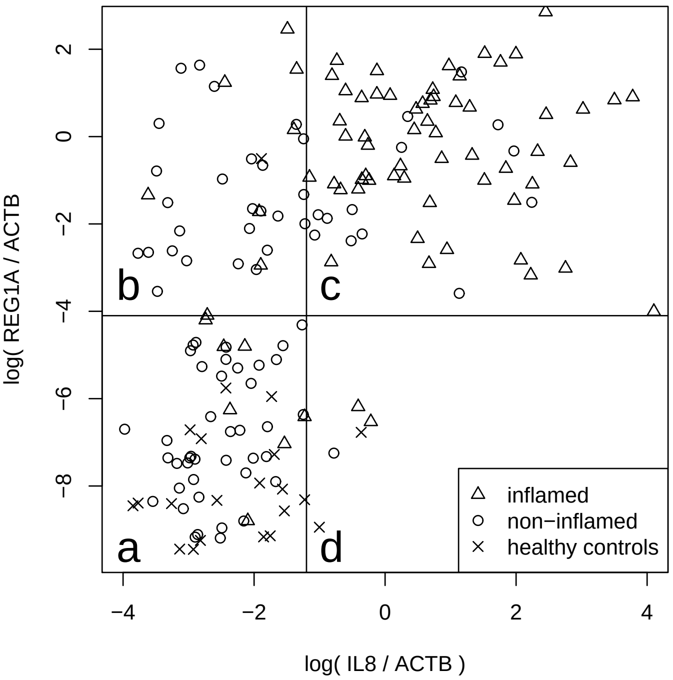


Figure 16: Expression profiles of IL8 and REG1A in mucosal biopsies of IBD patients and healthy controls. Expression values shown are log-transformed expression ratios of IL8 and REG1A to ACTB. Biopsy specimens from IBD and DSC patients have been rated by gastroenterologists during endoscopy as inflamed (triangle) or non-inflamed (circle). Biopsy specimens from healthy control subjects are marked by cross.

from IBD patients we found twenty-one splicing factors differentially regulated for each subgroup. Table 15 depicts the expression fingerprint of 149 splicing factors as observed in our experiments grouped by functional category.

### 3.2.4 IBD subtype influences regulation of splicing factors

To investigate the differences of splicing factor regulation in Crohn's disease and ulcerative colitis, we selected seven exemplary splicing factors from different functional categories for further verification by real time PCR in 195 individuals (table 3). A strong effect of inflammation on expression levels of splicing factors was identified in tissue of CD patients (fig. 17). In contrast to results from UC patients, all seven splicing factors under investigation showed down-regulated expression levels comparing non-inflamed to inflamed CD patient tissue. On the other hand, we found a strong disease effect on expression levels exclusively in UC patient tissue. Six of seven splicing factors were down-regulated in UC patient tissue when compared to healthy control tissue, independent of the inflammation status of the biopsy. As opposed to our findings in UC patient tissue, non-inflamed CD patient tissue did not show significant differences in five out of seven splicing factors, when compared to healthy control tissue.

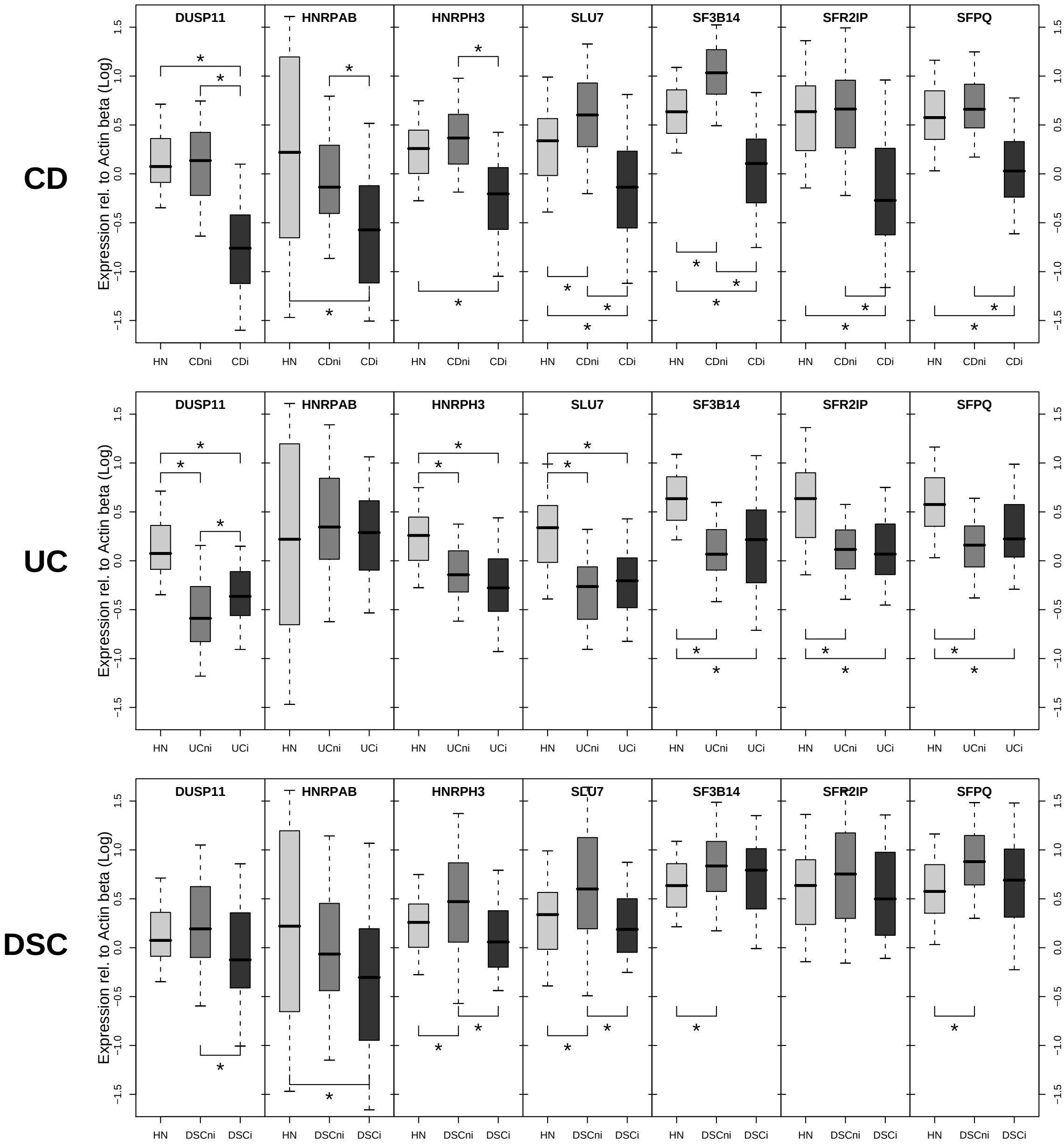


Figure 17: Expression profiles of seven splicing factors in biopsies of IBD patients and healthy controls. Expression values shown are log-transformed expression ratios of the genes of interest to ACTB. The expression profile of healthy control tissue (HN) was compared to Crohn's disease patient tissue (CD) or Ulcerative colitis patient tissue (UC) or Disease control patient tissue (DSC). Patient tissue could be inflamed (i) or non-inflamed (ni), healthy control tissue was non-inflamed. Comparisons found significant by Wilcoxon Rank Sum test ($p < 0.05$) are marked by asterix.

Additional to the different regulatory patterns found in UC or CD patients, *SF3B14* and *SFPQ* showed regulation in opposite directions, when comparing non-inflamed patient tissue to healthy control tissue. While in non-inflamed tissue *SF3B14* and *SFPQ* are up-regulated in CD patients, both splicing factors are down-regulated in UC patients.

### 3.2.5 Regulation of splicing factors is specific to IBD

In order to assess disease-specific regulation, disease specificity controls (DSC) suffering from colonic disease, but not IBD, were included in the verification cohort (table 3). The effect of inflammation, which was predominantly seen in CD tissue but not in UC tissue, was found only in three out of seven splicing factors in DSC patient tissue. Although the direction of regulation of *DUSP11*, *HNRPH3* and *SLU7* in inflammation is shared between CD and DSC patients, the extent of regulation is higher in CD patients. The disease effect, which was predominantly seen in UC samples but not in CD samples, was not found in DSC patients. *HNRPH3*, *SF3B14*, *SFPQ* and *SLU7* show up-regulation in non-inflamed DSC patient tissue as compared to healthy control tissue, while only *SF3B14* and *SFPQ* show the same pattern in CD patients but not in UC patients. Combining our findings *SFPQ* and *SFR2IP* showed the most disease (subtype) specific regulation.

### 3.2.6 Intronic sequence is expressed and regulated in IBD

Based on the findings of differential expression of splicing factors in the context of disease, we chose to investigate intron-retention as potentially pathogenic splicing outcome. To ensure the quality of our observations, we evaluated if the expression detected by our probes was intron-retention using the UCSC genome browser as data-source. Of 145 events, seventy-four (51%) showed strong, forty-eight (33%) intermediate and twenty-three (16%) weak evidence for intron-retention in the UCSC database (for the grading scheme see section 3.1.7). We qualitatively verified present intron-retention in six genes in a pooled cohort of IBD patients by PCR. The possibility of a hidden exon was ruled out in three of six cases, with only weak evidence for an unknown exon in two cases, however, intron-retention was detected in all six genes (figure 13).

Analyzing our microarray data, we found thirty-three intron-retention events with expression levels regulated due to inflammation and/or IBD (see table 16). Seventy-five percent of those regulated intron-retention events showed strong supporting evidence for intron-retention in the UCSC database.

Grouping the thirty-three genes with regulated intron-retention events (see table 16) identified in our initial screening according to gene function highlighted signal transduction (*DUSP3, PTK7, GRM4, ZNF76, ANKS1, FGD2, ITPR3, RDS*) among other groups like the secretory pathway (*YIPF3, MDGA1*), the immune system (*IER3, PARC, AGER, C6orf25*) or drug metabolism (*ABCC10*). In a next step *IER3*, *FGD2* and *PARC* were selected for detailed verification of intron-retention and its regulation in IBD and its subtypes.

### 3.2.7 Intron-retention in PARC, IER3 and FGD2 is regulated in IBD

Intron-retention was assessed by monitoring two forms per transcript: The intron-retained form (IR) and the correctly spliced form (the intron-excised form (IE)). Intron-retention of three genes (*PARC*, *IER3* and *FGD2*) was monitored by real-time PCR using specific primers designed to exclusively amplify one transcript form (IE or IR). The experimental setup addressed three major questions: disease- and/or inflammation-effects regulating intron-retention; regulatory patterns shared by the intron-retained and the intron-excised form and regulatory patterns specific to IBD subtype and/or IBD in general.

Inflammation influenced intron-retention in UC tissue for all three genes investigated in detail; however, in CD tissue, the effect of inflammation was only seen in *FGD2*. In both CD and UC tissue, inflammation led to an increase of the intron-retained form in contrast to DSC tissue, where no significant differences in expression levels due to neither inflammation or to disease were found. Disease-dependent down-regulation of intron-retention was found for *PARC* in UC non-inflamed tissue and for *IER3* in CD and DSC non-inflamed tissue and CD inflamed tissue. *FGD2* represents the only exception since the intron-retained form is up-regulated in UC inflamed tissue.

Comparing the regulatory patterns of the intron-retained to the intron-excised forms our observations can be grouped into two categories. The first group embodies the cases where the regulations of both forms are in agreement. This was found only for *FGD2* and *IER3* and only in UC tissue. The second group describes discordant

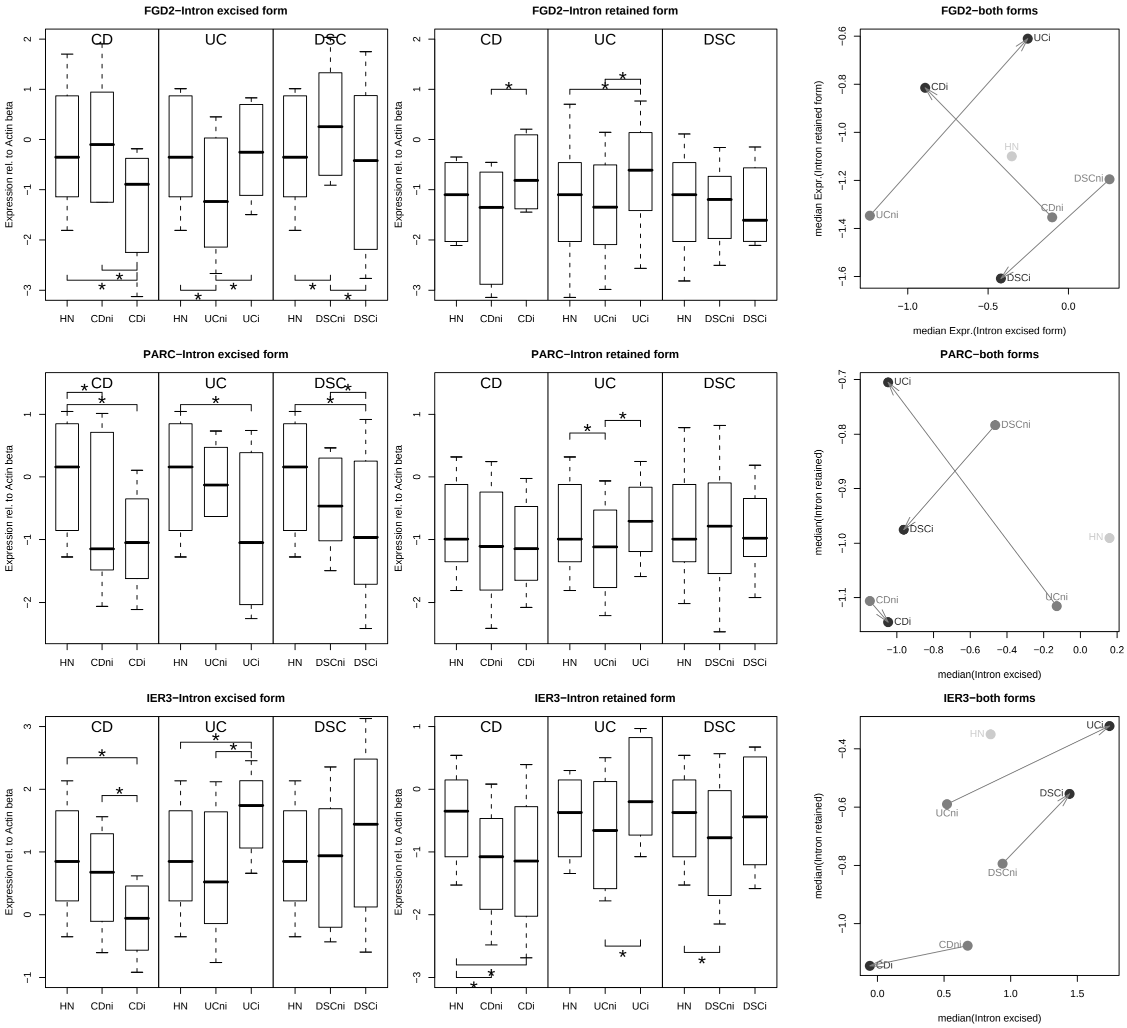


Figure 18: Expression profiles of two different splice-forms of three genes from IBD patients and healthy controls. The different splice-forms are grouped by column. The first column harbors the expression levels of the intron-excised forms, the second column of the intron-retained forms. The third column depicts concerted changes in expression for the intron-excised and the intron-retained forms. Arrows in the third column highlight the different concerted changes due to inflammation. Expression values shown are log-transformed expression ratios of the genes of interest to actin beta. The expression profile of healthy control tissue (HN) is compared to Crohn's disease patient tissue (CD) or ulcerative colitis patient tissue (UC) or disease control patient tissue (DSC). Patient tissue can be inflamed (i) or non-inflamed (ni), healthy control tissue was non-inflamed. Comparisons found significant by *Wilcoxon Rank Sum* test ($p < 0.05$) are marked with an asterix.

regulatory patterns of both forms. We found this group to be larger than the group of concordant regulation. For *FGD2* we found for instance the intron-excised form down-regulated in inflammation in CD tissue, while the intron-retained form was found up-regulated in inflammation in CD tissue. For *PARC* we found the same pattern (IE down- IR up-regulated in inflammation) but in UC tissue only. A slightly different example of discordant regulation is given by the expression levels of *IER3* in CD tissue.

We additionally calculated the ratio of the intron-retained form and the intron-excised form (see table 14). For healthy subjects the amount of the intron-excised form is about twice the amount of the intron-retained form in all three genes analyzed. In contrast ratios were found for IBD patients, where the intron-retained form dominated the intron-excised form. In UC patients for instance we found for the gene *PARC* a sharp drop of the ratio from a value close to healthy subjects, to a 1.2 fold excess of the intron-retained form over the intron-excised form. In general the patterns found were unique for IBD as compared to disease specificity controls and unique for each subtype of IBD. Discordant regulatory patters were found for example for *IER3*, where the ratio of the intron-excised to the intron-retained form was close to healthy subjects in CD inflamed and UC non-inflamed tissues, while the ratio is about 1.5 times higher than in healthy subjects in CD non-inflamed and UC inflamed and both DSC tissue groups.

Table 14: Log-ratio of intron-excised vs intron-retained form for FGD2, PARC and IER3.

| Log$_2$ intron-excised/ intron-retained | HN | CDni | CDi | UCni | UCi |
|---|---|---|---|---|---|
| FGD | 0.75 | 1.25 | -0.08 | 0.11 | 0.36 |
| PARC | 1.15 | -0.04 | 0.10 | 0.99 | -0.34 |
| IER | 1.20 | 1.75 | 1.09 | 1.11 | 2.06 |

Given our results presented above (figure 18 & table 14), it becomes clear that we found no general regulatory pattern, neither due to inflammation, nor due to disease. On the other hand, we deduce from this finding that the results described are specific for each subtype of IBD and specific to IBD in general. To emphasize the differences in concerted regulation found for the groups at hand, we plotted

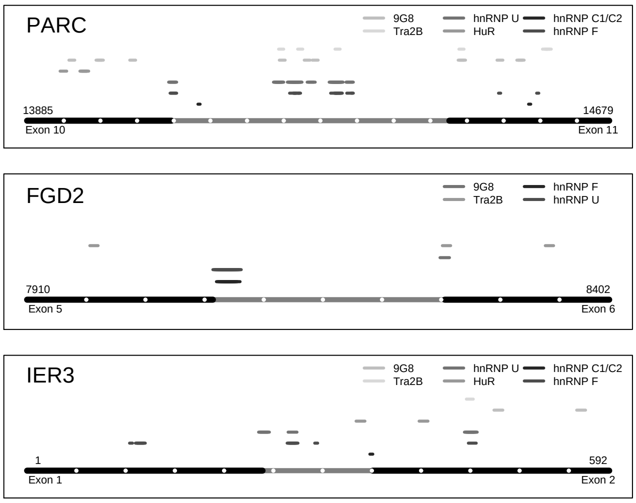


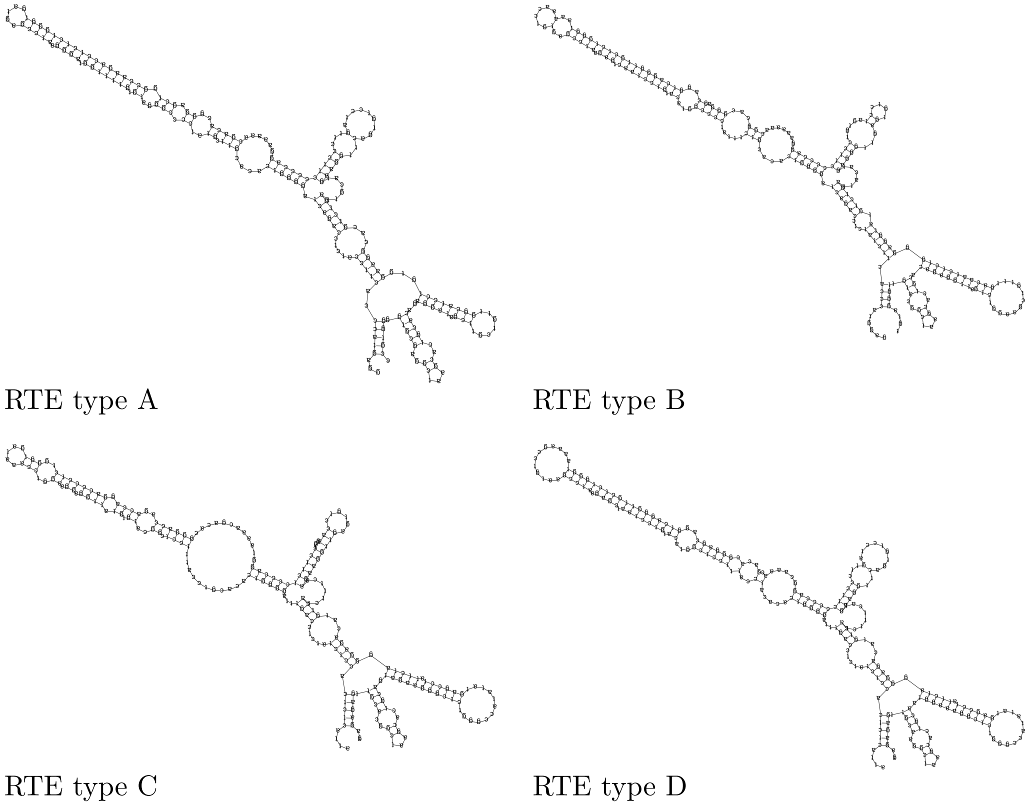
Figure 19: Splicing factor binding sites found *in silico* by the software *splicing rainbow*[197] in IER3, FGD2 and PARC. These splicing factors were regulated due to inflammation or IBD as found by microarray analysis ($p < 0.05$). Depicted is the intron sequence (grey) plus exon sequence (black) on each side. The localization of each sequence window within the pre-mRNA sequence is given in base-pairs. White dots every 50 bases are plotted to show the different scales.

the expression levels of the intron-excised form against the expression levels of the intron-retained form (right column in figure 18).

To summarize our findings, we found evidence for inflammation and/or disease-related regulation of intron-retention. Intron-retained and intron-excised forms in most cases showed different regulatory patterns, while in some cases they were co-regulated. Additionally, we present evidence that the concerted regulation of both forms is specific for each subtype of IBD and specific to IBD in general.

### 3.2.8 IER3, FGD2 and PARC contain binding sites for splicing factors regulated in IBD

For a number of splicing factors data on their binding sites has been published. *Splicing Rainbow*[197] a software tool to discover known binding sites was used in this study to map the binding sites of splicing factors found regulated in IBD. We analyzed the transcript sequence of *IER3*, *FGD2* and *PARC*, where we had discovered regulated intron-retention. *In silico* mapping of regulated splicing factors to transcripts with regulated intron-retention revealed several splicing factor binding sites located within or surrounding the introns of interest as is shown in figure 19. These findings represent a potential connection between splicing factor expression levels and subsequent splicing patterns in the context of IBD pathology.



RTE type A

RTE type B

RTE type C

RTE type D

Figure 20: Related RNA transport element (RTE) structures reproduced in this study by calculating an abstract consensus shape with RNAshapes software[198] and sequences from Smulevitch et al.[190]. The consensus structure contains four stem-loops, which have been found crucial for mRNA export from the nucleus.

### 3.2.9 IER3 mRNA can form a RNA transport element-like structure

*IER3* is different from *FGD2* and *PARC* in that there is no further intron which is crucial for nonsense-mediated decay (NMD) or regulated unproductive splicing and translation (RUST) to exert its action[127,231]. Since the intron-retained form of *IER3* is not spliced at all, it possibly contains no exon-junction-complex. The exon-junction complex however is thought to be crucial for export from the nucleus[110]. One can question, if the intron-retained form of *IER3* ever leaves the nucleus. We therefore investigated the possibility for *IER3* to contain a RNA transport element using an abstract shape approach implemented by the software RNAshapes[198].

The software RNA shapes requires an abstract shape model, which we deduced as consensus shape from four related sequences described by Smulevitch et al.[190].

Using an abstraction level of depth two the abstract shape found was

_[_[[]]_[[]]_[[_[[]][[[[[[[[]]]]]]]]_]]_]_

with unpaired regions represented by underscores and stacking regions represented by squared brackets.

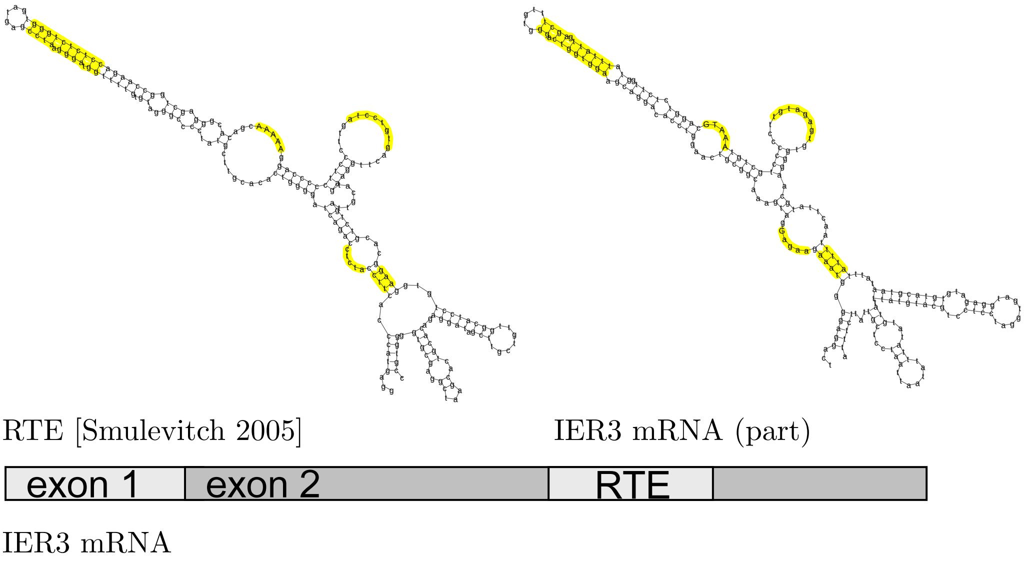


Figure 21: RNA transport element like structure found in mRNA of IER3. Using the abstract consensus shape model found for RNA transport elements described by Smulevitch et al.[190] we were able to identify a sequence within the second exon of IER3 that can fold into a similar structure. Software settings for RNAshapes were kept by default with a sequence window of 230 nucleotides. Structural features important for RTE function are highlighted.

Using this abstract shape we were able to reproduce the structures found by Smulevitch et al[190] and we identified a RTE-like structure within the second exon of *IER3*. The four stem-loops found crucial for RTE function are present in the RTE-like structure we identified in *IER3*. The overall conservation of sequence was very low in contrast to the four RTE-like structures described by Smulevitch et al.[190]; however, we found three of four conserved consecutive adenine residues within an internal loop of stem-loop IV. These conserved adenine residues are critical for RTE function[190].

Table 15: Splicing factor expression changes related to inflammation status and disease; $\log_2$-ratios of expression changes are depicted for inflamed and non-inflamed tissue of Inflammatory Bowel Disease patients (IBD) and and healthy control subjects. Comparisons found significant are marked by asterix ($p < 0.01$). $\text{Log}_2$-ratios have to be interpreted in the light of ratio quenching (quenching factor: ~3-4, see section 3.1.6).

| | *IBD* | *Healthy Control* | |
|---|---:|---:|---:|
| | inflamed vs<br>non-inflamed | vs IBD<br>inflamed | vs IBD<br>non-inflamed |
| *SR proteins* | | | |
| CROP | -0.01 | -0.05 | -0.04 |
| FLJ11021 | -0.04 | 0.10 | 0.14 |
| SFRS1 | -0.21 | -0.06 | 0.16 |
| SFRS10 | * 0.93 | * 0.95 | 0.02 |
| SFRS11 | 0.17 | * 0.32 | 0.15 |
| SFRS12 | -0.06 | -0.03 | 0.03 |
| SFRS14 | -0.07 | -0.06 | 0.00 |
| SFRS15 | 0.04 | -0.07 | -0.10 |
| SFRS16 | * 0.40 | 0.24 | -0.17 |
| SFRS2 | -0.13 | -0.12 | 0.01 |
| SFR2IP | -0.22 | * -0.30 | -0.08 |
| SFRS3 | 0.00 | -0.08 | -0.07 |
| SFRS4 | 0.01 | 0.16 | 0.15 |
| SFRS6 | -0.11 | -0.33 | -0.23 |
| SFRS7 | -0.12 | -0.17 | -0.04 |
| SFRS8 | 0.04 | -0.05 | -0.09 |
| SFRS9 | 0.16 | 0.03 | -0.14 |
| SR140 | -0.02 | 0.02 | 0.03 |
| SRRM2 | 0.07 | 0.01 | -0.06 |
| TRA2A | 0.00 | 0.04 | 0.04 |
| U2AF1L2 | 0.10 | -0.05 | -0.15 |

Continued on Next Page...

Table15 – Continued

| | *IBD* | *HealthyControl* | |
|---|---|---|---|
| | inflamed vs non-inflamed | vs IBD inflamed | vs IBD non-inflamed |
| *hnRNP proteins* | | | |
| HNRPA3 | -0.01 | 0.12 | 0.13 |
| HNRPAB | * -0.43 | * -0.53 | -0.10 |
| HNRPC | 0.10 | * 0.53 | * 0.42 |
| HNRPD | -0.08 | 0.09 | 0.18 |
| HNRPDL | * 0.21 | 0.12 | -0.10 |
| HNRNPE | -0.18 | 0.12 | 0.30 |
| HNRPF | -0.11 | 0.08 | 0.19 |
| HNRPH1 | -0.01 | -0.07 | -0.06 |
| HNRPH3 | * -0.28 | * -0.34 | -0.06 |
| HNRPM | 0.09 | 0.19 | 0.10 |
| HNRPR | 0.01 | * 0.25 | * 0.24 |
| HNRPU | * -0.19 | 0.02 | * 0.21 |
| *Non − snRNP spliceosome − assembly proteins* | | | |
| BCAS2 | * -0.22 | 0.02 | * 0.24 |
| C21orf66 | 0.03 | 0.06 | 0.02 |
| CDC5L | * -0.21 | -0.15 | 0.06 |
| CUGBP2 | -0.01 | 0.05 | 0.06 |
| DNAJC8 | 0.11 | -0.06 | -0.17 |
| FLJ20273 | * -0.31 | * -0.30 | 0.01 |
| FNBP3 | -0.04 | * 0.42 | * 0.46 |
| HPSE | -0.13 | -0.03 | 0.10 |
| IK | 0.15 | 0.06 | -0.09 |
| NCBP2 | 0.01 | -0.09 | -0.10 |
| PPARGC1B | -0.05 | 0.02 | 0.07 |
| PRP19 | -0.05 | 0.00 | 0.05 |
| RBM14 | 0.06 | -0.06 | -0.12 |
| RBM16 | -0.08 | 0.03 | 0.12 |

Continued on Next Page…

Table15 – Continued

| | *IBD* | *HealthyControl* | |
|---|---:|---:|---:|
| | inflamed vs non-inflamed | vs IBD inflamed | vs IBD non-inflamed |
| RBM17 | -0.01 | -0.07 | -0.06 |
| RBM18 | -0.04 | -0.12 | -0.08 |
| RBM21 | 0.25 | 0.03 | -0.22 |
| RBM6 | 0.20 | 0.14 | -0.06 |
| SART1 | 0.08 | -0.05 | -0.12 |
| SF1 | -0.01 | -0.05 | -0.04 |
| SF3B14 | * -0.26 | -0.16 | 0.10 |
| SF4 | 0.14 | -0.14 | -0.28 |
| SMN2 | * -0.28 | 0.02 | * 0.30 |
| SMNDC1 | -0.04 | 0.06 | 0.09 |
| SNAPC1 | 0.08 | 0.01 | -0.07 |
| SON | -0.12 | -0.08 | 0.04 |
| TCERG1 | -0.08 | 0.07 | 0.15 |
| TTF2 | -0.06 | 0.02 | 0.07 |
| U2AF1 | 0.00 | -0.18 | -0.17 |
| U2AF2 | -0.07 | -0.03 | 0.04 |
| ZNF239 | -0.03 | -0.05 | -0.02 |
| *U1 snRNP associated proteins* | | | |
| SNRP70 | * 0.65 | * 0.41 | -0.25 |
| SNRPC | -0.21 | * -0.39 | -0.18 |
| U1SNRNPBP | 0.03 | 0.07 | 0.03 |
| *U2 snRNP associated proteins* | | | |
| SF3A1 | 0.02 | -0.06 | -0.08 |
| SF3A2 | 0.19 | 0.24 | 0.05 |
| SF3A3 | 0.04 | 0.08 | 0.04 |
| SF3B1 | -0.17 | -0.02 | 0.16 |
| SF3B3 | 0.09 | 0.14 | 0.04 |
| SF3B4 | * 0.49 | -0.10 | * -0.59 |
| SF3B5 | 0.24 | * -0.43 | * -0.67 |

Continued on Next Page...

Table15 – Continued

| | *IBD* | *HealthyControl* | |
|---|---:|---:|---:|
| | inflamed vs non-inflamed | vs IBD inflamed | vs IBD non-inflamed |
| SNRPA1 | 0.07 | 0.01 | -0.06 |
| SNRPB2 | 0.10 | -0.16 | * -0.26 |
| *U5 snRNP associated proteins* | | | |
| ASCC3L1 | * -0.21 | 0.02 | * 0.23 |
| PRPF8 | -0.15 | * -0.48 | * -0.33 |
| *U4/U6 snRNP associated proteins* | | | |
| NHP2L1 | 0.14 | 0.00 | -0.13 |
| PRPF3 | 0.01 | -0.03 | -0.04 |
| PRPF31 | -0.03 | 0.10 | 0.14 |
| PRPF4 | 0.00 | 0.07 | 0.06 |
| *U4/U6.U5 tri – snRNP associated proteins* | | | |
| RY1 | -0.12 | -0.02 | 0.10 |
| *Sm/LSm core proteins* | | | |
| LSM2 | * 0.34 | 0.06 | * -0.28 |
| SNRPD2 | * 0.73 | 0.17 | * -0.56 |
| SNRPE | * 0.29 | * 0.33 | 0.04 |
| SNRPF | 0.17 | 0.00 | -0.16 |
| SNRPG | * 0.73 | * 0.68 | -0.05 |
| *Catalytic step II and late acting proteins* | | | |
| CDC40 | 0.01 | 0.09 | 0.08 |
| CDC5L | * -0.21 | -0.15 | 0.06 |
| DHX15 | * -0.48 | 0.14 | * 0.63 |
| DHX38 | 0.10 | -0.15 | -0.25 |
| PRPF18 | * 0.45 | 0.27 | -0.18 |
| PRPF8 | -0.15 | * -0.48 | * -0.33 |
| SLU7 | * -0.33 | * -0.33 | 0.00 |

Continued on Next Page…

Table15 – Continued

| | *IBD* | *HealthyControl* | |
|---|---:|---:|---:|
| | inflamed vs<br>non-inflamed | vs IBD<br>inflamed | vs IBD<br>non-inflamed |
| ZCCHC17 | 0.05 | 0.04 | -0.01 |
| *SplicedmRNP/EJC proteins* | | | |
| BAT1 | -0.05 | * 0.68 | * 0.74 |
| RNPS1 | -0.13 | 0.21 | 0.34 |
| *RRM − containing proteins* | | | |
| ACIN1 | 0.16 | 0.09 | -0.07 |
| CIRBP | 0.08 | 0.18 | 0.10 |
| PSEN1 | -0.10 | 0.01 | 0.11 |
| RBM15B | -0.04 | -0.11 | -0.07 |
| RBM5 | -0.02 | -0.05 | -0.04 |
| RNPC2 | -0.17 | 0.05 | * 0.22 |
| SPEN | 0.02 | 0.00 | -0.02 |
| *DExD box proteins* | | | |
| DDX1 | * -0.19 | * -0.17 | 0.01 |
| ASCC3L1 | * -0.21 | 0.02 | * 0.23 |
| DDX24 | 0.22 | 0.12 | -0.10 |
| DDX3X | 0.08 | 0.17 | 0.09 |
| DDX46 | -0.11 | -0.04 | 0.07 |
| DHX15 | * -0.48 | 0.14 | * 0.63 |
| DHX16 | -0.10 | -0.14 | -0.04 |
| DHX34 | 0.02 | -0.03 | -0.05 |
| DHX38 | 0.10 | -0.15 | -0.25 |
| DHX9 | -0.10 | 0.03 | 0.13 |
| SKIV2L | 0.10 | -0.15 | -0.25 |
| ZCCHC17 | 0.05 | 0.04 | -0.01 |
| *Proteins of secondary regulatory function* | | | |
| TNPO3 | -0.20 | -0.09 | 0.11 |

Continued on Next Page. . .

Table15 – Continued

| | *IBD* | *HealthyControl* | |
|---|---|---|---|
| | inflamed vs non-inflamed | vs IBD inflamed | vs IBD non-inflamed |
| AKAP1 | -0.14 | 0.09 | * 0.24 |
| CLK1 | * -0.25 | 0.01 | * 0.25 |
| CLK4 | 0.02 | 0.07 | 0.05 |
| FASTK | -0.04 | -0.04 | 0.00 |
| SRPK1 | 0.02 | -0.01 | -0.02 |
| SRPK2 | 0.04 | 0.06 | 0.01 |
| HRMT1L1 | -0.14 | -0.16 | -0.02 |
| HRMT1L2 | * 0.51 | 0.24 | -0.27 |
| DUSP11 | * -0.36 | * -0.46 | -0.10 |
| PPM1G | 0.12 | 0.09 | -0.04 |
| *Intronic/Exonic splice enhancers and silencers* | | | |
| NOVA1 | * 0.21 | 0.14 | -0.07 |
| SIAHBP1 | 0.02 | -0.06 | -0.08 |
| YBX2 | -0.02 | 0.05 | 0.07 |
| C1QBP | 0.14 | 0.15 | 0.02 |
| ELAVL1 | * -0.34 | -0.05 | 0.30 |
| ELAVL2 | -0.01 | -0.01 | 0.01 |
| ELAVL3 | 0.10 | 0.05 | -0.06 |
| ELAVL4 | 0.12 | 0.00 | -0.11 |
| KHSRP | 0.07 | 0.05 | -0.02 |
| RBM10 | 0.12 | -0.03 | -0.15 |
| SFPQ | * -0.27 | * -0.40 | -0.13 |
| *Cyclophilin – like proteins* | | | |
| DNAJA2 | -0.12 | * -0.26 | -0.14 |
| PPIA | 0.12 | 0.24 | 0.12 |
| PPIB | * 0.75 | 0.38 | -0.37 |
| PPIF | 0.21 | -0.03 | -0.23 |
| PPIG | -0.02 | -0.04 | -0.01 |
| PPIL1 | * -0.28 | * -0.27 | 0.01 |

Continued on Next Page…

Table15 – Continued

| | *IBD* | *HealthyControl* | |
|---|---:|---:|---:|
| | inflamed vs non-inflamed | vs IBD inflamed | vs IBD non-inflamed |
| PPIL2 | 0.06 | 0.02 | -0.03 |
| PPIL4 | -0.11 | 0.08 | 0.18 |
| PPIL5 | 0.06 | 0.07 | 0.01 |
| *Spliceosome proteins involved in mRNA export* | | | |
| BAT1 | -0.05 | * 0.68 | * 0.74 |
| CIRBP | 0.08 | 0.18 | 0.10 |
| NXF3 | 0.01 | -0.04 | -0.05 |
| NXT1 | * 0.39 | 0.25 | -0.14 |
| PSIP1 | 0.08 | 0.08 | 0.00 |
| RNPS1 | -0.13 | 0.21 | 0.34 |
| SFRS1 | -0.21 | -0.06 | 0.16 |
| SFRS3 | 0.00 | -0.08 | -0.07 |
| SFRS7 | -0.12 | -0.17 | -0.04 |
| STAU2 | * -0.40 | * -0.40 | 0.01 |
| STRBP | * -0.34 | -0.09 | * 0.24 |
| THOC1 | -0.07 | * -0.23 | -0.16 |
| THOC2 | -0.05 | * 0.24 | * 0.29 |
| ZCCHC17 | 0.05 | 0.04 | -0.01 |

Table 16: Intronic sequence expression changes ($Log_2$) related to inflammation status and disease. Comparisons found significant are marked by asterix ($p < 0.01$). $Log_2$-ratios have to be interpreted in the light of ratio quenching (quenching factor: ∼3-4, see section 3.1.6)

| | | *IBD* | *Healthy Control* | |
|---|---|---:|---:|---:|
| | | inflamed vs non-inflamed | vs IBD inflamed | vs IBD non-inflamed |
| HUGO | Supp.Sequence | | | |
| AARSL | BI859749 | 0.03 | -0.03 | -0.06 |
| ABCC10.1 | CA307443 | * 0.21 | * 0.16 | -0.04 |
| ABCC10.2 | AK024446 | -0.02 | -0.13 | -0.11 |
| AGER | AB061669 | 0.25 | * -0.47 | * -0.73 |
| AGPAT1 | AL133975 | 0.02 | -0.01 | -0.03 |
| AL832447 | AA411899 | * 0.90 | * 0.66 | -0.24 |
| AL833650 | AK128294 | -0.09 | * -0.19 | -0.11 |
| ALDH5A1 | AW592482 | 0.01 | -0.03 | -0.05 |
| ANKS1.1 | DA062959 | 0.21 | 0.30 | 0.08 |
| ANKS1.2 | AF038657 | -0.08 | -0.11 | -0.02 |
| ANKS1.3 | DB314115 | -0.08 | * -0.21 | -0.14 |
| APOBEC2 | AI992340 | -0.02 | -0.03 | 0.00 |
| BAT5 | AW968302 | 0.20 | 0.17 | -0.04 |
| BI828649 | AA284236 | -0.01 | -0.12 | -0.11 |
| BRPF3.1 | AF038424 | -0.08 | 0.02 | 0.10 |
| BRPF3.2 | BG004007 | 0.05 | -0.10 | -0.15 |
| C2 | BC029781 | 0.07 | -0.07 | -0.14 |
| C4A.1 | T64249 | 0.03 | 0.12 | 0.09 |
| C4A.2 | CB163642 | 0.02 | 0.03 | 0.01 |
| C6orf1 | AL557041 | 0.01 | 0.07 | 0.06 |
| C6orf106 | BF911464 | -0.10 | -0.21 | -0.12 |
| C6orf108 | NM.199184 | 0.14 | -0.17 | * -0.31 |
| C6orf125 | CN284202 | -0.10 | -0.03 | 0.06 |
| C6orf141 | CB115931 | 0.02 | 0.07 | 0.05 |
| C6orf153 | CN311648 | -0.01 | -0.01 | 0.00 |

Continued on Next Page. . .

Table16 – Continued

| | | $IBD$ | $HealthyControl$ | |
|---|---|---|---|---|
| | | inflamed vs non-inflamed | vs IBD inflamed | vs IBD non-inflamed |
| C6orf25 | AJ292264 | 0.12 | -0.23 | * -0.35 |
| C6orf48.1 | BM423352 | 0.10 | -0.05 | -0.15 |
| C6orf48.2 | BM982401 | -0.03 | -0.09 | -0.07 |
| C6orf49.1 | BC110459 | -0.10 | * -0.22 | -0.13 |
| C6orf49.2 | AA514248 | -0.01 | -0.06 | -0.05 |
| C6orf62.1 | BF748822 | -0.12 | -0.09 | 0.03 |
| C6orf62.2 | DA426118 | -0.06 | -0.13 | -0.07 |
| C6orf64 | NM.018322 | -0.07 | -0.17 | -0.10 |
| CCL15 | BC050647 | 0.03 | 0.06 | 0.03 |
| CD2AP | BG215751 | 0.00 | -0.04 | -0.03 |
| CDKAL1.1 | AI637916 | 0.05 | -0.03 | -0.08 |
| CDKAL1.2 | AA621866 | -0.14 | -0.16 | -0.02 |
| CFB | BQ317007 | 0.02 | 0.03 | 0.02 |
| CLIC5 | BP320577 | 0.15 | 0.04 | -0.11 |
| CMAH.1 | CR749466 | 0.14 | 0.09 | -0.05 |
| CMAH.2 | AK000716 | 0.03 | -0.08 | -0.10 |
| CREBL1 | R02494 | -0.14 | -0.08 | 0.06 |
| CROP | NM.016424 | -0.04 | -0.10 | -0.06 |
| CSNK1D | DB127690 | 0.00 | -0.08 | -0.07 |
| CSNK2B | AL537527 | 0.01 | -0.07 | -0.08 |
| CUL7 | BM273286 | 0.04 | -0.10 | -0.14 |
| DNAH8 | BF372166 | -0.03 | -0.11 | -0.08 |
| DOM3Z | BC071651 | 0.11 | 0.05 | -0.07 |
| DUSP3 | NM.004090 | * -0.29 | -0.20 | 0.10 |
| ENPP4 | AV726702 | 0.02 | -0.15 | -0.17 |
| FGD2.1 | BQ708371 | 0.12 | -0.05 | -0.17 |
| FGD2.2 | AK092732 | * 0.35 | 0.27 | -0.08 |
| FGD2.3 | BC062363 | -0.05 | -0.18 | -0.13 |
| FLOT1 | AW798231 | -0.12 | -0.22 | -0.10 |
| FOXP4.1 | EL582986 | -0.06 | -0.01 | 0.04 |

Continued on Next Page…

Table16 – Continued

| | | *IBD* | *HealthyControl* | |
|---|---|---|---|---|
| | | inflamed vs<br>non-inflamed | vs IBD<br>inflamed | vs IBD<br>non-inflamed |
| FOXP4.2 | BG997253 | -0.05 | -0.10 | -0.05 |
| GPR110 | BG433170 | -0.06 | -0.09 | -0.03 |
| GRM4 | AB209104 | * 0.46 | 0.04 | * -0.42 |
| GTPBP2 | DA446337 | -0.06 | -0.15 | -0.08 |
| HFE | DA957218 | 0.01 | -0.16 | -0.18 |
| HIST1H2AC | BC017379 | * -1.05 | * -0.78 | 0.27 |
| HLA-DRB1.1 | BF849061 | 0.17 | -0.01 | -0.18 |
| HLA-DRB1.2 | BF879238 | -0.01 | 0.00 | 0.01 |
| IER3 | AF039067 | * 0.26 | -0.03 | * -0.29 |
| ITPR3.1 | BI825723 | * 0.26 | -0.02 | * -0.28 |
| ITPR3.2 | AL577972 | -0.11 | -0.14 | -0.03 |
| ITPR3.3 | AL832807 | 0.03 | -0.07 | -0.10 |
| KIAA0240.1 | AV726402 | -0.11 | -0.16 | -0.05 |
| KIAA0240.2 | BE773025 | 0.01 | 0.03 | 0.02 |
| KNSL8 | BM989115 | 0.03 | -0.05 | -0.08 |
| LEMD2 | AV708966 | 0.12 | * 0.20 | 0.08 |
| LRRC16.1 | AV656596 | -0.06 | -0.05 | 0.01 |
| LRRC16.2 | BE173626 | 0.03 | 0.00 | -0.03 |
| MBOAT1 | BG400775 | -0.07 | -0.05 | 0.02 |
| MDGA1.1 | AA054635 | 0.24 | * 0.77 | 0.54 |
| MDGA1.2 | AK090677 | * 0.19 | 0.02 | * -0.17 |
| MDGA1.3 | AK091149 | -0.01 | 0.05 | 0.06 |
| MGC45491.1 | BX390676 | -0.06 | -0.08 | -0.03 |
| MGC45491.2 | BC032706 | * 1.82 | * 1.19 | -0.62 |
| MRPL2 | BM553982 | * -0.32 | * -0.32 | 0.00 |
| MRPS18A | DQ884400 | 0.00 | -0.11 | -0.11 |
| MRPS18B | AL050361 | * 0.18 | 0.05 | -0.13 |
| MRS2L | AW835456 | -0.04 | -0.03 | 0.01 |
| NFYA | AI688265 | 0.00 | -0.05 | -0.05 |
| OTTHUMG14664 | AW086174 | -0.02 | -0.11 | -0.09 |

Continued on Next Page. . .

Table16 – Continued

| | | *IBD* | *HealthyControl* | |
|---|---|---|---|---|
| | | inflamed vs non-inflamed | vs IBD inflamed | vs IBD non-inflamed |
| OTTHUMG14673 | AA203184 | -0.12 | -0.09 | 0.03 |
| PACSIN1 | BX094526 | 0.03 | -0.11 | -0.14 |
| PARC.1 | CR749511 | * 0.55 | * 0.38 | -0.17 |
| PARC.2 | DA262387 | 0.06 | -0.12 | -0.18 |
| PI16 | AK124589 | 0.04 | -0.05 | -0.09 |
| PPT2 | BX426906 | -0.12 | -0.10 | 0.02 |
| PRL | AV749457 | -0.05 | -0.07 | -0.02 |
| PRR3 | DA164435 | -0.06 | -0.04 | 0.02 |
| PRRT1.1 | BX112489 | -0.03 | -0.03 | 0.00 |
| PRRT1.2 | BM688858 | -0.03 | 0.04 | 0.07 |
| PRSS16 | BI756721 | 0.20 | 0.04 | -0.15 |
| PTK7.1 | AL538801 | -0.20 | -0.18 | 0.03 |
| PTK7.2 | AK131487 | * 0.41 | -0.01 | * -0.42 |
| RDS | BE266298 | -0.07 | * -0.30 | * -0.23 |
| RGL2 | AI279162 | 0.21 | 0.16 | -0.04 |
| RING1 | BX406293 | -0.15 | * -0.19 | -0.04 |
| RPL10A | BC013864 | * 0.25 | 0.18 | -0.06 |
| RPL7L1 | CX781752 | 0.00 | 0.05 | 0.05 |
| RPS18.1 | CN346578 | 0.07 | -0.17 | -0.25 |
| RPS18.2 | BX537589 | * -0.17 | -0.05 | 0.12 |
| RXRB.1 | AW860348 | -0.12 | -0.01 | 0.11 |
| RXRB.2 | BP366955 | 0.16 | 0.00 | -0.16 |
| SCUBE3 | BF746129 | 0.05 | -0.03 | -0.08 |
| SFRS3 | AI700741 | -0.02 | -0.10 | -0.09 |
| SLC17A1 | AL700121 | 0.22 | 0.22 | 0.00 |
| SLC17A4 | AW662505 | * -0.24 | * -0.47 | -0.23 |
| SLC29A1 | AA195199 | 0.05 | -0.01 | -0.07 |
| SRPK1 | DA638682 | -0.04 | 0.01 | 0.05 |
| STK19.1 | BG473149 | 0.01 | -0.15 | -0.16 |
| STK19.2 | BG473149 | -0.09 | -0.12 | -0.03 |

Continued on Next Page...

Table16 – Continued

| | | *IBD* | *HealthyControl* | |
|---|---|---|---|---|
| | | inflamed vs non-inflamed | vs IBD inflamed | vs IBD non-inflamed |
| STK38 | AA112431 | 0.00 | -0.17 | -0.17 |
| TAF11 | BX647568 | -0.03 | -0.06 | -0.03 |
| TAPBP | AA379282 | -0.02 | 0.14 | 0.17 |
| TBC1D22B | BF986668 | 0.01 | 0.00 | -0.01 |
| TCTE1 | DB450345 | * 0.39 | 0.29 | -0.10 |
| TDRG1 | BC042123 | -0.02 | -0.05 | -0.03 |
| TEAD3 | DB287000 | -0.04 | -0.11 | -0.07 |
| TNFRSF21 | AB209394 | -0.01 | -0.01 | 0.00 |
| TNRC5 | AK090425 | 0.01 | 0.01 | 0.01 |
| TRERF1.1 | AV695566 | 0.05 | -0.19 | -0.24 |
| TRERF1.2 | AK024851 | -0.09 | -0.10 | -0.01 |
| TRIM10 | NM.006778 | -0.06 | -0.12 | -0.06 |
| TRIM38 | CR936707 | -0.08 | -0.05 | 0.03 |
| TTBK1 | BG912008 | -0.04 | -0.15 | -0.11 |
| TULP1 | BQ640101 | -0.03 | -0.12 | -0.09 |
| UBR2 | BI715192 | -0.03 | -0.27 | -0.24 |
| URFB1 | BM722280 | -0.17 | * -0.25 | -0.08 |
| VARS | AI023520 | * -0.41 | * -0.61 | -0.20 |
| VARS2L | AK094483 | 0.16 | 0.14 | -0.02 |
| VPS52 | DB251678 | -0.10 | -0.05 | 0.04 |
| XPO5 | AK127513 | 0.03 | -0.02 | -0.05 |
| YIPF3 | BI822291 | * 0.25 | 0.13 | -0.12 |
| ZFAND3.1 | AA379379 | 0.03 | 0.00 | -0.03 |
| ZFAND3.2 | AV683446 | -0.11 | -0.14 | -0.03 |
| ZFAND3.3 | BF333057 | -0.08 | 0.17 | 0.26 |
| ZFAND3.4 | AA398665 | 0.03 | -0.05 | -0.08 |
| ZNF184 | AK123011 | -0.11 | * -0.22 | -0.11 |
| ZNF193 | AY261373 | -0.06 | -0.09 | -0.02 |
| ZNF391 | AK092633 | 0.03 | -0.03 | -0.06 |
| ZNF76 | CN276895 | * 0.61 | * 0.55 | -0.06 |

# 4 Discussion

The pathology of IBD is influenced by various factors, including genetic and environmental factors, which contribute to an inappropriate immune response in the gastrointestinal tract. To date, genetic linkage and association studies have identified sequence variants that confer disease susceptibility, with the strongest association being found in *CARD15* for CD. An extended stretch of ∼20Mb on chromosome 6 has been associated not only to IBD but also to psoriasis, atopic dermatitis and diabetes, to name but a few diseases. Within this region, which is of general importance to the human immune system, the HLA genes and others are genetically linked in extensive haplotype structures. These structural constraints complicate genetic association analysis. The first aim of this study was to construct a HLA-microarray to investigate the transcriptome of the HLA linkage region. The assumption that potentially disease causing genes might either exhibit aberrant expression or malignant splicing patterns, initiated this study.

## 4.1 Establishment of a HLA microarray

### 4.1.1 Merging databases

Even years after the release of the genomic sequence[105,216], there are still large uncertainties about the exact number of human genes. The number of known and hypothetical genes in current databases such as RefSeq and Ensembl is continuously growing. Over the years estimated gene count has shrunk to *less than 25000*[184], as compared to earlier predictions ranging from 35000[46] to 120000[117]. As Larsson et al. point out[107], despite all improvements, it is the near constant percentage of gene predictions that are only found in single databases, that keeps all databases in *incomplete* status. The significant differences between transcript databases concerning gene and transcript count are displayed in table 5. To enrich the sequence-space for research we decided to merge four different transcript databases. This strategy enabled us not only to address as many sequences as possible, but also to thoroughly check each sequence for possible cross-hybridization.

**Database choice** NCBI's Reference Sequence database, Ensembl core database and Sanger Center's Vega database are well established sequence databases, while the ESTgene database is of experimental status [Ensembl help team, personal communication]. Two databases (RefSeq and Vega) contain manually curated transcript models while the transcript models of the other two databases are created exclusively *in silico*. RefSeq, Ensembl core and Vega integrate *ab initio* gene predictions with supporting evidence, while ESTgene uses clusters of EST sequences to create transcript models[13,47,158,223]. While Ensembl core, Vega and ESTgene construct their transcript models using the *golden path* as sequence basis, RefSeq additionally creates transcript models, which are based solely on supporting evidence. One consequence is, that the transcript models of the former databases are well positioned in the human genome, while some transcript models of RefSeq do not easily map to the *golden path*.

While manually curated databases usually contain transcript models of high quality, the ESTgene database contains transcript models that are more unstable because of the high sequencing error rate of ESTs and the problem of genomic contamination, which has been estimated to affect 5-8% of all ESTs[6,225]. Nevertheless ESTs offer a considerable advantage in aiding the prediction of non-coding exons, especially those located within the 3'-UTR[79]. We therefore decided to include the ESTgene database due to its wealth of unknown transcript models. Our choice to include the ESTgene database is supported by the fact that manually curated databases have integrated many of the transcript models of the ESTgene database in the last few years. In contrast other transcript models have vanished and microarray probes have become orphans. But in science some good ideas emerge out of failure if the properties of alleged disasters are well-studied. In an act of serendipity, the skilled scientist stumbled upon the expression of sublime signals. In the line of *its not a bug, it's a feature* some orphaned microarray probes enabled us to measure intron-retention.

**Grouping transcripts** Creation of a complete set of known and hypothetical genes requires merging of existing datasets into a non-redundant set of genes. We developed an algorithm to merge the four chosen transcript databases: Ensembl core, Ensembl ESTgene, Vega and RefSeq. Such merging is often based on sequence alignment using a similarity threshold. Multiple alignments of biological nucleic acid sequences are one of the most commonly used techniques in sequence analysis. These techniques demand a large computational load and, as Hogenesch et al.[74] and Li et al.[116] pointed out, are prone to misjudgement, since it is difficult to find a threshold that always gives the correct result.

In contrast, our algorithm exploited the fact that the transcript models were already mapped to the human genome consensus sequence called *the golden path*. The algorithm increased computational efficiency by treating transcript models as sets of integer intervals and comparing them on a numeric basis. Thereby we circumvented the computationally demanding sequence alignment, while obtaining the same result – groups of related transcripts were obtained in a fraction of the time. However, it can be problematic to find good quality genomic alignments for transcript sequences that are not defined directly from the genomic sequence. For example, only 21073 of 22239 sequences of the RefSeq database (RefSeq 8) could be found in the Ensembl database (Ensembl 25.34). The transcripts without placement on the *golden path* were compared to our merged database using BLAST. If no hit with 99% sequence identity and 95% coverage was found, the transcript was added to the merged databases at the final stage.

**Specificity** The rules used to group transcripts were simple. Two transcripts were assigned to the same group if the genomic locations and strand of at least two exons overlapped and, consequential, the two transcripts shared exon sequence of at least two exons. One can question the stringency of our heuristic approach. From a sequence point of view, two transcripts need not to be related if they share some exon sequence. Our strategy, which did not use any information on reading frame nor splice-site selection, would probably not work if genes, as groups of transcripts, had to be found *a priori*. Nevertheless, in the context of mature transcript databases where stringent algorithms have been applied to determine transcripts and genes, this strategy works for most transcript models. Compared to the algorithm of Larsson et al.[107] who designated transcripts as related if they shared at least one base, our algorithm is of moderate stringency. Eyras et al.[47] also based their initial clustering of ESTs on simple overlap. Our strategy failed when the same exon sequence was used by two distinct genes in the same location on the same strand. In this case transcripts got assigned to more than one gene within one database scheme. These cases were solved manually with preference for the gene models created by the database instead of merging genes.

The ultimate goal of our database merging strategy was to determine the sequence quality of all transcripts, in order to design microarray probes for transcript-specific sequence. In this line our algorithm assured that the sequence quality of a transcript was not judged overly specific. We assume that the problem of double assignment will get smaller the more mature transcript databases get.

**Sensitivity** Given the simplicity of rules used to group transcripts, one can question if all related sequences could be found with our algorithm. The intention of the grouping algorithm was not to build final gene models but to enable the reduction of redundancy, which arose by the use of multiple databases. Nevertheless, to explore the limits of microarray technology, related transcripts were treated as if they belonged to one gene. If a transcript remained unassigned to another transcript by mistake, this would result in false negative judgment on the gene-specificity of both transcripts. In other words, both transcript sequences would be tagged to have ambiguous sequence, even though they have gene-specific sequence. The details of our algorithm reveal that this rare case would compromise at most sequence-specificity judgments for less than 200 bases of transcript models that do not share any other exon sequence. Again, using our strategy, sequence quality of a transcript was never judged overly specific.

### 4.1.2 Reducing redundancy

Pooling the data sources created redundancy which was down-sized by a second algorithm. We made a distinction between global and local redundancy and treated them separately.

**Global redundancy** The main source of evolutionary innovation is gene duplication. According to the classic scenario, one copy retains the original function and remains under strong selection, whereas its twin gene is free of selective constraints. Spontaneous mutations and positive selection may gradually shape the copy into a novel gene[147] or, as Rodin[167] points out in his review, may turn it into a pseudogene. Without manual sequence analysis of the gene copies it is difficult to reduce global redundancy. The different genetic environment of two identical gene models may lead to different expression between developmental stages or tissues because of epigenetic regulation of expression[168]. The expressed sequences of two identical gene models at different loci cannot be distinguished from each other. Given the sequence identity of two transcripts from different loci, the assumption is that their translation is the same. With respect to our aim, the grading of sequence quality of all transcripts, multiple identical copies of a transcript in the transcriptome database would estimate the sequence quality of all copies as ambiguous. We therefore decided that only one copy of two identical transcript models was kept in our database. Our criteria to down-size global redundancy were strict. A fuzzier algorithm would have lead to more reduction of global redundancy. However, in some databases at the time the algorithm was created, differing numbers of transcript models were found for the same genomic sequence in two genomic locations. This database inconsistency motivated our conservative approach.

**Local redundancy** Our algorithm to reduce local redundancy was applied to each group of related transcripts as found by our grouping algorithm. Again, we treated transcripts as sets of integer intervals and applied set operations to compare them. The criteria to reduce local redundancy were less strict than that used to reduce global redundancy. We used fuzzy exon boundaries to detect the potential redundancy of two transcript models. As proposed by Modrek et al.[134] we allowed for a difference in length of up to six bases per exon. While Modrek et al.[134] tried to screen out sequencing errors and alignment artifacts in using fuzzy exon boundaries, we followed this strategy because of the challenge to integrate the ESTgene database. As Eyras et al.[47] reported, the exon boundary comparison of the ESTgene sequences to manual annotations yielded a sensitivity of 39% and a specificity of 30%. While the fuzzy matching of transcripts limited the artificial increase of splice-variants due to the properties of the ESTgene database, it came at the cost of neglecting NAGNAG splice sites. This type of differential splicing affects the 5' boundary of exons[70]. To resolve this level of differential splicing in a microarray setting one would need to design a probe across the exon-junction, which is reasonable for at most ∼60% of exon-junctions (see section 8). At the time the microarray probes were designed, the databases themselves did not treat NAGNAG splice sites in a coherent way; therefore, we chose to accept the costs of our fuzzy matching algorithm.

While this part of the merging procedure helped to eliminate redundancy in our database it did not solve possible fragmentation of the transcript models. EST information is incomplete by nature and the Cluster-Merge algorithm of Eyras et al.[47]

did not try to predict complete gene structures beyond the available EST-sequence information. To reduce the number of incomplete transcript models, we compared all transcripts and searched for those which were subsets of another transcript concerning sequence and exon-intron structure. The shorter transcripts were tagged incomplete and omitted from our database. We argue, that the shorter transcripts, which have the same exon-intron boundaries as the longer transcripts, did not represent differential splicing with alternative 3' end, but rather incomplete transcript models. To err on the side of caution, transcripts from manually curated databases were kept in our database, even if they were subsets of longer transcript models. This strategy, however, lead to a bias in favor of long transcript models that were generated without manual curation. In the worst case scenario, failure of our strategies to reduce local redundancy, more transcript models would have been judged transcript-specific, but gene-specificity would have been uncompromised.

Finally, our intent was not to create an *once and for all* database merging algorithm, a monumentous task enganging many research groups around the world. Instead we chose to, accomplish the task using sound assumptions and minimal manual intervention for final quality control.

### 4.1.3 Microarray quality control

The manufactured microarray was tested in a series of quality controls. Random primer hybridization was used to grade the success of the spotting procedure. The typical drop out rate was found below five drop outs per thousand spots, well below the rate reported for cDNA microarrays of 10-20%[72]. Although transcript abundance in homogenous cell populations span over six orders of magnitude as was measured for yeast by Holland et al.[75] the typical linear dynamic range measured by microarray technology spans ∼3 orders of magnitude[10,23,162]. The linear dynamic range of 3 to 4 orders of magnitude found in our experiments is in concordance with these findings. The ratio controls on our microarray revealed quenching of ratios by a factor of ∼4-5. A comparison of our microarray results to real-time PCR data revealed a similar quenching factor of 4.1 (SD: 3.4). Although the underestimation of ratios is large, the direction of regulation was found to be correct in all cases. The quenching of ratios might be explained by the short hybridization time (∼16h) and washing conditions of moderate stringency as was pointed out by Sartor et al.[177]

and Korkola et al.[96]. Background correction methods also tend to induce a bias in lowering the absolute ratio[166]. Although all ratios controls were observed to show the right tendency after normalization and the underestimation of ratios did not affect the statistical analysis, these factors must be taken into account in the interpretation of the results.

### 4.1.4 Outlook on microarray technology

At present, our method of microarray analysis is standard for investigations on global gene expression. Sensitivity and specificity issues have been addressed and improved[80] and even the daunting tasks of standardization and statistical analysis are coming of age[1,42,156,193]. However this type of analysis is based on the hybridization of sequences, which embodies a number of problems. First, a relatively small stretch of sequence (the probe) is used to measure a much larger sequence (the transcript). Even if many probes are used to measure a transcript, no array results have been published which correctly identify the sequence of unknown transcripts. Whole genome tiling arrays theoretically allow the capture of much of the complexity of the transcriptome[11,28], but ignore splice junction information and are associated with high costs and difficulties in data analysis. Arrays that specifically detect alternative splicing events[22,150], are hampered by their dependence on *a priori* knowledge of exon junctions present and specificity issues.

The last years have witnessed a constant drop in costs for shotgun sequencing, which now renders transcriptome analysis by sequencing reasonable. The transcript-counting approaches for instance by Kim et al.[93], Sultan et al.[199] and Rosenkranz et al.[171] overcome many of the inherent limitations of array-based systems and bypass problems inherent to analog measurements, including complex normalization procedures, and limitations in detecting low abundance transcripts.

While microarray technology will still be useful for organisms that have not yet been sequenced, or for scientific questions within the limits of this technology, the future of transcriptome analysis in all its complexity will see sequencing as the source of data.

## 4.2 Splicing Factors and Intron-Retention in Inflammatory Bowel Disease

### 4.2.1 Experimental setup

In the present study, primary tissues were used in order to emulate the situation in the human body as closely as possible. The complexity of transcriptomal networks emphasizes the need for semi-quantitative assessment of its components in patients. For investigating the role of splicing factors and intron-retention in chronic inflammation, CD and UC provide excellent conditions: CD and UC are related, but clearly distinct disorders in which the diseased organ is large enough to be inspected visually via endoscopy, and a relatively large amount of tissue can be obtained for ex vivo and in vitro studies (as opposed to rheumatoid arthritis, for example, where the limited amount of inflamed joint tissue hampers large-scale parallel investigations). The interaction between splicing events, splicing factors and disease is still poorly understood. The second part of the thesis presents a systematic analysis of the differential expression of 149 splicing factors and 145 intronic sequences in IBD.

### 4.2.2 IBD subtype-specific regulation of splicing factors

As a result of the study, we found that the altered expression of splicing factors in inflamed mucosa of IBD patients is specific to IBD subtype and IBD in general. In Crohn's disease patients, the biggest differences were found comparing inflamed to non-inflamed tissue, while in ulcerative colitis patients the biggest difference was observed comparing patient tissue (without regard to inflammation) to healthy control subject tissue. We can rule out the possibility that undetected inflammation was the cause for our findings in non-inflamed UC tissue, since the inflammation status was confirmed by the molecular inflammation markers, *IL8* and *REG1A*. Therefore, we conclude that for the seven splicing factors investigated, inflammation drives differential expression of the splicing factors in CD, but not in UC. However, our findings on *SFPQ* and *SF3B14* that were up-regulated in CD non-inflamed tissue as compared to healthy control tissue demonstrate that splicing factor expression changes in CD tissue cannot be explained by inflammation alone.

Several studies[57,99,170,175,185,202,221,224] reported an inflammation-dependent regulation of splicing; however, only one study by Cattaruzza et al.[25] describes the regulatory network of splicing factors in inflammation. Our study aimed to assess whether our findings on the seven splicing factors represented a common inflammatory pattern or an IBD-specific change in splicing factor expression levels. Our experimental setup was designed to provide the statistical power to detect subgroup specific differences. The regulation pattern observed for the seven splicing factors investigated was IBD-specific, as opposed to solely being inflammation-dependent. This offers new options to the researcher in the field of IBD, since *DUSP11*, *HNRPAB*, *HNRPH3*, *SF3B14*, *SFPQ*, *SFR2IP* and *SLU7* are splicing factors that not only show IBD-specific effects of inflammation, but might also be capable of distinguishing Crohn's disease from ulcerative colitis.

### 4.2.3 IBD subtype-specific regulation of intron-retention

Based on our findings for differential expression of splicing factors in IBD, selected intron-retention events were further investigated as potentially pathogenesis-associated splicing mechanisms. The screening approach provided evidence for inflammation and/or disease dependent effects that regulated intron-retention in *IER3*, *FGD2* and *PARC*. A general up- or down-regulation of intron-retention in response to disease or inflammation could not be observed. This suggests a complex regulatory network where the individual splicing events are associated to different pathophysiological processes. Additionally, verification with disease specificity controls demonstrated that the co-expression of intron-retained and intron-excised forms is specific for each subtype of IBD and specific to IBD in general. Comparing the expression patterns of the seven splicing factors investigated in detail with the patterns of intron-retention in *FGD2*, *PARC* and *IER3* did not reveal a common regulatory motif. This indicates that beside the seven splicing factors investigated, additional splicing factors have to be monitored to explain the complex splicing patterns of *FGD2*, *PARC* and *IER3*.

### 4.2.4 Expression of splicing factors and intron-retention

For several model organisms, such as *drosophila* and *mus musculus*, there is evidence that splicing factor concentration regulates and guides the splicing process[55,83,130,135,160,215]. In this study, an *in silico* approach using the software *splicing rainbow*[197] was used to predict splicing factor binding sites surrounding the introns of interest in *IER3*, *FGD2* and *PARC*. This analysis identified several binding sites for splicing factors that were found regulated in IBD by microarray analysis. Although a computational approach has its limitations, our findings present a potential connection between splicing factor expression levels and subsequent splicing patterns in the context of IBD pathology.

### 4.2.5 Intron-retention in light of nonsense-mediated decay

Differential splicing influences the function, the location or the expression level of a protein, and therefore has a functional impact on the fate of a cell[15]. To prevent translation of ill-spliced transcripts, several mechanisms either eliminate such transcripts via nonsense mediated decay (NMD) or prevent their export from the nucleus[33,111,196]. The introns in question of *FGD2* and *PARC* are surrounded by several other introns up- and down-stream. Since an exon junction downstream of a stop codon is mandatory for NMD, both genes are potential candidates for nonsense mediated decay[110]. This hypothesis is supported by fact that the introns under investigation contain stop codons in all three reading frames.

Comparable amounts of co-expression of the intron-excised and the intron-retained forms for *PARC* and *FGD2* suggests that unproductive splicing is taking place. This can rapidly reduce the amount of translated RNA by nonsense mediated decay, a regulatory process recently termed RUST (regulated unproductive splicing and translation)[114]. In light of RUST, our data proposed synergistic down-regulation of *FGD2* in CD tissue and *PARC* in UC tissue due to inflammation. Both cases show down-regulation of the productive splice-form and up-regulation of the unproductive splice-form. While the ratio of the productive to the unproductive splice-form stayed about the same for *FGD2* in UC tissue and for *PARC* in CD tissue, it was nevertheless shifted in favor of the unproductive form as compared to healthy subjects.

In contrast to *FGD2* and *PARC*, *IER3* lacks a further intron which is crucial for NMD and RUST to exert its action[127,231]. Since the intron-retained form of *IER3* is not spliced at all, it may not contain an exon-junction-complex. The exon-junction complex, however, is thought to be crucial for export from the nucleus[110]. It is therefore likely that the mRNA of the intron-retained form of *IER3* cannot be exported the nucleus in this manner, rendering it as unproductive as the intron-retained forms of *FGD2* and *PARC*. However, there are other mechanisms by which intron-retained transcripts can be exported from the nucleus. Studies by Taddeo et al.[203] detected the non-spliced form of *IER3* in the cytosol of cells infected with *herpes simplex virus 1*. To export their non-spliced RNAs from the nucleus, viruses like *HIV1* or *HSV1* express proteins (*Rev*[30] and *icp27*[189]) that direct non-spliced RNAs to the nuclear export complex. Other simpler viruses, such as *simian type D retrovirus* (*SRV*/*MPMV*), depend on structural elements in their RNA that bind to the nuclear export receptors for mRNAs *TAP/NXF1* and *p15/NXT*[62] and act as constitutive transport elements[120,141]. We therefore investigated the presence of a RNA transport element-like structure was present in *IER3*.

### 4.2.6 RNA transport element-like structure in IER3

Non-Spliced RNA usually does not leave the nucleus[110]. One way to circumvent this barrier was discovered in viral RNAs, which contain structural elements (termed constitutive transport elements (CTE) or RNA transport elements (RTE)) that connect to the nuclear export complex. We investigated if the intron-retained form of *IER3* contained RNA transport element (RTE)-like structures using the abstract shape finding software RNAshapes[198]. Intriguingly, we found *in silico* evidence for a RTE-like structure described by Smulevitch et al.[190]. The structure found in *IER3* contains the four stem-loops described as being crucial for RNA export[190]. RTE-like structures have been grouped into classes A to D. One of these structures, RNA transport element D, was found within *IER3*. The overall conservation of sequence was very low in contrast to the four RTE-like structures described by Smulevitch et al.[190]; however, we found three of four conserved consecutive adenine residues within an internal loop of stem-loop IV. These conserved adenine residues are critical for RTE function[190].

The expression and nuclear export of non-spliced *IER3* forms could have novel implications in IBD pathology; therefore, the function of the RTE-like structure in *IER3* remains an interesting candidate for further research.

### 4.2.7 FGD2, PARC and IER3 in IBD

To generalize our findings, the amount of translatable mRNA for *FGD2* and *PARC* is most likely reduced in inflamed IBD compared to healthy subjects. In contrast, *IER3* was found up-regulated in UC tissue and down-regulated in CD tissue due to inflammation.

**FGD2** *Rho* GTPases control many aspects of cell behavior through the regulation of multiple signal transduction pathways[64,213]. Evidence has accumulated to show that in all eukaryotic cells, *Rho* GTPases are involved in most, if not all, actin-dependent processes such as those involved in migration, adhesion, morphogenesis, axon guidance, and phagocytosis[29,85,122]. In addition to their well-established roles in controlling the actin cytoskeleton, *Rho* GTPases regulate the microtubule cytoskeleton, cell polarity, gene expression, cell cycle progression, and membrane transport pathways like a binary switch alternating between GDP- or GTP-bound form[37,45,213]. The cell controls this binary switch by regulating the interconversion and accessibility of these two forms. Guanine nucleotide exchange factors (GEFs) stimulate the exchange of GDP for GTP to generate the activated form and are regulated themselves in a variety of ways.

*FGD2*, together with at least *FGD1*, *FGD3*, *FRG* and *Frabin*, is a member of a family of *CDC42*-specific guanine nucleotide exchange factors[152]. These proteins all have multiple phosphoinositide-binding domains, including two PH domains and a FYVE or FERM domain[181]. It is likely that they couple the actin cytoskeleton with the plasma membrane. GEF activities of *FGD2*, *FGD3*, *FGD5* and *FGD6* have not been reported, but the sequence similarities suggest that they also function as *CDC42*-specific GEFs. Although little has been published on *FGD2* itself, one can nevertheless try to deduce its function from its protein family members. *FGD1* has an essential role in embryonic development and stimulates the GDP-GTP exchange of the isoprenylated form of *CDC42*, a small GTPase of the *Rho*-subfamily . Hayakawa et al.[66] found *FGD1* to be regulated by proteasomal degradation and to stimulate cell motility. Similarily, *FGD3* is regulated by proteasomal degradation, but forms different morphological structures and inhibits cell motility in contrast to *FGD1*[66]. *FRG* was implicated in forming adherens junctions[52]. These findings demonstrate that the highly homologous GEFs – *FGD1*, *FGD3* and *FRG* – play different roles to regulate cellular functions, which makes it difficult to deduce specific *FGD2* functions in the context of IBD. The presumable target of *FGD2*, *CDC42*, can nevertheless be linked to the scheme, which will be drawn for *PARC* and *IER3*. *CDC42* is linked to the *ERK* pathway[104,232] for actions as diverse as granule mobilization and the control of cell death, which is accompanied by activation of *p53*[174].

The implications of GEFs in actin structure reorganization proposes an impact of *FGD2* expression levels on epithelial barrier function in IBD. Wether the down-regulation of *FGD2* in CD inflamed tissue and UC non-inflamed tissue leads directly to an impaired barrier integrity or indirectly alters susceptibility to apoptosis needs to be elucidated.

**PARC** Tumor suppressor *p53*, the guardian of genome integrity, is involved in a network of regulatory factors, one of which, *PARC*, was found regulated in our study on IBD patients. *PARC* is a member of the *Cullin* gene family, which act as scaffolds for ubiquitin ligases (E3). E3 ubiquitin ligases regulate an extensive number of dynamic cellular processes, including multiple aspects of the cell cycle, transcription, signal transduction, and development. *PARC* has been suggested to suppress *p53* activity by binding *p53* in the cytosolic compartment and thereby preventing further action[145]. This view has recently been challenged by *PARC*$^{-/-}$ mice without apparent phenotype, rejecting a prominent effect of *PARC* on *p53* regulation[187]. Next to unchanged transactivation activity of *p53* in radiation challenged cells, no deviation concerning the distribution ratio between nucleus and cytosol was found for *PARC* deficient embryonic stem cells. While binding of *p53* to *PARC* was confirmed by multiple studies in vivo, *PARC* was also found to bind to *CUL7*, a evolutionary descendant of *PARC*, with 60% sequence homology[187]. In contrast to *PARC*, *CUL7* deletion was found lethal during mouse development with severe intra-uterine growth retardation and vascular defects[2]. *CUL7*, *PARC* and *p53* were found to bind each other in multiple combinations in vivo, but only mono- or di-ubiquitination was reported in vitro, with contrasting evidence in vivo, discarding models of *CUL7*/*PARC* mediated *p53* degradation[188]. Down-regulation of *CUL7* augmented *p53*-mediated inhibition of cell cycle progression and ectopic expression of *CUL7* inhibited *p53* levels, leading to a reduced transactivation activity of *p53*[84]. DNA damaging agents together with ectopic expression of *CUL7* increased apoptosis levels, possibly due to the inability of cells to stably arrest. Attempts to deduce complementary function of *PARC* and *CUL7* were turned down by indirect evidence[188]. Although the viability of *PARC*$^{-/-}$ mice and the neonatal lethality of *CUL7*$^{-/-}$ mice suggest that the interaction of *CUL7* and *PARC* is not essential for normal mouse development, specific interactions between *PARC* and

*CUL7* enable *PARC* to regulate *CUL7*. The binding of *PARC* to *CUL7* was found mutually exclusive to *CUL7* binding to *FBX29*, an F-box protein that confers substrate specificity to *Cullin* family proteins[188]. This interaction might allow *PARC* to serve as scaffold protein and define substrate specificity of *CUL7*.

These findings can be integrated into three models. **1)** *PARC* might determine *p53* oligomerization status and influence its interaction partners and/or migration of *p53* into the nucleus. **2)** *PARC* association with *p53* in the cytoplasm might control *p53* translocation to the mitochondria during apoptotic signaling[133] through an indirect mechanism by inhibiting substrate specificity of *CUL7*, which, in turn, modulates the effect of *CUL7* on *p53*. **3)** *PARC* might be a downstream effector of *p53* signaling, which is regulated by *p53* binding.

Within the context of inflammatory bowel disease we found *PARC* to be down-regulated in CD tissue independent from inflammation and down-regulated by inflammation in UC tissue. Down-regulation in UC tissue was probably accelerated by increased intron-retention, followed by nonsense mediated decay of the RNA transcript. The decline of *PARC* expression in IBD might increase the susceptibility to epithelial apoptosis by means of potential abrogated inhibition of *p53* at the mitochondria[133]. In this scenario the down-regulation of *PARC* as potential modulator of *CUL7* signaling might exaggerate the effect of the ~2-fold increased expression of *CUL7*, found in IBD inflammation by microarray analysis [data not shown]. Skaar et al.[187] proposed that the ratio of *CUL7* and *PARC* expression might be important for their combined action, as this ratio is found to differ in a number of tissues. Increased levels of epithelial apoptosis have recently been shown to direct an IBD-like phenotype in mice[144]. Taken together, it becomes clear that the role of *PARC* in IBD will remain elusive unless viewed in a broader context.

**IER3** Recent years have seen tremendous progress in determining *IER3* function. Although its function in detail is still controversial, *IER3* activity has been linked to *ERK* pathway regulation, the ubiquitin pathway and, one of the most important signal transducers in immune system regulation, *NF-κB*.

Despite its involvement with many signal transduction pathways, *IER3* has not yet been examined in detail in IBD. In-depth analysis of our large patient cohort by real-time PCR revealed opposing regulation in CD and UC associated to inflammation. *IER3* was found down-regulated in inflamed CD tissue and up-regulated in inflamed UC tissue.

To disentangle the different functions and connect *IER3* to inflammatory bowel diseases the implications of *IER3* in the *ERK* pathway will be discussed first followed by the ubiquitin pathway and *NF-κB*.

The *ERK* pathway is implicated in diverse cellular processes including proliferation, differentiation and survival. This variety of biological responses is determined by the cell-specific combination of downstream substrates and by differences in the magnitude and kinetics of *ERK* signaling[128,139]. Letourneux et al.[112] point out that the net *ERK* activation is also dependent on the ratio between kinases and phosphatases. Garcia et al.[56] have recently isolated *IER3* as a substrate of *ERK* and attributed a dual role in *ERK* signaling. In response to many stimuli, for instance a combination of *TNFα* and cycloheximide, *IER3* was phosphorylated by *ERK* and found to decrease cell death in different cell types. Additionally, *IER3* was found by Letourneux et al.[112] as a positive regulator of *ERK* signaling, sustaining *ERK* activity by inhibiting the dephosphorylation of *ERK* by *PP2A*. Activation of *p53* by phosphorylation is in part regulated by *ERK*[118] signaling, the activity of which might be prolonged by *IER3*. On the other hand Huang et al. reported that *NF-κB* mediated activation of *IER3* expression was synergized by *p53*[78]. This positive loop might account for the anti-apoptotic effect of *IER3*.

The ubiquitin-proteasome pathway plays a pivotal role in signal transduction[36,220]. By mediating the tightly regulated turnover of a great variety of regulatory proteins, including cyclins, kinases, phosphatases or transcription factors, the ubiquitin-proteasome pathway essentially contributes to the orchestration of multiple cellular processes such as differentiation, proliferation and apoptosis[27,36,39,142,220,230]. The synthesis of proteasomal proteins, as well as the formation and maturation of the 20S and 19S subunits and the assembly of the 26S proteasome itself is a tightly regulated process[8,132]. Any deregulation in proteasome activity leads to a loss in cellular homeostasis and to severe cellular dysfunctions often leading to disease[38,67,109,137].

One of the most prominent signal transduction pathways that are under control of the ubiquitin-proteasome module is the *IκBα* pathway. In response to *NF-κB*-inducing conditions, poly-ubiquitination and subsequent proteasomal degradation of *IκBα* release *NF-κB* from its sequestration in the cytoplasm and promotes its transcriptional activity in the nucleus. As previously shown by Arlt et al.[3], the induced *IκBα* degradation is under control of *IER3*, which was found to inhibit phospho-*I$\kappa$B$\alpha$* degradation by the 26S proteasome[35,226]. It was demonstrated that *IER3* itself is a target gene of *NF-$\kappa$B* and might therefore be part of the *NF-$\kappa$B* self-termination process by preventing *I$\kappa$B$\alpha$* degradation in the presence of stimuli like *IL1$\beta$* or *TNF$\alpha$*[3]. Furthermore, pro-apoptotic stimuli like anti-cancer drugs were shown to down-regulate the activity of the 26S proteasome and executor caspases like *CASP3* cleave parts of the regulatory 19S subunits of the proteasome[200]. The findings on down-regulated proteasomal components have been linked to *IER3*. Arlt et al.[4] demonstrated that *IER3* is able to down-regulate proteasomal proteins independent of caspase activity. Thereby *IER3* acts additive to the decreasing effect of apoptotic stimuli on protein levels of proteasomal components, mainly those of the 19S regulatory subunits. As a consequence, *IER3* prevents the *NF-$\kappa$B*-dependent protection from apoptosis.

While *IER3* in part regulates *NF-$\kappa$B* activity by the proteasomal pathway, a recent study by Arlt et al.[5] demonstrated that *IER3* directly interacts with *NF-$\kappa$B*, possibly as transcriptional co-repressor. The C-terminal domain of *IER3* and the nuclear localization signal were found to be indispensable for *NF-$\kappa$B* binding and inhibition.

Addressing the controversy about the effects of *IER3* on apoptotic signaling, Arlt et al.[5] propose that conflicting results may be attributed by the fact that the studies, which argue for anti-apoptotic function of *IER3*, use quite distinct apoptotic conditions in which *RelA/p65* transactivation is not induced.

To explain how an apparently common theme, inflammation, can be associated to up-regulation of *IER3* in UC tissue and down-regulation of *IER3* in CD tissue, one can take the distinct cytokine profiles of CD and UC into account. While CD is observed to have a Th1-biased inflammatory response, with secretion of *IL2*, *IL12*, *IFN$\gamma$* and/or *TNF$\alpha$*, UC is associated to a Th2-biased inflammation, with secretion of *IL4*, *IL5*, *IL10* and/or *IL13*[149]. Given that inflammatory pathways are tightly interconnected, the different cytokine background of UC and CD, might provide an explanation as to why apparently contradictory regulations of *IER3* can lead to inflammation and disease in both subtypes of IBD.

The traditional paradigm for the pathogenesis of inflammatory bowel diseases holds that bacterial products, such as LPS, trigger acute inflammatory responses because of defects in epithelial barrier function. Kinugasa et al.[94] showed, that regulation of the intestinal barrier is mediated by *IL17* through the *ERK*1/2 pathway.[76] *IL17* signaling via the *ERK* pathway induced the formation of tight junctions and correlated with claudin expression. Down-regulation of *IER3* in CD might shift the balance of *ERK* signaling towards signal termination and thereby challenge barrier integrity.[94] In the same line, gastrointestinal homeostasis is in part restored by mucus cell-secreted proteins of the trefoil gene family, which are also under control of *ERK*1/2 signaling pathways and might be deficient in CD[86,205]. Intestinal epithelial cell differentiation has also been shown to be regulated by the *ERK* pathway[204]. A change in *ERK* activity was found to alter the differentiation status of intestinal epithelial cells[204], which might also affect the susceptibility to cancer, a complication found for inflammatory bowel disease[43,44]

Exposure to food allergens by the oral route can trigger immediate local hypersensitivity reactions in the intestine followed by a late-phase inflammatory response. Uptake of allergen-IgE complexes mediated by *CD23* triggered up-regulation of *IL8* and *CCL20* in epithelial cells via *ERK* and *JNK* pathways[115], a reaction which might be severed through up-regulated *IER3* in UC by extending *ERK* activation.

Recent research by Nenci et al.[144] demonstrated that conditional ablation of *NEMO* in intestinal epithelial cells terminated the *NF-$\kappa$B* pathway and led to severe chronic inflammation of the whole intestine in mice. *NF-$\kappa$B* pathway deficiency led to apoptosis of colonic epithelial cells, impaired expression of antimicrobial peptides and translocation of bacteria into the mucosa. Up-regulated *IER3* in UC tissue fits well into this model, since inhibition of *NF-$\kappa$B*, either via the proteasomal pathway or by direct interaction, increases the vulnerability to apoptosis under inflammatory conditions. In this light, up-regulation of the intron-retained form as found in inflamed UC tissue might reflect an attempt to reduce the levels of *IER3* in the inflamed areas. Intriguingly, the intron-retained mRNA of *IER3*, if translated, would lead to a protein that lacks the C-terminal domain, which was found crucial for *NF-$\kappa$B* inhibition. Therefore regulated intron-retention is potentially able to counteract the function of *IER3* in both scenarios.

Next to the potential impact on mucosal epithelial cells, increased expression of *IER3* has recently been shown to initiate an auto-immune disease in transgenic mice, due to a deficiency in apoptosis of activated T-cells[230]. Furthermore, the gene of *IER3* is located within the HLA region on chromosome 6, which has been associated to susceptibility to inflammatory bowel disease by multiple studies.

The intestinal mucosa represents a key barrier between the hostile environment of

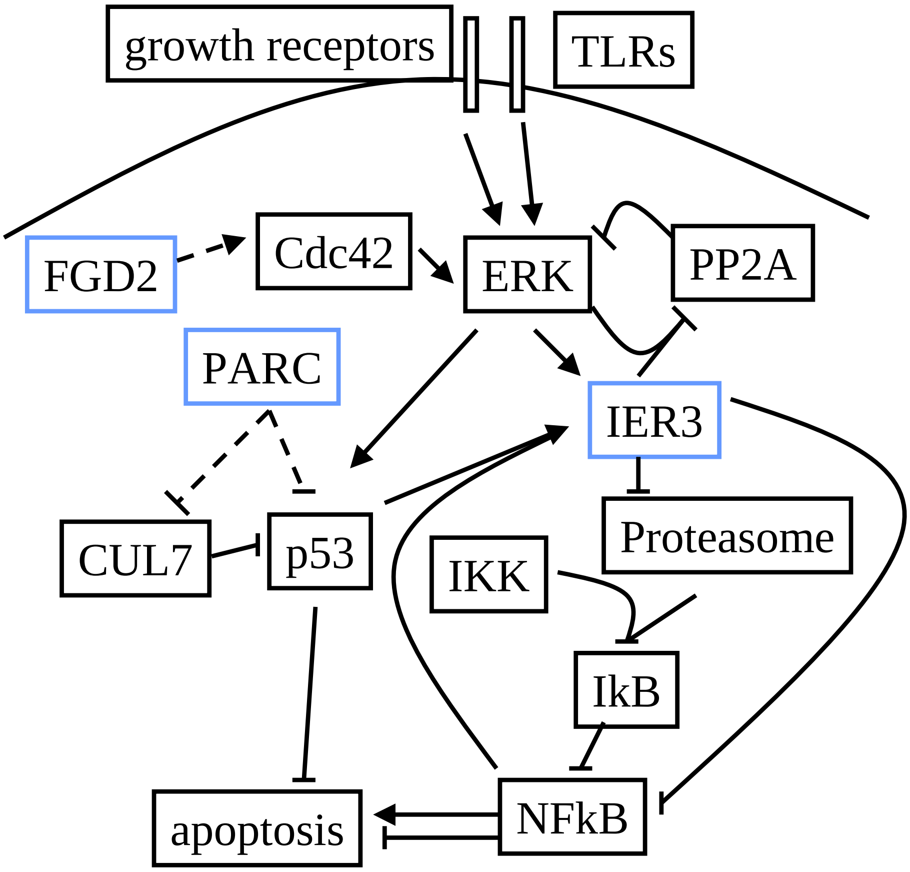


Figure 22: Signaling pathways which involve FGD2, PARC and IER3. The diagram shows a highly simplified model of the pathways involved. Only connections with respect to FGD2, PARC and IER3 have been depicted. Connections leading to activation are plotted as arrows. Connections leading to inhibition are drawn as lines with bar ends. Speculative connections are highlighted by dotted lines.

pathogenic luminal antigens and internal tissues. It is no surprise that the IBD mucosa, compared to normal controls, exhibits differential regulation of many immune-related genes. Figure 22 depicts the interactions of *FGD2*, *PARC* and *IER3* in the pathways mentioned above. Our findings in the light of recent literature render, in increasing importance, *FGD2*, *PARC* and *IER3* as interesting candidates for research to further elucidate their molecular interactions in general and in particular in IBD.

# 5 Summary

Inflammatory bowel disease (IBD) is a relapsing chronic inflammatory disorder of the gastrointestinal tract. IBD and its two major sub-forms, namely Crohn's disease (CD) and ulcerative colitis (UC), are typical complex diseases thought to be caused by the intricate interplay of genetic, environmental and immunological factors . Due to the extensive haplotype structures in the HLA region on chromosome 6, it is impossible to disentangle the genetic association found for IBD and to single out potentially disease causing genes. The present study was based on the premise that disease-associated genes exhibit aberrant expression or splicing patterns.

The first aim was to construct an HLA-microarray to investigate the transcriptome of the HLA linkage region. To gather as many genes as possible, including known and unknown genes, we developed an algorithm to merge four different transcript databases (Ensembl core, Ensembl ESTgene, Vega and RefSeq). We then developed a second algorithm to down-size the local and global redundancy created by pooling the data sources. This unique enrichment strategy tripled and quadrupled the transcript-count in the HLA region compared to the RefSeq database and the Affymetrix HU 133A microarrays, respectively. Microarray probe placement is crucial to discern expression levels of different transcripts of a gene. Therefore, in order to develop the optimal probe placement strategy, we investigated the limits of microarray technology in accordance to recent work on specificity. At most 51% of transcripts and 13% of exons contain a transcript-specific stretch to design oligonucleotide probes of length 50 nt. The chosen strategy designed two probes: on probe to catch the maximum number of transcript models and one probe for each transcript model with transcript-specific sequence. This strategy was shown to be the most efficient among four strategies compared. The HLA chip was manufactured at the Max Planck Institute for Molecular Genetics. Control measures on drop out rate, dynamic range and ratio correctness confirmed the chips' high quality for subsequent expression experiments.

The second objective of this study was to investigate differential gene expression in IBD patient samples (n = 24) and normal controls (n = 6). Specific focus was placed on differential expression of splicing factors and intron-retention in IBD, since aberrant splicing had recently been shown to contribute to a number of human diseases. To date, the role of splicing factors and intron-retention events has not

been described in the context of IBD. This study is the first to identify 47 splicing factors and 33 intron-retention events that were differentially regulated in mucosal tissue of IBD patients.

In the third part of this work, we verified our findings on seven splicing factors (*DUSP11*, *HNRPAB*, *HNRPH3*, *SLU7*, *SFR2IP*, *SFPQ*, *SF3B14*) and three intron-retention events (*IER3*, *FGD2*, *PARC*) in a patient cohort of 195 individuals. Stratification by disease type and inflammation status revealed that most of our results were indeed specific for IBD and its subtypes. *In silico* analysis on *FGD2*, *PARC* and *IER3* identified several binding sites for splicing factors that are regulated in IBD. These results suggest a link between splicing factors and intron-retention in the context of IBD pathology. Additionally, a structural analysis of *IER3* mRNA revealed a RNA transport element-like structure, which could allow the *IER3* intron-retention transcript to travel into the cytoplasm, where it can interfere with *NF-$\kappa$B* pathways and influence IBD pathology.

This study demonstrates for the first time the potential impact of regulated splicing factors and regulated intron-retention in the pathogenesis of chronic inflammation in barrier tissues, exemplified by IBD.

# 6 Zusammenfassung

Morbus Crohn (MC) und Colitis Ulcerosa (CU); Hauptformen der chronisch entzündlichen Darmerkrankungen (CED); weisen steigende Patientenzahlen in industrialisierten Ländern auf. Nach heutigem Verständnis ist ein Zusammenwirken verschiedener Faktoren (Genetik, Umwelt, Immunsystem) wesentlich für die Pathogenese der CED. Aufgrund der genetischen Kopplung innerhalb der HLA-Region auf Chromosom 6 ist es schwierig die genetischen Assoziationen zu CED einzelnen Genen zuzuordnen. Die Annahme, dass pathogenese-relevante Gene ein verändertes Expressions- oder Splice-Muster aufweisen initiierte diese Arbeit.

Erste Aufgabe war die Erstellung eines Microarray zur Untersuchung der Genaktivität der HLA Region. Um den Sequenzraum zu erweitern wurden ein Algorithmus entwickelt, der die Integration von vier Sequenz-Datenbanken (Ensembl core, Ensembl ESTgene, Vega and RefSeq) über die *golden path* Koordinaten ermöglicht. Die lokale und globale Redundanz, die bei Integration von vier Datenbanken entsteht, wurde durch einen weiteren Algorithmus minimiert. Die Anzahl der HLA-Transkripte wurde durch diese neue Strategie gegenüber der RefSeq Datenbank verdreifacht, gegenüber dem Affymetrix Microarray HU 133 vervierfacht. Um die Expressions-Aktivität auf Transkriptebene zu untersuchen, wurde eine Strategie zur Kombination von Microarray Sequenzen (Proben) entwickelt. In diesem Zusammenhang wurden die Spezifitätsgrenzen der Microarray-Technologie ermittelt und die Ergebnisse genutzt um die Effizienz verschiedene Kombinations-Strategien zu vergleichen. Da maximal 50% der Transkripte und 13% der Exons eine transkriptspezifische Sequenz ($>$ 50nt) aufweisen wurde die effizienteste von vier Strategie verwand, welche eine Probe für alle Transkripte eines Gens erstellt neben zusätzlichen Proben pro Transcript mit transkript-spezifischer Sequenz. Der HLA-Chip wurde am MPI für Molekulare Genetik hergestellt. Interne Kontrollen zu dynamischem Umfang sowie Veränderungen der Messwerte (Ratio) zeigten die gute Funktionalität des HLA Microarray.

Der zweite Teil dieser Arbeit analysiert die Expressionsmustern in Darmepithel-Biopsien. 24 Patienten mit CED, sowie 6 Probanten ohne CED wurden mit Hilfe von 60 Microarray Experimenten untersucht. Erstmalig für CED wurden die Expressionsmuster von Splicefaktoren sowie Intron-Sequenzen analysiert, da veränderte Splice-Muster kürzlich mit der Pathogenese verschiedener Erkrankungen assoziiert wurden. 47 der 149 Splicefaktoren und 33 der 145 Intron-Sequenzen zeigten veränderte Expressionsmuster in Krankheit und/oder Entzündung.

Im dritten Teil wurden die Ergebnisse für sieben Splicefaktoren (*DUSP11*, *HNRPAB*, *HNRPH3*, *SLU7*, *SFR2IP*, *SFPQ*, *SF3B14*) und drei Intron-Sequenzen (*IER3*, *FGD2*, *PARC*) in einer Kohorte von 195 Individuen per Real-time-PCR Analyse verifiziert und stratifiziert. Die Resultate zu Splicefaktoren und Intron-Sequenzen erwiesen sich als spezifisch für CED sowie deren Untergruppen. Als Bindeglied zwischen differentieller Expression der Splicefaktoren sowie den Intron-Sequenzen konnten Splicefaktor-Bindestellen für in CED regulierte Splicefaktoren durch "in Silico" Analyse von *FGD2*, *PARC* und *IER3* festgestellt werden. Die mRNA Sequenz von *IER3* wurde zusätzlich einer Strukturanalyse unterzogen und ein Abschnitt identifiziert dessen Struktur der eines RNA Transport Elements ähnelt.

Diese Studie demonstriert erstmalig die Bedeutung der Regulation von Splicefaktoren und Intron-Sequenzen in chronisch entzündlichen Darmerkrankungen.

# 7 Acknowledgments

I am very grateful to Bernd Timmermann, Nancy Mah, Robert Häsler for their helpful and inspiring comments to my thesis, to Dorina Ölsner, Anne Zergiebel, Antje Glöckler, Ninette von der Dellen, Katja Tamms, Brigitte Mauracher for their expert technical assistance, to Claus Hultschig, Thomas Nietzsche, Tom Privlitschig for their expertise in microarray production, to Tim Lu, Peter Croucher, Silvia Mascheretti-Croucher, Andreas Dahl and Prof. Christian Kaltschmidt for inspiring discussions and for encouraging me so many times to move on despite all challenges. I want to thank Wilfried Nietfeld for his lab, his advice and support and finally I want to thank Professor Schreiber, Professor Lehrach and Christine Costello for the initial idea and funding, which have been the foundations of my work.

# References

[1] , Leming Shi, Laura H Reid, Wendell D Jones, Richard Shippy, Janet A Warrington, Shawn C Baker, Patrick J Collins, Francoise de Longueville, Ernest S Kawasaki, Kathleen Y Lee, Yuling Luo, Yongming Andrew Sun, James C Willey, Robert A Setterquist, Gavin M Fischer, Weida Tong, Yvonne P Dragan, David J Dix, Felix W Frueh, Frederico M Goodsaid, Damir Herman, Roderick V Jensen, Charles D Johnson, Edward K Lobenhofer, Raj K Puri, Uwe Schrf, Jean Thierry-Mieg, Charles Wang, Mike Wilson, Paul K Wolber, Lu Zhang, Shashi Amur, Wenjun Bao, Catalin C Barbacioru, Anne Bergstrom Lucas, Vincent Bertholet, Cecilie Boysen, Bud Bromley, Donna Brown, Alan Brunner, Roger Canales, Xiaoxi Megan Cao, Thomas A Cebula, James J Chen, Jing Cheng, Tzu-Ming Chu, Eugene Chudin, John Corson, J Christopher Corton, Lisa J Croner, Christopher Davies, Timothy S Davison, Glenda Delenstarr, Xutao Deng, David Dorris, Aron C Eklund, Xiao-hui Fan, Hong Fang, Stephanie Fulmer-Smentek, James C Fuscoe, Kathryn Gallagher, Weigong Ge, Lei Guo, Xu Guo, Janet Hager, Paul K Haje, Jing Han, Tao Han, Heather C Harbottle, Stephen C Harris, Eli Hatchwell, Craig A Hauser, Susan Hester, Huixiao Hong, Patrick Hurban, Scott A Jackson, Hanlee Ji, Charles R Knight, Winston P Kuo, J Eugene LeClerc, Shawn Levy, Quan-Zhen Li, Chunmei Liu, Ying Liu, Michael J Lombardi, Yunqing Ma, Scott R Magnuson, Botoul Maqsodi, Tim McDaniel, Nan Mei, Ola Myklebost, Baitang Ning, Natalia Novoradovskaya, Michael S Orr, Terry W Osborn, Adam Papallo, Tucker A Patterson, Roger G Perkins, Elizabeth H Peters, Ron Peterson, Kenneth L Philips, P Scott Pine, Lajos Pusztai, Feng Qian, Hongzu Ren, Mitch Rosen, Barry A Rosenzweig, Raymond R Samaha, Mark Schena, Gary P Schroth, Svetlana Shchegrova, Dave D Smith, Frank Staedtler, Zhenqiang Su, Hongmei Sun, Zoltan Szallasi, Zivana Tezak, Danielle Thierry-Mieg, Karol L Thompson, Irina Tikhonova, Yaron Turpaz, Beena Vallanat, Christophe Van, Stephen J Walker, Sue Jane Wang, Yonghong Wang, Russ Wolfinger, Alex Wong, Jie Wu, Chunlin Xiao, Qian Xie, Jun Xu, Wen Yang, Liang Zhang, Sheng Zhong, Yaping Zong, and William Slikker. The microarray quality control (maqc) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nature biotechnology*, 24(9):1151–1161, 2006.

[2] Takehiro Arai, Jocelyn S Kasper, Jeffrey R Skaar, Syed Hamid Ali, Chiaki Takahashi, and James A DeCaprio. Targeted disruption of p185/cul7 gene results in abnormal vascular morphogenesis. *Proceedings of the National Academy of Sciences of the United States of America*, 100(17):9855–9860, 2003.

[3] Alexander Arlt, Marie-Luise Kruse, Maike Breitenbroich, Andre Gehrz, Bülent Koc, Jörg Minkenberg, Ulrich R Fölsch, and Heiner Schäfer. The early response gene iex-1 attenuates nf-kappab activation in 293 cells, a possible counter-regulatory process leading to enhanced cell death. *Oncogene*, 22(21):3343–3351, 2003.

[4] Alexander Arlt, Jörg Minkenberg, Marie-Luise Kruse, Frauke Grohmann, Ulrich R Fölsch, and Heiner Schäfer. Immediate early gene-x1 interferes with 26 s proteasome

activity by attenuating expression of the 19 s proteasomal components s5a/rpn10 and s1/rpn2. *The Biochemical journal*, 402(2):367–375, 2007.

[5] Alexander Arlt, Philip Rosenstiel, Marie-Luise Kruse, Frauke Grohmann, Jörg Minkenberg, Neil D Perkins, Ulrich R Fölsch, Stefan Schreiber, and Heiner Schäfer. Iex-1 directly interferes with rela/p65 dependent transactivation and regulation of apoptosis. *Biochimica et biophysica acta*, 1783(5):941–952, 2008.

[6] L C Bailey, D B Searls, and G C Overton. Analysis of est-driven gene annotation in human genomic sequence. *Genome research*, 8(4):362–376, 1998.

[7] Xavier Bailly, Gilles Béna, Vanina Lenief, Philippe de Lajudie, and Jean-Christophe Avarre. Development of a lab-made microarray for analyzing the genetic diversity of nitrogen fixing symbionts sinorhizobium meliloti and sinorhizobium medicae. *Journal of microbiological methods*, 67(1):114–124, 2006.

[8] W Baumeister, J Walz, F Zühl, and E Seemüller. The proteasome: paradigm of a self-compartmentalizing protease. *Cell*, 92(3):367–380, 1998.

[9] Simon W Beaven and Maria T Abreu. Biomarkers in inflammatory bowel disease. *Current opinion in gastroenterology*, 20(4):318–327, 2004.

[10] Kenneth B Beckman, Kathleen Y Lee, Tamara Golden, and Simon Melov. Gene expression profiling in mitochondrial disease: assessment of microarray accuracy by high-throughput q-pcr. *Mitochondrion*, 4(5-6):453–470, 2004.

[11] Piotr Berman, Paul Bertone, Bhaskar Dasgupta, Mark Gerstein, Ming-Yang Kao, and Michael Snyder. Fast optimal genome tiling with applications to microarray design and homology search. *Journal of computational biology : a journal of computational molecular cell biology*, 11(4):766–785, 2004.

[12] I Biemond, W R Burnham, J D'Amaro, and M J Langman. Hla-a and -b antigens in inflammatory bowel disease. *Gut*, 27(8):934–941, 1986.

[13] Ewan Birney, Michele Clamp, and Richard Durbin. Genewise and genomewise. *Genome research*, 14(5):988–995, 2004.

[14] Tanja Birrenbach and Ulrich Böcker. Inflammatory bowel disease and smoking: a review of epidemiology, pathophysiology, and therapeutic implications. *Inflammatory bowel diseases*, 10(6):848–859, 2004.

[15] Douglas L Black. Mechanisms of alternative pre-messenger rna splicing. *Annual review of biochemistry*, 72, 2003.

[16] Levente Bodrossy and Angela Sessitsch. Oligonucleotide microarrays in microbial diagnostics. *Current opinion in microbiology*, 7(3):245–254, 2004.

[17] G Bouma, B Xia, J B Crusius, G Bioque, I Koutroubakis, B M Von Blomberg, S G Meuwissen, and A S Peña. Distribution of four polymorphisms in the tumour necrosis factor (tnf) genes in patients with inflammatory bowel disease (ibd). *Clinical and experimental immunology*, 103(3):391–396, 1996.

[18] Gerd Bouma and Warren Strober. The immunological and genetic basis of inflammatory bowel disease. *Nature reviews. Immunology*, 3(7):521–533, 2003.

[19] Rose Boutros, Angela M Bailey, Sarah H D Wilson, and Jennifer A Byrne. Alternative splicing as a mechanism for regulating 14-3-3 binding: interactions between hd53 (tpd52l1) and 14-3-3 proteins. *Journal of molecular biology*, 332(3), 2003.

[20] F Bretz, J Landgrebe, and E Brunner. Design and analysis of two-color microarray experiments using linear models. *Methods of information in medicine*, 44(3), 2005.

[21] Sarah L Brown, Terrence E Riehl, Monica R Walker, Michael J Geske, Jason M Doherty, William F Stenson, and Thaddeus S Stappenbeck. Myd88-dependent positioning of ptgs2-expressing stromal cells maintains colonic epithelial proliferation during injury. *The Journal of clinical investigation*, 117(1):258–269, 2007.

[22] John A Calarco, Arneet L Saltzman, Joanna Y Ip, and Benjamin J Blencowe. Technologies for the global discovery and analysis of alternative splicing. *Advances in experimental medicine and biology*, 623:64–84, 2007.

[23] Roger D Canales, Yuling Luo, James C Willey, Bradley Austermiller, Catalin C Barbacioru, Cecilie Boysen, Kathryn Hunkapiller, Roderick V Jensen, Charles R Knight, Kathleen Y Lee, Yunqing Ma, Botoul Maqsodi, Adam Papallo, Elizabeth Herness Peters, Karen Poulter, Patricia L Ruppel, Raymond R Samaha, Leming Shi, Wen Yang, Lu Zhang, and Federico M Goodsaid. Evaluation of dna microarray results with quantitative gene expression platforms. *Nature biotechnology*, 24(9):1115–1122, 2006.

[24] Luca Cartegni and Adrian R Krainer. Disruption of an sf2/asf-dependent exonic splicing enhancer in smn2 causes spinal muscular atrophy in the absence of smn1. *Nature genetics*, 30(4), 2002.

[25] Marco Cattaruzza, Katrin Schäfer, and Markus Hecker. Cytokine-induced downregulation of zfm1/splicing factor-1 promotes smooth muscle cell proliferation. *The Journal of biological chemistry*, 277(8), 2002.

[26] Charles E Chalfant, Kristin Rathman, Ryan L Pinkerman, Rachel E Wood, Lina M Obeid, Besim Ogretmen, and Yusuf A Hannun. De novo ceramide regulates the alternative splicing of caspase 9 and bcl-x in a549 lung adenocarcinoma cells. dependence on protein phosphatase-1. *The Journal of biological chemistry*, 277(15), 2002.

[27] Zhijian J Chen. Ubiquitin signalling in the nf-kappab pathway. *Nature cell biology*, 7(8):758–765, 2005.

[28] Jill Cheng, Philipp Kapranov, Jorg Drenkow, Sujit Dike, Shane Brubaker, Sandeep Patel, Jeffrey Long, David Stern, Hari Tammana, Gregg Helt, Victor Sementchenko, Antonio Piccolboni, Stefan Bekiranov, Dione K Bailey, Madhavan Ganesh, Srinka Ghosh, Ian Bell, Daniela S Gerhard, and Thomas R Gingeras. Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. *Science (New York, N.Y.)*, 308(5725):1149–1154, 2005.

[29] G Chimini and P Chavrier. Function of rho family proteins in actin dynamics during phagocytosis and engulfment. *Nature cell biology*, 2(10):E191–6, 2000.

[30] Alan W Cochrane, Mark T McNally, and Andrew J Mouland. The retrovirus rna trafficking granule: from birth to maturity. *Retrovirology*, 3:18, 2006.

[31] Jörn Coers, Christina Ranft, and Radek C Skoda. A truncated isoform of c-mpl with an essential c-terminal peptide targets the full-length receptor for degradation. *The Journal of biological chemistry*, 279(35), 2004.

[32] ML. Cohen. Changing patterns of infectious disease. *Nature*, 406(6797), 2000.

[33] Elena Conti and Elisa Izaurralde. Nonsense-mediated mrna decay: molecular insights and mechanistic variations across species. *Current opinion in cell biology*, 17(3):316–325, 2005.

[34] Mark Cuff, Jane Dyer, Mark Jones, and Soraya Shirazi-Beechey. The human colonic monocarboxylate transporter isoform 1: its potential importance to colonic tissue homeostasis. *Gastroenterology*, 128(3):676–686, 2005.

[35] L Dahan, C Lepage, M Ouaissi, B Sastre, L Bedenne, and J-F Seitz. [postoperative follow-up in patients with colorectal cancers who have undergone curative resection: Intensive or conventional follow-up?]. *Gastroenterologie clinique et biologique*, 2008.

[36] William S Dalton. The proteasome. *Seminars in oncology*, 31(6 Suppl 16):3–9; discussion 33, 2004.

[37] H Daub, K Gevaert, J Vandekerckhove, A Sobel, and A Hall. Rac/cdc42 and p65pak regulate the microtubule-destabilizing protein stathmin through phosphorylation at serine 16. *The Journal of biological chemistry*, 276(3):1677–1680, 2001.

[38] Anny Devoy, Tim Soane, Rebecca Welchman, and R John Mayer. The ubiquitin-proteasome system and cancer. *Essays in biochemistry*, 41:187–203, 2005.

[39] Sarath C Dhananjayan, Ayesha Ismail, and Zafar Nawaz. Ubiquitin and control of transcription. *Essays in biochemistry*, 41:69–80, 2005.

[40] F Diehl, S Grahlmann, M Beier, and J D Hoheisel. Manufacturing dna microarrays of high spot homogeneity and reduced background signal. *Nucleic acids research*, 29(7), 2001.

[41] R Duchmann, I Kaiser, E Hermann, W Mayet, K Ewe, and K H Meyer zum Büschenfelde. Tolerance exists towards resident intestinal flora but is broken in active inflammatory bowel disease (ibd). *Clinical and experimental immunology*, 102(3):448–455, 1995.

[42] Alain Dupuy and Richard M Simon. Critical review of published microarray studies for cancer outcome and guidelines on statistical analysis and reporting. *Journal of the National Cancer Institute*, 99(2):147–157, 2007.

[43] A Ekbom, C Helmick, M Zack, and H O Adami. Increased risk of large-bowel cancer in crohn's disease with colonic involvement. *Lancet*, 336(8711):357–359, 1990.

[44] A Ekbom, C Helmick, M Zack, and H O Adami. Ulcerative colitis and colorectal cancer. a population-based study. *The New England journal of medicine*, 323(18):1228–1233, 1990.

[45] S Etienne-Manneville and A Hall. Integrin-mediated activation of cdc42 controls cell polarity in migrating astrocytes through pkczeta. *Cell*, 106(4):489–498, 2001.

[46] B Ewing and P Green. Analysis of expressed sequence tags indicates 35,000 human genes. *Nature genetics*, 25(2):232–234, 2000.

[47] Eduardo Eyras, Mario Caccamo, Val Curwen, and Michele Clamp. Estgenes: alternative splicing from ests in ensembl. *Genome research*, 14(5):976–987, 2004.

[48] Nuno André Faustino and Thomas A Cooper. Pre-mrna splicing and human disease. *Genes & development*, 17(4), 2003.

[49] C Fiocchi. Inflammatory bowel disease: etiology and pathogenesis. *Gastroenterology*, 115(1):182–205, 1998.

[50] Sheila A Fisher, Jochen Hampe, Andrew J S Macpherson, Alastair Forbes, John E Lennard-Jones, Stefan Schreiber, Mark E Curran, Christopher G Mathew, and Cathryn M Lewis. Sex stratification of an inflammatory bowel disease genome search shows male-specific linkage to the hla region of chromosome 6. *European journal of human genetics : EJHG*, 10(4):259–265, 2002.

[51] K Fujita, S Naito, N Okabe, and T Yao. Immunological studies in crohn's disease. i. association with hla systems in the japanese. *Journal of clinical & laboratory immunology*, 14(2):99–102, 1984.

[52] Tatsuro Fukuhara, Kazuya Shimizu, Tomomi Kawakatsu, Taihei Fukuyama, Yukiko Minami, Tomoyuki Honda, Takashi Hoshino, Tomohiro Yamada, Hisakazu Ogita, Masato Okada, and Yoshimi Takai. Activation of cdc42 by trans interactions of the cell adhesion molecules nectins through c-src and cdc42-gef frg. *The Journal of cell biology*, 166(3):393–405, 2004.

[53] I J Fuss, M Neurath, M Boirivant, J S Klein, C de la Motte, S A Strong, C Fiocchi, and W Strober. Disparate cd4+ lamina propria (lp) lymphokine secretion profiles in inflammatory bowel disease. crohn's disease lp cells manifest increased secretion of ifn-gamma, whereas ulcerative colitis lp cells manifest increased secretion of il-5. *Journal of immunology (Baltimore, Md. : 1950)*, 157(3):1261–1270, 1996.

[54] H Galkowska, L O Waldemar, and U Wojewodzka. Reactivity of antibodies directed against human antigens with surface markers on canine leukocytes. *Veterinary immunology and immunopathology*, 53(3-4):329–334, 1996.

[55] Huirong Gao, William J Gordon-Kamm, and L Alexander Lyznik. Asf/sf2-like maize pre-mrna splicing factors affect splice site utilization and their transcripts are alternatively spliced. *Gene*, 339, 2004.

[56] Josefina Garcia, Yunbin Ye, Valérie Arranz, Claire Letourneux, Guillaume Pezeron, and Francoise Porteu. Iex-1: a new erk substrate involved in both erk survival activity and erk activation. *The EMBO journal*, 21(19):5151–5163, 2002.

[57] Katrina Gee, Marko Kryworuchko, and Ashok Kumar. Recent advances in the regulation of cd44 expression and its role in inflammation and autoimmune diseases. *Archivum immunologiae et therapiae experimentalis*, 52(1), 2004.

[58] UCSC genome browser. Annotation track refseq genes. http://hgdownload.cse.ucsc.edu/goldenPath/hg17/database/mrnaRefseq.txt.gz, 2005.

[59] Robert C Gentleman, Vincent J Carey, Douglas M Bates, Ben Bolstad, Marcel Dettling, Sandrine Dudoit, Byron Ellis, Laurent Gautier, Yongchao Ge, Jeff Gentry, Kurt Hornik, Torsten Hothorn, Wolfgang Huber, Stefano Iacus, Rafael Irizarry, Friedrich Leisch, Cheng Li, Martin Maechler, Anthony J Rossini, Gunther Sawitzki, Colin Smith, Gordon Smyth, Luke Tierney, Jean Y H Yang, and Jianhua Zhang. Bioconductor: open software development for computational biology and bioinformatics. *Genome biology*, 5(10), 2004.

[60] A T Gewirtz, T A Navas, S Lyons, P J Godowski, and J L Madara. Cutting edge: bacterial flagellin activates basolaterally expressed tlr5 to induce epithelial proinflammatory gene expression. *Journal of immunology (Baltimore, Md. : 1950)*, 167(4):1882–1885, 2001.

[61] M H Gleeson, J S Walker, J Wentzel, J A Chapman, and R Harris. Human leucocyte antigens in crohn's disease and ulcerative colitis. *Gut*, 13(6):438–440, 1972.

[62] P Grüter, C Tabernero, C von Kobbe, C Schmitt, C Saavedra, A Bachi, M Wilm, B K Felber, and E Izaurralde. Tap, the human homolog of mex67p, mediates cte-dependent rna export from the nucleus. *Molecular cell*, 1(5):649–659, 1998.

[63] Stefan A Haas, Marc Hild, Anthony P H Wright, Torsten Hain, Driss Talibi, and Martin Vingron. Genome-scale design of pcr primers and long oligomers for dna microarrays. *Nucleic acids research*, 31(19):5576–5581, 2003.

[64] A Hall. Rho gtpases and the actin cytoskeleton. *Science (New York, N.Y.)*, 279(5350):509–514, 1998.

[65] M Harren, G Schönfelder, M Paul, I Horak, E O Riecken, B Wiedenmann, and M John. High expression of inducible nitric oxide synthase correlates with intestinal inflammation of interleukin-2-deficient mice. *Annals of the New York Academy of Sciences*, 859:210–215, 1998.

[66] Makio Hayakawa, Hideo Kitagawa, Keiji Miyazawa, Masatoshi Kitagawa, and Kiyomi Kikugawa. The fwd1/beta-trcp-mediated degradation pathway establishes a 'turning off switch' of a cdc42 guanine nucleotide exchange factor, fgd1. *Genes to cells : devoted to molecular & cellular mechanisms*, 10(3):241–251, 2005.

[67] T Hayashi and D L Faustman. Implications of altered apoptosis in diabetes mellitus and autoimmune disease. *Apoptosis : an international journal on programmed cell death*, 6(1-2):31–45, 2001.

[68] Barbara A Hendrickson, Ranjana Gokhale, and Judy H Cho. Clinical aspects and pathophysiology of inflammatory bowel disease. *Clinical microbiology reviews*, 15(1):79–94, 2002.

[69] M L Hermiston and J I Gordon. Inflammatory bowel disease and adenomas in mice expressing a dominant negative n-cadherin. *Science (New York, N.Y.)*, 270(5239):1203–1207, 1995.

[70] Michael Hiller, Klaus Huse, Karol Szafranski, Niels Jahn, Jochen Hampe, Stefan Schreiber, Rolf Backofen, and Matthias Platzer. Widespread occurrence of alternative splicing at nagnag acceptors contributes to proteome plasticity. *Nature genetics*, 36(12):1255–1257, 2004.

[71] David C. Hoaglin, Frederick Mosteller, and John Wilder Tukey. *Understanding robust and exploratory data analysis*. Wiley series in probability and mathematical statistics. Applied probability and statistics,. Wiley, New York, 1983. 82008528 edited by David C. Hoaglin, Frederick Mosteller, John W. Tukey. Includes index. Bibliography: p. 427-429.

[72] E Hoffmann. Expression profiling–best practices for data generation and interpretation in clinical trials. *Nature reviews. Genetics*, 5(3):229–237, 2004.

[73] Y Hofmann, C L Lorson, S Stamm, E J Androphy, and B Wirth. Htra2-beta 1 stimulates an exonic splicing enhancer and can restore full-length smn expression to survival motor neuron 2 (smn2). *Proceedings of the National Academy of Sciences of the United States of America*, 97(17), 2000.

[74] J B Hogenesch, K A Ching, S Batalov, A I Su, J R Walker, Y Zhou, S A Kay, P G Schultz, and M P Cooke. A comparison of the celera and ensembl predicted gene sets reveals little overlap in novel genes. *Cell*, 106(4):413–415, 2001.

[75] Michael J Holland. Transcript abundance in yeast varies over six orders of magnitude. *The Journal of biological chemistry*, 277(17):14363–14366, 2002.

[76] Veera Hölttä, Paula Klemetti, Taina Sipponen, Mia Westerholm-Ormio, Guillermo Kociubinski, Harri Salo, Laura Räsänen, Kaija-Leena Kolho, Martti Färkkilä, Erkki Savilahti, and Outi Vaarala. Il-23/il-17 immunity as a hallmark of crohn's disease. *Inflammatory bowel diseases*, 2008.

[77] R. Horton, L. Wilming, V. Rand, RC. Lovering, EA. Bruford, VK. Khodiyar, MJ. Lush, S. Povey, CC. Jr. Talbot, MW. Wright, HM. Wain, J. Trowsdale, A. Ziegler, and S. Beck. Gene map of the extended human mhc. *Nature reviews. Genetics.*, 5(12), 2004.

[78] Yan-Hong Huang, Jim Yujin Wu, Yujin Zhang, and Mei X Wu. Synergistic and opposing regulation of the stress-responsive gene iex-1 by p53, c-myc, and multiple nf-kappab/rel complexes. *Oncogene*, 21(44):6819–6828, 2002.

[79] T Hubbard, D Barker, E Birney, G Cameron, Y Chen, L Clark, T Cox, J Cuff, V Curwen, T Down, R Durbin, E Eyras, J Gilbert, M Hammond, L Huminiecki, A Kasprzyk, H Lehvaslaiho, P Lijnzaad, C Melsopp, E Mongin, R Pettett, M Pocock, S Potter, A Rust, E Schmidt, S Searle, G Slater, J Smith, W Spooner, A Stabenau, J Stalker, E Stupka, A Ureta-Vidal, I Vastrik, and M Clamp. The ensembl genome database project. *Nucleic acids research*, 30(1), 2002.

[80] T R Hughes, M Mao, A R Jones, J Burchard, M J Marton, K W Shannon, S M Lefkowitz, M Ziman, J M Schelter, M R Meyer, S Kobayashi, C Davis, H Dai, Y D He, S B Stephaniants, G Cavet, W L Walker, A West, E Coffey, D D Shoemaker, R Stoughton, A P Blanchard, S H Friend, and P S Linsley. Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer. *Nature biotechnology*, 19(4):342–347, 2001.

[81] J P Hugot, M Chamaillard, H Zouali, S Lesage, J P Cézard, J Belaiche, S Almer, C Tysk, C A O'Morain, M Gassull, V Binder, Y Finkel, A Cortot, R Modigliani, P Laurent-Puig, C Gower-Rousseau, J Macry, J F Colombel, M Sahbatou, and G Thomas. Association of nod2 leucine-rich repeat variants with susceptibility to crohn's disease. *Nature*, 411(6837):599–603, 2001.

[82] Jean-Pierre Hugot, Corinne Alberti, Dominique Berrebi, Edouard Bingen, and Jean-Pierre Cézard. Crohn's disease: the cold chain hypothesis. *Lancet*, 362(9400):2012–2015, 2003.

[83] H Jumaa and P J Nielsen. The splicing factor srp20 modifies splicing of its own mrna and asf/sf2 antagonizes this regulation. *The EMBO journal*, 16(16), 1997.

[84] Peter Jung, Berlinda Verdoodt, Aaron Bailey, John R Yates, Antje Menssen, and Heiko Hermeking. Induction of cullin 7 by dna damage attenuates p53 function. *Proceedings of the National Academy of Sciences of the United States of America*, 104(27):11388–11393, 2007.

[85] K Kaibuchi, S Kuroda, and M Amano. Regulation of the cytoskeleton and cell adhesion by the rho family gtpases in mammalian cells. *Annual review of biochemistry*, 68:459–486, 1999.

[86] M Kanai, C Mullen, and D K Podolsky. Intestinal trefoil factor induces inactivation of extracellular signal-regulated protein kinase in intestinal epithelial cells. *Proceedings of the National Academy of Sciences of the United States of America*, 95(1):178–182, 1998.

[87] M D Kane, T A Jatkoe, C R Stumpf, J Lu, J D Thomas, and S J Madore. Assessment of the sensitivity and specificity of oligonucleotide (50mer) microarrays. *Nucleic acids research*, 28(22):4552–4557, 2000.

[88] D Karolchik, R Baertsch, M Diekhans, T S Furey, A Hinrichs, Y T Lu, K M Roskin, M Schwartz, C W Sugnet, D J Thomas, R J Weber, D Haussler, W J Kent, and . The ucsc genome browser database. *Nucleic acids research*, 31(1):51–54, 2003.

[89] W James Kent. Blat–the blast-like alignment tool. *Genome research*, 12(4):656–664, 2002.

[90] W James Kent, Charles W Sugnet, Terrence S Furey, Krishna M Roskin, Tom H Pringle, Alan M Zahler, and David Haussler. The human genome browser at ucsc. *Genome research*, 12(6):996–1006, 2002.

[91] M K Kerr and G A Churchill. Experimental design for gene expression microarrays. *Biostatistics (Oxford, England)*, 2(2):183–201, 2001.

[92] K Kett, T O Rognum, and P Brandtzaeg. Mucosal subclass distribution of immunoglobulin g-producing cells is different in ulcerative colitis and crohn's disease of the colon. *Gastroenterology*, 93(5):919–924, 1987.

[93] Jae Bum Kim, Gregory J Porreca, Lei Song, Steven C Greenway, Joshua M Gorham, George M Church, Christine E Seidman, and J G Seidman. Polony multiplex analysis of gene expression (pmage) in mouse hypertrophic cardiomyopathy. *Science (New York, N.Y.)*, 316(5830):1481–1484, 2007.

[94] T Kinugasa, T Sakaguchi, X Gu, and H C Reinecker. Claudins regulate the intestinal barrier in response to immune mediators. *Gastroenterology*, 118(6):1001–1011, 2000.

[95] Eyal Klement, Regev V Cohen, Jonathan Boxman, Aviva Joseph, and Shimon Reif. Breastfeeding and risk of inflammatory bowel disease: a systematic review with meta-analysis. *The American journal of clinical nutrition*, 80(5):1342–1352, 2004.

[96] James E Korkola, Anne L Estep, Sunanda Pejavar, Sandy DeVries, Ronald Jensen, and Frederic M Waldman. Optimizing stringency for expression microarrays. *BioTechniques*, 35(4):828–835, 2003.

[97] Joshua R Korzenik. Past and current theories of etiology of ibd: toothpaste, worms, and refrigerators. *Journal of clinical gastroenterology*, 39(4 Suppl 2):S59–65, 2005.

[98] M M Kosiewicz, C C Nast, A Krishnan, J Rivera-Nieves, C A Moskaluk, S Matsumoto, K Kozaiwa, and F Cominelli. Th1-type responses mediate spontaneous ileitis in a novel murine model of crohn's disease. *The Journal of clinical investigation*, 107(6):695–702, 2001.

[99] A T Kotsimbos and Q Hamid. Il-5 and il-5 receptor in asthma. *Memórias do Instituto Oswaldo Cruz*, 92 Suppl 2, 1997.

[100] Kosuke Kozaiwa, Kazuhiko Sugawara, Michael F Smith, Virginia Carl, Vladimir Yamschikov, Brian Belyea, Sherri B McEwen, Christopher A Moskaluk, Theresa T Pizarro, Fabio Cominelli, and Marcia McDuffie. Identification of a quantitative trait locus for ileitis in a spontaneous mouse model of crohn's disease: Samp1/yitfc. *Gastroenterology*, 125(2):477–490, 2003.

[101] Thomas A Kraus, Adam Cheifetz, Lisa Toy, Jonathan B Meddings, and Lloyd Mayer. Evidence for a genetic defect in oral tolerance induction in inflammatory bowel disease. *Inflammatory bowel diseases*, 12(2):82–8; discussion 81, 2006.

[102] Thomas A Kraus, Lisa Toy, Lisa Chan, Joseph Childs, and Lloyd Mayer. Failure to induce oral tolerance to a soluble protein in patients with inflammatory bowel disease. *Gastroenterology*, 126(7):1771–1778, 2004.

[103] Ken-ichi Kucho, Hidekatsu Yoneda, Manabu Harada, and Masahiro Ishiura. Determinants of sensitivity and specificity in spotted dna microarrays with unmodified oligonucleotides. *Genes & genetic systems*, 79(4):189–197, 2004.

[104] J M Kyriakis and J Avruch. Protein kinase cascades activated by stress and inflammatory cytokines. *BioEssays : news and reviews in molecular, cellular and developmental biology*, 18(7):567–577, 1996.

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

J D McPherson, M A Marra, E R Mardis, L A Fulton, A T Chinwalla, K H Pepin, W R Gish, S L Chissoe, M C Wendl, K D Delehaunty, T L Miner, A Delehaunty, J B Kramer, L L Cook, R S Fulton, D L Johnson, P J Minx, S W Clifton, T Hawkins, E Branscomb, P Predki, P Richardson, S Wenning, T Slezak, N Doggett, J F Cheng, A Olsen, S Lucas, C Elkin, E Uberbacher, M Frazier, R A Gibbs, D M Muzny, S E Scherer, J B Bouck, E J Sodergren, K C Worley, C M Rives, J H Gorrell, M L Metzker, S L Naylor, R S Kucherlapati, D L Nelson, G M Weinstock, Y Sakaki, A Fujiyama, M Hattori, T Yada, A Toyoda, T Itoh, C Kawagoe, H Watanabe, Y Totoki, T Taylor, J Weissenbach, R Heilig, W Saurin, F Artiguenave, P Brottier, T Bruls, E Pelletier, C Robert, P Wincker, D R Smith, L Doucette-Stamm, M Rubenfield, K Weinstock, H M Lee, J Dubois, A Rosenthal, M Platzer, G Nyakatura, S Taudien, A Rump, H Yang, J Yu, J Wang, G Huang, J Gu, L Hood, L Rowen, A Madan, S Qin, R W Davis, N A Federspiel, A P Abola, M J Proctor, R M Myers, J Schmutz, M Dickson, J Grimwood, D R Cox, M V Olson, R Kaul, C Raymond, N Shimizu, K Kawasaki, S Minoshima, G A Evans, M Athanasiou, R Schultz, B A Roe, F Chen, H Pan, J Ramser, H Lehrach, R Reinhardt, W R McCombie, M de la Bastide, N Dedhia, H Blöcker, K Hornischer, G Nordsiek, R Agarwala, L Aravind, J A Bailey, A Bateman, S Batzoglou, E Birney, P Bork, D G Brown, C B Burge, L Cerutti, H C Chen, D Church, M Clamp, R R Copley, T Doerks, S R Eddy, E E Eichler, T S Furey, J Galagan, J G Gilbert, C Harmon, Y Hayashizaki, D Haussler, H Hermjakob, K Hokamp, W Jang, L S Johnson, T A Jones, S Kasif, A Kaspryzk, S Kennedy, W J Kent, P Kitts, E V Koonin, I Korf, D Kulp, D Lancet, T M Lowe, A McLysaght, T Mikkelsen, J V Moran, N Mulder, V J Pollara, C P Ponting, G Schuler, J Schultz, G Slater, A F Smit, E Stupka, J Szustakowski, D Thierry-Mieg, J Thierry-Mieg, L Wagner, J Wallis, R Wheeler, A Williams, Y I Wolf, K H Wolfe, S P Yang, R F Yeh, F Collins, M S Guyer, J Peterson, A Felsenfeld, K A Wetterstrand, A Patrinos, M J Morgan, P de Jong, J J Catanese, K Osoegawa, H Shizuya, S Choi, Y J Chen, J Szustakowki, and . Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, 2001.

[106] Jobst Landgrebe, Frank Bretz, and Edgar Brunner. Efficient two-sample designs for microarray experiments with biological replications. *In silico biology*, 4(4), 2004.

[107] Thomas P Larsson, Christian G Murray, Tobias Hill, Robert Fredriksson, and Helgi B Schiöth. Comparison of the current refseq, ensembl and est databases for counting genes and gene discovery. *FEBS letters*, 579(3):690–698, 2005.

[108] D A Lashkari, J L DeRisi, J H McCusker, A F Namath, C Gentile, S Y Hwang, P O Brown, and R W Davis. Yeast microarrays for genome wide parallel genetic and gene expression analysis. *Proceedings of the National Academy of Sciences of the United States of America*, 94(24):13057–13062, 1997.

[109] Robert Layfield, James Lowe, and Lynn Bedford. The ubiquitin-proteasome system and neurodegenerative disorders. *Essays in biochemistry*, 41:157–171, 2005.

[110] H Le Hir, D Gatfield, E Izaurralde, and M J Moore. The exon-exon junction complex provides a binding platform for factors involved in mrna export and nonsense-mediated mrna decay. *The EMBO journal*, 20(17), 2001.

[111] Fabrice Lejeune and Lynne E Maquat. Mechanistic links between nonsense-mediated mrna decay and pre-mrna splicing in mammalian cells. *Current opinion in cell biology*, 17(3):309–315, 2005.

[112] Claire Letourneux, Géraldine Rocher, and Francoise Porteu. B56-containing pp2a dephosphorylate erk and their activity is controlled by the early gene iex-1 and erk. *The EMBO journal*, 25(4):727–738, 2006.

[113] Jaroslaw Letowski, Roland Brousseau, and Luke Masson. Designing better probes: effect of probe size, mismatch position and number on hybridization in dna oligonucleotide microarrays. *Journal of microbiological methods*, 57(2):269–278, 2004.

[114] Benjamin P Lewis, Richard E Green, and Steven E Brenner. Evidence for the widespread coupling of alternative splicing and nonsense-mediated mrna decay in humans. *Proceedings of the National Academy of Sciences of the United States of America*, 100(1), 2003.

[115] Hongxing Li, Mirna Chehade, Weicheng Liu, Huabao Xiong, Lloyd Mayer, and M Cecilia Berin. Allergen-ige complexes trigger cd23-dependent ccl20 release from human intestinal epithelial cells. *Gastroenterology*, 133(6):1905–1915, 2007.

[116] Shuyu Li, Gene Cutler, Jane Jijun Liu, Timothy Hoey, Liangbiao Chen, Peter G Schultz, Jiayu Liao, and Xuefeng Bruce Ling. A comparative analysis of hgsc and celera human genome assemblies and gene sets. *Bioinformatics (Oxford, England)*, 19(13):1597–1605, 2003.

[117] F Liang, I Holt, G Pertea, S Karamycheva, S L Salzberg, and J Quackenbush. Gene index analysis of the human genome estimates approximately 120,000 genes. *Nature genetics*, 25(2):239–240, 2000.

[118] T Lin, N K Mak, and M S Yang. Mapk regulate p53-dependent cell death induced by benzo[a]pyrene: involvement of p53 phosphorylation and acetylation. *Toxicology*, 247(2-3):145–153, 2008.

[119] E Lindberg, C Tysk, K Andersson, and G Järnerot. Smoking and inflammatory bowel disease. a case control study. *Gut*, 29(3):352–357, 1988.

[120] Susan Lindtner, Barbara K Felber, and Jørgen Kjems. An element in the 3' untranslated region of human line-1 retrotransposon mrna binds nxf1(tap) and can function as a nuclear export element. *RNA (New York, N.Y.)*, 8(3):345–356, 2002.

[121] D J Lockhart, H Dong, M C Byrne, M T Follettie, M V Gallo, M S Chee, M Mittmann, C Wang, M Kobayashi, H Horton, and E L Brown. Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature biotechnology*, 14(13):1675–1680, 1996.

[122] L Luo. Rho gtpases in neuronal morphogenesis. *Nature reviews. Neuroscience*, 1(3):173–180, 2000.

[123] Y Ma, J D Ohmen, Z Li, L G Bentley, C McElree, S Pressman, S R Targan, N Fischel-Ghodsian, J I Rotter, and H Yang. A genome-wide search identifies potential new susceptibility loci for crohn's disease. *Inflammatory bowel diseases*, 5(4):271–278, 1999.

[124] M Mähler, I J Bristol, E H Leiter, A E Workman, E H Birkenmeier, C O Elson, and J P Sundberg. Differential susceptibility of inbred mouse strains to dextran sulfate sodium-induced colitis. *The American journal of physiology*, 274(3 Pt 1):G544–51, 1998.

[125] Michael Mähler, Claudia Most, Sybille Schmidtke, John P Sundberg, Renhua Li, Hans Jürgen Hedrich, and Gary A Churchill. Genetics of colitis susceptibility in il-10-deficient mice: backcross versus f2 results contrasted by principal component analysis. *Genomics*, 80(3):274–282, 2002.

[126] Peter J Mannon, Ivan J Fuss, Lloyd Mayer, Charles O Elson, William J Sandborn, Daniel Present, Ben Dolin, Nancy Goodman, Catherine Groden, Ronald L Hornung, Martha Quezado, Zhiqiong Yang, Markus F Neurath, Jochen Salfeld, Geertruida M Veldman, Ullrich Schwertschlag, Warren Strober, and . Anti-interleukin-12 antibody for active crohn's disease. *The New England journal of medicine*, 351(20):2069–2079, 2004.

[127] L E Maquat and X Li. Mammalian heat shock p70 and histone h4 transcripts, which derive from naturally intronless genes, are immune to nonsense-mediated decay. *RNA (New York, N.Y.)*, 7(3), 2001.

[128] C J Marshall. Specificity of receptor tyrosine kinase signaling: transient versus sustained extracellular signal-regulated kinase activation. *Cell*, 80(2):179–185, 1995.

[129] J F Mayberry and J Rhodes. Epidemiological aspects of crohn's disease: a review of the literature. *Gut*, 25(8):886–899, 1984.

[130] A Mayeda and A R Krainer. Regulation of alternative pre-mrna splicing by hnrnp a1 and splicing factor sf2. *Cell*, 68(2), 1992.

[131] M A Meijssen, S L Brandwein, H C Reinecker, A K Bhan, and D K Podolsky. Alteration of gene expression by intestinal epithelial cells precedes colitis in interleukin-2-deficient mice. *The American journal of physiology*, 274(3 Pt 1):G472–9, 1998.

[132] Silke Meiners, Dirk Heyken, Andrea Weller, Antje Ludwig, Karl Stangl, Peter-M Kloetzel, and Elke Krüger. Inhibition of proteasome activity induces concerted expression of proteasome genes and de novo formation of mammalian proteasomes. *The Journal of biological chemistry*, 278(24):21517–21525, 2003.

[133] Motohiro Mihara, Susan Erster, Alexander Zaika, Oleksi Petrenko, Thomas Chittenden, Petr Pancoska, and Ute M Moll. p53 has a direct apoptogenic role at the mitochondria. *Molecular cell*, 11(3):577–590, 2003.

[134] B Modrek, A Resch, C Grasso, and C Lee. Genome-wide detection of alternative splicing in expressed sequences of human genes. *Nucleic acids research*, 29(13):2850–2859, 2001.

[135] M Molin and G Akusjärvi. Overexpression of essential splicing factor asf/sf2 blocks the temporal shift in adenovirus pre-mrna splicing and reduces virus progeny formation. *Journal of virology*, 74(19), 2000.

[136] G Monteleone, L Biancone, R Marasco, G Morrone, O Marasco, F Luzza, and F Pallone. Interleukin 12 is expressed and actively released by crohn's disease intestinal lamina propria mononuclear cells. *Gastroenterology*, 112(4):1169–1178, 1997.

[137] John D Mountz. Significance of increased circulating proteasome in autoimmune disease. *The Journal of rheumatology*, 29(10):2027–2030, 2002.

[138] AJ. Mungall, SA. Palmer, SK. Sims, CA. Edwards, JL. Ashurst, L. Wilming, MC. Jones, R. Horton, SE. Hunt, CE. Scott, JG. Gilbert, ME. Clamp, G. Bethel, S. Milne, R. Ainscough, JP. Almeida, KD. Ambrose, TD. Andrews, RI. Ashwell, AK. Babbage, CL. Bagguley, J. Bailey, R. Banerjee, DJ. Barker, KF. Barlow, K. Bates, DM. Beare, H. Beasley, O. Beasley, CP. Bird, S. Blakey, S. Bray-Allen, J. Brook, AJ. Brown, JY. Brown, DC. Burford, W. Burrill, J. Burton, C. Carder, NP. Carter, JC. Chapman, SY. Clark, G. Clark, CM. Clee, S. Clegg, V. Cobley, RE. Collier, JE. Collins, LK. Colman, NR. Corby, GJ. Coville, KM. Culley, P. Dhami, J. Davies, M. Dunn, ME. Earthrowl, AE. Ellington, KA. Evans, L. Faulkner, MD. Francis, A. Frankish, J. Frankland, L. French, P. Garner, J. Garnett, MJ. Ghori, LM. Gilby, CJ. Gillson, RJ. Glithero, DV. Grafham, M. Grant, S. Gribble, C. Griffiths, M. Griffiths, R. Hall, KS. Halls, S. Hammond, JL. Harley, EA. Hart, PD. Heath, R. Heathcott, SJ. Holmes, PJ. Howden, KL. Howe, GR. Howell, E. Huckle, SJ. Humphray, MD. Humphries, AR. Hunt, CM. Johnson, AA. Joy, M. Kay, SJ. Keenan, AM. Kimberley, A. King, GK. Laird, C. Langfordm, S. Lawlor, DA. Leongamornlert, M. Leversha, CR. Lloyd, DM. Lloyd, JE. Loveland, J. Lovell, S. Martin, M. Mashreghi-Mohammadi, GL. Maslen, L. Matthews, OT. McCann, SJ. McLaren, K. McLay, A. McMurray, MJ. Moore, JC. Mullikin, D. Niblett, T. Nickerson, KL. Novik, K. Oliver, EK. Overton, Larty, A. Parker, R. Patel, AV. Pearce, AI. Peck, B. Phillimore, S. Phillips, RW. Plumb, KM. Porter, Y. Ramsey, SA. Ranby, CM. Rice, MT. Ross, SM. Searle, HK. Sehra, E. Sheridan, CD. Skuce, S. Smith, M. Smith, L. Spraggon, SL. Squares, CA. Steward, N. Sycamore, G. Tamlyn-Hall, J. Tester, AJ. Theaker, DW. Thomas, A. Thorpe, A. Tracey, A. Tromans, B. Tubby, M. Wall, JM. Wallis, AP. West, SS. White, SL. Whitehead, H. Whittaker, A. Wild, DJ. Willey, TE. Wilmer, JM. Wood, PW. Wray, JC. Wyatt, L. Young, RM. Younger, DR. Bentley, A. Coulson, R. Durbin, T. Hubbard, JE. Sulston, I. Dunham, J. Rogers, and Beck S. The dna sequence and analysis of human chromosome 6. *Nature*, 425(6960), 2003.

[139] Leon O Murphy, Sallie Smith, Rey-Huei Chen, Diane C Fingar, and John Blenis. Molecular interpretation of erk signal duration by immediate early gene products. *Nature cell biology*, 4(8):556–564, 2002.

[140] A Nakajima, N Matsuhashi, T Kodama, Y Yazaki, M Takazoe, and A Kimura. Hla-linked susceptibility and resistance genes in crohn's disease. *Gastroenterology*, 109(5):1462–1467, 1995.

[141] F Nappi, R Schneider, A Zolotukhin, S Smulevitch, D Michalowski, J Bear, B K Felber, and G N Pavlakis. Identification of a novel posttranscriptional regulatory element by using a rev- and rre-mutated human immunodeficiency virus type 1 dna proviral clone as a molecular trap. *Journal of virology*, 75(10):4558–4569, 2001.

[142] Cord Naujokat and Stephan Hoffmann. Role and function of the 26s proteasome in proliferation and apoptosis. *Laboratory investigation; a journal of technical methods and pathology*, 82(8):965–980, 2002.

[143] Maria E Nelson, Paul J Thurmes, James D Hoyer, and David P Steensma. A novel 5' atrx mutation with splicing consequences in acquired alpha thalassemia-myelodysplastic syndrome. *Haematologica*, 90(11), 2005.

[144] Arianna Nenci, Christoph Becker, Andy Wullaert, Ralph Gareus, Geert van Loo, Silvio Danese, Marion Huth, Alexei Nikolaev, Clemens Neufert, Blair Madison, Deborah Gumucio, Markus F Neurath, and Manolis Pasparakis. Epithelial nemo links innate immunity to chronic intestinal inflammation. *Nature*, 446(7135):557–561, 2007.

[145] Anatoly Y Nikolaev, Muyang Li, Norbert Puskas, Jun Qin, and Wei Gu. Parc: a cytoplasmic anchor for p53. *Cell*, 112(1):29–40, 2003.

[146] Malka Nissim-Rafinia and Batsheva Kerem. The splicing machinery is a genetic modifier of disease severity. *Trends in genetics : TIG*, 21(9), 2005.

[147] S. Ohno. *Evolution by Gene Duplication*. Springer, Berlin, 1970.

[148] M Orholm, V Binder, T I Sørensen, L P Rasmussen, and K O Kyvik. Concordance of inflammatory bowel disease among danish twins. results of a nationwide study. *Scandinavian journal of gastroenterology*, 35(10):1075–1081, 2000.

[149] F Pallone and G Monteleone. Interleukin 12 and th1 responses in inflammatory bowel disease. *Gut*, 43(6):735–736, 1998.

[150] Qun Pan, Ofer Shai, Christine Misquitta, Wen Zhang, Arneet L Saltzman, Naveed Mohammad, Tomas Babak, Henry Siu, Timothy R Hughes, Quaid D Morris, Brendan J Frey, and Benjamin J Blencowe. Revealing global regulatory features of mammalian alternative splicing using a quantitative microarray platform. *Molecular cell*, 16(6):929–941, 2004.

[151] P Parronchi, P Romagnani, F Annunziato, S Sampognaro, A Becchio, L Giannarini, E Maggi, C Pupilli, F Tonelli, and S Romagnani. Type 1 t-helper cell predominance and interleukin-12 expression in the gut of patients with crohn's disease. *The American journal of pathology*, 150(3):823–832, 1997.

[152] N G Pasteris and J L Gorski. Isolation, characterization, and mapping of the mouse and human fgd2 genes, faciogenital dysplasia (fgd1; aarskog syndrome) gene homologues. *Genomics*, 60(1):57–66, 1999.

[153] P G Persson, A Ahlbom, and G Hellers. Diet and inflammatory bowel disease: a case-control study. *Epidemiology (Cambridge, Mass.)*, 3(1):47–52, 1992.

[154] S E Plevy, S R Targan, H Yang, D Fernandez, J I Rotter, and H Toyoda. Tumor necrosis factor microsatellites define a crohn's disease-associated haplotype on chromosome 6. *Gastroenterology*, 110(4):1053–1060, 1996.

[155] D K Podolsky. Inflammatory bowel disease (1). *The New England journal of medicine*, 325(13):928–937, 1991.

[156] Stan Pounds and Cheng Cheng. Statistical development and evaluation of microarray gene expression data filters. *Journal of computational biology : a journal of computational molecular cell biology*, 12(4):482–495, 2005.

[157] C S Probert, V Jayanthi, D S Rampton, and J F Mayberry. Epidemiology of inflammatory bowel disease in different ethnic and religious groups: limitations and aetiological clues. *International journal of colorectal disease*, 11(1):25–28, 1996.

[158] Kim D Pruitt, Tatiana Tatusova, and Donna R Maglott. Ncbi reference sequences (refseq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic acids research*, 35(Database issue):D61–5, 2007.

[159] J Purrmann, S Korsten, J Bertrams, B Miller, B Lapsien, H Münch, H E Reis, and G Strohmeyer. [hla haplotype study in familial crohn disease]. *Zeitschrift für Gastroenterologie*, 23(8):432–437, 1985.

[160] Junlin Qi, Shihuang Su, M Elaine McGuffin, and William Mattox. Concentration dependent selection of targets by an sr splicing regulator results in tissue-specific rna processing. *Nucleic acids research*, 34(21), 2006.

[161] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2006. ISBN 3-900051-07-0.

[162] Dilip Rajagopalan. A comparison of statistical methods for analysis of high density oligonucleotide array data. *Bioinformatics (Oxford, England)*, 19(12):1469–1476, 2003.

[163] Miguel Regueiro, Kevin E Kip, Onki Cheung, Refaat A Hegazi, and Scott Plevy. Cigarette smoking and age at diagnosis of inflammatory bowel disease. *Inflammatory bowel diseases*, 11(1):42–47, 2005.

[164] S Reif, I Klein, F Lubin, M Farbstein, A Hallak, and T Gilat. Pre-illness dietary factors in inflammatory bowel disease. *Gut*, 40(6):754–760, 1997.

[165] Angela Relógio, Christian Schwager, Alexandra Richter, Wilhelm Ansorge, and Juan Valcárcel. Optimization of oligonucleotide-based dna microarrays. *Nucleic acids research*, 30(11):e51, 2002.

[166] Matthew E Ritchie, Jeremy Silver, Alicia Oshlack, Melissa Holmes, Dileepa Diyagama, Andrew Holloway, and Gordon K Smyth. A comparison of background correction methods for two-colour microarrays. *Bioinformatics (Oxford, England)*, 23(20):2700–2707, 2007.

[167] S N Rodin, D V Parkhomchuk, and A D Riggs. Epigenetic changes and repositioning determine the evolutionary fate of duplicated genes. *Biochemistry*, 70(5):559–567, 2005.

[168] Sergei N Rodin and Arthur D Riggs. Epigenetic silencing may aid evolution by gene duplication. *Journal of molecular evolution*, 56(6):718–729, 2003.

[169] W E Roediger, A Duncan, O Kapaniris, and S Millard. Reducing sulfur compounds of the colon impair colonocyte nutrition: implications for ulcerative colitis. *Gastroenterology*, 104(3):802–809, 1993.

[170] Stefan Rose-John, Jürgen Scheller, Greg Elson, and Simon A Jones. Interleukin-6 biology is coordinated by membrane-bound and soluble receptors: role in inflammation and cancer. *Journal of leukocyte biology*, 80(2), 2006.

[171] Ruben Rosenkranz, Tatiana Borodina, Hans Lehrach, and Heinz Himmelbauer. Characterizing the mouse es cell transcriptome with illumina sequencing. *Genomics*, 2008.

[172] J I Rotter. Inflammatory bowel disease. *Lancet*, 343(8909):1360, 1994.

[173] S. Rozen and H. Skaletsky. Primer3 on the www for general users and for biologist programmers. *Methods Mol Biol*, 132:365–86, 2000. 1064-3745 (Print) Journal Article.

[174] Wilfrid Rul, Olivier Zugasti, Pierre Roux, Carole Peyssonnaux, Alain Eychene, Thomas F Franke, Philippe Lenormand, Philippe Fort, and Ursula Hibner. Activation of erk, controlled by rac1 and cdc42 via akt, is required for anoikis. *Annals of the New York Academy of Sciences*, 973:145–148, 2002.

[175] S Russwurm, M Wiederhold, M Oberhoffer, I Stonans, P F Zipfel, and K Reinhart. Molecular aspects and natural source of procalcitonin. *Clinical chemistry and laboratory medicine : CCLM / FESCC*, 37(8), 1999.

[176] B Sadlack, H Merz, H Schorle, A Schimpl, A C Feller, and I Horak. Ulcerative colitis-like disease in mice with a disrupted interleukin-2 gene. *Cell*, 75(2):253–261, 1993.

[177] Maureen Sartor, Jennifer Schwanekamp, Danielle Halbleib, Ismail Mohamed, Saikumar Karyala, Mario Medvedovic, and Craig R Tomlinson. Microarray results improve significantly as hybridization approaches equilibrium. *BioTechniques*, 36(5):790–796, 2004.

[178] A Saxon, F Shanahan, C Landers, T Ganz, and S Targan. A distinct subset of antineutrophil cytoplasmic antibodies is associated with inflammatory bowel disease. *The Journal of allergy and clinical immunology*, 86(2):202–210, 1990.

[179] M Schena, D Shalon, R W Davis, and P O Brown. Quantitative monitoring of gene expression patterns with a complementary dna microarray. *Science (New York, N.Y.)*, 270(5235):467–470, 1995.

[180] W Scheppach, S U Christl, H P Bartram, F Richter, and H Kasper. Effects of short-chain fatty acids on the inflamed colonic mucosa. *Scandinavian journal of gastroenterology. Supplement*, 222:53–57, 1997.

[181] Anja Schmidt and Alan Hall. Guanine nucleotide exchange factors for rho gtpases: turning on the switch. *Genes & development*, 16(13):1587–1609, 2002.

[182] M Schultz, S L Tonkonogy, R K Sellon, C Veltkamp, V L Godfrey, J Kwon, W B Grenther, E Balish, I Horak, and R B Sartor. Il-2-deficient mice raised under germfree conditions develop delayed mild focal intestinal inflammation. *The American journal of physiology*, 276(6 Pt 1):G1461–72, 1999.

[183] Christian Schwerk and Klaus Schulze-Osthoff. Regulation of apoptosis by alternative pre-mrna splicing. *Molecular cell*, 19(1), 2005.

[184] Xinwei She, Zhaoshi Jiang, Royden A Clark, Ge Liu, Ze Cheng, Eray Tuzun, Deanna M Church, Granger Sutton, Aaron L Halpern, and Evan E Eichler. Shotgun sequence assembly and recent segmental duplications within the human genome. *Nature*, 431(7011):927–930, 2004.

[185] Futoshi Shibata. [the role of rat cytokine-induced neutrophil chemoattractants (cincs) in inflammation]. *Yakugaku zasshi : Journal of the Pharmaceutical Society of Japan*, 122(4), 2002.

[186] S Shivananda, J Lennard-Jones, R Logan, N Fear, A Price, L Carpenter, and M van Blankenstein. Incidence of inflammatory bowel disease across europe: is there a difference between north and south? results of the european collaborative study on inflammatory bowel disease (ec-ibd). *Gut*, 39(5):690–697, 1996.

[187] Jeffrey R Skaar, Takehiro Arai, and James A DeCaprio. Dimerization of cul7 and parc is not required for all cul7 functions and mouse development. *Molecular and cellular biology*, 25(13):5579–5589, 2005.

[188] Jeffrey R Skaar, Laurence Florens, Takeya Tsutsumi, Takehiro Arai, Adriana Tron, Selene K Swanson, Michael P Washburn, and James A DeCaprio. Parc and cul7 form atypical cullin ring ligase complexes. *Cancer research*, 67(5):2006–2014, 2007.

[189] R W P Smith, P Malik, and J B Clements. The herpes simplex virus icp27 protein: a multifunctional post-transcriptional regulator of gene expression. *Biochemical Society transactions*, 33(Pt 3):499–501, 2005.

[190] Sergey Smulevitch, Daniel Michalowski, Andrei S Zolotukhin, Ralf Schneider, Jenifer Bear, Patricia Roth, George N Pavlakis, and Barbara K Felber. Structural and functional analysis of the rna transport element, a member of an extensive family present in the mouse genome. *Journal of virology*, 79(4), 2005.

[191] G. K. Smyth and T. Speed. Normalization of cdna microarray data. *Methods*, 31(4):265–73, 2003. 1046-2023 (Print) Journal Article.

[192] Gordon K Smyth. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical applications in genetics and molecular biology*, 3, 2004.

[193] Gordon K Smyth, Yee Hwa Yang, and Terry Speed. Statistical issues in cdna microarray data analysis. *Methods in molecular biology (Clifton, N.J.)*, 224:111–136, 2003.

[194] E Southern, K Mir, and M Shchepinov. Molecular interactions on microarrays. *Nature genetics*, 21(1 Suppl):5–9, 1999.

[195] S M Srinivasula, M Ahmad, J H Lin, J L Poyet, T Fernandes-Alnemri, P N Tsichlis, and E S Alnemri. Clap, a novel caspase recruitment domain-containing protein in the tumor necrosis factor receptor pathway, regulates nf-kappab activation and apoptosis. *The Journal of biological chemistry*, 274(25), 1999.

[196] Stefan Stamm, Shani Ben-Ari, Ilona Rafalska, Yesheng Tang, Zhaiyi Zhang, Debra Toiber, T A Thanaraj, and Hermona Soreq. Function of alternative splicing. *Gene*, 344:1–20, 2005.

[197] Stefan Stamm, Jean-Jack Riethoven, Vincent Le Texier, Chellappa Gopalakrishnan, Vasudev Kumanduri, Yesheng Tang, Nuno L Barbosa-Morais, and Thangavel Alphonse Thanaraj. Asd: a bioinformatics resource on alternative splicing. *Nucleic acids research*, 34(Database issue), 2006.

[198] Peter Steffen, Björn Voss, Marc Rehmsmeier, Jens Reeder, and Robert Giegerich. Rnashapes: an integrated rna analysis package based on abstract shapes. *Bioinformatics (Oxford, England)*, 22(4):500–503, 2006.

[199] Marc Sultan, Marcel H Schulz, Hugues Richard, Alon Magen, Andreas Klingenhoff, Matthias Scherf, Martin Seifert, Tatjana Borodina, Aleksey Soldatov, Dmitri Parkhomchuk, Dominic Schmidt, Sean O'Keeffe, Stefan Haas, Martin Vingron, Hans Lehrach, and Marie-Laure Yaspo. A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science (New York, N.Y.)*, 2008.

[200] Xiao-Ming Sun, Michael Butterworth, Marion MacFarlane, Wolfgang Dubiel, Aaron Ciechanover, and Gerald M Cohen. Caspase activation inhibits proteasome function during apoptosis. *Molecular cell*, 14(1):81–93, 2004.

[201] Leslie C Sutherland, Nina D Rintala-Maki, Ryan D White, and Cory D Morin. Rna binding motif (rbm) proteins: a novel family of apoptosis modulators? *Journal of cellular biochemistry*, 94(1), 2005.

[202] Björn Szotowski, Silvio Antoniak, and Ursula Rauch. Alternatively spliced tissue factor: a previously unknown piece in the puzzle of hemostasis. *Trends in cardiovascular medicine*, 16(5), 2006.

[203] Brunella Taddeo, Audrey Esclatine, Weiran Zhang, and Bernard Roizman. The stress-inducible immediate-early responsive gene iex-1 is activated in cells infected with herpes simplex virus 1, but several viral mechanisms, including 3' degradation of its rna, preclude expression of the gene. *Journal of virology*, 77(11):6178–6187, 2003.

[204] D Taupin and D K Podolsky. Mitogen-activated protein kinase activation regulates intestinal epithelial differentiation. *Gastroenterology*, 116(5):1072–1080, 1999.

[205] D Taupin, D C Wu, W K Jeon, K Devaney, T C Wang, and D K Podolsky. The trefoil gene family are coordinately expressed immediate-early genes: Egf receptor- and map kinase-dependent interregulation. *The Journal of clinical investigation*, 103(9):R31–8, 1999.

[206] The SSAT, AGA, ASLD, ASGE, AHPBA Consensus Panel. Ulcerative colitis and colon carcinoma: epidemiology, surveillance, diagnosis, and treatment. the society for surgery of the alimentary tract, american gastroenterological association american society for liver diseases, american society for gastrointestinal endoscopy, american hepato-pancreato-biliary association. *Journal of gastrointestinal surgery : official journal of the Society for Surgery of the Alimentary Tract*, 2(4):305–306, 1998.

[207] N P Thompson, R Driscoll, R E Pounder, and A J Wakefield. Genetics versus environment in inflammatory bowel disease: results of a british twin study. *BMJ (Clinical research ed.)*, 312(7023):95–96, 1996.

[208] J. L. Tiwari and P. I. Terasaki. *HLA and Disease Associations.* New York: Springer-Verlag, 1985.

[209] H Toyoda, S J Wang, H Y Yang, A Redford, D Magalong, D Tyan, C K McElree, S R Pressman, F Shanahan, and S R Targan. Distinct associations of hla class ii genes with inflammatory bowel disease. *Gastroenterology*, 104(3):741–748, 1993.

[210] J. Trowsdale. Hla genomics in the third millennium. *Curr Opin Immunol.*, 5(17), 2005.

[211] C Tysk, E Lindberg, G Järnerot, and B Flodérus-Myrhed. Ulcerative colitis and crohn's disease in an unselected population of monozygotic and dizygotic twins. a study of heritability and the influence of smoking. *Gut*, 29(7):990–996, 1988.

[212] Markus Utech, Andrei I Ivanov, Stanislav N Samarin, Matthias Bruewer, Jerrold R Turner, Randall J Mrsny, Charles A Parkos, and Asma Nusrat. Mechanism of ifn-gamma-induced endocytosis of tight junction proteins: myosin ii-dependent vacuolarization of the apical plasma membrane. *Molecular biology of the cell*, 16(10):5040–5052, 2005.

[213] L Van Aelst and C D'Souza-Schorey. Rho gtpases and signaling networks. *Genes & development*, 11(18):2295–2322, 1997.

[214] E M van den Berg-Loonen, B J Dekker-Saeys, S G Meuwissen, L E Nijenhuis, and C P Engelfriet. Histocompatibility antigens and other genetic markers in ankylosing spondylitis and inflammatory bowel diseases. *Journal of immunogenetics*, 4(3):167–175, 1977.

[215] Julian P Venables, Cyril F Bourgeois, Caroline Dalgliesh, Liliane Kister, James Stevenin, and David J Elliott. Up-regulation of the ubiquitous alternative splicing factor tra2beta causes inclusion of a germ cell-specific exon. *Human molecular genetics*, 14(16), 2005.

[216] J C Venter, M D Adams, E W Myers, P W Li, R J Mural, G G Sutton, H O Smith, M Yandell, C A Evans, R A Holt, J D Gocayne, P Amanatides, R M Ballew, D H Huson, J R Wortman, Q Zhang, C D Kodira, X H Zheng, L Chen, M Skupski, G Subramanian, P D Thomas, J Zhang, G L Gabor Miklos, C Nelson, S Broder, A G Clark, J Nadeau, V A McKusick, N Zinder, A J Levine, R J Roberts, M Simon, C Slayman, M Hunkapiller, R Bolanos, A Delcher, I Dew, D Fasulo, M Flanigan, L Florea, A Halpern, S Hannenhalli, S Kravitz, S Levy, C Mobarry, K Reinert, K Remington, J Abu-Threideh, E Beasley, K Biddick, V Bonazzi, R Brandon, M Cargill, I Chandramouliswaran, R Charlab, K Chaturvedi, Z Deng, V Di Francesco, P Dunn, K Eilbeck, C Evangelista, A E Gabrielian, W Gan, W Ge, F Gong, Z Gu, P Guan, T J Heiman, M E Higgins, R R Ji, Z Ke, K A Ketchum, Z Lai, Y Lei, Z Li, J Li, Y Liang, X Lin, F Lu, G V Merkulov, N Milshina, H M Moore, A K Naik, V A Narayan, B Neelam, D Nusskern, D B Rusch, S Salzberg, W Shao, B Shue, J Sun, Z Wang, A Wang, X Wang, J Wang, M Wei, R Wides, C Xiao, C Yan, A Yao, J Ye, M Zhan, W Zhang, H Zhang, Q Zhao, L Zheng, F Zhong, W Zhong, S Zhu, S Zhao, D Gilbert, S Baumhueter, G Spier, C Carter, A Cravchik, T Woodage, F Ali, H An, A Awe,

D Baldwin, H Baden, M Barnstead, I Barrow, K Beeson, D Busam, A Carver, A Center, M L Cheng, L Curry, S Danaher, L Davenport, R Desilets, S Dietz, K Dodson, L Doup, S Ferriera, N Garg, A Gluecksmann, B Hart, J Haynes, C Haynes, C Heiner, S Hladun, D Hostin, J Houck, T Howland, C Ibegwam, J Johnson, F Kalush, L Kline, S Koduru, A Love, F Mann, D May, S McCawley, T McIntosh, I McMullen, M Moy, L Moy, B Murphy, K Nelson, C Pfannkoch, E Pratts, V Puri, H Qureshi, M Reardon, R Rodriguez, Y H Rogers, D Romblad, B Ruhfel, R Scott, C Sitter, M Smallwood, E Stewart, R Strong, E Suh, R Thomas, N N Tint, S Tse, C Vech, G Wang, J Wetter, S Williams, M Williams, S Windsor, E Winn-Deen, K Wolfe, J Zaveri, K Zaveri, J F Abril, R Guigó, M J Campbell, K V Sjolander, B Karlak, A Kejariwal, H Mi, B Lazareva, T Hatton, A Narechania, K Diemer, A Muruganujan, N Guo, S Sato, V Bafna, S Istrail, R Lippert, R Schwartz, B Walenz, S Yooseph, D Allen, A Basu, J Baxendale, L Blick, M Caminha, J Carnes-Stine, P Caulk, Y H Chiang, M Coyne, C Dahlke, A Mays, M Dombroski, M Donnelly, D Ely, S Esparham, C Fosler, H Gire, S Glanowski, K Glasser, A Glodek, M Gorokhov, K Graham, B Gropman, M Harris, J Heil, S Henderson, J Hoover, D Jennings, C Jordan, J Jordan, J Kasha, L Kagan, C Kraft, A Levitsky, M Lewis, X Liu, J Lopez, D Ma, W Majoros, J McDaniel, S Murphy, M Newman, T Nguyen, N Nguyen, M Nodell, S Pan, J Peck, M Peterson, W Rowe, R Sanders, J Scott, M Simpson, T Smith, A Sprague, T Stockwell, R Turner, E Venter, M Wang, M Wen, D Wu, M Wu, A Xia, A Zandieh, and X Zhu. The sequence of the human genome. *Science (New York, N.Y.)*, 291(5507):1304–1351, 2001.

[217] TJ. Vyse and JA. Todd. Genetic analysis of autoimmune disease. *Cell*, 3(85), 1996.

[218] J V Weinstock, R Summers, and D E Elliott. Helminths and harmony. *Gut*, 53(1):7–9, 2004.

[219] Joel V Weinstock, Arthur Blum, Ahmed Metwali, David Elliott, Nigel Bunnett, and Razvan Arsenescu. Substance p regulates th1-type colitis in il-10 knockout mice. *Journal of immunology (Baltimore, Md. : 1950)*, 171(7):3762–3767, 2003.

[220] Rebecca L Welchman, Colin Gordon, and R John Mayer. Ubiquitin and ubiquitin-like proteins as multifunctional signals. *Nature reviews. Molecular cell biology*, 6(8):599–609, 2005.

[221] Christine A Wells, Timothy Ravasi, and David A Hume. Inflammation suppressor genes: please switch out all the lights. *Journal of leukocyte biology*, 78(1), 2005.

[222] M Wills-Karp, J Santeliz, and C L Karp. The germless theory of allergic disease: revisiting the hygiene hypothesis. *Nature reviews. Immunology*, 1(1):69–75, 2001.

[223] L G Wilming, J G R Gilbert, K Howe, S Trevanion, T Hubbard, and J L Harrow. The vertebrate genome annotation (vega) database. *Nucleic acids research*, 36(Database issue):D753–60, 2008.

[224] Bianca M Wittig, Andreas Stallmach, Martin Zeitz, and Ursula Günthert. Functional involvement of cd44 variant 7 in gut immune response. *Pathobiology : journal of immunopathology, molecular and cellular biology*, 70(3), 2003.

[225] T G Wolfsberg and D Landsman. A comparison of expressed sequence tags (ests) to human genomic sequences. *Nucleic acids research*, 25(8):1626–1632, 1997.

[226] M X Wu. Roles of the stress-induced gene iex-1 in regulation of cell death and oncogenesis. *Apoptosis : an international journal on programmed cell death*, 8(1):11–18, 2003.

[227] Guo-Xi Xie, Yuka Yanagisawa, Emi Ito, Kazuo Maruyama, Xiaokang Han, Ki Jun Kim, Kyung Ream Han, Kumi Moriyama, and Pamela Pierce Palmer. N-terminally truncated variant of the mouse gaip/rgs19 lacks selectivity of full-length gaip/rgs19 protein in regulating orl1 receptor signaling. *Journal of molecular biology*, 353(5), 2005.

[228] H Yang, K D Taylor, and J I Rotter. Inflammatory bowel disease. i. genetic epidemiology. *Molecular genetics and metabolism*, 74(1-2):1–21, 2001.

[229] Dongmei Ye, Iris Ma, and Thomas Y Ma. Molecular mechanism of tumor necrosis factor-alpha modulation of intestinal epithelial tight junction barrier. *American journal of physiology. Gastrointestinal and liver physiology*, 290(3):G496–504, 2006.

[230] Huang-Ge Zhang, Jianhua Wang, Xinwen Yang, Hui-Chen Hsu, and John D Mountz. Regulation of apoptosis proteins in cancer cells by ubiquitin. *Oncogene*, 23(11):2009–2015, 2004.

[231] J Zhang, X Sun, Y Qian, J P LaDuca, and L E Maquat. At least one intron is required for the nonsense-mediated decay of triosephosphate isomerase mrna: a possible link between nuclear splicing and cytoplasmic translation. *Molecular and cellular biology*, 18(9), 1998.

[232] Bin Zhong, Kun Jiang, Danielle L Gilvary, Pearlie K Epling-Burnette, Connie Ritchey, Jinhong Liu, Rosalind J Jackson, Elizabeth Hong-Geller, and Sheng Wei. Human neutrophils utilize a rac/cdc42-dependent mapk pathway to direct intracellular granule mobilization toward ingested microbial pathogens. *Blood*, 101(8):3240–3248, 2003.

[233] Z. Zhou, L. J. Licklider, S. P. Gygi, and R. Reed. Comprehensive proteomic analysis of the human spliceosome. *Nature*, 419(6903):182–5, 2002. 0028-0836 (Print) Journal Article.

# A Materials

Table 17: Materials

| Material | Provider |
|---|---|
| 10 X MOPS | Sigma; München, Germany |
| 100 bp DNA ladder | Invitrogen; Karlsruhe, Germany |
| 20 X SSC | Sigma; München, Germany |
| 2-Mercaptoethanol | Sigma; München, Germany |
| 37% Formaldehyde | Sigma; München, Germany |
| 384-deep-well storage plate | ABgene, Epsom, UK |
| ABI PRISM© 96-Well Optical Reaction Plate | Applied Biosystems; Weiterstadt, Germany |
| Agarose | Eurogentec; Köln, Germany |
| AmpliTaq© DNA Polymerase | Applied Biosystems; Weiterstadt, Germany |
| AmpliTaq© Gold DNA Polymerase | Applied Biosystems; Weiterstadt, Germany |
| Cryotubes (2ml) | Greiner Bio-One; Frickenhausen, Germany |
| Diethyl pyrocarbonate (DEPC) | Sigma; München, Germany |
| dNTP set (100mM solutions 100$\mu$M each ) | Amersham Biosciences; Freiburg, Germany |
| Easy peel heat seal foil | ABgene, Epsom, UK |
| EDTA | Sigma; München, Germany |
| Ethanol p. a. | Merck; Darmstadt, Germany |
| Ethidium bromide solution (10mg/ml) | Invitrogen; Karlsruhe, Germany |
| GeneAmp© PCR buffer system 10x | Applied Biosystems; Weiterstadt, Germany |
| Hybridization blocking reagent | Roche Diagnostics, Mannheim, Germany |
| Isopropanol | Merck; Darmstadt, Germany |
| MicroAmp© optical 96 well reaction plate | Applied Biosystems; Weiterstadt, Germany |
| MicroAmp© single strips | Applied Biosystems; Weiterstadt, Germany |
| Microtiter 384 well plates | Sarstedt; Nürnberg, Germany |
| Microtiter 96 well plates | Costar Corning Incorporated; Cambridge, |
| MultiScribe Reverse Transcriptase | Applied Biosystems; Weiterstadt, Germany |
| Oligotex mRNA Mini Kit | Qiagen, Hilden, Germany |
| Phosphate buffered saline (PBS) | Invitrogen/Gibco; Karlsruhe, Germany |
| Pipette tips with filter (10 / 200 / 1000 $\mu$l) | Sarstedt; Nürnberg, Germany |
| Power SYBR© Green PCR Master Mix | Applied Biosystems; Weiterstadt, Germany |
| Primers | Biotez, Berlin, Germany |
| Primers and Probes | Applied Biosystems; Weiterstadt, Germany |
| RNA Ladder | Invitrogen/Gibco; Karlsruhe, Germany |
| RNase-Free DNase Set | Qiagen, Hilden, Germany |
| RNeasy mini RNA extraction kit | Qiagen, Hilden, Germany |
| Salmon sperm DNA | Sigma; München, Germany |
| SmartLadder DNA marker | Eurogentec; Köln, Germany |
| Sodium acetate | Sigma; München, Germany |
| Sodium carbonate | Sigma; München, Germany |
| Sodium chloride | Merck; Darmstadt, Germany |
| Sodium citrate | Sigma; München, Germany |
| Sodium dodecyl sulfate | Sigma; München, Germany |
| Sodium hydrogen carbonate | Sigma; München, Germany |

Continued on Next Page...

Table – Continued

| Material | Provider |
|---|---|
| Sodium hydroxide | Merck; Darmstadt, Germany |
| Sodium phosphate | Sigma; München, Germany |
| SuperScript II$^{TM}$reverse transcriptase | Invitrogen, Carlsbad, USA |
| TAE Buffer 25x Ready | Pack Amresco; Solon, OH, USA |
| TaqMan© Universal PCR Master Mix | Applied Biosystems; Weiterstadt, Germany |
| Tris | Merck; Darmstadt, Germany |
| Tubes (0.5 / 1.5 / 2.0 mL) | Eppendorf; Köln, Germany |
| Tubes, sterile (15 mL) | Sarstedt; Nürnberg, Germany |
| Tubes, sterile (50 mL) | BD Biosciences; Heidelberg, Germany |
| Amino Allyl MessageAmp$^{TM}$Kit | Ambion, Cambridge, UK |
| 10x DNA gel loading buffer | 50% v/v glycerol, 0.1% bromophenol blue w/v |
| RNA gel loading buffer | 40 % (v/v) formaldehyde, 40% (v/v) formamide , 0.9 X MOPS, 0.3$\mu$g/$\mu$L ethidium bromide, 1 mM EDTA (pH 7.5), 1 X DNA loading buffer |
| TE (pH 7.5, 8.0) | 10 mM Tris-HCl, 1 mM EDTA |
| DEPC treated water | 1 mL DEPC in 1 L ddH$_2$O, autoclaved |

# B Equipment

Table 18: Equipment

| | |
|---|---|
| ABI 9700 Sequencer | Applied Biosystems, Foster City, USA |
| ABI Prism 7900HT | Applied Biosystems, Foster City, USA |
| Agilent 2100 Bioanalyzer | Agilent Technologies, Palo Alto, USA |
| GeneAmp PCR system 9700 | Applied Biosystems, Foster City, USA |
| GenePix 3000 scanner | Molecular Devices, Sunnyvale, USA |
| Centrifuge (Labofuge 400R) | Heraeus, Osterode, Germany |
| Homogenizer for microfuge tubes | Teflon head Omnilab, Bremen, Germany |
| Microfuge (Biofuge fresco) | Heraeus Instruments, Osterode, Germany |
| Shaking incubator (GFL 3033) | Gesells. f. Labortechnik, Burgwedel, Germany |
| Thermomixer compact | Eppendorf, Hamburg, Germany |
| Ultrospec 3100 pro spectrophotometer | Amersham Biotech, Uppsala, Sweden |
| Vortex-genie 2 | Scientific Industries, Bohemia, USA |