School of Medicine

MSc Bioinformatics

Masters Thesis

Identification of gene expression signatures and patterns of disease and treatments in rheumatoid arthritis

Author:
Maria Tsochatzidou

Supervisor:
Christoforos Nikolaou

ΠΑΝΕΠΙΣΤΗΜΙΟ ΚΡΗΤΗΣ
UNIVERSITY OF CRETE

Contents

List of Figures

Abstract

This work is an attempt of identification of gene signatures and patterns in rheumatoid arthritis and widely used anti-TNF treatments. The experimental procedure included wild-type (healthy), huTNF-transgenic (diseased) and huTNF-transgenic mice treated with one anti-TNF treatments: infliximab, adalimumab, etanercept or certolizumab pegol with a total of 63 samples. Total RNA was isolated from aqueous extracts of the animals' whole ankle joints which where analyzed with Affy Mouse Gene 2.0 standard array. After data pre-processing two biclustering algorithms, Plaid (Turner et al., 2005) and ISA (Csárdi et al., 2010), were applied, in order to locate sub-matrices of the initial dataset with distinct expression patterns between different combinations of samples. Plaid is a distribution parameter identification method which was combined with an ensemble method to take into account strict and looser thresholds (Kaiser and Leisch, 2008, Kasim et al., 2016) as well as multiple initialization seeds. ISA is a greedy algorithm which is initiated from an input seed that corresponds to a set of genes or samples. The set is improved at each iteration by adding and/or removing genes and/or samples until convergence is reached with a stable set that is evaluated through correlation of rows and columns. Goal of this procedure was to link treatments and conditions through common expression patterns of the genes that participated in each group and extract interesting functions from these modules. The results showed that mostly mild patterns were identifiable in our dataset. Some interesting functions also emerged, but sample representation was inadequate in most of these modules either due to the strict analysis applied or due to unsuitability of the method for the current data.

Our next goal, was to identify a unique group of genes among each of the 4 treatment and the two conditions (Tg and Wt) that would maximize the distance between Tg samples' expression profiles and huTNF-transgenic treated samples while minimizing the distance between the latter and the Wt samples. For this reason, an implementation of a genetic algorithm (GA) (Holland, 1975) approach was designed, which forms subs-groups of possible solutions each one called "chromosome", evaluates them through a fitness function and produces new solutions based on the Darwinian concept of evolution. The advantage of this procedure is the flexibility of the fitness function. In our case, a version of Dunn index was used which is widely applied as a measure of clustering quality. Its advantage is that takes into account distances from both centroids (Wt and Tg in our case). After multiple independent runs, we concluded that the algorithm converges to multiple sub-optimal solutions that are similar only in terms of fitness and not in terms of genes that participate. Thus, this approach was not dependable when applied on the current dataset.

Chapter 1

Introduction: Rheumatoid arthritis and anti-TNF treatments

1.1 Rhematoid arthritis

Rheumatoid rthritis (RA) is a chronic,inflammatory, autoimmune disease that results in progressive articular destruction and associated conditions in vascular, metabolic, bone, and psychological domains. RA affects about 1% of the population, can develop at any age, and is more prevalent in women than in men by an approximate ratio of 4 to 1. The primary pathogenesis is thought to be caused by autoimmune dysfunction (McInnes and Schett, 2017)

Among the main characteristics of the disease are: synovial inflammation and hyperplasia, autoantibody production (rheumatoid factor and anti–citrullinated protein antibody [ACPA]), cartilage and bone destruction, and systemic features, including cardiovascular, pulmonary, psychological, and skeletal disorders. (McInnes and Schett, 2011)

Pathogenesis of RA

The aetiopathogenesis of RA may be attributed to multiple causes, including environmental and genetic factors. Genetically, variations in human leukocyte antigen(HLA)-DRB1 alleles and the major histocomatibility complex (MHC) class II antigen driven HLA-DR4 alleles, have been accounted for susceptibility in RA, with greater prevalence in patients positive for Rhematoid Factors and anti-citrullinated protein antibody (ACPA). (Zampeli et al., 2015) The are numerous environmental conditions that act synergistic with the genetic background and promote the disease. Genetically predisposed individuals increase the risk of disease in combiation with smoking, exposure to chemichals (such as silica dust), lack of vitamin D or obesity. The condition of the microbiota has also been shown to affect long-term immune regulation. Especially,

the presence of bacteria such as Porphyromonas gingivalis or Aggregatibacter actinomycetemcomitans in the oral mucosa has been linked with the promotion of the disease by perturbating citrullination. (McInnes and Schett, 2017) The simultaneous existence of these factors leads to early immune perturbations in both the innate and adaptive compartments and subsequent chronic inflammation. In particular, studies of sera from predisease onset human cohorts reveal the presence of auto-antibodies against citrullinated self-proteins and rheumatoid factors that predate disease onset by up to 10 years. (Asquith et al., 2009)

Epidemiological and genetic analyses, as well as clinical observations, suggest that RA pathogenesis can be divided into three distinct stages: the first phase of an early, non-specific inflammatory and immune response, a second phase that constitutes the clinical onset characterized by an inflammatory attack to the joints and a third phase, most frequently targeted by treatments, that concerns a chronic joint inflammation and extensive tissue damage. ((Holmdahl et al., 2014)

At the state of *chronic inflammation*, an inflammatory response of the synovial membrane that includes hyperplasia, increased vascularity and antigen-driven CD4+ T-cells form the characteristics of RA onset. The CD4+ cells, produce different cytokines such as interferon-γ and TNFα through contact with other cells and thus it is no surprise that many modern treatment strategies target these overproduced cytokines. Monocytes, macrophages and synovial fibroblasts are then overactivated and produce increased proinflammatory cytokines IL-1, IL-6. Other soluble signals like IL-17, IL-18, IL-15 and angiogenic factors activate transcription factors leading to upregulation of genes responsible for inflammation and tissue destruction, such as the extracellular matrix metalloproteinases (MMPs). At a later stage, osteoclast production is induced by cathepsin K secretion and RANK signalling leading to the degradation of bone tissue. (Firestein and McInnes, 2017)

1.1.1 Treatment strategies. Anti-TNF treatments

No treatment can cure RA but modern therapeutic strategies are successful in minimizing disease activity to the extent that joint damage as well as other risks such as heart disease or stroke are prevented or controlled. However, it is vital that the disease is diagnosed at an early state in order to achieve preservation of joint function (Ma and Xu, 2013) and in this respect, strategies for early diagnosis through the discovery of novel biomarkers are of great importance.

The availability of treatments for rheumatoid arthritis has grown tremendously in the past 30 years. Currently, three main categories of therapeutic choices exist: non-steroidal anti-inflammatory drugs (NSAIDs), glucocorticoids (GC), and disease-modifying anti-rheumatic drugs (DMARDs). DMARDs may be a: conventional DMARDs of synthetic origin, such as methotrexate or targeted DMARDs, such as janus kinase [JAK]-inhibitors or b: biological

DMARDs, such as tumour necrosis factor [TNF]-inhibitors, interleukin-6-inhibitors, and B-cell targeting drugs. (Quan et al., 2008)

NSAIDS (Aspirin, Ibuprofen) and corticosteroids (Prednisone, Cortisone) can relief the fist painful symptoms of the disease caused by inflammation such as fever and stiffness, and are primarily used during the first weeks of the disease. However, it has not been shown that NSAIDs can slow down the progression of the disease. NSAIDs primarily target the cyclooxygenase (COX) gene that has a significant role in the metabolism of arachidonic acid derived from the cell membrane that forms proinflammatory prostangladines. The isoform of COX, COX-1 is found in most physiological homeostatic procedures of most tissues, such as the defence from hydrochloric acid in the gastric tube. COX-2, is found to be over-expressed in specific tissues such as brain and kidney tissues when there is inflammation or injury. (O'Dell, 2004) Treatments of this kind may be effective but can cause a variety of side effects when used long-term such as nausea or ulcers and increase the risk for thrombosis and heart attack and renal deficiency.

DMARDs include traditional synthetic therapies, such as methotrextate and cyclosporine, biological DMARDS and novel potential small molecules. Biological DMARDs are a big group of drugs which act through different targets and mechanisms. There are DMARDS that target: cytokines/TNFα (Infliximab, Etanceprt, Adalimumab), IL-6 (Tocilizumab), Il-1 (Anakira),Il 17A and 17F (Secukinumab),JAK pathway(Tofacitinib),Syk (Spleen-tyrosine Kinase) (Fostamatinib). (Zampeli et al., 2015) For this study we will focus on the actions of DMARDs that act as TNF inhibitors.

TNFα inhibitors: TNFα belongs to the cytokine family and acts through TNF receptors 1 and 2 (p55 and p75) provoking an inflammatory response after the activation of monocytes, macrophages and T-lymphocytes. As a result, key pathways of inflammation are activated such as NF-kB pathway, RANKL signaling, extracellular signal-regulated kinase (ERK) signaling pathway, tumor progression locus 2 (TPL2) pathway, and proapoptotic signaling. (Hayden and Ghosh, 2014) TNFα is found to be overexpressed in synovial fluid of RA patients.
The first DMARD developed was a TNFα blockade, *Infliximab* (IFX), a chimeric monoclonal antibody composed from a murine variable region and a human constant region with affinity to TNFα.
Etanercept, also a TNFα inhibitor, is a recombinant fusion protein constructed by two TNF receptor 2 extracellular domains and a Fc fragment of the human immunoglobulin (Ig) G1 class. Etanercept restrains soluble TNFα blocking its binding to the receptor, resulting to the neutralization of its activity. (Zampeli et al., 2015)
Adalimumab, is a fully human IgG1 monoclonal antibody that binds specifically to TNF thus preventing its binding to the receptors. (Furst et al., 2003)
Certolizumab pegol is a Fc-free, PEGylated, anti-TNFα monoclonal antibody. The parent

	Description	Mechanism
Etarncept (Enbrel)	Fully humanised fusion protein of Fc fragment of the human Ig and TNF p75 receptor	Decoy receptor that binds to TNF (soluble receptor)
Infliximab (Remicade)	Human and murine chimeric anti-TNFa antibody	Binds soluble/membrane TNF
Certolizumab pegol (Cimzia)	Human and murine PEGylated,anti-TNFa monoclonal antibody	Binds soluble/membrane TNF
Adalimumab (Humira)	IgG1 monoclonal antibody	Binds soluble/membrane TNF, prohibits cytokine production

antibody was selected from a screen of hybridomas for human TNFα binding. The complementary determining regions from the murine antibody were then inserted into a human Fab IgG framework, along with several other framework residues of the variable domain that were essential for maintenance of affinity. The Certolizumab Fab' was subsequently PEGylated via the site-specific attachment of a 40 kDa polyethylene glycol (PEG) moiety. In vitro, it has been shown that Certolizumab can inhibit signalling through both receptors by binding to soluble and transmembrane TNF. (Goel and Stephens, 2010)

1.1.2 Mouse models in the investigation of RA

The multi-factorial nature of RA becomes more complex when trying to elucidate the molecular mechanisms leading to the disease and the action of potential drug targets or propose new preventive and therapeutic strategies. For this reason, mouse models have been widely used to study such mechanisms in vivo. (Caplazi et al., 2015, Lindqvist et al., 2002)

Despite the great number of animal models for RA developed, none of them represents a universal manifestation of the disease because of its complex nature, as previously mentioned. However, there are models that accurately represent different subtypes or pathways of the disease and have provided insight for target discovery.(Kollias et al., 2011). Mouse models in RA may either originate from induced arthritis conditions (induced models) or genetically modified mouse strains (spontaneous models). There are numerous ways to induce arthritis including collagen and antigen induced arthritis. (Kannan et al., 2005) Collagen-induced arthritis (CIA) belongs to methods of active immunization with collagen II and is the most commonly used model of induced RA. CIA models develop an acute to subacute mono-phasic erosive polyarthritis with a number of human-like RA symptoms including the presence of rheumatoid factor or anticitrullinated peptide antibody. (Caplazi et al., 2015) A rapidly progressing form of CIA can be induced by using a cocktail of monoclonal antibodies targeting various collagen II arthritogenic epitopes. Other kinds of induced arthritis include, streptococcal cell wall arthritis in mice, immune-complex induced arthritis (Kannan et al., 2005), Staphylococcus aureus-induced

arthritis, pristane-induced arthritis and anti-GPI antibody transfer-induced arthritis (Lindqvist et al., 2002)

Genetically engineered mice serve a dual role in the investigation of RA according to (Kannan et al., 2005). Firstly, insertion or deletion of important genes such as receptor molecules or cytokines lead to conclusions regarding their role in disease mechanisms. Secondly, these interventions may lead to spontaneous inflammatory responses giving insight on immune and inflammation mechanisms. As of now, a number of transgenic mice that develop spontaneous arthritis has been developed. Most common are the transgenic mice with a human TNFα transgene modified in the 3' region, which over produce TNF, first introduced by the Tg-huTNF model. The importance of Tg-huTNF is outlined by the crucial role of TNF in RA and is a great tool in TNF-driven mechanisms investigation. Transgenic mice present a polyarthritis whose symptoms greatly mimic those of human RA. (Kollias et al., 2011)

Another example, are mice deficient in IL-1 receptor antagonist (IL-1ra–/–), which is the natural occurring inhibitor of IL-1 and competes for the receptor. Another example is the K/BxN mouse that expresses both the T cell receptor (TCR) transgene KRN and the MHC class II molecule A(g7). This model develops develop severe inflammatory arthritis, and serum from these mice induces a similar condition to a wide range of mouse strains and is important for investigation of the role of antibodies in the inflammatory response. (Monach et al.)

1.2 Gene signatures and patterns

Why is it important to identify them? What information can they provide?

When conducting a gene expression profiling experiment, the result is the simultaneous monitoring of the expression at a genome scale. The development of cost-effective and dependable techniques such as microarrays has given the opportunity to produce gene profiles for multiple conditions and samples.

Patterns of expression:A series of differential accumulations of a gene product in a subset of cells can be defined as a gene expression pattern for a given gene.(Tomancak et al., 2007)

A first step for the organization of expression data is to group together genes with similar expression patterns as they could be linked at a functional level, be co-regulated or even encode proteins that interact and participate in common processes. Co-regulation, is a common phenomenon in gene expression. Small sets of genes can be co-expressed or co-regulated in a number of conditions and have no link or be inactive in other conditions.

Given the great influx of information derived from gene profiling experiments, it is possible to study patterns of expression in order to gain insight on the dynamic behaviour of genes and

pathways linked with a biological process. Analysis of patterns can lead to the discovery of groups of genes that participate in similar biological processes or functions. This possibility can prove significant in drug design, as it may result in possible molecular targets that interact with the drugs. (Roy et al.) Most methods for pattern identification in gene expression data are based on clustering algorithms that intend to group together samples of different condition such as healthy and diseased or treated or origin such as samples form different organs or organisms based on the profile each sample that is formed from all the genes provided by the researcher.

Gene signatures: Groups of genes that exhibit specific and very characteristic patterns of expression hat may be seen as representative of a biological condition -pathogenic or natural, of the cell.

Responses in physiological cell processes or other stimuli include the activation of pathways or cascades of signal transduction that result in alteration of the gene expression levels which could be characterized as the gene signatures of this process or stimuli. The identification of these characteristic perturbations has been extensively studied in the past decade as it promises useful applications in clinical practice. Gene signatures are already used for the prognosis of individual patients,("*prognostic*"), or the prediction of response in a treatment, ("*predictive*" or "treatment-effect modifiers"). A signature is called predictive (of the treatment effect) if the relative treatment benefit varies according to signature values. The importance of identification of such gene sets is clear: it can give clinicians a way to choose a therapeutic intervention that is more appropriate to the profile of each individual and gives greater possibility of effectiveness and positive response. In the literature there are hundreds of candidate predictive biomarkers only some of them have passed the validation for clinical routine practice.

1.2.1 Gene expression patterns and signatures of RA and anti-TNF factors in literature

Currently, five disease modifying agents exist, approved as clinical medication infliximab, etarncept, adalimumab, certolizumab pegol and golimumab with the first three being most frequently used. Despite their success, anti-TNF agents are reported to present great variance in treated patient response with a 30-40% showing inadequate response. Their use is linked with varying side effects or risk of infection. However, data about their differences and the clinical response are still inconclusive. (van Baarsen et al., 2010) The underlying causes of these differences are not yet elucidated but because of their frequency of use, anti-TNF factors are profoundly studied in pharmacogenomics studies (van Baarsen et al., 2010), clinical response studies or signature/biomarker identification studies. (Szekanecz et al., 2013, Wijbrandts and Tak, 2017)

Expression patterns of multiple genes and genetic signatures have been studied in association with responses to the above agents using microarray platforms. Most studies have been conducted on infliximab and etarncept. Unfortunately, these studies yielded very inconclusive results. These differences could stem from various reason of biological and technical inconsistency; For example, differing signatures emerged from studies that examined samples of multiple sources (e.g. synovial tissue/cells and blood samples) or different microarray platforms.

Infliximab is the most well studied treatment in terms of signatures. Multiple analyses have been conducted with samples of different origins. From synovial tissue biopsies, Lindberg et al. (2006) compared the expression profiles of patients that were characterized as good responders to infliximab treatment and TNF-a positive participants. The analysis lead to 1,058 genes that were differentially expressed. Among them were the genes CXCL3 and CXCL1 that appeared for the first time in signatures. GO terms that emerged concerned immune system response, cell communication, signal transduction and chemotaxis. (Lindberg et al., 2006) However four years later, Lindberg et al. (2010) could not identify significant differences in expression profiles of synovial biopsies from responders and non-responders but showed differences in gene expression between patients with or without lymphocyte aggregates that were functionally linked with leukocyte differentiation,immune response, hemopoietic or lymphocyte organ development. (Lindberg et al., 2010) Another study analyzed peripheral blood samples before and after the start of infliximab. The results revealed an increase in type I IFN response gene activity only in a group of RA patients who showed a poor clinical outcome. Functions were related to immune – stimulatory terms: activation of myeloid dendritic cells, chemoines, chemokine receptors, co-stimulatory molecules, humoral responses and immune suppressive activities: Th2 cell skewing, antiproliferative and proapoptotic effects. (van Baarsen et al., 2010) Koczan et al. (2008), performed microarray analysis of peripheral blood before and after the first treatment with etarncept. An nteresting DE gene set included NFKBIA, CCL4, IL8, IL1B, TNFAIP3, PDE4B, PPP1R15A and ADM. Pathways and processes such as TNFα signalling via NFκB, NFκB-independent signalling via cAMP, and the regulation of cellular and oxidative stress response. (Koczan et al., 2008)

In a more "computational spirit", analysis of combined gene expression datasets of peripheral blood samples from patients treated with adalimumab,etarncept and infliximab separated responders from non-responders. These datasets originated from studies that attempted to find gene signatures of treatment response, but their results were highly inconsistent with only two genes appearing in more than one study. Using a meta-analysis approach only one of the two genes, GOS2 was identified as predictive of the response to anti-TNF factors. GOS2 is linked with cell proliferation, apoptosis, inflammation, metabolism, and carcinogenesis. However, the role of G0S2 in RA has not been yet studied except for one study that reported higher expression of G0S2 in both the bone marrow-derived mononuclear cells (BMMCs) and peripheral blood mononuclear cells (PBMCs) of RA patients (Kim et al., 2014).

1.3 Motivation & goal of thesis

In this work we used data from a gene expression profiling experiment of healthy mouse models, Tg-huTNF transgenic mice that develop polyarthritis and transgenic mice that were treated with one of the anti-TNF mentioned above which we will from now on refer to with their commercial names: Remicade (infliximab), Humira (adalimumab), Enbrel (etarnecept) and Cimzia (certolizumab pegol). A brief description of the experiment is found in Chapter 2 (Methods)

At the first part two biclustering algorithms were applied using two discrete pipelines described in Methods, in order to extract modules of co-regulated genes among the three types of samples. All available samples were used but the analysis was restricted to genes that were differentially expressed in at least one sample (among wild-type, transgenic or treated). Our goal, through this application was dual:

(i) Observe the similarities and dissimilarities of treatments and conditions. The construction of biclusters could give novel information about the mechanism of action of the drugs, reveal sets co-regulated genes that are "responsible" for side effects or for effectiveness of the treatment(s). The acquisition of unambiguous patterns could serve as a precursor for the mining of novel signatures in human data.

(ii)Comment on: (a) the use of the concept of biclustering in a theoretical framework and the interpretation of the results, (b) the ease of use of two popular biclustering algorithms both implemented in R packages. As mentioned later, biclustering has seen growth in the past decade through the development of novel algorithms, but it has seen little practical application by biologists.

In the second part, a simple implementation of a genetic algorithm was applied, for each subset of data containing samples of healthy, transgenic and one of the anti-TNF treatments in order to select the genes that lead to a better clustering of our samples. The rationale of this implementation was a simple idea: to *locate among all the DE genes of the treated, the Wt and Tg animals, the ones that maximize the similarity with the healthy and simultaneously maximize the dissimilarity with the diseased animal.*

The work presented herein does not comprise a direct effort to extract gene signatures ready to be clinically validated, as it does not include response data. Instead, it is more of an highly investigational effort to identify computationally "common ground" or "points of departure" between conditions, using algorithms and pipelines designed to identify patterns in datasets. The identification of modules with discrete patterns could provide useful information about the mechanism of action of the treatments and act as a indication of where novel signatures

or bimarkers could be "hidden". Biclustering, as opposed to traditional clustering, allows for the selection, simultaneously, of specific samples using only subsets of the given genes. This constitutes the basic advantage of the method as it enables to focus on particular subsets of both samples of features disregarding others that are either problematic or not informative. Consequently, the use of the genetic algorithm was intended to satisfy a double condition of maximal Wt-similarity/Tg-dissimilarity which could prove effective in the identification of gene signatures when asking the question: " *Which of the genes related with the disease are perturbed in terms of expression in a way that is responsible for the effectiveness of the treatment?*"

Chapter 2

Methods: Biclustering and GA for the identification of patterns and genetic signatures

2.1 Description of experiment

The experimental procedure included wild-type (healthy), huTNF-transgenic (diseased) and huTNF-transgenic mice that were treated with four different anti-TNF treatments: infliximab, adalimumab, etanercept and certolizumab pegol. Total RNA was isolated from aqueous extracts of whole ankle joints of the animals.

All treatments were performed at a therapeutic stage of intervention which corresponded to administration of the drug in transgenic mice of six weeks of age. All therapeutic interventions were carried out in 10 biological replicates. Wild-type and transgenic mice were analyzed in replicates of 10 and 13 respectively for a total sum of 63 profiles. All animals were sacrificed at the age of 11 weeks.

2.1.1 Data pre-processing and exploratory analysis

2.1.1.1 Data Normalization

All samples were hybridized on Affy Mouse Gene 2.0 standard array. Raw data normalization was carried out using the standard process. Log-transformed expression measurements were then converted to gene space by calculating mean probeset values referring to the same genes. This was done in order to minimize the complexity of alternative transcript abundance, which

we considered, at this level, to be minimal. Only values from genes that were measured in all samples were finally included in the dataset, which consisted of 18704 common measured genes in all 63 samples. Measurements were normalized across samples through quantile normalization.

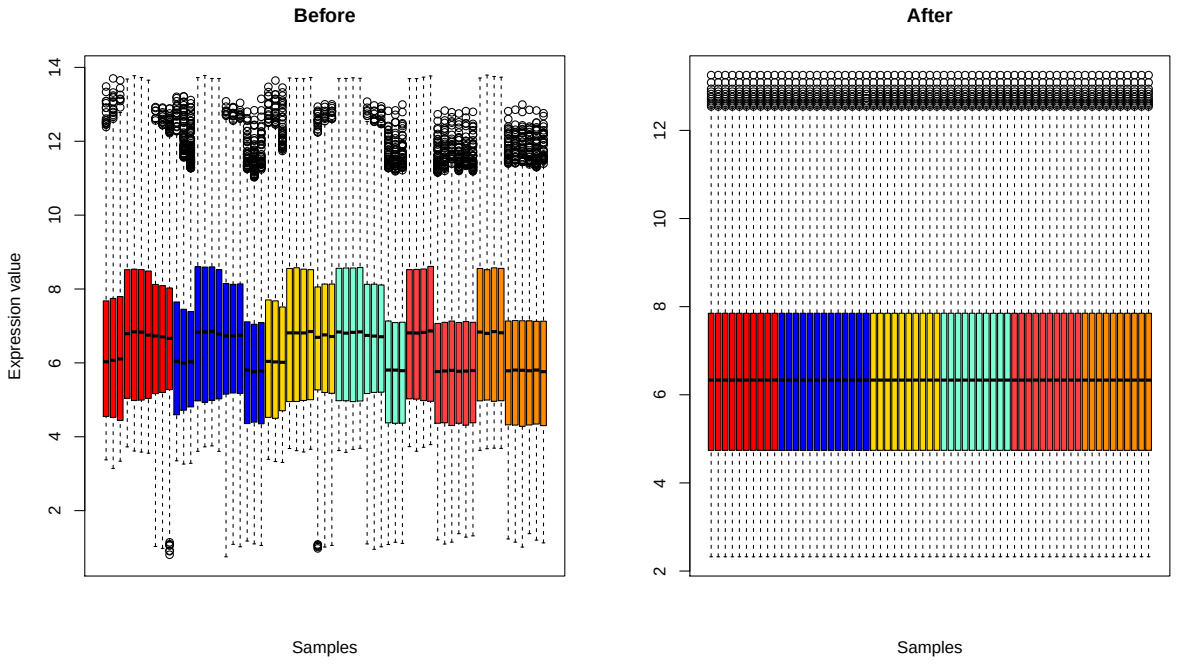


FIGURE 2.1: Samples' distributions before and after quantile normalization

2.1.1.2 Dunnet's analysis of Differentially expressed genes (DEGs)

Having in our disposition two conditions that could be considered as baseline controls, the healthy, wild-type samples and the untreated transgenic samples, we calculated the differential expression against both conditions. We applied a differential expression analysis using AVOVA to calculate log2Fold-Change (log2FC) values followed by Dunnett's test for multiple comparisons using either Wild-type (Wt) or Transgenic (Tg) samples as control conditions respectively standard thresholds of $| \, log2FC \, | \geq 1$ and an adjusted p-value ≤ 0.05 were applied for the definition of differentially expressed genes. Figure2.2 shows the number of DEGs for each comparison. The data-set used in further analyses included genes which were DE in at least one comparison. Its final size was 1178 genes x 63 samples.

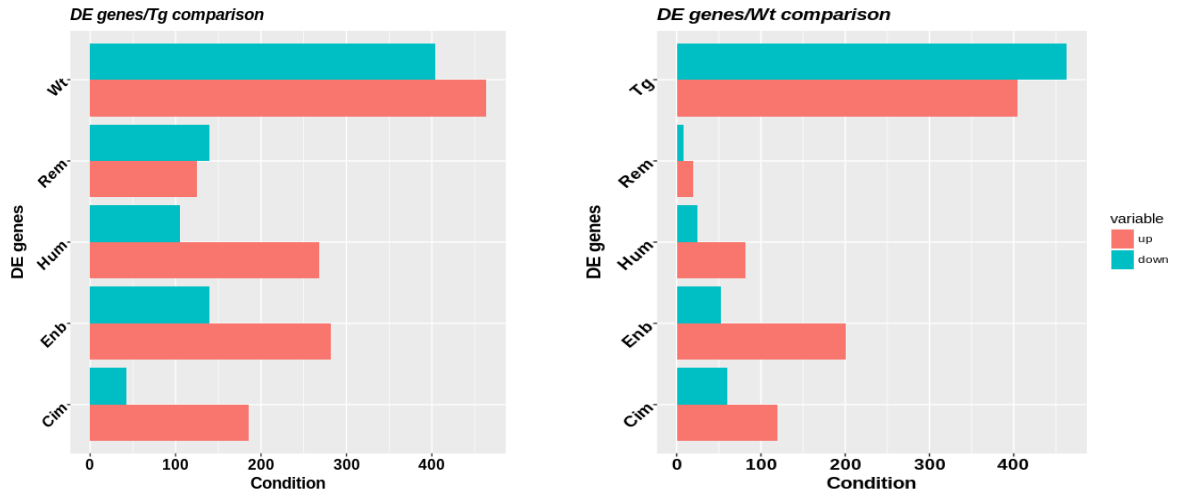


FIGURE 2.2: Barplots of DEGs from Dunnet's analysis

2.2 Biclustering Background

2.2.1 Biclustering definition

Given a two-dimensional matrix M of m rows and n columns, traditional clustering algorithms such as k-means or hierarchical clustering algorithms, would either cluster rows or columns. For example, in a gene expression matrix, clustering approaches would group together samples (or respectively genes) based on the assumption that all genes (or respectively conditions) of the group behave similarly.(Padilha and Campello, 2017) Clustering has been widely used for purposes of gene expression data analysis such as, tissue classification and functional annotation. However, this approach also entails some drawbacks: first of all, each gene or experimental condition must be assigned in exactly one cluster i.e. clusters cannot overlap and data cannot be left out. This fact presents a limitation concerning gene expression data as a gene may participate in more than one function therefore resulting in a specific regulation pattern in one context and a different pattern in another, or may not provide important information about the current conditions. (Eren et al., 2013)

A possible solution to these limitations was given with the development of methods that simultaneously cluster rows and columns in order to identify sub-matrices of the initial matrix where the features present a coordinated behavior across a group of samples. This approach, alternatively known as two-way clustering, results to the creation of certain numbers of overlapping sub-matrices called bi-clusters and hence the process is called bi-clustering. (Padilha and Campello, 2017)

The first algorithm able to cluster simultaneously rows and columns was first introduced in 1972. (Hartigan, 1972) However, it is only in the past two decades that biclustering approaches

have been widely investigated. During this period, several biclustering algorithms have been developed mostly for gene expression data, each with a different mathematical background and computational approach. In the next section, the basic algorithms are briefly described as well as the type and structure of the biclusters that can be identified.

2.2.2 Bicluster strictures and types of algorithms

Types of biclusters

Biclusters are identified by the type of values the contain (Fig.2.1) :

- *constant*: All features present the same value over all samples
- *coherent rows or columns*: The bicluster contains features (or samples) that have the exact same constant expression values under a given subset of conditions (or features respectively), but their expression levels differ among features (samples).
- *coherent*: The features present a coherence in their values. Each row of the b/c is a multiple (multiplicative coherent) or sum (additive coherent) of a specific row added/multiplied with a constant.
- coherent evolution: A qualitative form of coherence.

Different algorithms can identify biclusters with different structures. :
1. Single bicluster (Figure 2.1 (a)),
2. Exclusive rows and columns group of biclusters (Figure 2.1 (c)),
3. Exhaustive non-overlapping group of biclusters with checkerboard structure (Figure 2.1 (f)),
4. Exclusive rows group of biclusters (Figure 2.1 (d)),
5. Exclusive columns group of biclusters (Figure 2.1 (g)),
6. Overlapping group of biclusters with hierarchical structure (Figure 2.1 (e)),
7. Or, arbitrarily positioned overlapping group of biclusters (Figure 2.1 (b)).

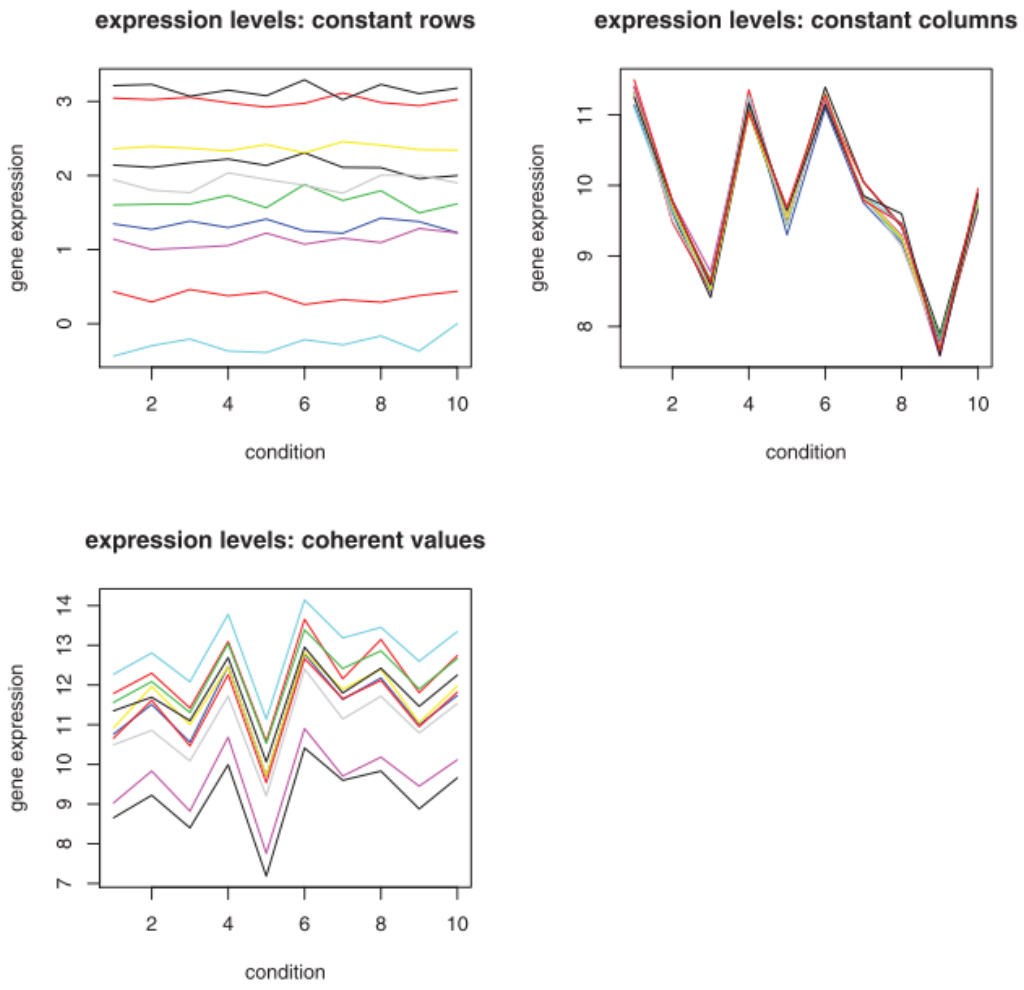


FIGURE 2.3: Examples of three types of biclusters.(Kasim et al., 2016)

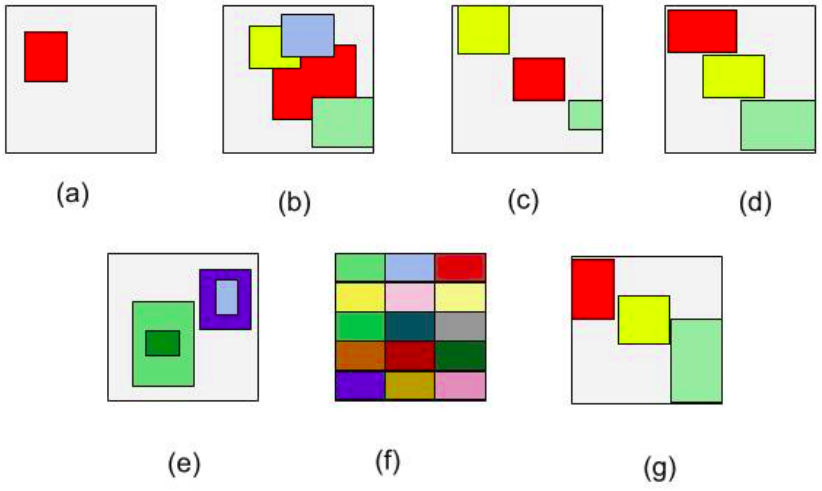


FIGURE 2.4: Structure of biclusters

Biclustering algorithms are divided according to the type of heuristic that they are based on [5]: greedy, divide-and-conquer, exhaustive enumeration or distribution parameter identification.

(Madeira and Oliveira, 2004)

Greedy algorithms

Greedy algorithms make the optimal choice at each step as they attempt to find the overall optimal way to solve the entire problem. A greedy biclustering heuristic was introduced by Cheng and Church with the *Cheng-Church algorithm* (CC). The CC algorithm is based on the addition/removal of rows and columns of the data-set in order to minimize the mean squared residue measure (Padilha and Campello, 2017). Examples of this approach include the *Iterative Signature Algorithm (ISA)*, which is initiated with random sets of genes and conditions, all genes and conditions are scored according to the conditions and genes of the sets respectively in order to evaluate their coherence and they are kept or discarded according to user-defined thresholds. The entire procedure is repeated until the set of genes and the set of samples converge. (Bergmann et al., 2003). *Qualitative Biclustering (QUBIC)*, converts the input data into a discrete form and builds a graph where each node corresponds to a gene, each edge has a weight equal to the number of experimental conditions for which two genes have the same nonzero integer values, and searches for biclusters corresponding to heavy sub-graphs (Li et al., 2009)

Distribution parameter identification algorithms

Distribution parameter identification methods assume that the data structure follows a statistical model and attempt to iteratively identify the parameters by minimizing a certain statistic. *Plaid* is an approach that defines the expression levels as a sum of layers, constructed as biclusters and tries to fit each layer to a model while iteratively minimizing the binary least squares to update the cluster membership parameters (Lazzeroni and Owen, 2002) (Turner et al., 2005). *Spectral* uses singular value decomposition to simultaneously cluster genes and experimental conditions in order to find checkerboard patterns in the data matrix. Only biclusters with variance lower than a given threshold are returned. (Kluger et al., 2003) *Factor Analysis for Bicluster Acquisition* (FABIA), assumes a multiplicative model and uses a factor analysis approach together with an expectation maximization algorithm to fit it to the data.(Hochreiter et al., 2010)

Exhaustive enumeration algorithms These algorithms attempt to construct all possible row and column combinations in order to identify the best possible sub-matrices and reduce running time by limiting the size of the biclusters. *Statistical-Algorithmic Method for Bicluster Analysis* (SAMBA), converts the input data set into a bipartite-graph, where one set of nodes corresponds to genes and the other is related to the experimental conditions, and finds complete bipartite sub-graphs composed of gene nodes with bounded degree. (Tanay et al., 2002)

Divide-and-conquer algorithms The *Binary Inclusion-Maximal Biclustering Algorithm* (Bimax) also discretizes the input data matrix which is recursively divided into a checker board

format. The goal of the Bimax method is to find maximal inclusion biclusters. This means a Bimax bicluster spans a submatrix of 1's which cannot be part of a larger submatrix of 1's.(Prelić et al., 2006)

2.2.3 Applications in biomedical data analysis

In the past two decades biological sciences have experienced a massive increase in the production of data derived from high-throughput measurement technologies. At first, with microarray data it was possible to measure simultaneously tens of thousands of genes in various conditions and obtain datasets rich in information at very low cost. Later the "revolution" of next-generation sequencing technologies gave the opportunity of higher accuracy and bigger resolution in the analysis of biological sequences providing a much flexible way to study a variety of conditions and organisms. The result was a great influx of large volume and high complexity data originating from different sources.

Biclustering appeared in 1972 as mentioned before, but was applied on microarray data for the first time by Cheng and Church in 2000. Ever since, there has been a development of a great number and different mathematical implementations of biclustering but very little application on real studies. The reasons for this imbalance of novel algorithmic works versus practical applications are important but will be further discussed later on.
However, one can intuitively understand that the advantage of biclustering as a method that can group simultaneously rows and columns and identify sub-matrices that present a pattern, could make the work of biologists easier in various issues. (Xie et al., 2018), have collected all cases that biclustering was applied in the past years and categorized them in 5 topics:

- *Functional annotation of unannotated genes*
  Annotated genes can be linked with functions or mechanisms if their interacting genes are known. In these cases biologists take advantage of the "guilt-by-association" concept in order to link genes with their physiological purpose. For this assessment, one needs known functional annotations (i.e. from databases like KEGG or GO) and measures of interactions between the target gene and its possible interactors that can be provided by databases such as RegulonDB

- *Analysis of modules*
  Biclustering can contribute to the identification of modules/groups of physically or functionally linked molecules that work together for a purpose. It has been applied to identify different types of modules, which could be groups of interacting molecules such as miRNA-mRNA modules (Bryan et al., 2014), functionally related genes/proteins or any

other manually defined clusters depending on the target modules, different inputs and strategies are needed. For example, (i) scRNA-seq gene expression data were used to identify molecularly distinct subtypes of cells that contribute different brain functions (Zeisel et al., 2015); time series expression data are often used to identify temporal transcriptional modules that consist of activated genes at consecutive time points (Madeira et al., 2010)

- *Network elucidation*
  Biclustering has been used as a module based method of identifying groups of genes/samples with similar expression profiles in order to infer networks. Tanay et al, used an approach like the aforementioned to identify modules and then constructed networks were nodes corresponded to modules and edges to intersections between their genes leading to a clustering of small modules into a bigger transcriptional network with hierarchical structure. (Tanay et al., 2002)

- *Classification of subjects/patients into disease subtypes*
  Identification of novel biomarkers in groups of stratified patients can be the result of an an accurate identification of disease sub-types. Clinical features, molecular characteristics or gene expression data can provide useful information for this purpose, as the disease could reasonably emerge from different pathways that are activated in different patient subpopulations. De novo identification of biclusters could group together genes with similar expression patterns only in one or in a group of sub-types, hence cluster patient groups that could be further evaluated by linking with clinical characteristics.

- *Identification of gene biomarkers, gene signatures and patterns of expression*
  Basically, given a gene expression matrix, biclustering can identify subgourps of genes that are being co-regulated among a number of conditions. These could range from disease sub-types to treated or untreated patients. If the genes that participate in the biclusters are differentially expressed between the conditions, they could stand as potential gene biomarkers or signatures. Because of the heterogeneity of tumor cells, many cases of biclustering for biomarkers have been applied in the search and analysis of cancer subtypes.

- *Identification of gene-drug associations*
  Biclustering has also been used as a computational alternative for drug repositioning. Through this method it is possible to reveal co-expression patterns in drug-perturbed responses using genome-scale drug-treated gene expression data. Drug induced modules can subsequently be functionally analyzed to reveal their biological relevance.

As it was described in Chapter 1, subsection 1.3 Motivation & goal of thesis, the current work aims to mine information that is most probably categorized as an effort for pattern and signature

identification or gene-associations as our gene expression data contain not only drug-treated samples but also healthy and diseased. Thus they provide an opportunity for more thorough investigation of the associations between the conditions.

2.3 Biclustering algorithms

2.3.1 Description of the algorithms

2.3.1.1 PLAID

The plaid model is an additive biclustering method (Lazzeroni and Owen, 2002, Turner et al., 2005). The characteristic that differentiates PLAID algorithm from other biclustering methods, is that Plaid considers a normal expression level for each gene and then tries to identify biclusters of genes that have similarly unusual expression levels in the bicluster samples. (Turner et al., 2005)

The plaid model consists of a series of additive layers intended to capture the underlying structure of a gene expression matrix. The model includes a background layer containing all the genes and samples, to account for global effects in the data. Any subsequent layers represent additional effects corresponding to biclusters that exhibit a strong pattern not explained by the general model. The background layer is firstly fitted for all samples and genes, then bicluster-specific layers are added, one at a time, until a pre-specified number is reached or no more significant layers can be found, as determined by a permutation test.

In the plaid model the expression level, Y_{ij}, i =1...n; j =1...p; of the ith gene in the jth sample is modelled by:

$$Y_{ij} = \Theta_{ij0} + \sum_{k=1}^{K} \Theta_{ijk}\rho_{ik}\kappa_{jk} + \varepsilon_{ij}$$

$$= (\mu_0 + \alpha_{i0} + \beta_{j0}) + \sum_{k=1}^{K} (\mu_k + \alpha_{ik} + \beta_{jk})\rho_{ik}\kappa_{jk} + \varepsilon_{ij},$$

where k is a layer index starting at zero for the background layer running to K, the number of biclusters; Θ_{ijk} is the sum of mean, gene and sample effects in layer k, which corresponds to the identification of coherent biclusters (see 2.2.1.1 and Fig.) ρ_{ik} is a binary cluster membership parameter defined for k $\geq$ 1 and equal to one if the i_{th} gene is in the k_{th} bicluster, zero otherwise; κ_{jk} similarly indicates cluster membership for the j_{th} sample, and ε_{ij} is a Gaussian error with mean zero and variance σ^2. Variants of the model may be obtained by simplifying Θ_{ijk}, but the form presented here is most suitable for analyzing microarray data. Θ_{ij0} is the background,

overall effect which is not necessarily constant and can be equal to $\mu_0 + \alpha_{i0} + \beta_{j0}$. (Turner et al., 2005)

The Plaid algorithm uses a minimization of binary least-squares approach in order to estimate the cluster membership parameters. For a given number of biclusters K the residual sum of squares is given by

$$Q = \frac{1}{2}\Sigma_{i=1}^{N}\Sigma_{j=1}^{M}\left(Y_{ij} - \theta_{ij0} - \Sigma_{k=1}^{K}\theta_{ijk}\rho_{ik}\kappa_{jk}\right)^2$$

Let us assume that l -1 biclusters are fitted (including the background bicluster). Given the estimates for the model parameters and the membership parameters, the residuals are given by

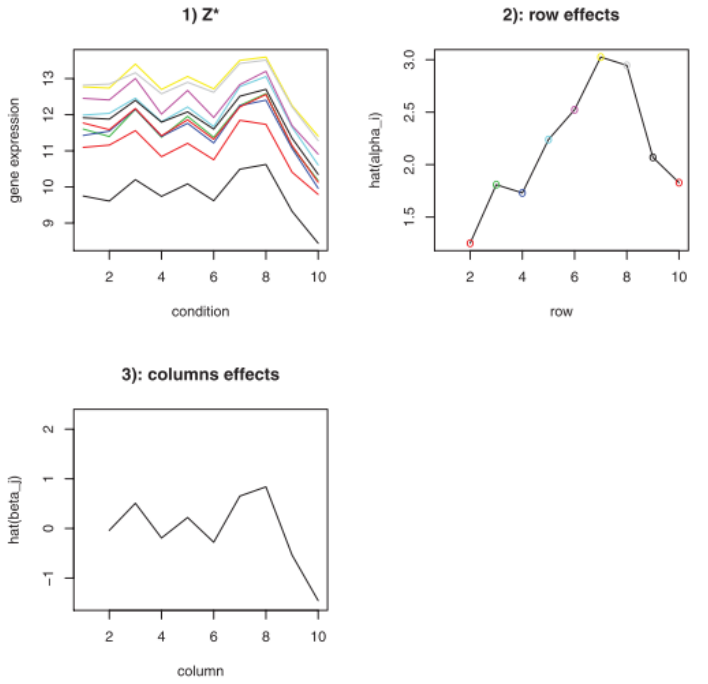
$$\hat{Z}_{ij} = Y_{ij} - \hat{\theta}_{ij0} - \Sigma_{k=1}^{\ell-2}\hat{\theta}_{ijk}\hat{\rho}_{ik}\hat{\kappa}_{jk},$$

The current residual matrix *Z is the input data matrix for the search of the lth bicluster. To simplify the notion we omit the bicluster's index. Thus, we assume Z^{ij} :

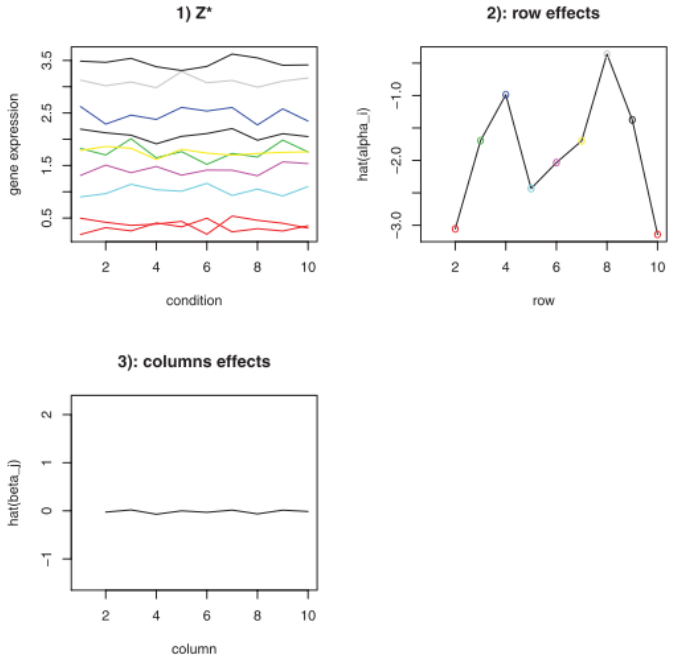
$$\hat{Z}_{ij} = (\mu + \alpha_i + \beta_j)\rho_i\kappa_j + \varepsilon_{ij}.$$

The aim of the algorithm is to estimate the bicluster effect μ, α_i and β_j and the membership vectors ρ_i and κ_j. This can be done by minimizing the residual sum of squares for the lth bicluster given. The estimation of the unknown parameters in (6.9) is done in an iterative procedure in which one set of parameters is the estimated condition of the other set.

Let $\hat{\rho}_i$ and $\hat{\kappa}_j$ be the current values of the membership parameters and let Z^* be a submatrix of Z containing the rows and columns for which the membership parameters are equal to 1. In this case Q is the residual sum of squares of a two-way ANOVA model with one observation per cell defined on the entries of Z^*. The parameter estimates for the bicluster effects, $\hat{\mu}$, $\hat{\alpha}_i$ and $\hat{\beta}_j$ are the usual maximum likelihood estimators for the overall, row and column effects in Z^* (and they are identical to the minimising values of Q). elaborate on the row and column effects obtained by ANOVA, giving examples of the latter in the cases of coherent and constant biclusters. (Kasim et al., 2016)



(a)



(b)

FIGURE 2.5: Row and column effect for (a) coherent additive bicluster and (b) for constant bicluster. (Kasim et al., 2016)

2.3.1.2 ISA

Iterative signature approach is implemented in R packages "isa2" and "eisa". (Csárdi et al., 2010) For the current work "isa2" was used. ISA is an iterative algorithm. It takes as input a matrix E_{mxn} and is initialized by a random seed vector r_0 of binary values and length m. Every "1" in the vector means the presence of corresponding gene in the current gene set. Then the transposed of E,Eis multiplied by r_0 and the result is compared to a threshold. The threshold application is an important step of ISA, and its absence would lead the algorithm to a simple SVD application. The user can set the thresholds to high or low values and thus make the analysis softer or more strict. Currently, the threshold is given from the mean and standard deviation of the binary vector and the elements kept are the ones that their difference from the mean corresponds to a given number of standard deviations. The user can decide to keep genes that are higher or lower than the mean or both. The thresholded vector c_0 is the (sample) signature of r_0. Then the (gene) signature of c_0 is calculated, E is multiplied by c_0 and then thresholded to get r_1. This iteration is performed until converges, i.e.r_{i-1} and r_i are close, and c_{i-1} and c_i are also close. The convergence criteria the closeness of r_{i-1} and r_i, is by default defined by high Pearson correlation. (Bergmann et al., 2003)

ISA FUNCTIONS: (Csárdi et al., 2010)

ISAIterate(): Executes ISA algorithm for a given number of times with many combinations of row and columns thresholds and multiple random initializations.

ISAunique(): In ISA there it is highly possible that two seeds will finally result to the same vector thus returning the same module more than once. This functions compares the modules and keeps the ones with low Pearson correlation.

ISArobust(): Runs ISA on a scrambled input matrix with same thresholds and drops every module found from the real matrix that has lower scores than the highest score acquired from scrambled matrix's modules.

2.3.2 Approach and choice of parameters

2.3.2.1 PLAID: Ensemble methods and hierarchical tree construction

The Plaid algorithm as described by (Turner et al., 2005) is implemented in R package "biclust". (Kaiser and Leisch, 2008) However, considering that the Plaid model requires initialization of binary memberships, it is expected that, depending on the seed number used for initialization, when multiple runs are considered, different biclusters are discovered. Also, the results can vary according to the choice of column and row thresholds that are set by the user. For this reason,

an ensemble method was in order to discover as many robust biclusters as possible. (Kaiser and Leisch, 2008, Kasim et al., 2016)

Ensemble method

Ensemble methods use multiple runs of one or more algorithms on slightly modified conditions, either parameter modifications, initialization seed (stochastic starting points), or data sampling. The result of multiple runs must be combined using an aggregation method (e.g., weighting scheme) in order to retrieve a useful result. (Kasim et al., 2016) In this work, the ensemble method was used from the "biclust" package along with a process for further processing biclusters in order to reduce high overlaps and obtain aggregated structures. The procedure is described in Fig.2.4

- Pre-process data: Described in 2.1
- Ensemble method: The ensemble method implemented in "biclust" was executed for Plaid algorithm. The user can set column and row thresholds, number of shuffles, number of repetitions and many other parameters.
- Compute Jaccard index: In this step, biclusters should be compared to each other. The Jaccard index for all pairs of biclusters obtained from ensemble method is computed by the "similarity" module from "superbiclust" R package (Kamiakova et al.,2015). The Jaccard index is given by
$$Ja = \frac{|A \cap B|}{|A \cup B|}$$
- Construct hierarchical tree of distances (1- Jaccard index) for biclusters with module "HCLtree" from "superbiclust". "HCLtree" implements a complete linkage method based on the notion that biclusters within a given group should all have a similarity larger than the user-specified threshold,
- Cut tree constructed in the previous step at an arbitrary set similarity.
- Merge grouped biclusters (we used only the groups that contained more than one member as they contained b/c that were found more than once throughout the total runs). Grouped biclusters are highly overlapping but they do not all contain exactly the same samples/genes.
- Filtering of the resulting biclusters to keep the ones of bigger size. Arbitrary thresholds of 5-6 samples, 20-30 genes.
- Further analysis can include functional enrichment,visualization, validation etc/

  The intersection approach was the most appropriate since it returned biclusters of manageable size, it is considered highly strict as it leads to small,core biclusters. On the other

hand, the union returned biclusters of very high size and with high percentages of overlaps between them. (Kasim et al., 2016) propose various was to score biclusters according to their similarity or correlation and aggregate them.

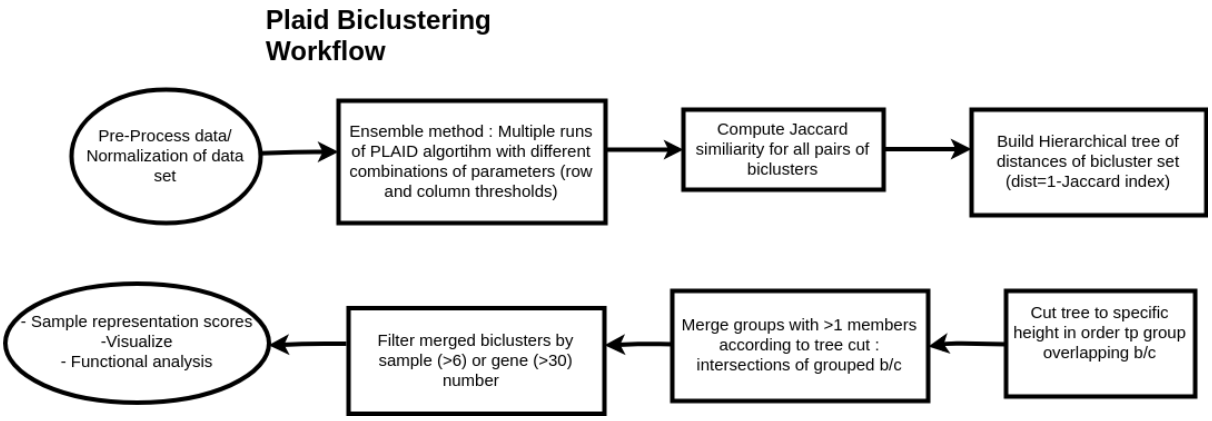


FIGURE 2.6: Step-by-step acquisition of biclusters with plaid model and further analysis to obtain final set of b/cs

2.3.2.2 Application of ISA with 'isa2'

In the current work we applied ISA in the dataset of 1178 genes and 63 samples. We used the quantile normalized across samples matrix as in PLAID application. We applied the algorithm with isa.iterate function using starter seeds: 5,10,15,20,25,35,45,50,60,100,125,250 for row seeds and 2,5,6,8,9,10 for column seeds as is is proposed that smaller starting seeds yield better results. The thresholds used ranged 0.5 to 3 with a step of 0.2 for both column and row thresholds. The algorithm was executed 20.0000 times for each threshold combination. The modules were merged, filtered for uniqueness with a Pearson correlation limit equal to 0.8 and checked for robustness with isa.filter.robust() and then filtered for size as in PLAID results.

2.4 Genetic algorithms for selection of optimal gene set

2.4.1 Background

In 1975 John Holland first proposed a heuristic solution-search or optimization algorithm, based on the Darwinian principle of evolution through selection, called Genetic Algorithms (GAs). GAs implement a process that is similar to an abstract evolutionary process in order to locate a solution to a given problem. The algorithm operated on a population of artificial "chromosomes". (Holland, 1975)

In GAs, the term *chromosome* typically refers to a candidate solution to a problem, often encoded as a bit string. The "genes" are either single bits or short blocks of adjacent bits that encode a particular element of the candidate solution. An allele in a bit string is either 0 or 1; for larger alphabets more alleles are possible at each locus. (Mitchell, 1998)

Each chromosome represents a solution and each solution gets evaluated according to its fitness which is a number that measures how good is this particular solution for the given problem. The algorithm performs a fitness-based selection, recombination and mutation of parent chromosomes in order to finally produce an offspring generation. In the recombination step two parent chromosomes are selected and a crossover takes place in order to produce child chromosomes that will participate as members of the offspring population. As this process is iterated, a sequence of successive generations evolves and the average fitness of the chromosomes tends to increase until some stopping criterion is reached. In this way, a GA "evolves" a best solution to a given problem. (McCall, 2005)

2.4.2 Implementation and choice of parameters

For a given problem a simple version of a genetic algorithm follows the steps below: (Mitchell, 1998)

1. Initialization with a randomly generated population of N chromosomes/candidate solutions with n bits. Set number of k offsprings, l size of elitism, r number of random chromosomes that participate in the population

2. For each chromosome calculate its fitness using the given fitness function

3. Repeat the following steps until k offsprings have been created:

    a. Choose two parents from current population with probability of selection defined by their fitness. Put them back in the population

    b. With probability p (crossover probability, the probability that two parents will cross at a single point) cross the pair of chromosomes at a randomly chosen point. If no crossover takes place, form two offspring that are exact copies of their respective parents.

    c. Perform mutation at each offspring with probability p_m (probability of mutation) and set the resulting chromosomes as member of the new generation. In the case of binary vectors, mutation means that a bit of 1 is replaced with 0 (and vice versa)

4. If elitism is applied and its size is l, keep the l best chromosomes from current population according to their fitness.

5. Produce r random chromosomes

6. Merge k, l, r chromosomes into a new population of N chromosomes and return to Step 2

Each iteration of this process is called a *generation* and is repeated for a number of runs. The execution of a GA can end when a certain number of generations is reached or when convergence is reached resulting to a best-fitness solution. The entire set of generations is called a *run*

The aforementioned characteristics comprise the basis for a simple implementation of a GA but there are many details to be defined later. These are: the type of selection, the type of crossover, the type of mutation and parameters such as the mutation probability, size of elitism, the probability of crossover. The functionality of the algorithm is highly depending on their choice. There are also more complicated versions of GAs (e.g., GAs that work on representations other than strings or GAs that have different types of crossover and mutation operators). (Mitchell, 1998)

Types of selection

One main issue in the implementation of GAs is how to perform selection i.e choose individuals from the population that will create offsprings for the next generation. Selection is intended to give an advantage to best fitted individuals with producing even fitter individuals through their recombination. Chromosomes with higher fitness should have a greater chance of selection than those with lower fitness, thus creating a selective pressure towards more highly fit solutions. (Holland, 1975) The selection process must be balanced in order to prevent too-strong selection (sub-optimal highly fit individuals will take over the population, reducing the diversity needed for further change and progress) or too-weak selection that will result in too-slow evolution.(Mitchell, 1998) The most popular selection types,also used in the current work, are described below:

Fitness proportionate selection or Roulette Wheel (RW) selection

The traditional selection method used is Roulette Wheel (or fitness proportional) selection. Each chromosome is assigned a probability of being selected proportional to its relative fitness. The relative fitness is defined as a proportion of the fitness and the sum of all fitness values. (**?**) It is clear that in this type of selection fitter individuals have greater possibility of selection or great piece of the pie as shown in Fig. 2.5.

Tournament selection

Tournament Selection first selects a group on n individuals from the population and the one with the best fitness is chosen. N is called tournament size and can be used to vary the selection

pressure because as it gets bigger individuals with higher fitness values might get chosen. The process is repeated until all offsprings are produced. The procedure is outlined in Fig. 2.5.

Linear Rank selection For rank selection the individuals are sorted according their fitness values, a rank N is assigned to the best individual and a rank 1 to the worst individual. The selection probability is linearly assigned to the individuals according to their rank instead of their fitness:

$$\text{ExpVal}(i,t) = Min + (Max - Min)\frac{\text{rank}(i,t) - 1}{N - 1},$$

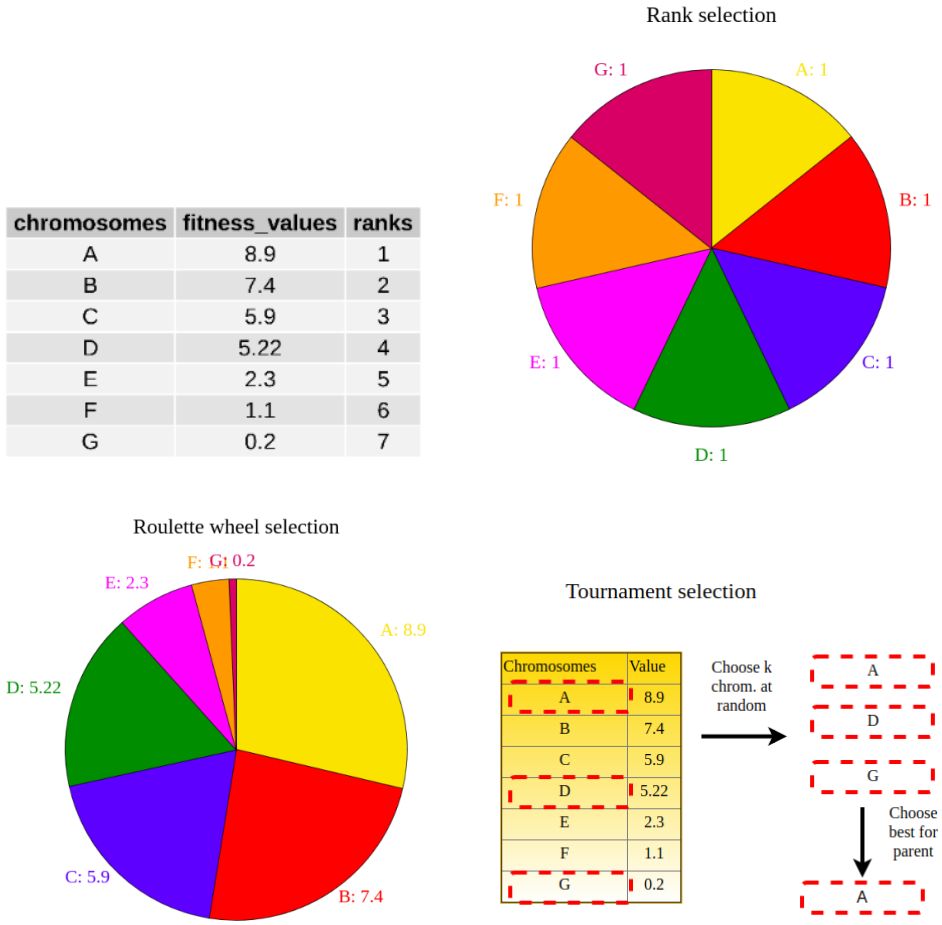
Rank selection has a possible disadvantage: slowing down selection pressure means that the GA will in some cases be slower in finding highly fit individuals. However, in many cases the increased preservation of diversity that results from ranking leads to more successful search than the quick convergence that can result from fitness proportionate selection. (Mitchell, 1998) Rank selection is mostly used when the individuals in the population have very close fitness values as it leads to each individual having an almost equal share of the pie (like in case of fitness proportionate selection). Hence each individual no matter how fit relative to each other has an approximately same probability of getting selected as a parent. This in turn leads to a loss in the selection pressure towards fitter individuals, making the GA to make poor parent selections in such situations

2.4.3 Application on current data set

2.4.3.1 Goal of GA: Choice of fitness function

In every GA implementation the most significant part in the achievement of the optimization is the design and implementation of the fitness function. In our case the choice of fitness function was defined by the goal of our approach.

We applied a version of GA in order to locate the optimal set of genes that maximizes the differences between a specific treatment and the diseased samples and minimizes the difference between the same treatment and the healthy mouse. We hypothesized that the GA would choose random sets of genes some of them would serve better this double conditions and in each iteration through the processes of selection and recombination, the gene set would get better. In order to evaluate every subset of the population according this double condition we designed the implementation as a *clustering optimization problem with fixed groups: the Wt samples together with the anti-TNF samples would form the first group and the Tg samples alone would form the second group.* As in each iteration our goal was to evaluate the gene set, this could be done using any metric for clustering evaluation such as silhouette score. We used a modified version



chromosomes	fitness_values	ranks
A	8.9	1
B	7.4	2
C	5.9	3
D	5.22	4
E	2.3	5
F	1.1	6
G	0.2	7

FIGURE 2.7: Three types of selection for hypothetical chromosomes A, B, C, D, E, F, G with fitness values shown in table (up-left). Pie charts show the partitions of possibilities of selection among the chromosomes according to actual fitness values (in RW) and ranks (in rank).

of Dunn index implemented in "fpc" R package defined as the minimum average dissimilarity between two clusters as a proportion of the maximum average within cluster dissimilarity

2.4.3.2 Design of the GA

There are several evolutionary schemes that can be used, depending on the extent to which chromosomes from the source population are allowed to pass unchanged into the successor population. These range from complete replacement, where all members of the successor population are generated through selection and recombination to steady state, where the successor population is created by generating one new chromosome at each generation and using it to replace a less-fit member of the source population. The choice of evolutionary scheme is an important aspect of GA design and will depend on the nature of the solution space being searched. A widely used scheme is replacement-with-elitism. This is almost complete replacement except

that the best one or two individuals from the source population are preserved in the successor population. This scheme prevents solutions of the highest relative fitness from being lost from the next generation through the non-deterministic selection process.

Our approach was formed by 3 practical goals:

(a)Choosing parameters: Observe the course of fitness values across generations in three selection types with existence or absence of elitism.

(b) Choose a set of parameters that seems to optimize and execute multiple to times to acquire a set of genes that appear in most runs

(c) Use this set of genes to classify samples and compare with models made with all DE genes in each case.

We decided to tune only the hyper-parameters of selection and elitism and kept the parameters of mutation rate and crossover rate stable at 0.1 and 0.8 respectively.

Chapter 3

Results

3.1 PLAID and ISA biclustering results

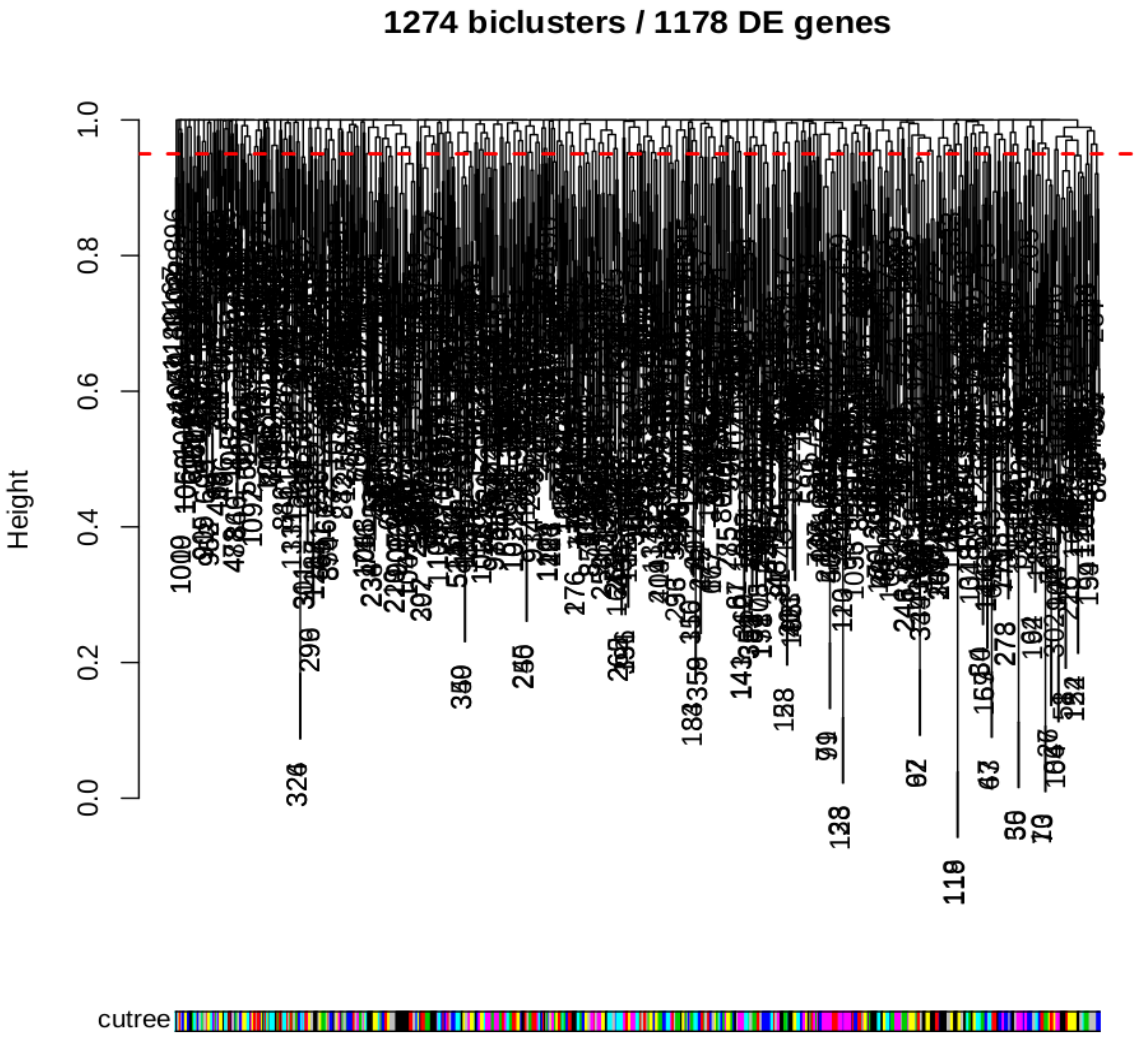
After executing PLAID and ISA as described in Chapter 2 we obtained the results shown in Table 1.

	PLAID	ISA
RUNS	20	2000
DATASET	1.178 genes x 63 samples	1.178 genes x 63 samples
NORMALIZATION	Quantile normalized	Z-score & across samples and genes (separate tables)
INITIAL # OF B/Cs	1274/261 (after merging)	17 (analysis performed internally)
ANALYSIS	Hierarchical tree of 1- Jaccard → cut tree in h=0.95 → intersection of grouped biclusters	Isa.unique(): Pearson correlation to discard similar, isa.robust(): permutation tests
SIZE FILTERS	genes <= 30, samples <= 6	genes <20, samples <5
FINAL # OF B/Cs	13	12
MEAN & SD: SAMPLES	10.6 & 2.7	12.5 & 6.1
MEAN & SD: GENES	51.3 & 18.5	85.04 & 89.9

Table 3.1: Table 1

3.1.1 Hierarchical tree of distances for PLAID biclusters

As mentioned earlier, the analysis of PLAID biclusters leads to the construction of a hierarchical tree of distances using as dissimilarity measure 1-Jaccard index for each pair of biclusters.



3.1.2 Overlap analysis and and sample representation

We performed an overlap analysis by comparing the rows (genes) and columns (samples) of each pair of biclusters. The results are shown in Fig.3.1. We observe that after filtering for similar biclusters there are modest overlaps in genes, which is acceptable because it is one of the main advantages of biclustering. Generally, both analyses are considered strict: PLAID biclustering because of the intersection of grouped biclusters which leads to many of them being filtered because of small size and ISA because of strict thresholds for convergence and similarity. As of that, it was expected that the acquired biclusters would be numbered, small-sized and with

minimum overlaps. Overlaps in samples are greater which is interesting as it seems that there are biclusters that share the same samples but different genes participate in them. This could be an indication of different functions bringing together the current samples through different kinds of patterns.

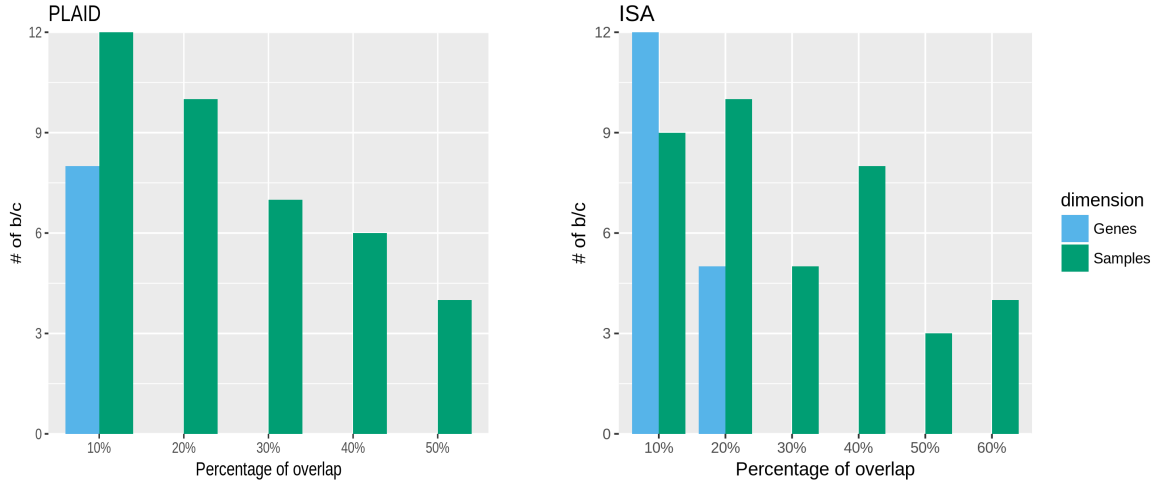


FIGURE 3.1: Overlaps for genes and samples in final biclustering results of both algorithms

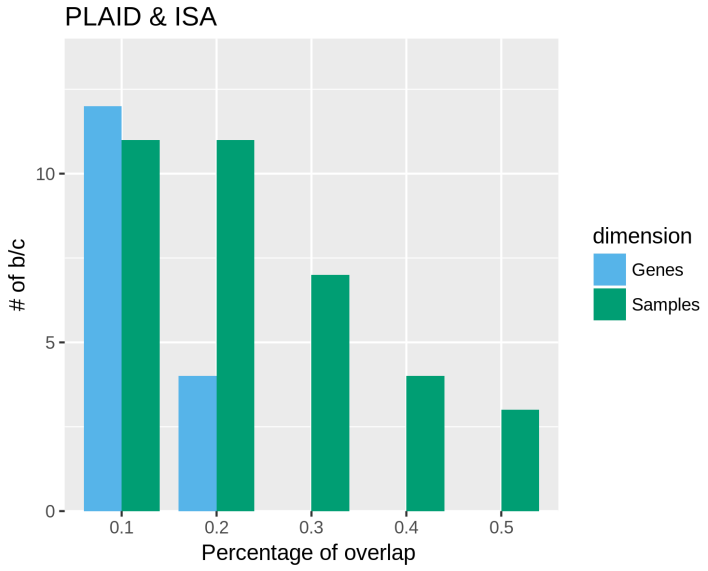


FIGURE 3.2: Overlaps for genes and samples in final biclustering results between the two algorithms

The representation scores where calculated as a ratio of the samples that were found for each class in a bicluster and the total number of samples that exist for the respective class. There scores are rounded at the first decimal, thus main lines do not sum to 1. We observe that there are only 3 biclusters in both matrices of Figure 3.2 that contain proportions higher than 0.5: 3,7,13 for PLAID and 2,4,6,10 for ISA. In many cases, we see conditions that are represented by one or two samples i.e the algorithm hasn't identified the current pattern in the rest of the

samples. This characteristic is far from encouraging as it would be reasonable to expect that a pattern is preserved across the majority of samples for it to be reliable. It could also be a sign of batch effects in our dataset, which could be validated if we observed samples originating from the same experimental batch appearing together in biclusters.

Bicluster	Genes	Samples	Wt	Tg	Rem	Hum	Enb	Cim
1	32	7	0	0	0	0.3	0.4	0.3
2	37	8	0.2	0.2	0	0.2	0.1	0.1
3	41	9	0	0.7	0	0	0.2	0.1
4	45	12	0	0.2	0.3	0	0.2	0.3
5	36	8	0	0.4	0	0.2	0.4	0
6	36	13	0.2	0.2	0.1	0.2	0.1	0.2
7	38	11	0	0.1	0	0.3	0.5	0.2
8	41	16	0.1	0.2	0.1	0.2	0.2	0.2
9	75	9	0.3	0.2	0.2	0.2	0	0
10	54	14	0.1	0.2	0.1	0.2	0.1	0.2
11	74	8	0.2	0	0	0.4	0.2	0.1
12	75	12	0.1	0.1	0.2	0.1	0.2	0.4
13	83	11	0	0.1	0	0.1	0.4	0.5

(A) PLAID

Bicluster	Genes	Samples	Wt	Tg	Rem	Hum	Enb	Cim
1	34	25	0.3	0	0.3	0.2	0.2	0.1
2	76	5	0.6	0	0.2	0.2	0	0
3	50	13	0	0.2	0	0.2	0.2	0.4
4	77	8	0	0.8	0.2	0	0	0
5	92	6	0.3	0	0.3	0.2	0.2	0
6	55	10	0.3	0	0	0.1	0.1	0.5
7	64	12	0.2	0.2	0.2	0.2	0	0
8	57	22	0.1	0	0.1	0.1	0.3	0.3
9	53	23	0	0.4	0.1	0.1	0.2	0.2
10	215	6	0.3	0	0.2	0	0	0.5
11	248	23	0.1	0	0	0.3	0.3	0.3
12	438	23	0	0.4	0.1	0.1	0.3	0.1

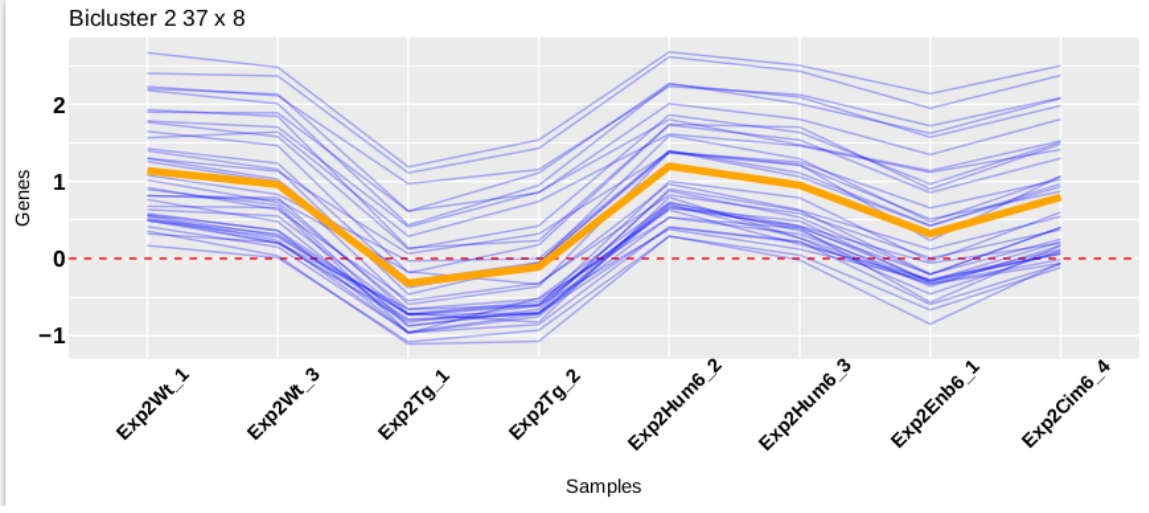
(B) ISA

FIGURE 3.3: Sample representations scores for both algorithms

3.1.3 Visualization of final biclusters

Visualization of the biclusters was attempted in three different ways in order to obtain as much information as they can provide in a comprehensible way. The main forms of visualization for single biclusters proposed in literature are: heatmaps and parallel coordinates i.e. common ways to visualized gene expression matrices in general.

Parallel coordinates plots



The parallel coordinate (PC) technique gives the opportunity of visualization and analysis of high dimensional data in a two dimensional framework. In this technique, each gene profile g_i is considered as a vector of m dimensions: $g_i = (a_{i1}, a_{i2}, ... a_{im})$ where a_{ik} is the transcript measurement of g_i under condition c_k. Conditions are represented as vertical lines of equal distances between them, along the x-axis. Each gene profile g_i is displayed as a line of m points(x_{ck}, y_k), with y_k proportional to a_{ik}. Despite the fact that the orthogonal property is destroyed, geometric structure can still be preserved by the PC plot. This means that the scheme described, includes the n-dimensional coordinate axis (n = number of genes) drawn as parallel instead of vertical as it is difficult to represent more than three orthogonal coordinate axis in a 3-D world. However, this Cartesian representation does not alter the geometrical relationships. (Wegman, 1990) In our implementation, the thick orange line represents the the mean expression levels. PC plots are very useful in cases of groups of genes (or biclusters in our case) rather than whole matrices, in which case one would only see clustered lines. In contrast, in groups of genes they provide patterns and slopes that researchers can comprehend intuitively. Changing the order of the plot axis can reveal more informative arrangements and thus, patterns between the samples. Also, the numerical representation gives better perception of the expression levels in contrast with color scale especially in the case of constant and coherent biclusters.

Heatmaps

Biologists are very familiar with gene expression data being represented in heatmaps, 2D representations where rows correspond to genes, columns refer to samples and expression levels are color-coded according to a given scale. Heatmaps are often accompanied by information about the structure of the expression matrix through the reordering of columns/ rows to reveal groups identified by hierarchical clustering. In Fig. 3.3 and 3.4 we observe how the reordering of the same heatmaps from rows to both rows and columns alters the overall observations we can make on the same bicluster

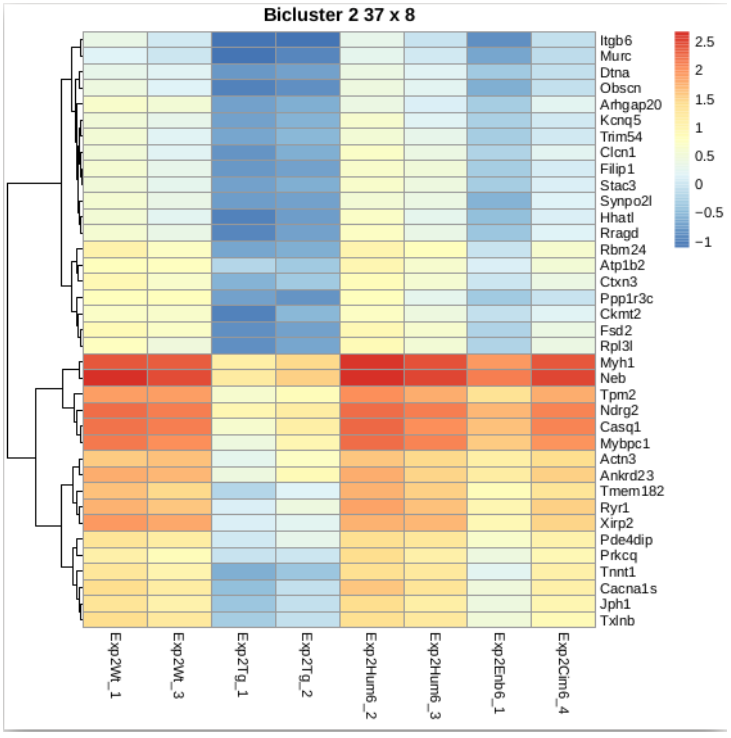


FIGURE 3.4: Example of bicluster heatmaps with rows reordered

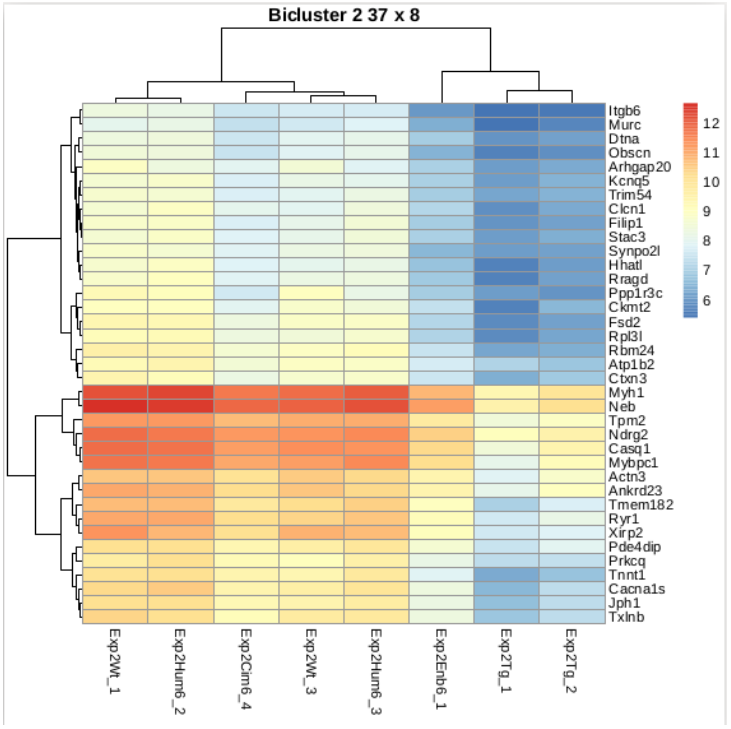


FIGURE 3.5: Example of bicluster heatmaps with rows and columns reordered

Boxplots of Log(FC)

We created another set of plots that uses the log(FC) value from Dunnet's analysis for each gene and all samples. They provide a quick overview of the behaviour of the genes in all samples when

compared to the healthy and the diseased samples. These plots could reveal information about overall differences or similarities between the samples and the state of the differential expression state of the genes (up- or down- regulated) that participate in contrast with the healthy and diseased controls. The downside is that information on individual genes is lost since we are now looking at general distributions of gene subsets.

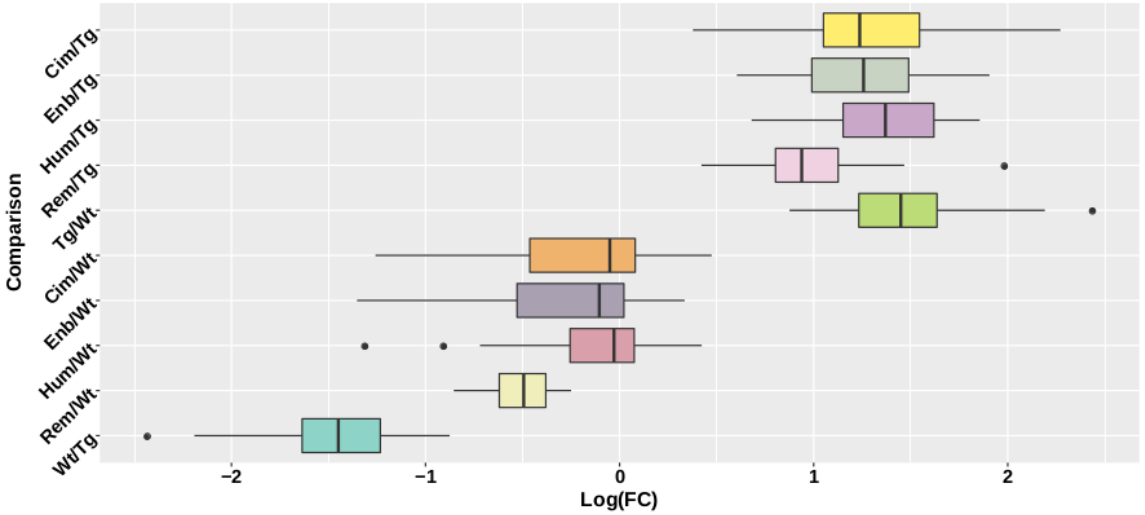


FIGURE 3.6: Log(FC) boxplots that results from the genes of the current bicluster

3.1.4 Functional enrichment of biclusters

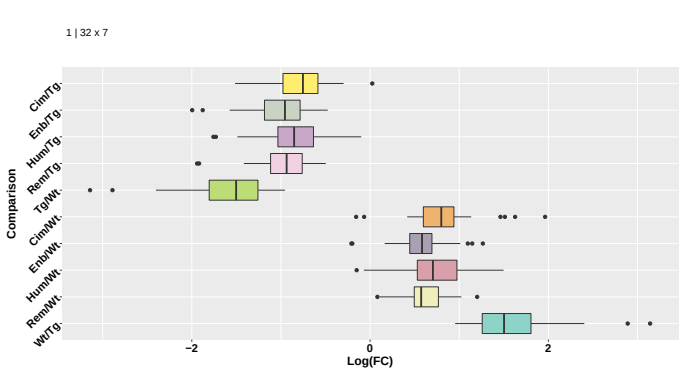
We performed a functional analysis of the biclusters using the "gProfiler" R package. For convenience in visualization and interpretation we combined the Log(FC) boxplots described above, the top-10 functional terms according to the significance of their enrichment and the samples representation scores in a merged version.

3.1.5 Modules

The basic results of the buclustering analysis are presented in the following section: Each module is represented in all three forms described above. The graphical representations are accompanied by the functional analysis of the participating genes and a sample representation table.

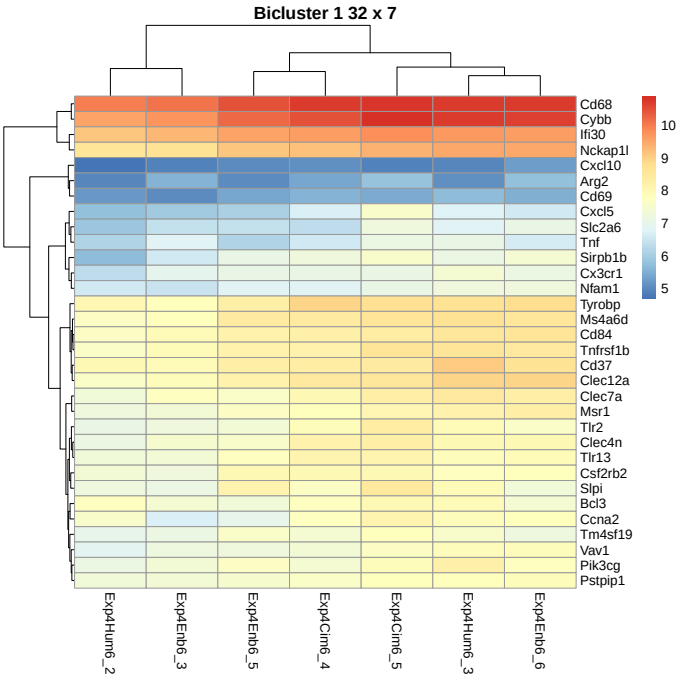
PLAID Bicluster #:

1. Down-regulated gene group disease. Overcompensated by all drugs but pattern of similar behavior is mostly found between among Cimzia, Enbrel, Humira. Cytokine meiated signaling pathway pathway enriched.



adj. P−value	Term
1.99e−08	response to peptide hormone
4.11e−08	regulation of neutrophil mediated killing of gram−negative bacterium
8.4e−08	regulation of membrane protein ectodomain proteolysis
8.4e−08	cytokine−mediated signaling pathway
2.41e−07	response to bacterial lipoprotein
9.51e−07	microglial cell activation involved in immune response
1.19e−06	phosphatidylinositol−4,5−bisphosphate 3−kinase activity
1.67e−06	perinuclear region of cytoplasm
1.71e−06	myoblast differentiation
2.01e−06	positive regulation of DNA binding transcription factor activity

Class	R.score
Wt	0
Tg	0
Rem	0
Hum	0.3
Enb	0.4
Cim	0.3



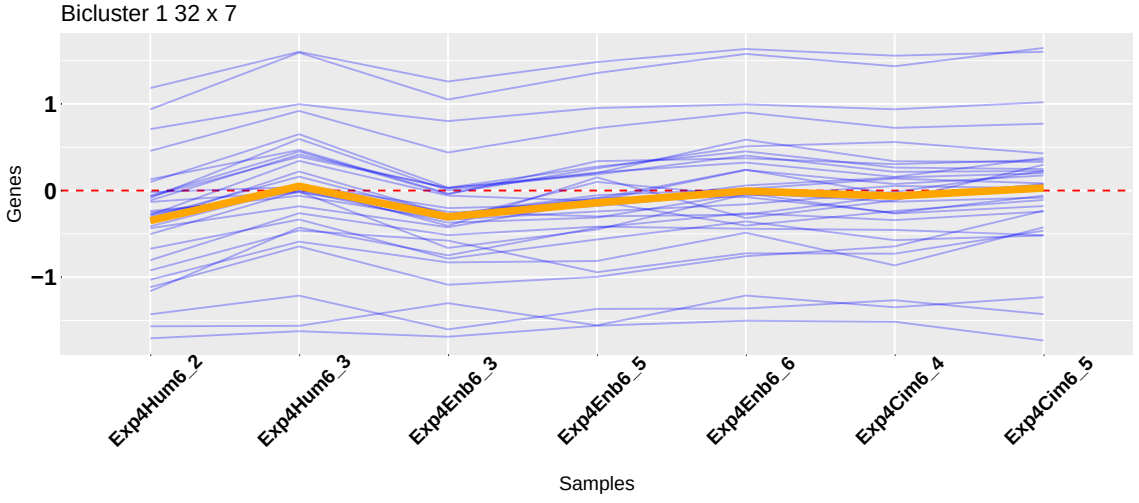
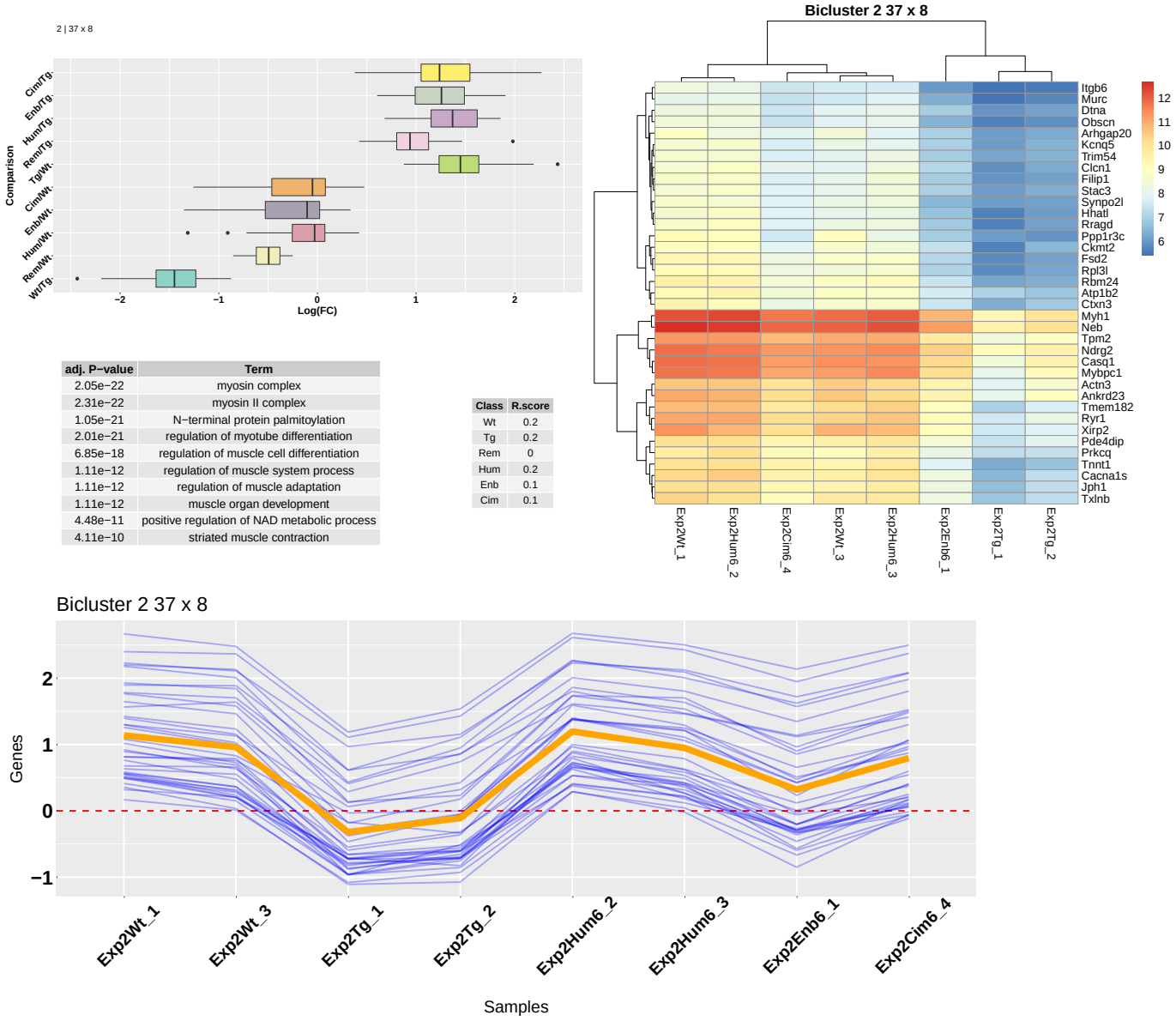


FIGURE 3.7: PLAID Bicluster 1

2. Up-regulated in disease, suppressed in healthy samples. Restored by Cimzia, Enbrel, Humira mostly. Well characterized in terms of functions as most of them concern muscular activity: myosin complex, regulation of muscle cell differentiation and muscle cell system process, muscle organ development.

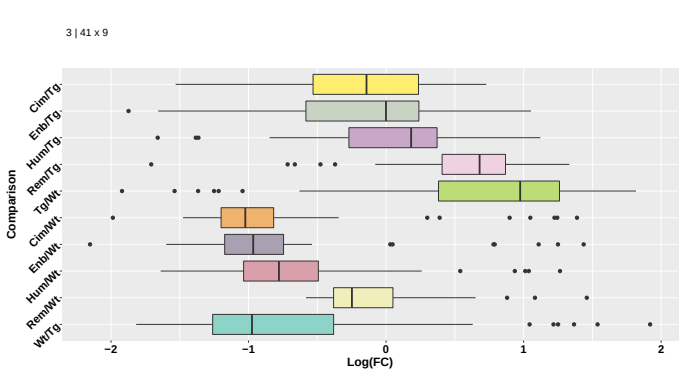


adj. P−value	Term
2.05e−22	myosin complex
2.31e−22	myosin II complex
1.05e−21	N−terminal protein palmitoylation
2.01e−21	regulation of myotube differentiation
6.85e−18	regulation of muscle cell differentiation
1.11e−12	regulation of muscle system process
1.11e−12	regulation of muscle adaptation
1.11e−12	muscle organ development
4.48e−11	positive regulation of NAD metabolic process
4.11e−10	striated muscle contraction

Class	R.score
Wt	0.2
Tg	0.2
Rem	0
Hum	0.2
Enb	0.1
Cim	0.1

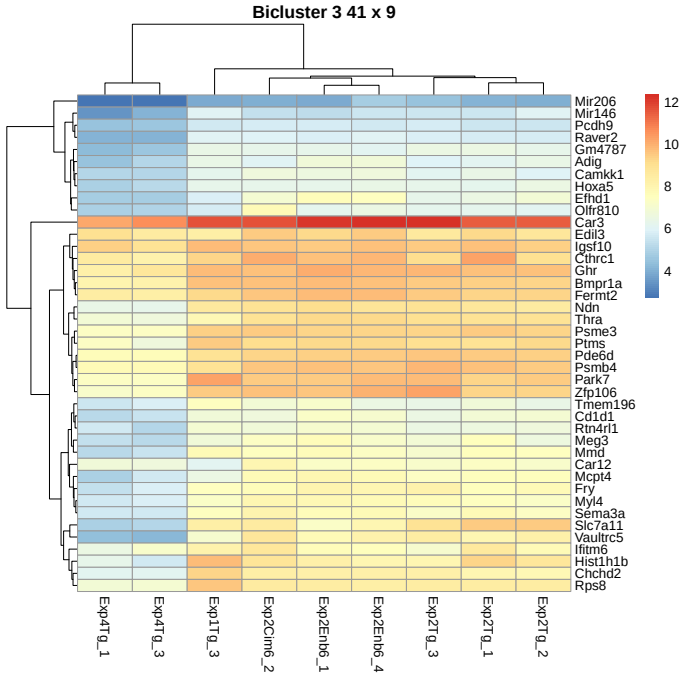
FIGURE 3.8: PLAID Bicluster 2

3. Great interquartile range for all samples and many outliers. Pattern is mostly identified between two experimental Tg batches (Exp4 & Exp2).



adj. P−value	Term
0.0149	neural plate mediolateral regionalization
0.0432	anatomical structure arrangement
0.0432	mesoderm structural organization
0.0432	paraxial mesoderm structural organization
0.0432	positive regulation of oxidative phosphorylation uncoupler activity
0.0432	regulation of pyrroline−5−carboxylate reductase activity
0.0432	positive regulation of pyrroline−5−carboxylate reductase activity
0.0432	regulation of L−dopa decarboxylase activity
0.0432	positive regulation of L−dopa decarboxylase activity
0.0432	one−carbon metabolic process

Class	R.score
Wt	0
Tg	0.7
Rem	0
Hum	0
Enb	0.2
Cim	0.1



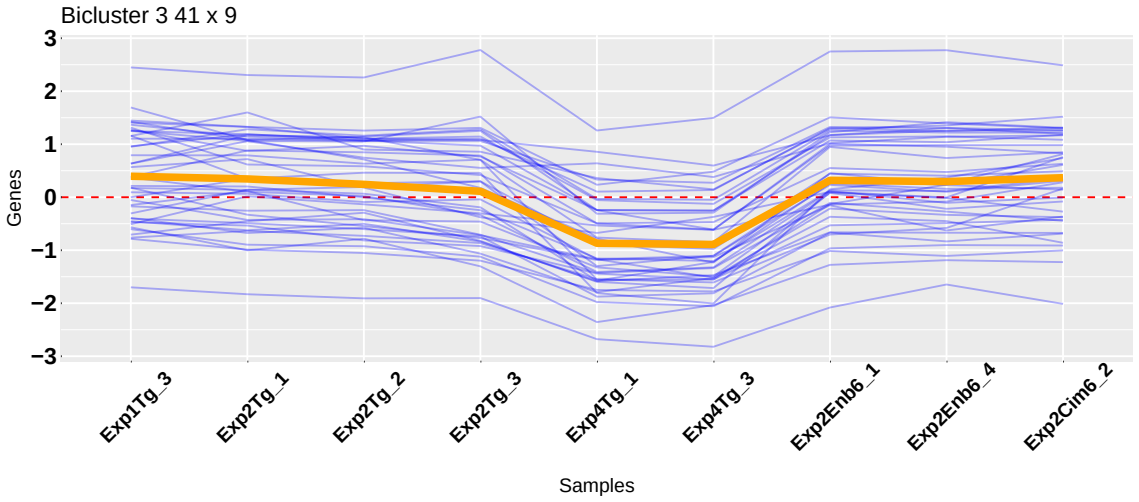
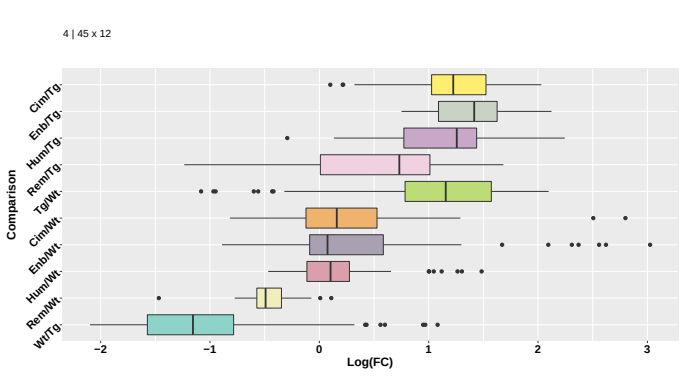


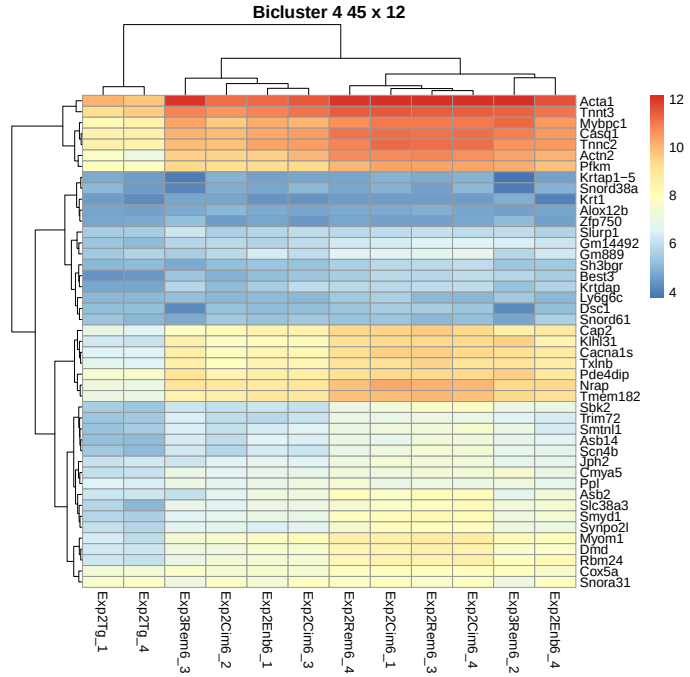
FIGURE 3.9: PLAID Bicluster 3

4. In this bicluster we observe a group of genes up-regulated in disease whose function is restored by 3 out of the 4 treatments and overcompensated by Remicade. There is good sample representation or Remicade and Cimzia and inadequate of Tg and Enbrel samples.



adj. P−value	Term
5.22e−16	phospholipase C−activating angiotensin−activated signaling pathway
5.22e−16	positive regulation of calcium ion transmembrane transport
1.47e−13	regulation of sodium ion transmembrane transport
2.2e−13	positive regulation of glutamine transport
1.34e−10	regulation of sodium ion transmembrane transporter activity
1.34e−10	regulation of sequestering of calcium ion
1.34e−10	lumenal side of membrane
3.55e−10	intermediate filament
4.38e−09	asparagine transport
4.47e−08	matrix side of mitochondrial inner membrane

Class	R.score
Wt	0
Tg	0.2
Rem	0.3
Hum	0
Enb	0.2
Cim	0.3



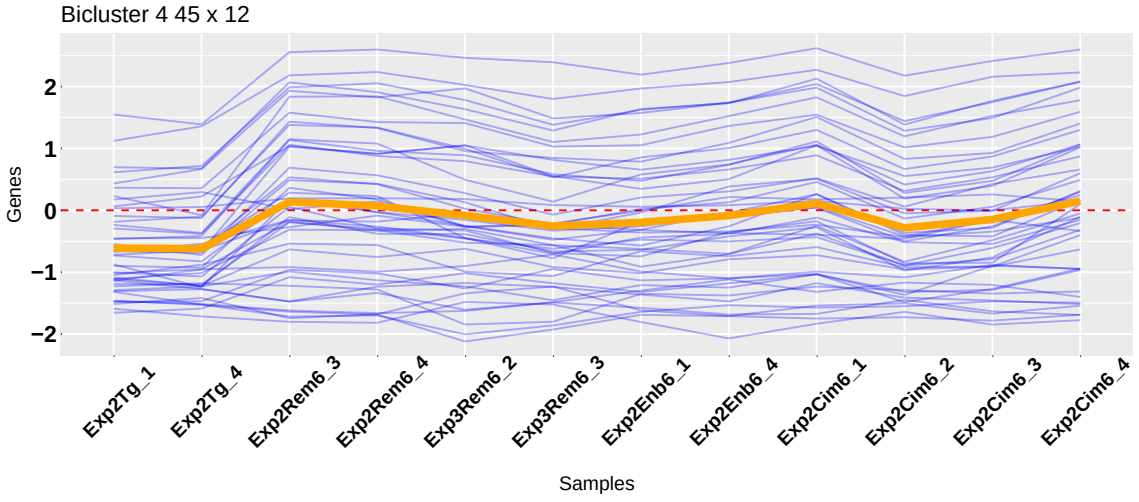
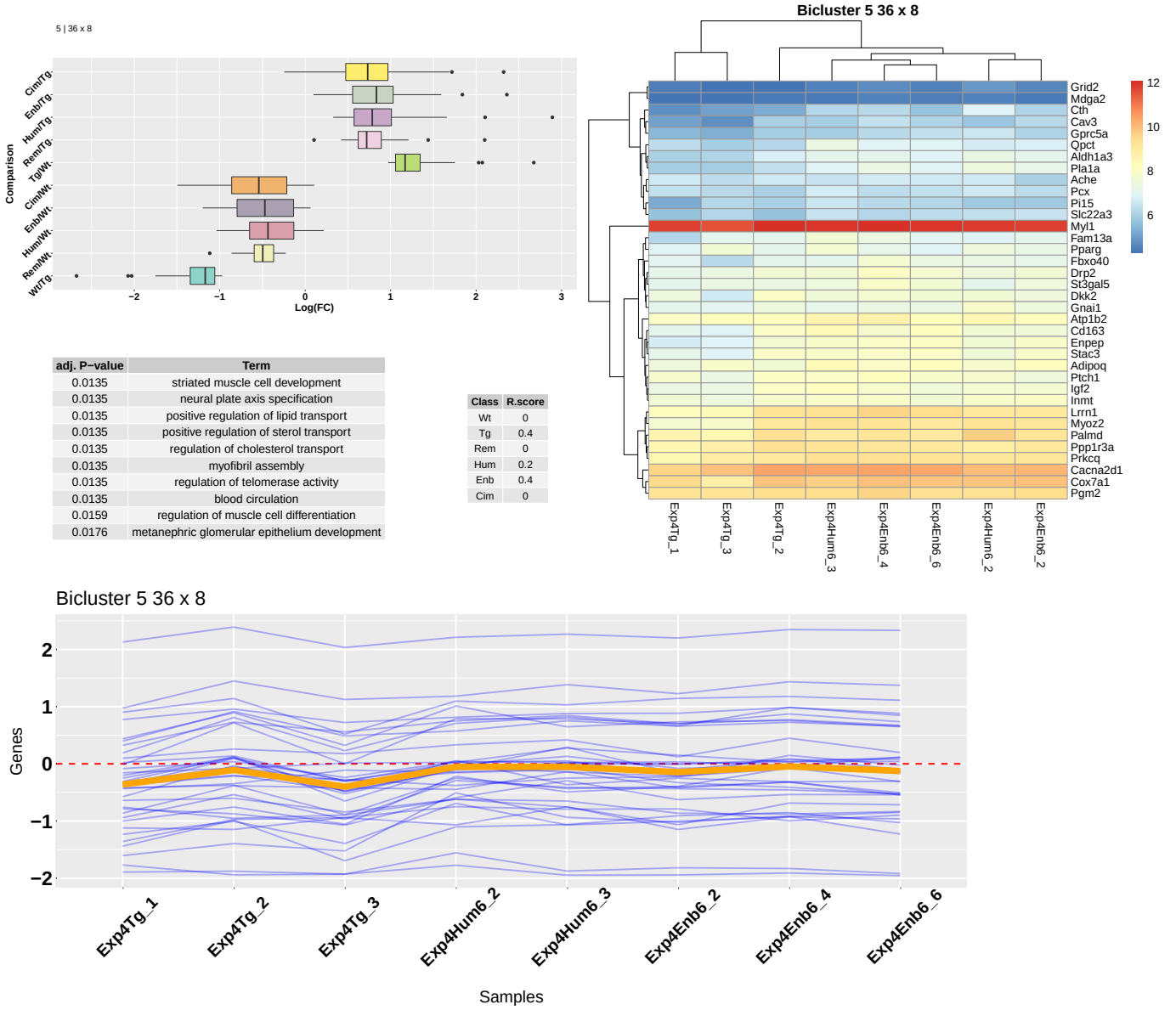


FIGURE 3.10: PLAID Bicluster 4

5. Gene set up-regulated in disease. Overcompensation by all drugs. Functions concern mostly to muscle activity.



adj. P−value	Term
0.0135	striated muscle cell development
0.0135	neural plate axis specification
0.0135	positive regulation of lipid transport
0.0135	positive regulation of sterol transport
0.0135	regulation of cholesterol transport
0.0135	myofibril assembly
0.0135	regulation of telomerase activity
0.0135	blood circulation
0.0159	regulation of muscle cell differentiation
0.0176	metanephric glomerular epithelium development

Class	R.score
Wt	0
Tg	0.4
Rem	0
Hum	0.2
Enb	0.4
Cim	0

FIGURE 3.11: PLAID Bicluster 5

6. Down-regulation in disease and overcompensation but no functional enrichment.

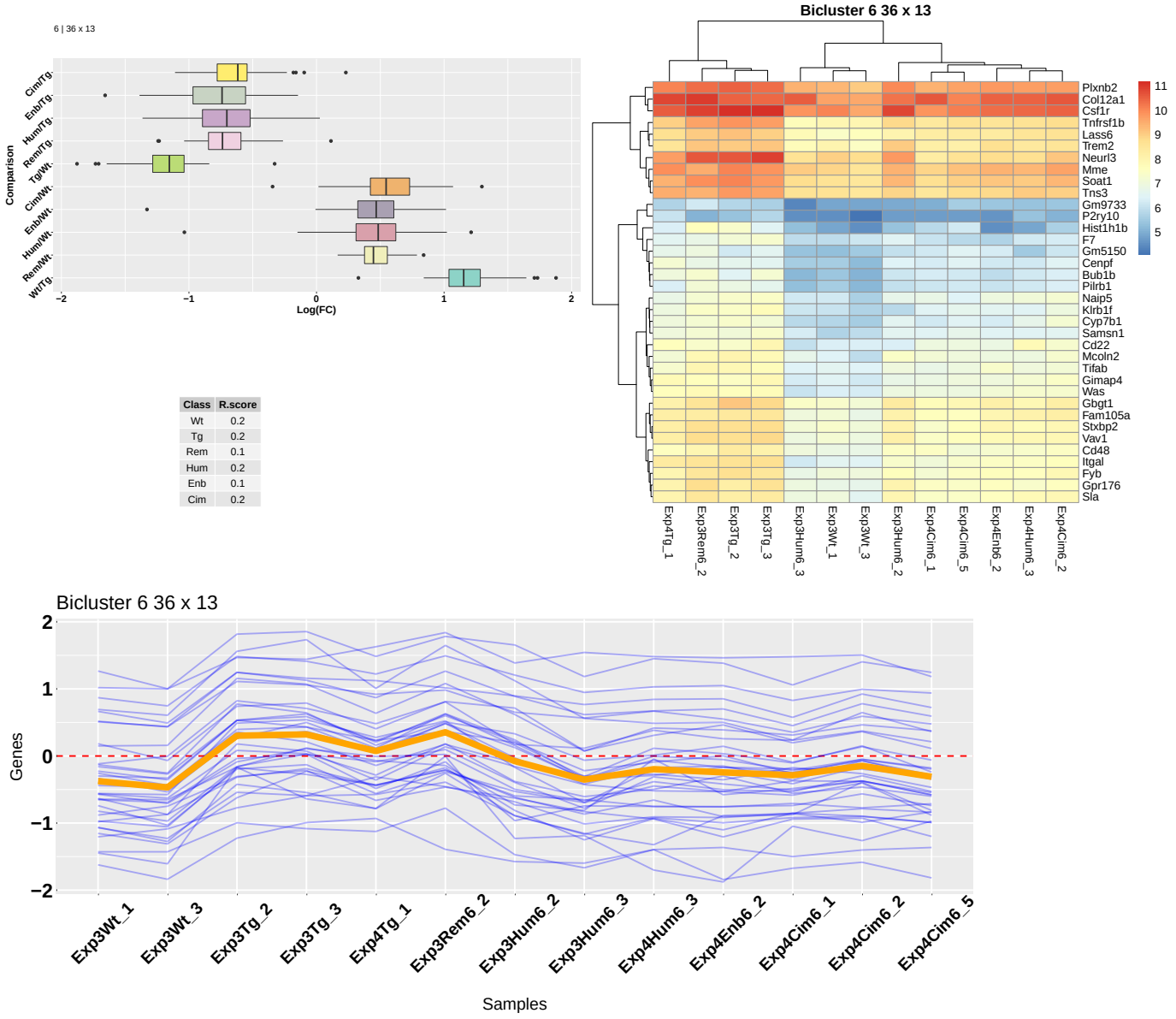
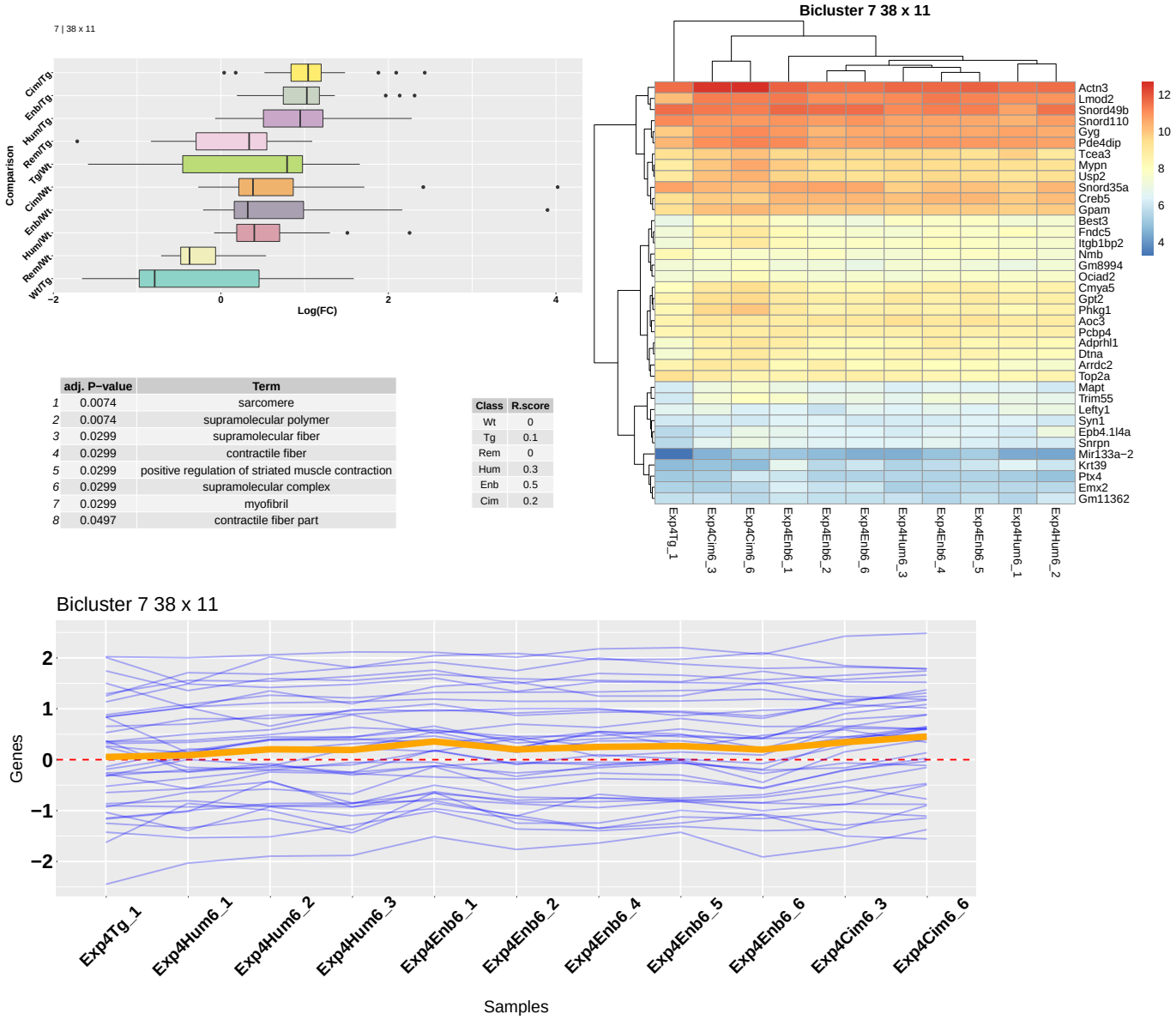


FIGURE 3.12: PLAID Bicluster 6

7. This group is characterized by mild up-regulation in disease but with great range. Overcompensation by all drugs but less with Remicade.

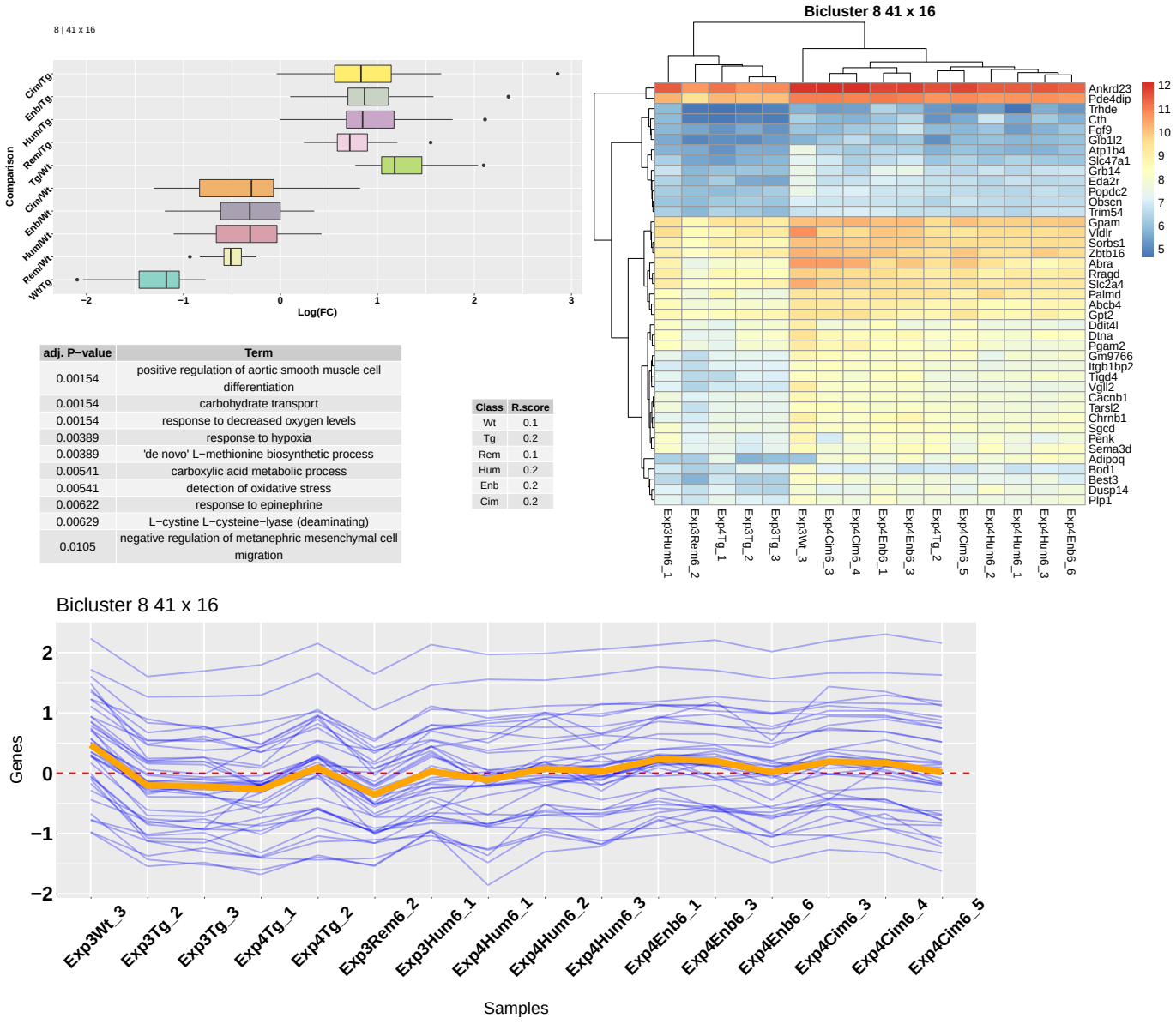


	adj. P−value	Term
1	0.0074	sarcomere
2	0.0074	supramolecular polymer
3	0.0299	supramolecular fiber
4	0.0299	contractile fiber
5	0.0299	positive regulation of striated muscle contraction
6	0.0299	supramolecular complex
7	0.0299	myofibril
8	0.0497	contractile fiber part

Class	R.score
Wt	0
Tg	0.1
Rem	0
Hum	0.3
Enb	0.5
Cim	0.2

FIGURE 3.13: PLAID Bicluster 7

8. Up-regulated gene sets in disease. Restored by all drugs but Remicade succeeds best. Functions concern metabolic and biosynthetic processes as well as detection of oxidative stress.

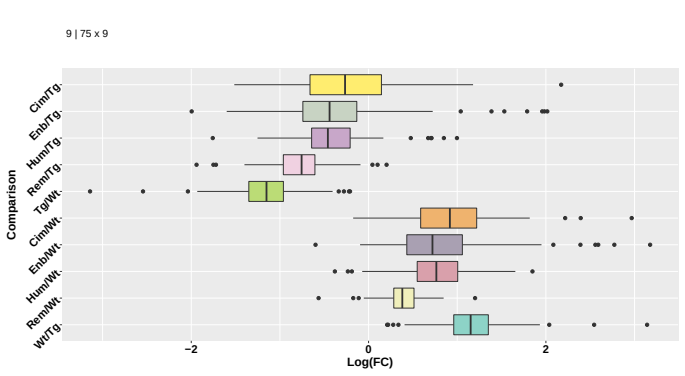


adj. P−value	Term
0.00154	positive regulation of aortic smooth muscle cell differentiation
0.00154	carbohydrate transport
0.00154	response to decreased oxygen levels
0.00389	response to hypoxia
0.00389	'de novo' L−methionine biosynthetic process
0.00541	carboxylic acid metabolic process
0.00541	detection of oxidative stress
0.00622	response to epinephrine
0.00629	L−cystine L−cysteine−lyase (deaminating)
0.0105	negative regulation of metanephric mesenchymal cell migration

Class	R.score
Wt	0.1
Tg	0.2
Rem	0.1
Hum	0.2
Enb	0.2
Cim	0.2

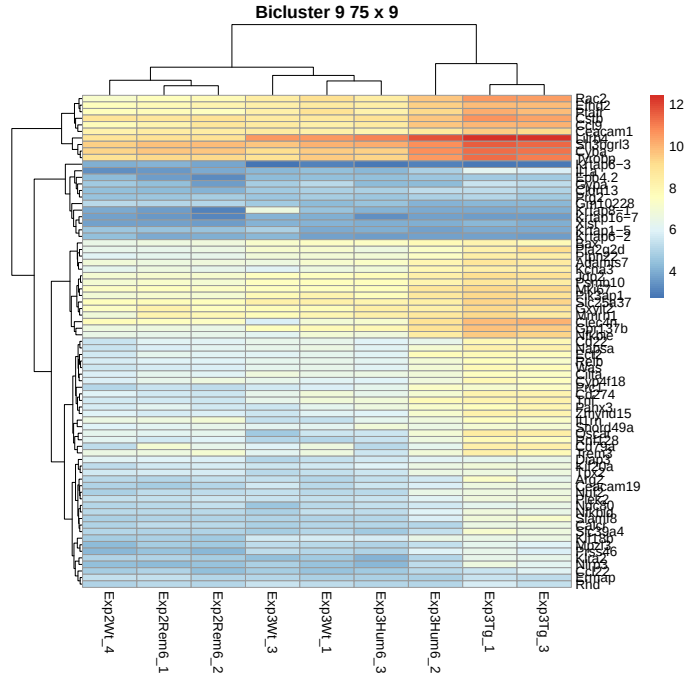
FIGURE 3.14: PLAID Bicluster 8

9. Down-regulated gene set in disease less overcompensated by Remicade. Functional enrichment concerns mostly immune system processes such as cytotoxic T cell degranulation, leukocute adhesion and differentiation and signaling pathways of the immune system.



adj. P−value	Term
9.41e−06	cytotoxic T cell degranulation
9.41e−06	cellular response to abiotic stimulus
9.41e−06	positive regulation of leukocyte adhesion to arterial endothelial cell
9.41e−06	myeloid leukocyte differentiation
9.41e−06	B cell receptor signaling pathway
1.62e−05	cellular response to 2−O−acetyl−1−O−hexadecyl−sn−glycero−3−phosphocholine
3.12e−05	innate immune response−activating signal transduction
3.12e−05	negative regulation of secretion by cell
3.12e−05	peptide secretion
3.6e−05	regulation of defense response to bacterium

Class	R.score
Wt	0.3
Tg	0.2
Rem	0.2
Hum	0.2
Enb	0
Cim	0



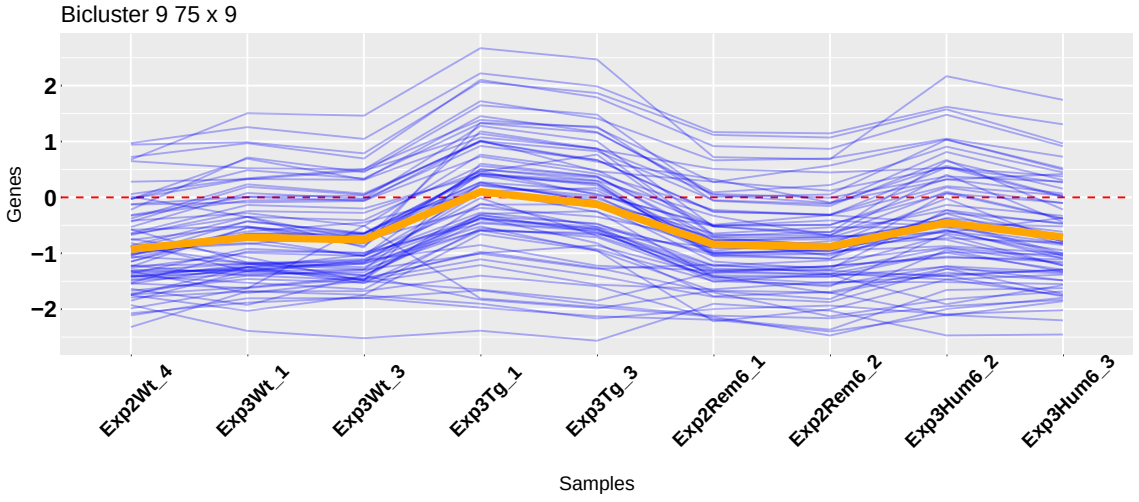
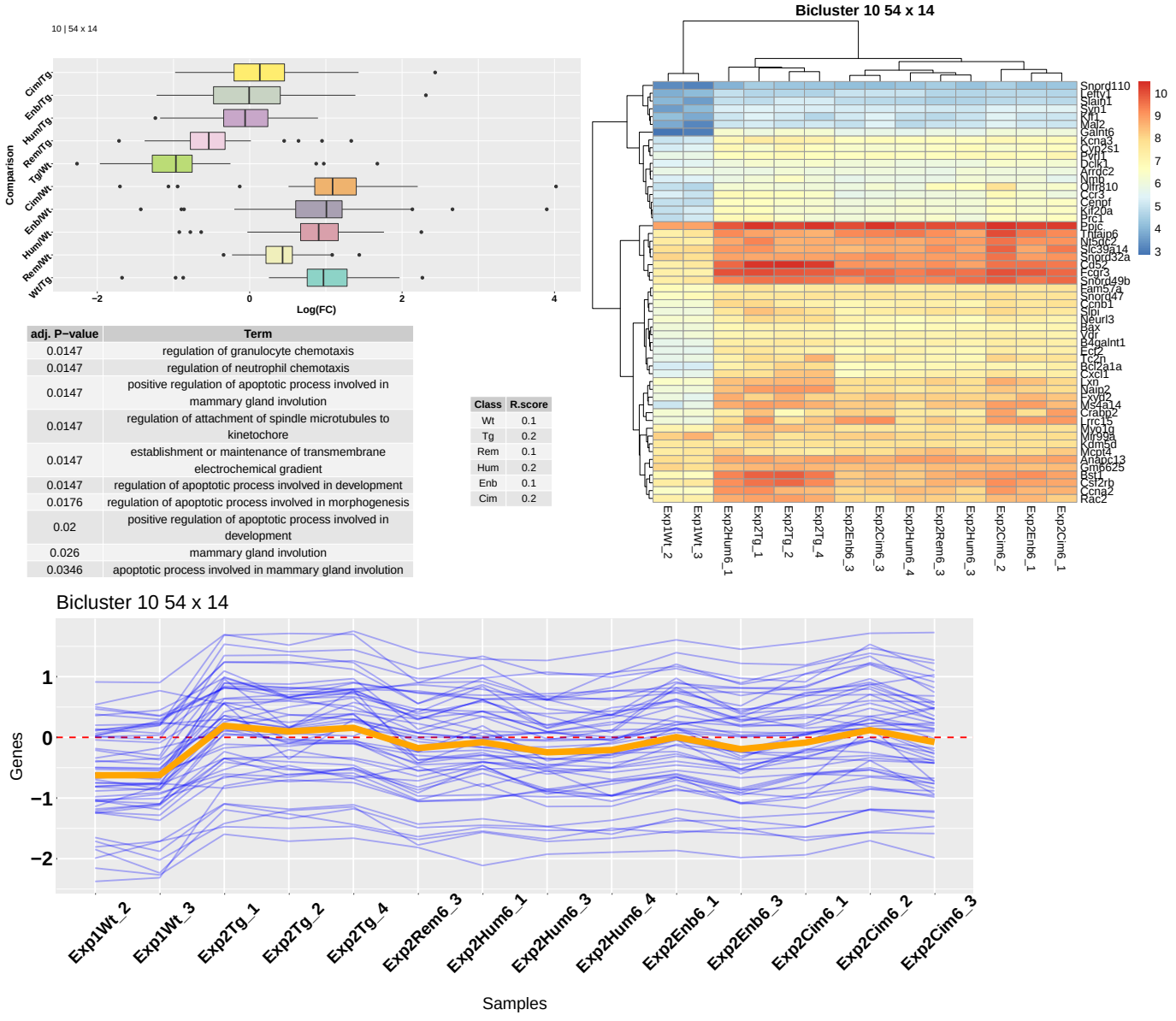


FIGURE 3.15: PLAID Bicluster 9

10. Mild down-regulation in disease. Over-compensated by all drugs but Remicade performs better. Functions concern regulation of apoptotic processes involved in different stages of development and regulation of chemotaxis.

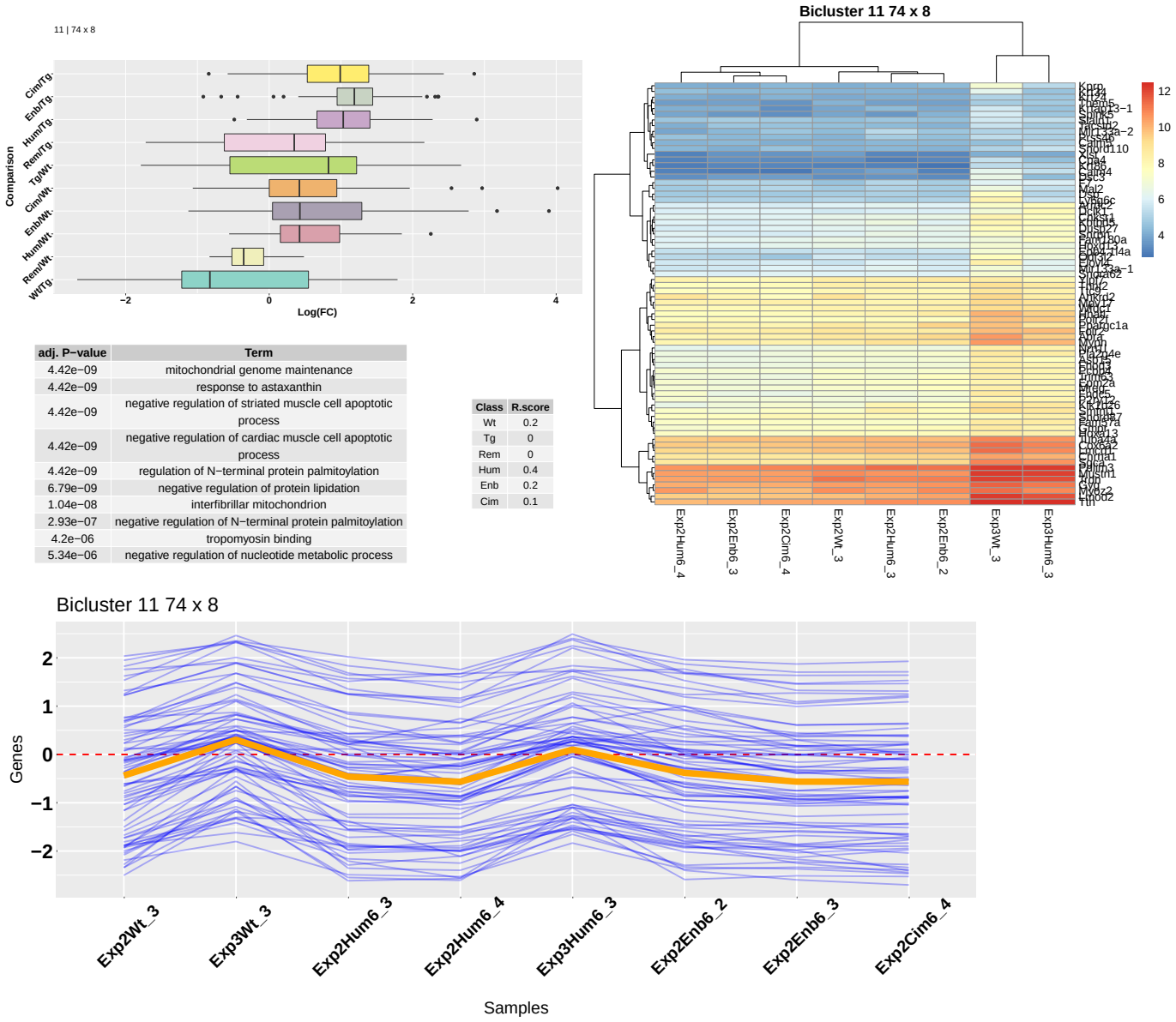


adj. P−value	Term
0.0147	regulation of granulocyte chemotaxis
0.0147	regulation of neutrophil chemotaxis
0.0147	positive regulation of apoptotic process involved in mammary gland involution
0.0147	regulation of attachment of spindle microtubules to kinetochore
0.0147	establishment or maintenance of transmembrane electrochemical gradient
0.0147	regulation of apoptotic process involved in development
0.0176	regulation of apoptotic process involved in morphogenesis
0.02	positive regulation of apoptotic process involved in development
0.026	mammary gland involution
0.0346	apoptotic process involved in mammary gland involution

Class	R.score
Wt	0.1
Tg	0.2
Rem	0.1
Hum	0.2
Enb	0.1
Cim	0.2

FIGURE 3.16: PLAID Bicluster 10

11. Up-regulation in disease mostly restored by Remicade. Great ranges in other treatments.

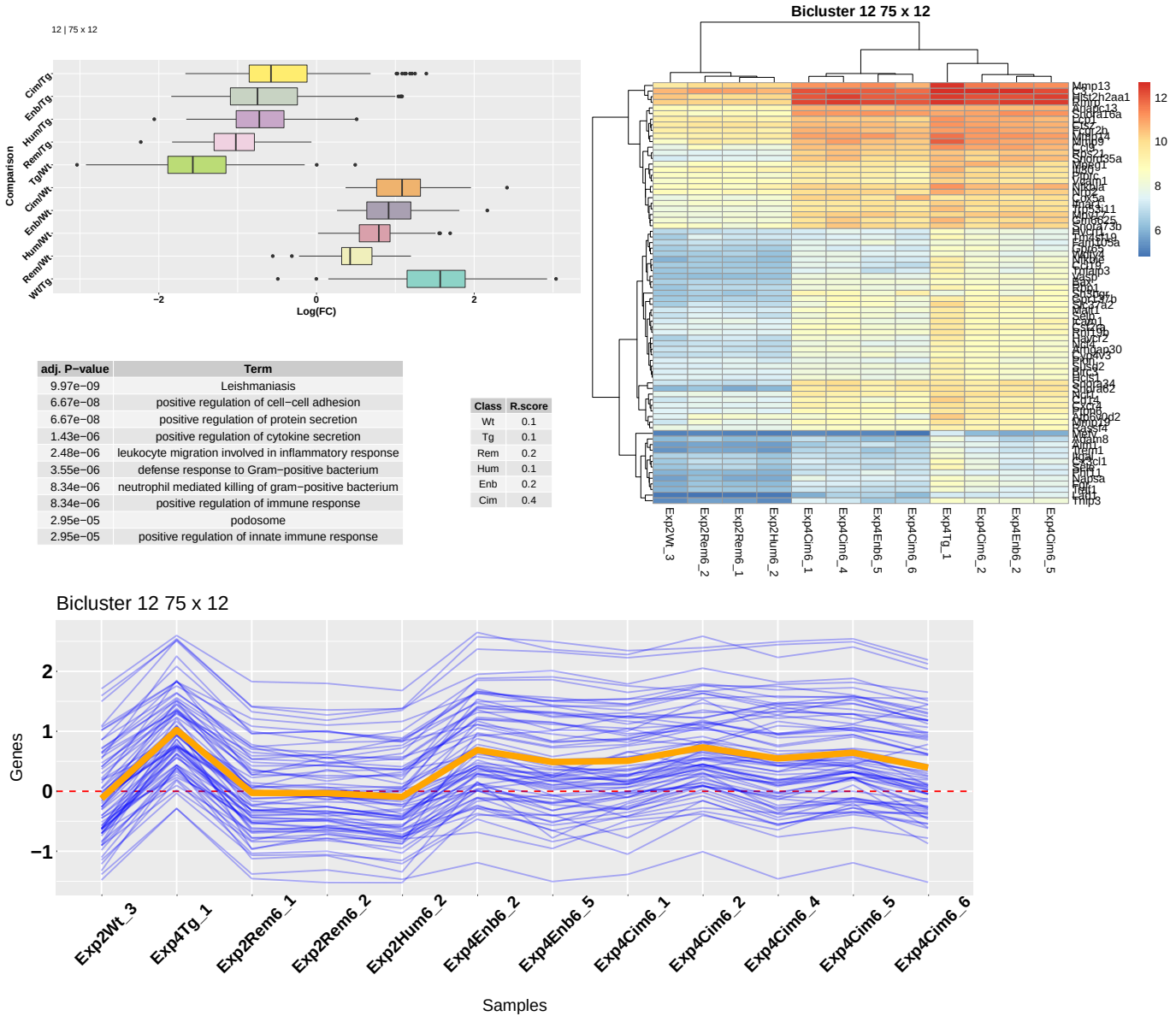


adj. P−value	Term
4.42e−09	mitochondrial genome maintenance
4.42e−09	response to astaxanthin
4.42e−09	negative regulation of striated muscle cell apoptotic process
4.42e−09	negative regulation of cardiac muscle cell apoptotic process
4.42e−09	regulation of N−terminal protein palmitoylation
6.79e−09	negative regulation of protein lipidation
1.04e−08	interfibrillar mitochondrion
2.93e−07	negative regulation of N−terminal protein palmitoylation
4.2e−06	tropomyosin binding
5.34e−06	negative regulation of nucleotide metabolic process

Class	R.score
Wt	0.2
Tg	0
Rem	0
Hum	0.4
Enb	0.2
Cim	0.1

FIGURE 3.17: PLAID Bicluster 11

12. Down-regulation in disease which is mostly restored by Remicade and gradually get worse for the rest of the treatments. Function refer to positive regulation of cytokine secretion, leukocyte migration involved in inflammatory response and other immune-related functions.

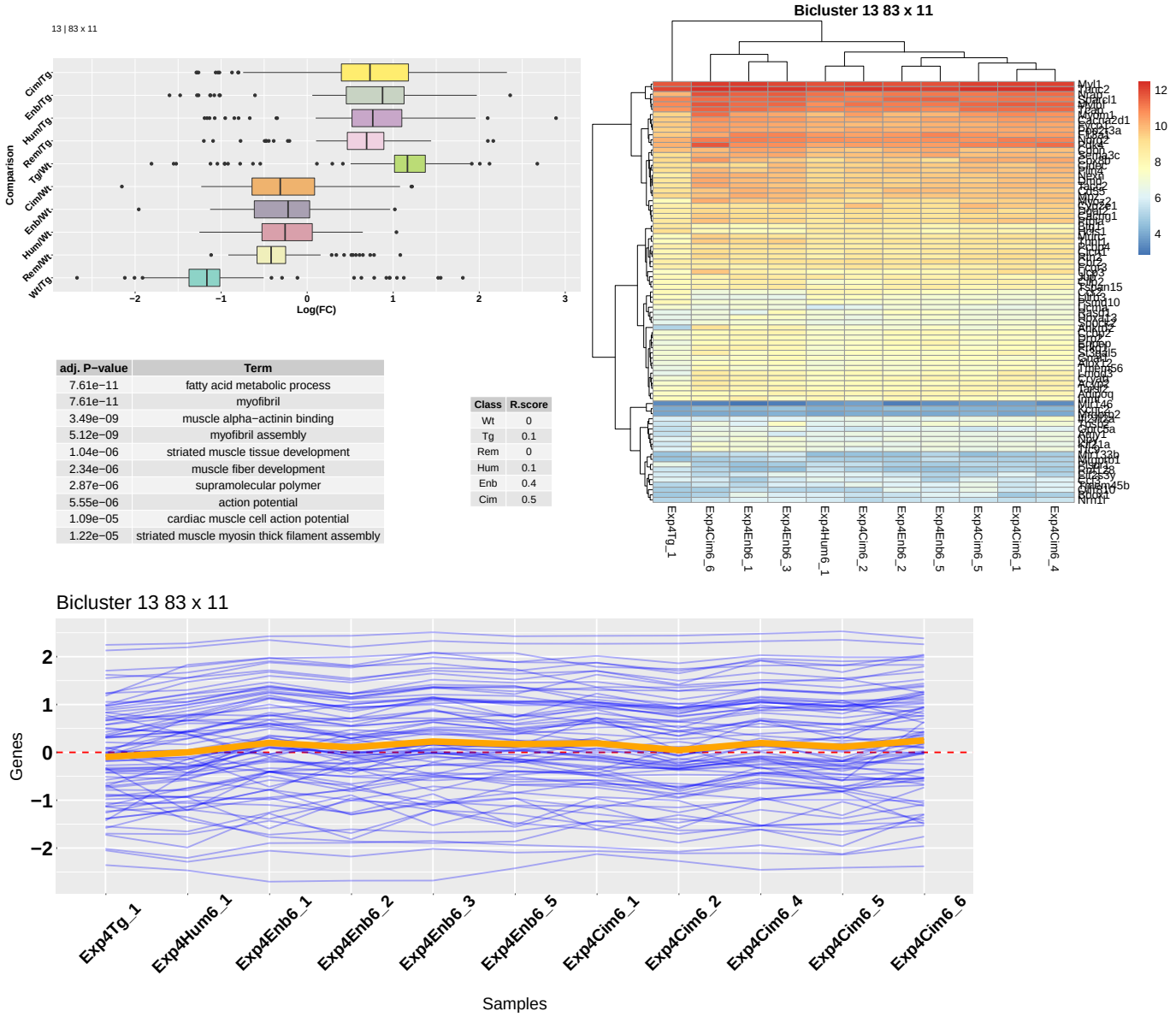


adj. P−value	Term
9.97e−09	Leishmaniasis
6.67e−08	positive regulation of cell−cell adhesion
6.67e−08	positive regulation of protein secretion
1.43e−06	positive regulation of cytokine secretion
2.48e−06	leukocyte migration involved in inflammatory response
3.55e−06	defense response to Gram−positive bacterium
8.34e−06	neutrophil mediated killing of gram−positive bacterium
8.34e−06	positive regulation of immune response
2.95e−05	podosome
2.95e−05	positive regulation of innate immune response

Class	R.score
Wt	0.1
Tg	0.1
Rem	0.2
Hum	0.1
Enb	0.2
Cim	0.4

FIGURE 3.18: PLAID Bicluster 12

13. Up-regulation in disease mostly restored by Remicade. Many outliers. Functional enrichment about muscle development.



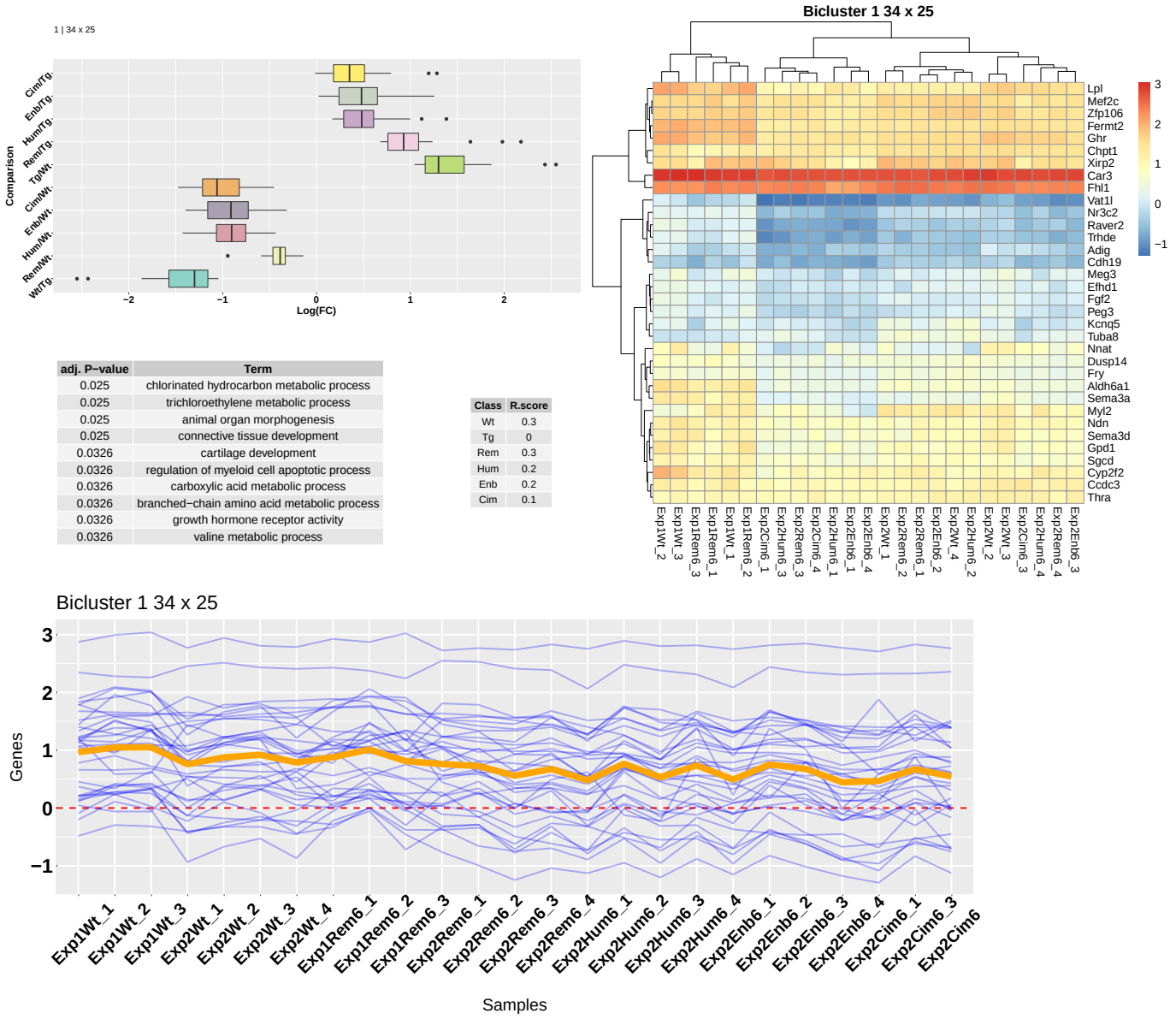
adj. P−value	Term
7.61e−11	fatty acid metabolic process
7.61e−11	myofibril
3.49e−09	muscle alpha−actinin binding
5.12e−09	myofibril assembly
1.04e−06	striated muscle tissue development
2.34e−06	muscle fiber development
2.87e−06	supramolecular polymer
5.55e−06	action potential
1.09e−05	cardiac muscle cell action potential
1.22e−05	striated muscle myosin thick filament assembly

Class	R.score
Wt	0
Tg	0.1
Rem	0
Hum	0.1
Enb	0.4
Cim	0.5

FIGURE 3.19: PLAID Bicluster 13

ISA Bicluster #:

1. Gene set up-regulated in disease. Not sufficient restoration by Humira, Cimzia, Enbrel but successful by Remicade. In this module, we have a good sample representation of Wt, Remicade, Humira and Enbrel classes. Also, the genes that participate are mostly related to metabolic processes as shown by the functional enrichment.

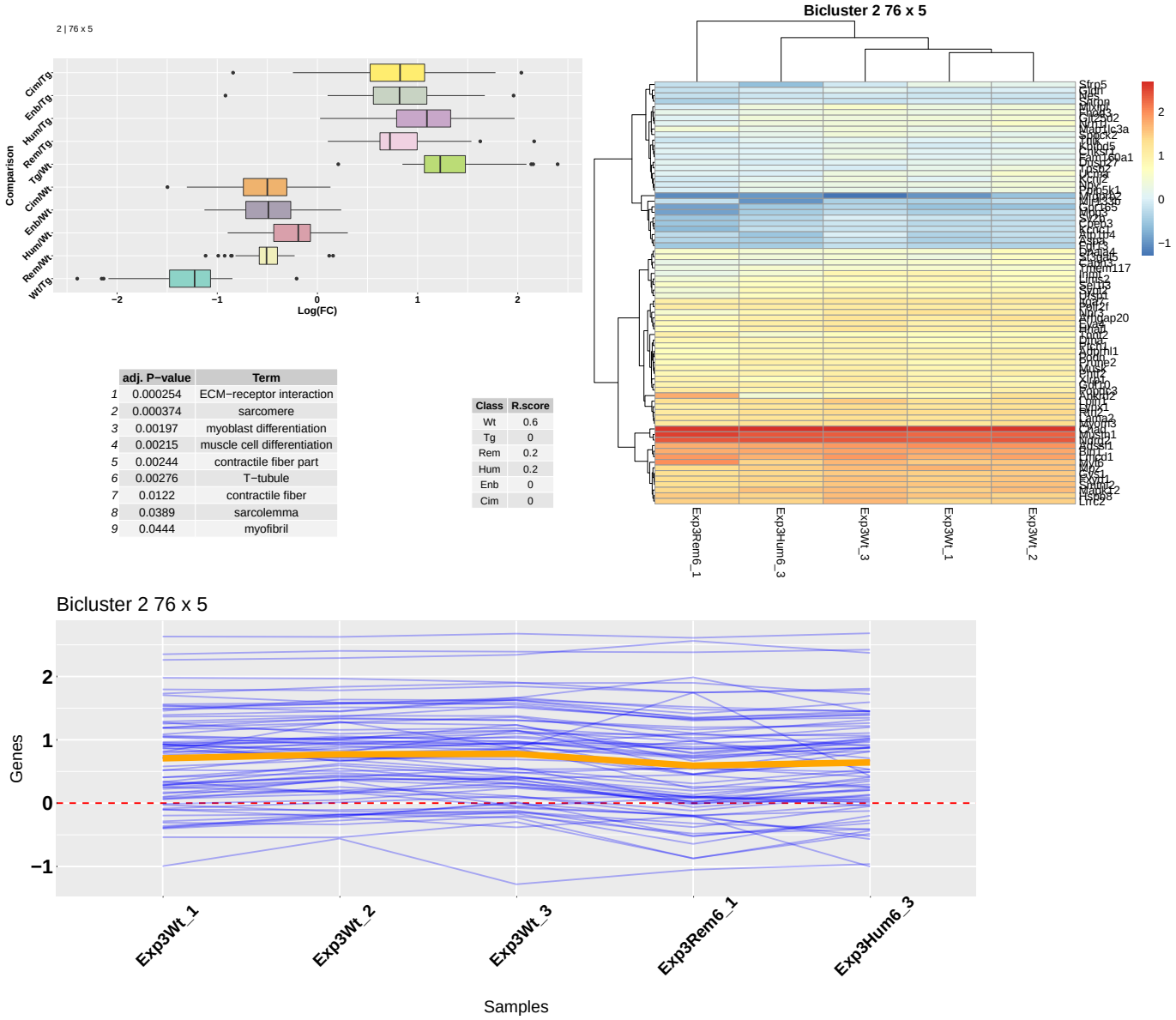


adj. P−value	Term
0.025	chlorinated hydrocarbon metabolic process
0.025	trichloroethylene metabolic process
0.025	animal organ morphogenesis
0.025	connective tissue development
0.0326	cartilage development
0.0326	regulation of myeloid cell apoptotic process
0.0326	carboxylic acid metabolic process
0.0326	branched−chain amino acid metabolic process
0.0326	growth hormone receptor activity
0.0326	valine metabolic process

Class	R.score
Wt	0.3
Tg	0
Rem	0.3
Hum	0.2
Enb	0.2
Cim	0.1

FIGURE 3.20: ISA Bicluster 1

2. Gene set up-regulated in disease. Over-compensated by all treatments. Functions related with muscle activity involved.



	adj. P−value	Term
1	0.000254	ECM−receptor interaction
2	0.000374	sarcomere
3	0.00197	myoblast differentiation
4	0.00215	muscle cell differentiation
5	0.00244	contractile fiber part
6	0.00276	T−tubule
7	0.0122	contractile fiber
8	0.0389	sarcolemma
9	0.0444	myofibril

Class	R.score
Wt	0.6
Tg	0
Rem	0.2
Hum	0.2
Enb	0
Cim	0

FIGURE 3.21: ISA Bicluster 2

3. Mild down-regulation in disease. Restored mostly by Remicade but not functionally enriched.

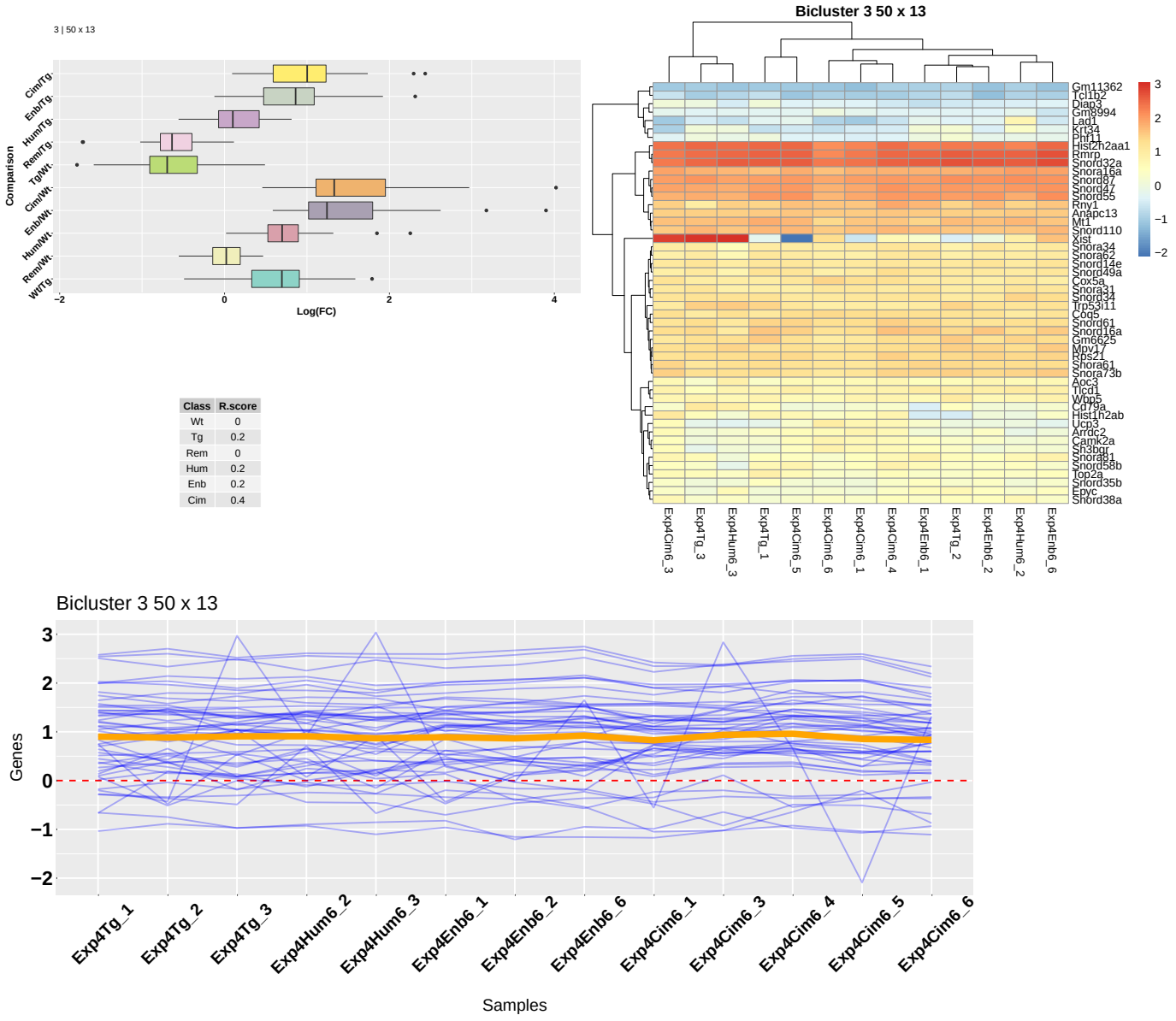
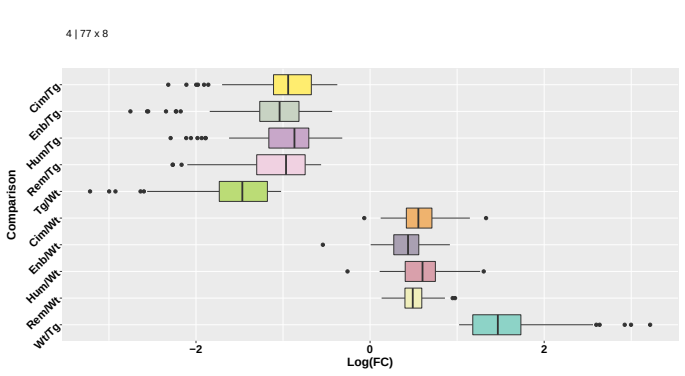


FIGURE 3.22: ISA Bicluster 3

4. Down-regulation in disease. Overcompensated successfully by all treatments. Function related among others with: regulation of toll-like receptor 13 pathway, regulation of interleukin-17 secretion and T cell apoptotic process.



adj. P−value	Term
6.03e−13	regulation of toll−like receptor 13 signaling pathway
4.73e−10	exocytosis
1.22e−08	ion homeostasis
1.06e−07	apical plasma membrane
1.06e−07	regulation of monocyte chemotaxis
3.41e−07	ATP hydrolysis coupled proton transport
3.41e−07	endocytic vesicle membrane
3.41e−07	regulation of T cell apoptotic process
3.41e−07	toll−like receptor signaling pathway
3.79e−07	regulation of interleukin−17 secretion

Class	R.score
Wt	0
Tg	0.8
Rem	0.2
Hum	0
Enb	0
Cim	0

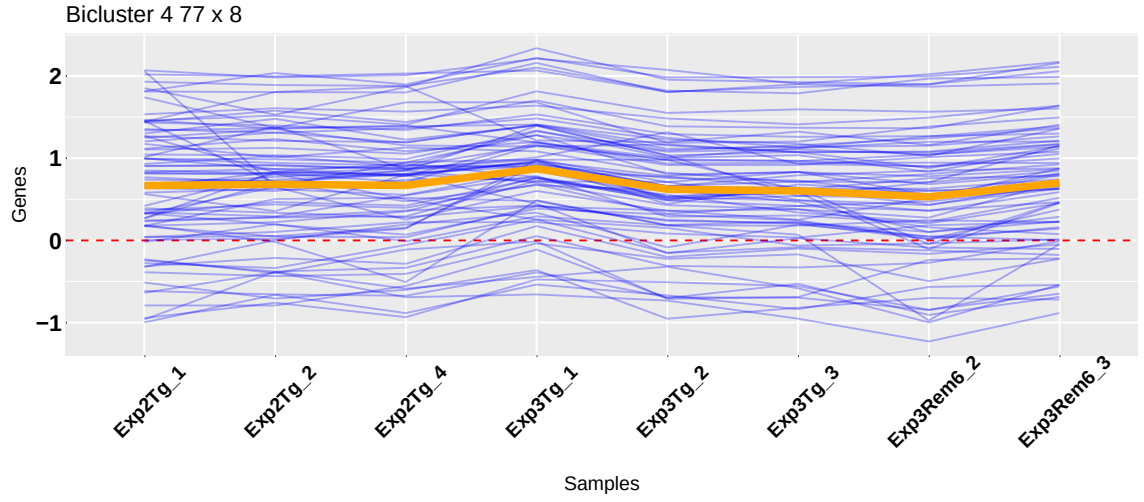
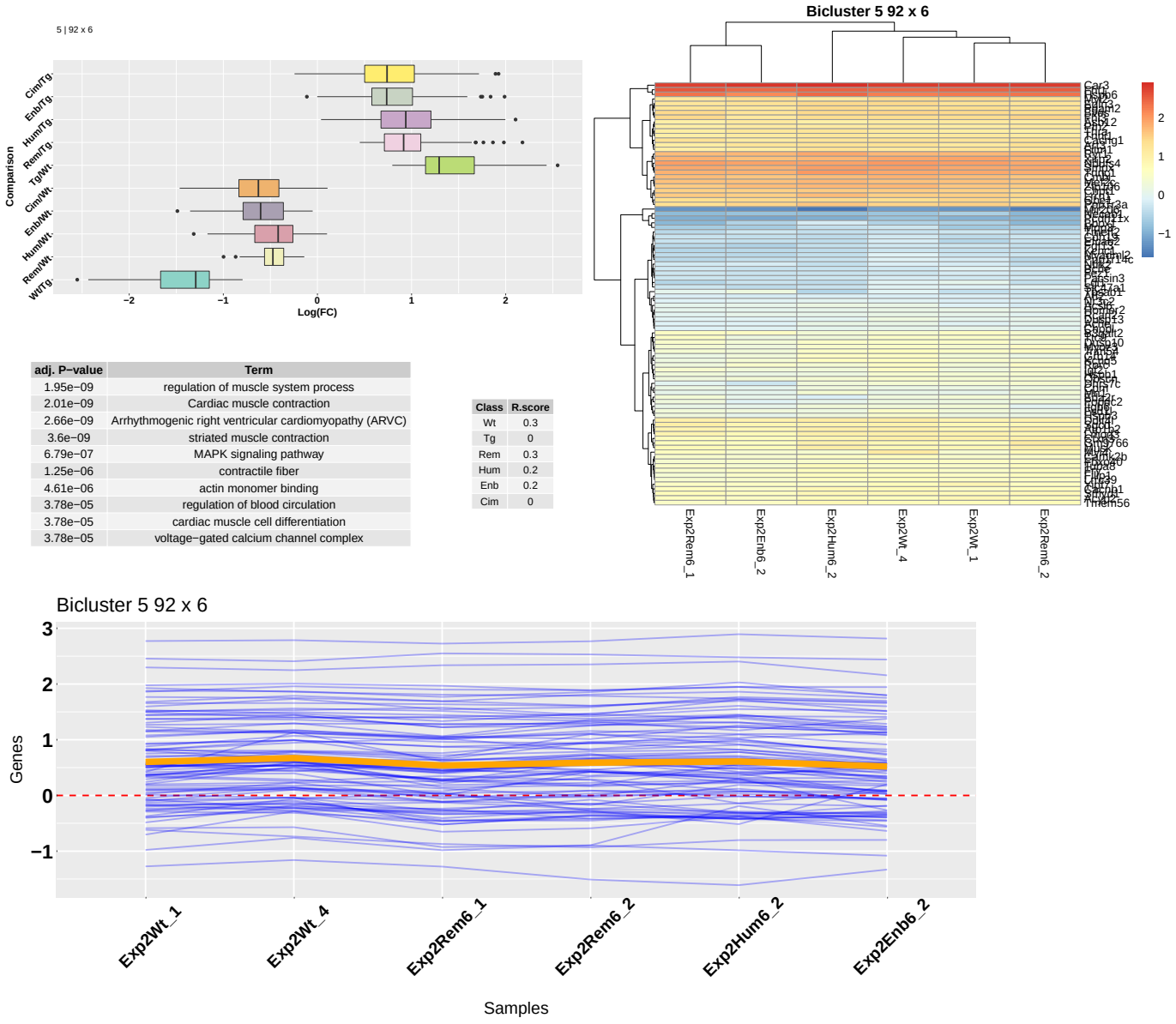


FIGURE 3.23: ISA Bicluster 4

5. Up-regulation in disease. Over-compensated by all drugs. Function related with cardiac muscle processes such as contraction and cell differentiation, MAPK signalling and blood circulation.

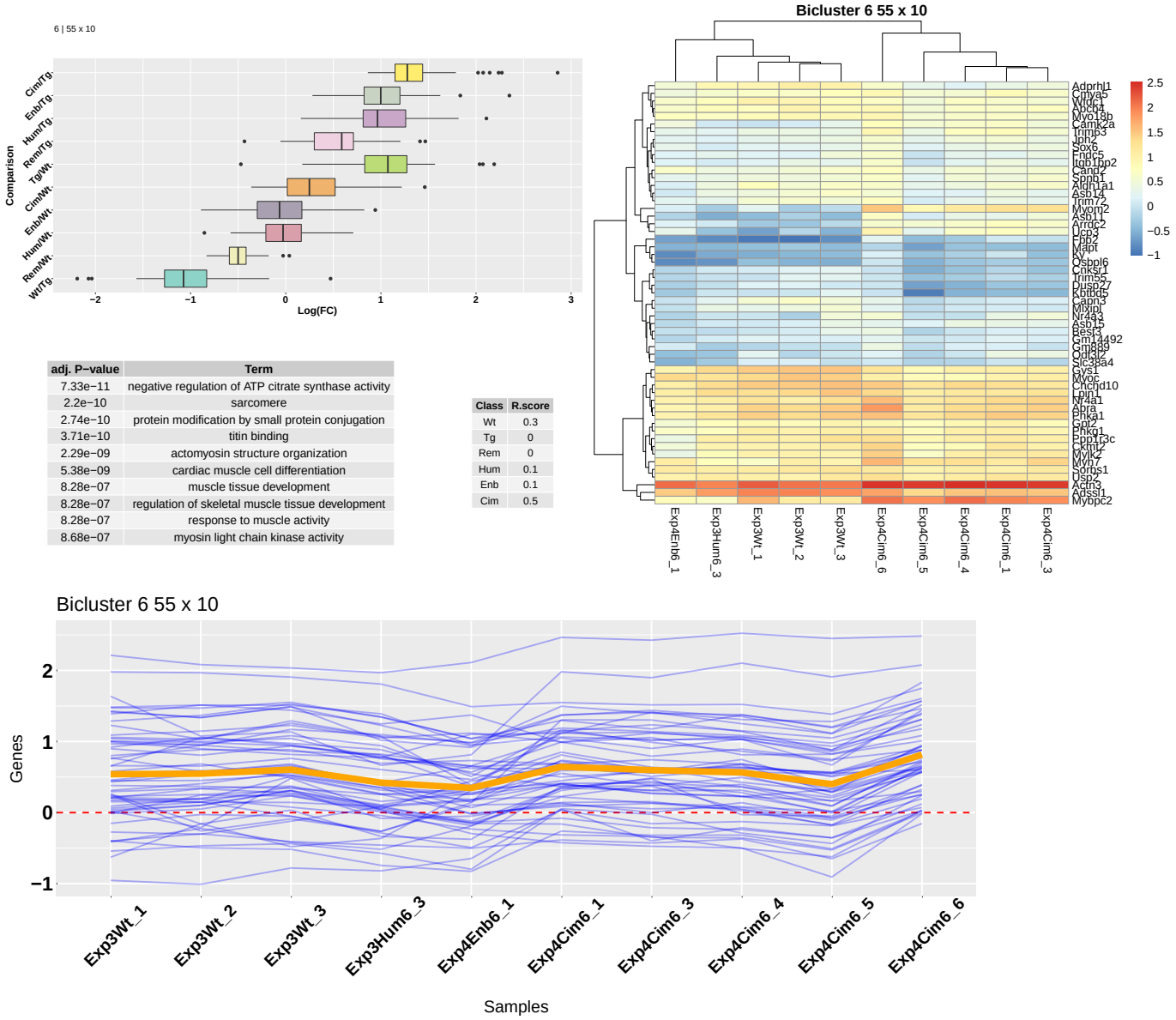


adj. P−value	Term
1.95e−09	regulation of muscle system process
2.01e−09	Cardiac muscle contraction
2.66e−09	Arrhythmogenic right ventricular cardiomyopathy (ARVC)
3.6e−09	striated muscle contraction
6.79e−07	MAPK signaling pathway
1.25e−06	contractile fiber
4.61e−06	actin monomer binding
3.78e−05	regulation of blood circulation
3.78e−05	cardiac muscle cell differentiation
3.78e−05	voltage−gated calcium channel complex

Class	R.score
Wt	0.3
Tg	0
Rem	0.3
Hum	0.2
Enb	0.2
Cim	0

FIGURE 3.24: ISA Bicluster 5

6. Up-regulation in disease. Over-compensated by Humira, Cimzia and Enbrel but insufficiently by Remicade. Functions related with muscle activity.



adj. P−value	Term
7.33e−11	negative regulation of ATP citrate synthase activity
2.2e−10	sarcomere
2.74e−10	protein modification by small protein conjugation
3.71e−10	titin binding
2.29e−09	actomyosin structure organization
5.38e−09	cardiac muscle cell differentiation
8.28e−07	muscle tissue development
8.28e−07	regulation of skeletal muscle tissue development
8.28e−07	response to muscle activity
8.68e−07	myosin light chain kinase activity

Class	R.score
Wt	0.3
Tg	0
Rem	0
Hum	0.1
Enb	0.1
Cim	0.5

FIGURE 3.25: ISA Bicluster 6

7. Gene set with great interquartile ranges in all comparisons and no functional enrichment.

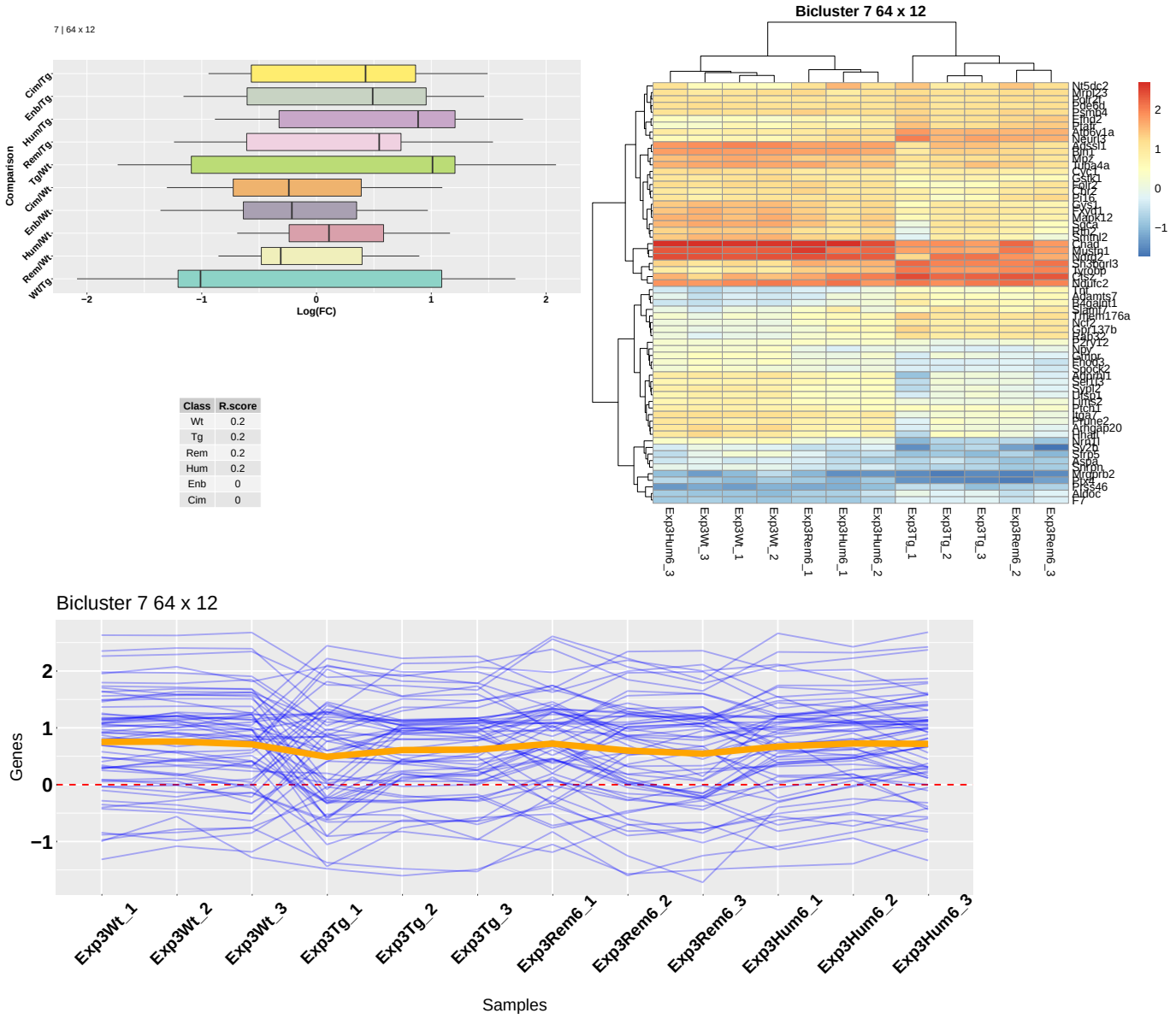
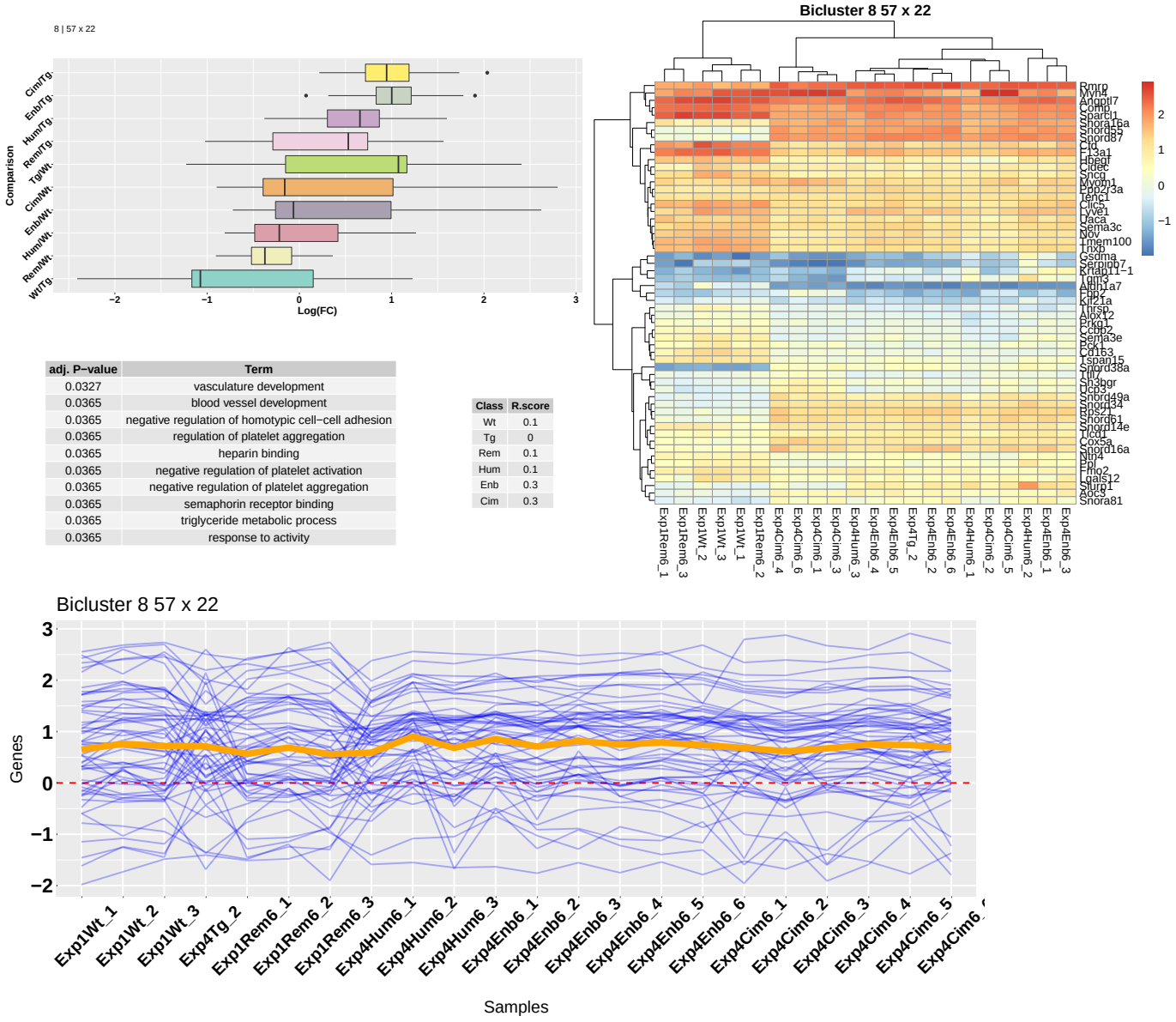


FIGURE 3.26: ISA Bicluster 7

8. Mild up-regulation in disease which is also mildly restored by all treatments.

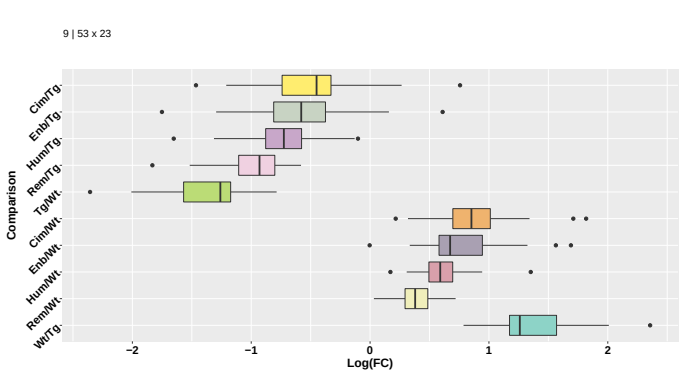


adj. P−value	Term
0.0327	vasculature development
0.0365	blood vessel development
0.0365	negative regulation of homotypic cell−cell adhesion
0.0365	regulation of platelet aggregation
0.0365	heparin binding
0.0365	negative regulation of platelet activation
0.0365	negative regulation of platelet aggregation
0.0365	semaphorin receptor binding
0.0365	triglyceride metabolic process
0.0365	response to activity

Class	R.score
Wt	0.1
Tg	0
Rem	0.1
Hum	0.1
Enb	0.3
Cim	0.3

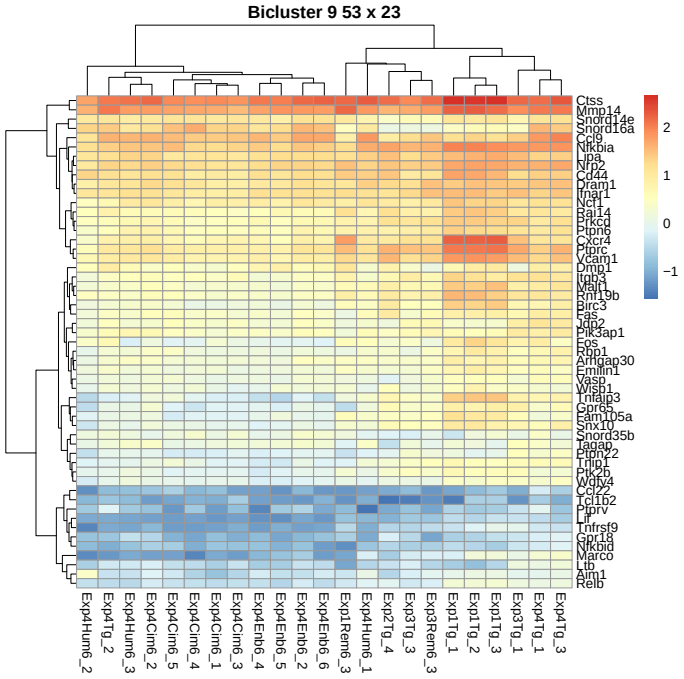
FIGURE 3.27: ISA Bicluster 8

9. Down-regulation in disease. Restored by all treatments but more sufficiently by Remicade. Functions: toll like receptor 5 and 3 signaling pathways, tumor necrosis factoractivated receptor activity, vascular endothelial growth factoractivated receptor. This module has a main sample representation of the Tg, Cimzia and Enbrel classes and mild representation of the Humira class. The genes that participate are related to toll-like receptor signaling pathways which have a central role in RA pathogenesis. (Goh and Midwood, 2011, Thwaites et al., 2014)



adj. P−value	Term
2.31e−10	toll−like receptor 5 signaling pathway
1.26e−09	regulation of granuloma formation
1.86e−09	tumor necrosis factor−activated receptor activity
3.72e−09	positive regulation of deoxyribonuclease activity
6.79e−09	fibrillar center
5.09e−08	insulin−like growth factor binding
6.82e−08	toll−like receptor 3 signaling pathway
1.01e−07	response to molecule of bacterial origin
3.19e−07	vascular endothelial growth factor−activated receptor activity
3.29e−07	nucleotide−binding domain, leucine rich repeat containing receptor signaling pathway

Class	R.score
Wt	0
Tg	0.4
Rem	0.1
Hum	0.1
Enb	0.2
Cim	0.2



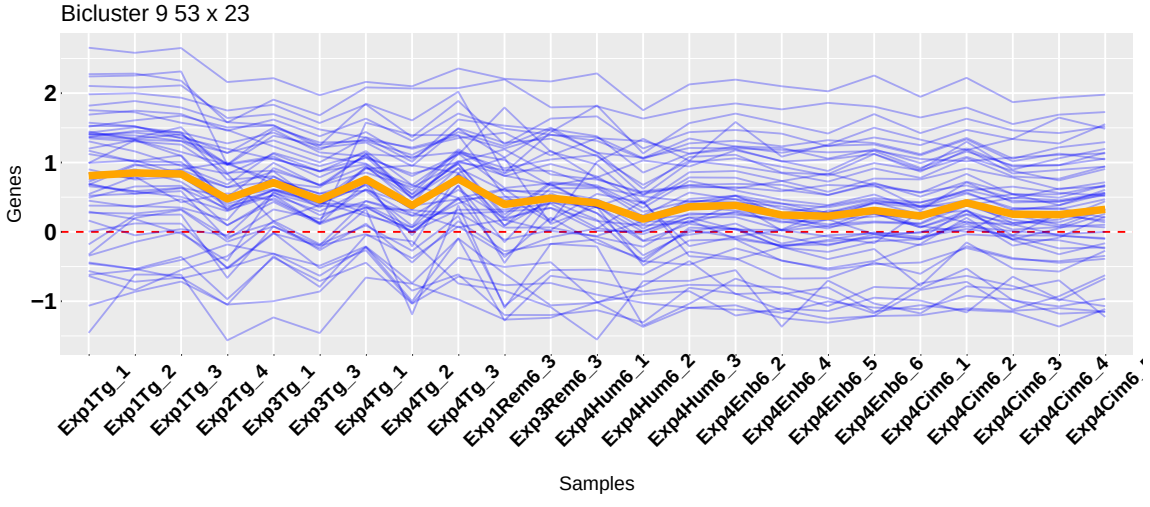


FIGURE 3.28: ISA Bicluster 9

10. Up-regulation in disease but many outliers in all comparisons. Best overcompensation by Remicade.

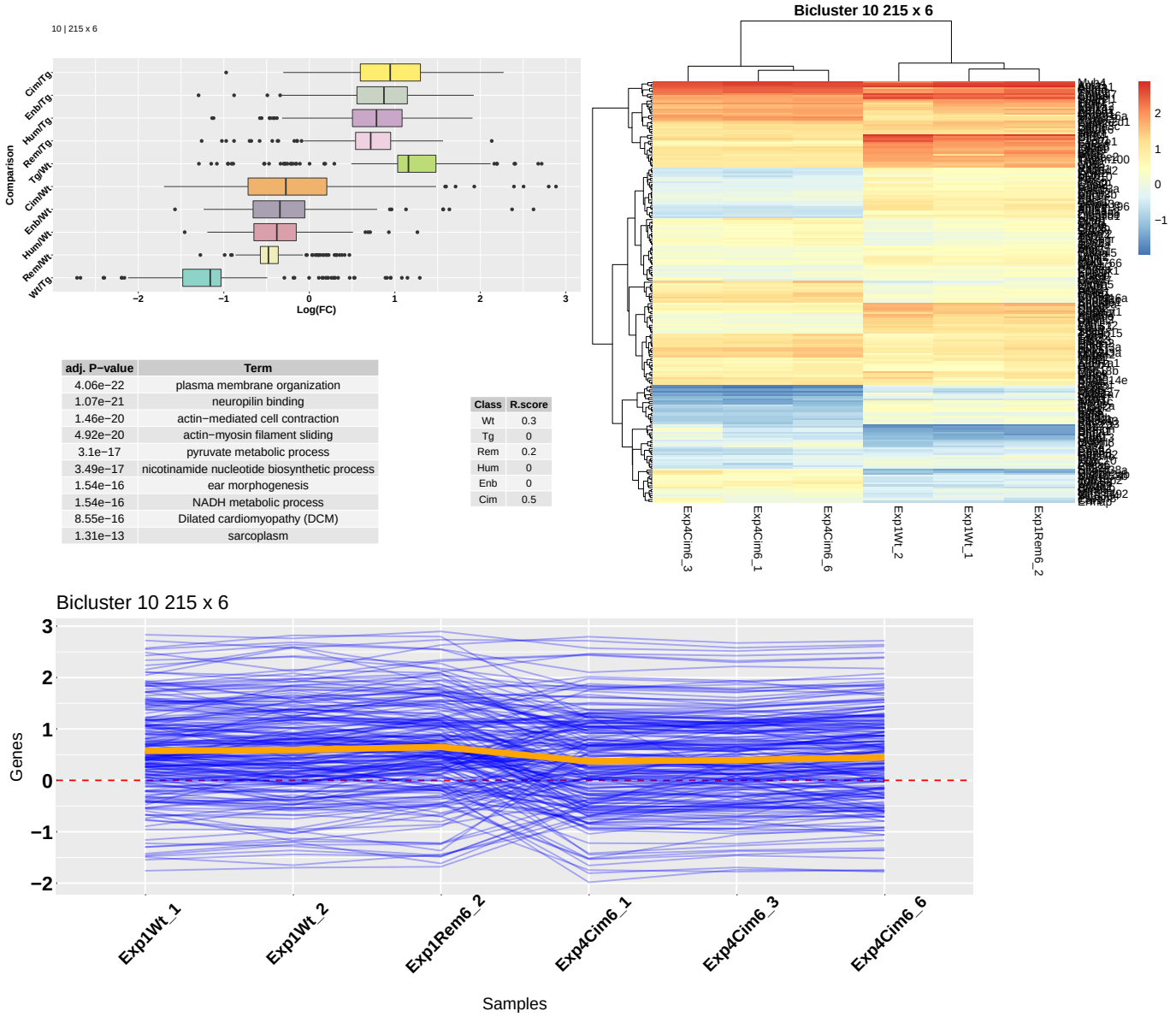
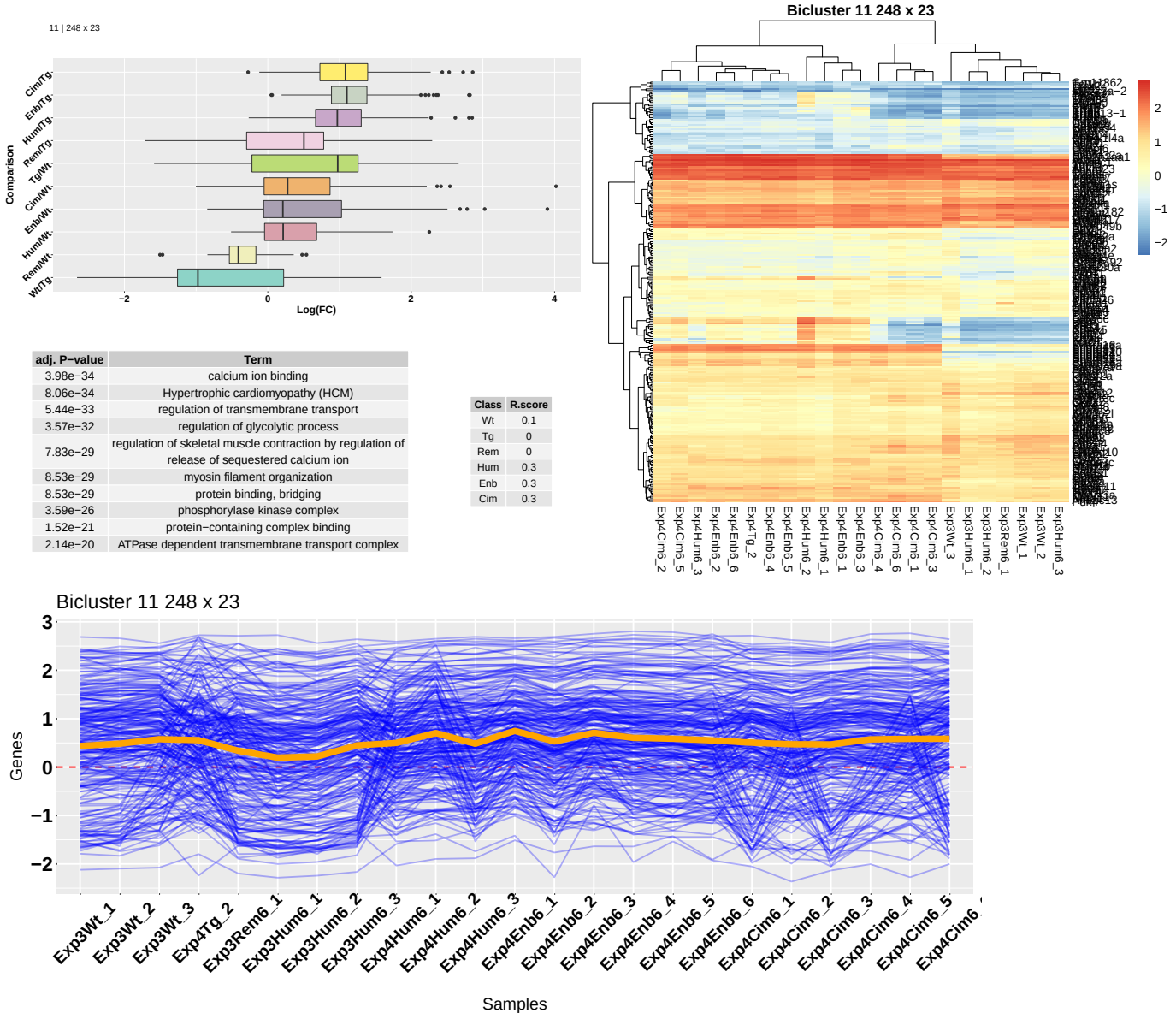


FIGURE 3.29: ISA Bicluster 10

11. Mild down-regulation in disease.



adj. P−value	Term
3.98e−34	calcium ion binding
8.06e−34	Hypertrophic cardiomyopathy (HCM)
5.44e−33	regulation of transmembrane transport
3.57e−32	regulation of glycolytic process
7.83e−29	regulation of skeletal muscle contraction by regulation of release of sequestered calcium ion
8.53e−29	myosin filament organization
8.53e−29	protein binding, bridging
3.59e−26	phosphorylase kinase complex
1.52e−21	protein−containing complex binding
2.14e−20	ATPase dependent transmembrane transport complex

Class	R.score
Wt	0.1
Tg	0
Rem	0
Hum	0.3
Enb	0.3
Cim	0.3

FIGURE 3.30: ISA Bicluster 11

3.2 GA results

3.2.1 General characteristics of GA application

Course of fitness throughout an execution

We executed a single run of the GA separately for each drug i.e. in each execution we used 10 samples of Wt,10 samples of treatment and 13 samples of Tg as described in Chapter 2. Our intention was to observe the course of the highest fitness values across all generations for

all types of selection with or without the existence of elitism, which introduces a high selective power for individuals that appear to be of high fitness. All other parameters where stable during the current and all further parts of the analysis. The plots appear in Appendix A.

We observe a common pattern, that exists for all executions and its main characteristics are:

- Roulette wheel and linear rank selection without elitism lead to a very unstable course of highest fitness values. We observe the presence of intense fluctuations that may or may not lead to a gradual increase of the value. In addition, the final gene sets acquired after 3000 generations is stable and equal to the starting number of gene sets (i.e. almost half of the DE genes used as choice of genes is random)
- Roulette wheel and rank selection with the existence of elitism leads to a gradual increase of the highest fitness value while the mean of each generation follows modest fluctuations. In addition, there is a gradual removal of genes that participate in chromosomes as we observe that the final gene sets are of very small size. The existence of elitism seems to lead in a form of feature selection in our dataset but it is still unknown if these gene sets are optimal solutions that serve the purpose of the fitness solutions or sub-optimal ones.
- Tournament selection appears to have the same effect on the course of the fitness value and the gene set size with or without the existence of elitism.

Course of final gene set size in multiple generation sizes

In order to observe the course of the final gene set size in each case we executed 5 independent runs of the GA each one with a different number of generations: 500, 1000, 1500, 2000, 2500. We observe the expected: in all executions starting gene sets are more or less around the half of the given gene set for each treatment, and in cases where elitism is equal to 1 (solid lines) there is gradual decrease of the final gene sets size whereas in the absence of elitism the size is more or less stable. In tournament selection on the other hand there is a decrease of the gene sets size independent of elitism. This is most probably a result of the process of selection: tournament implements an elitistic-like approach in random subgroup of individuals as described in Chapter 2.

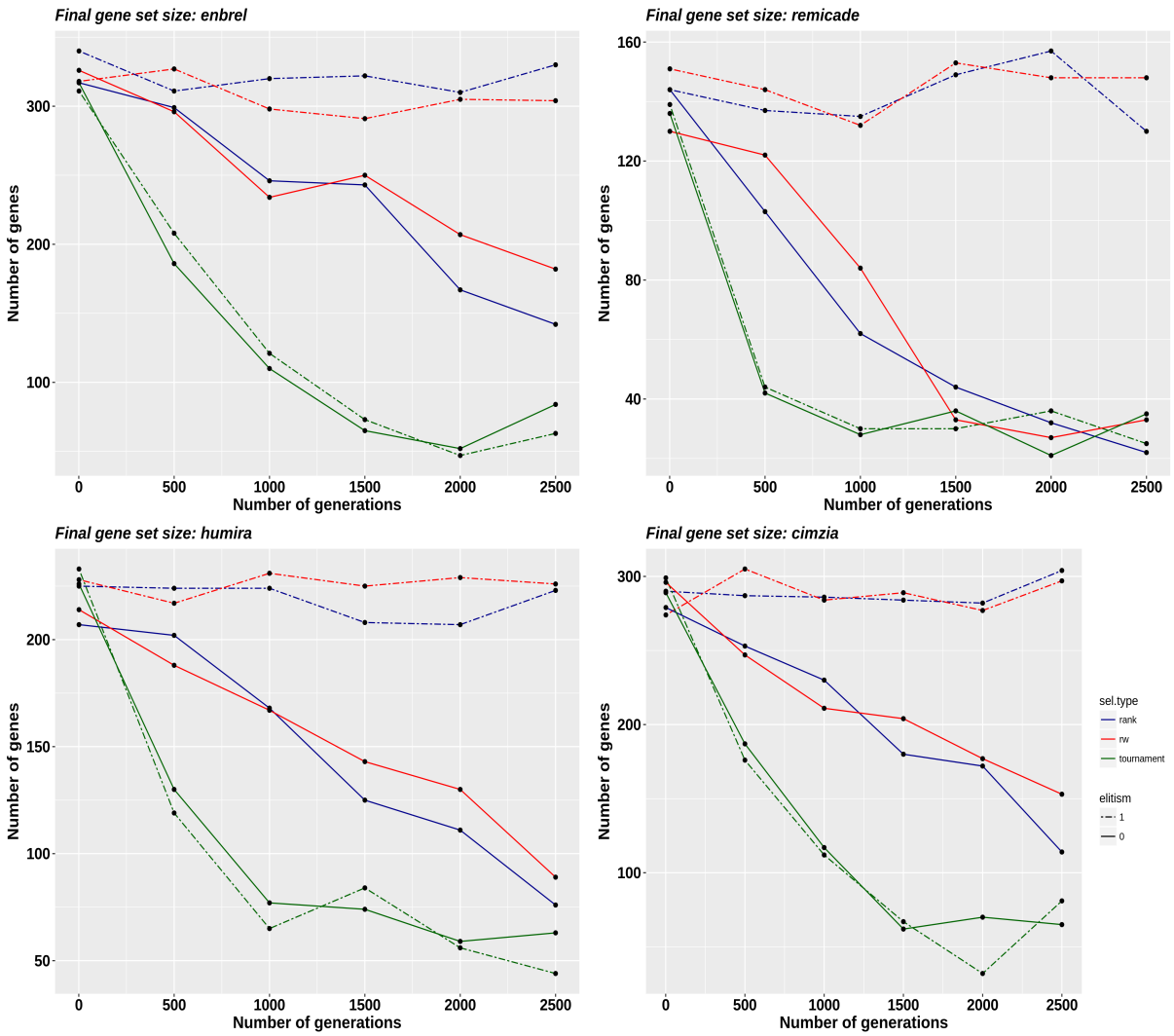


FIGURE 3.31: Course of final gene sets size across 5 independent GA runs with different numbers of generations

Repeated runs of GA to find most frequently occurring genes

For this third part of the analysis, our desire was to look further into the tendency of selection with elitism to lead in decreased gene sets size. At first, we wanted to observe if the results of an execution of 3000 generations preserved the characteristic of reproducibility i.e. if there are genes inside the final set that were observed in most or all the independent runs. We did not expect an overall consistency as the non-deterministic nature of the algorithm would introduce a randomness in the appearance of genes. For this reason, we executed 1000 independent runs and kept for each run the final acquired dataset, merged all the sets and counted the appearance of each gene.

Plots in Fig. 3.7 depict the distribution of gene appearances in the the merged result of the final gene sets. We observe in all four cases a similar tendency: most final sets contain genes that

appear only once and only very few genes appear in all 1000 of the runs or in big majority of the runs. These results show that the independent runs do not tend to bring consistent results i.e. similar gene sets. Similarly, in plots of final fitness value the are perturbations that agree with the lack of reproducibility in plots 3.7. We do not observe an approximately stable final fitness value as we would if the final gene sets would be similar. However, we observe that a high fitness value was returned multiple times during the runs. We hypothesize that the algorithm tends to converge to multiple and similar in terms of fitness sub-optimal solutions.

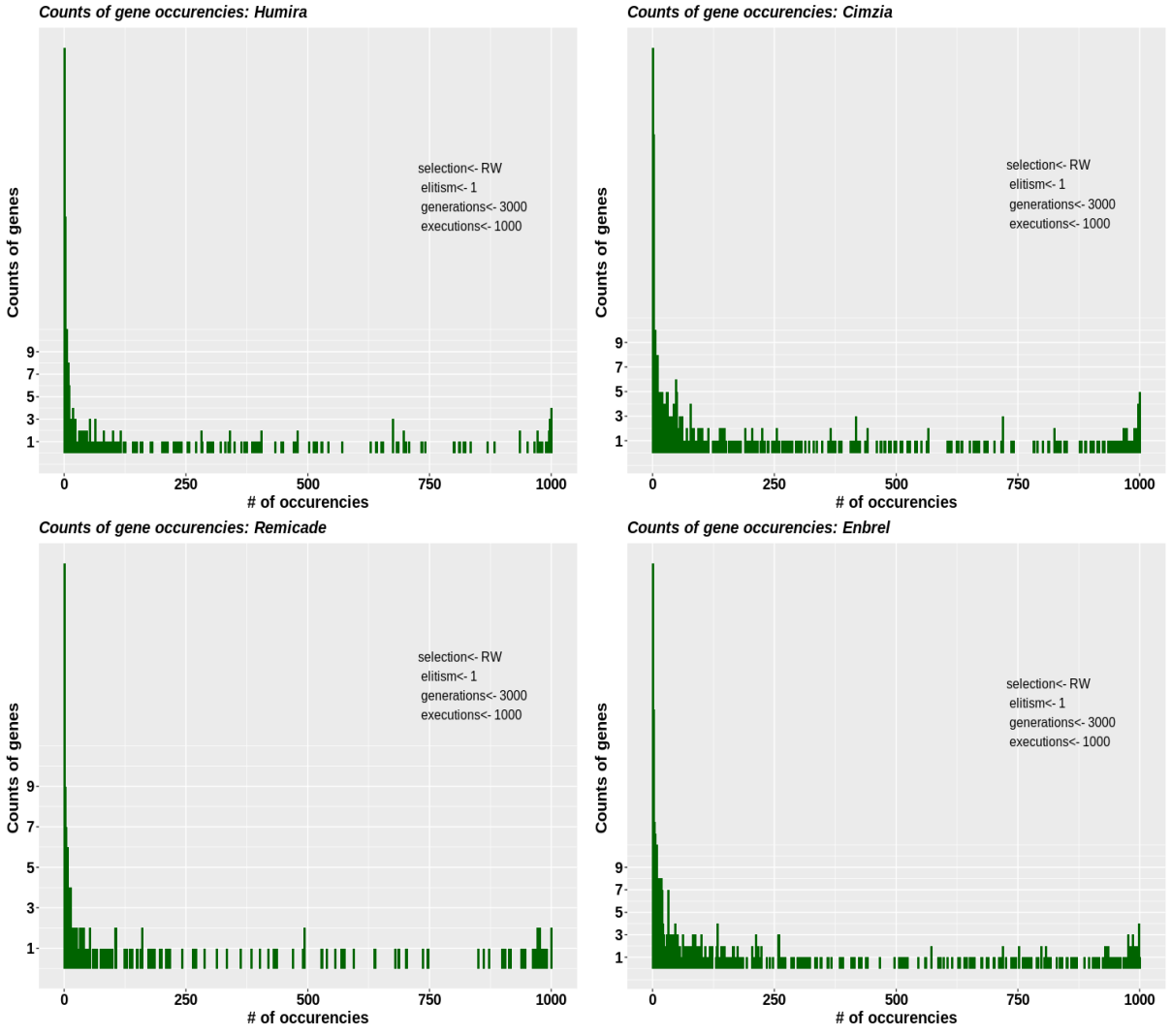


FIGURE 3.32: Distribution of gene appearances in final gene sets given over 1000 executions

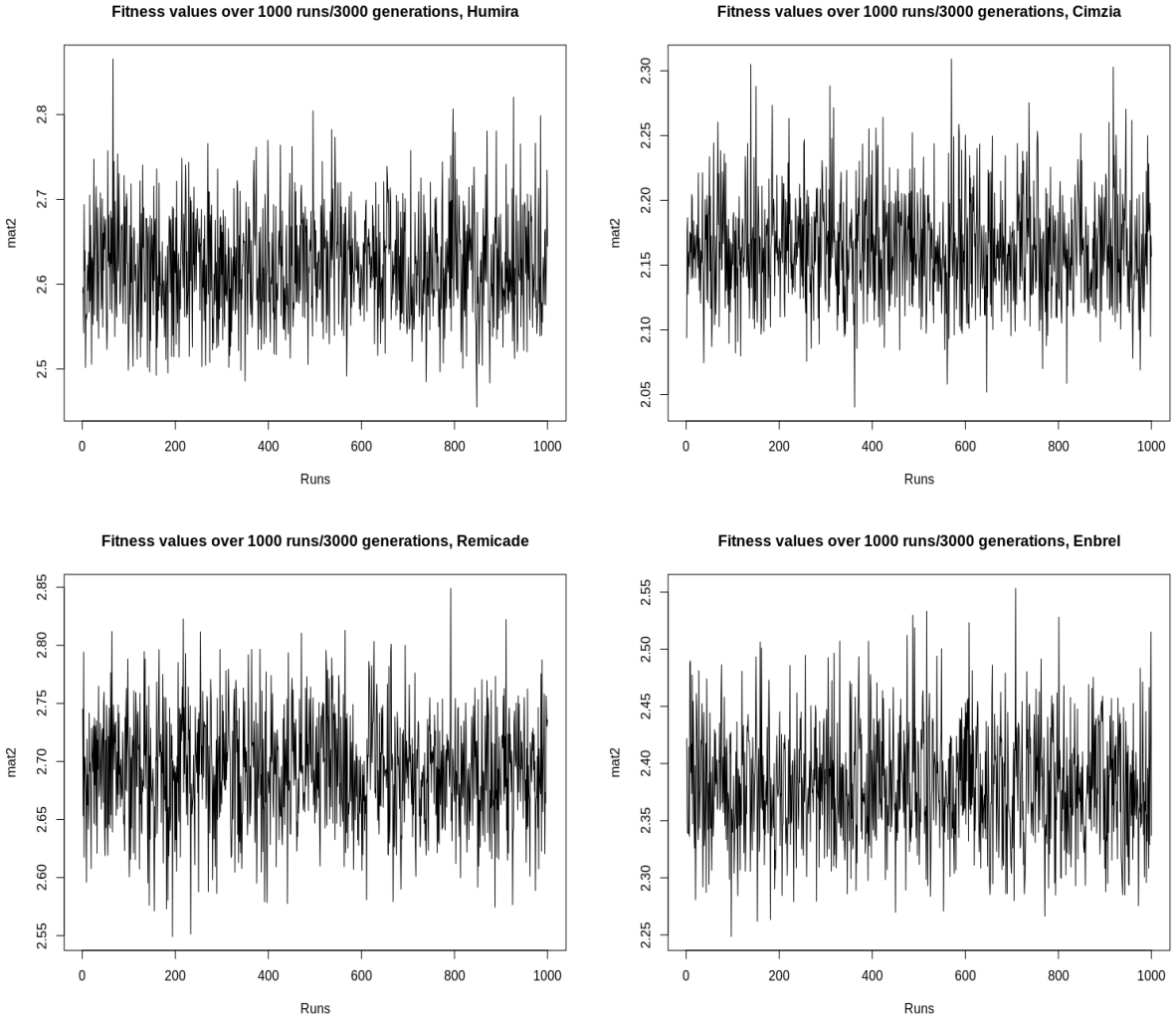


FIGURE 3.33: Distribution of fitness values of final gene sets given over 1000 executions

Chapter 4

Conclusions

4.1 Bicluster analysis - General observations

(i) They can reveal the presence of batch effects between samples of the same class. (Fig. 4.1) For example in Plaid bicluster No. 3 we observe that a batch effect may exist. In this particular module, there are 6 samples of the Tg class that originate from two experimental batches and are grouped hierarchically in remote groups.

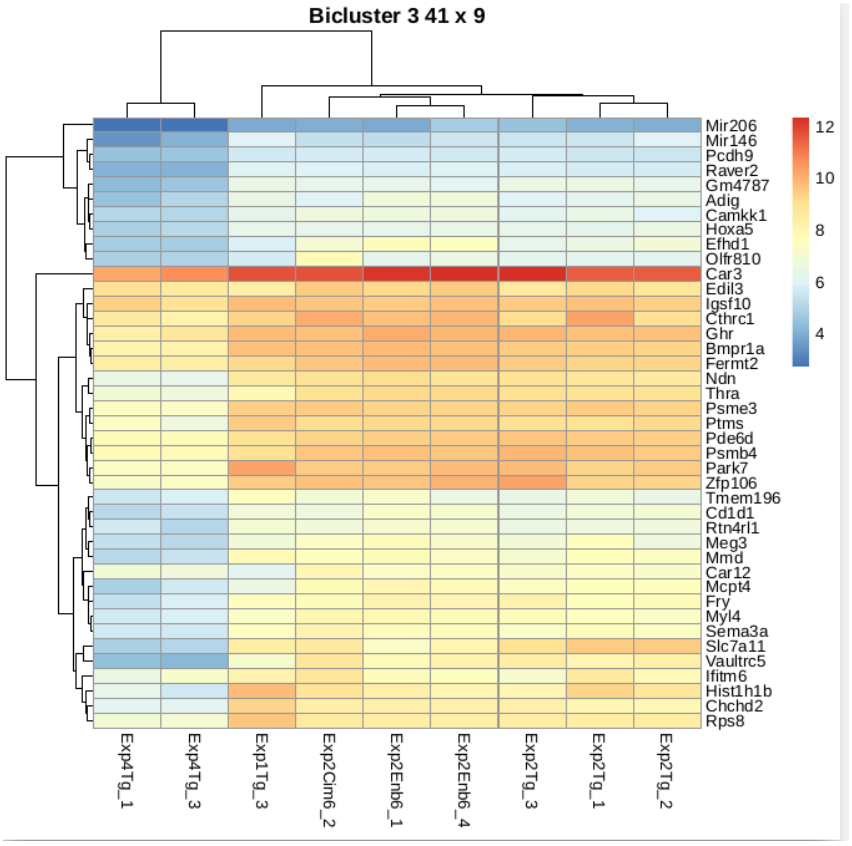


FIGURE 4.1: Example of a bicluster that mostly represents a batch effect in the datasets rather than a important pattern

(ii) Some biclustering algorithms can locate good coherence between the samples. ISA compared to PLAID has returned modules with an overall better sample representation, as shown in tables (A) and (B) in Fig. 3.2. However, besides the better coherence, the patterns revealed by ISA are pretty smooth, meaning that the differences are mostly observed among rows (genes) rather than columns (samples). An example is shown in Fig. 4.2

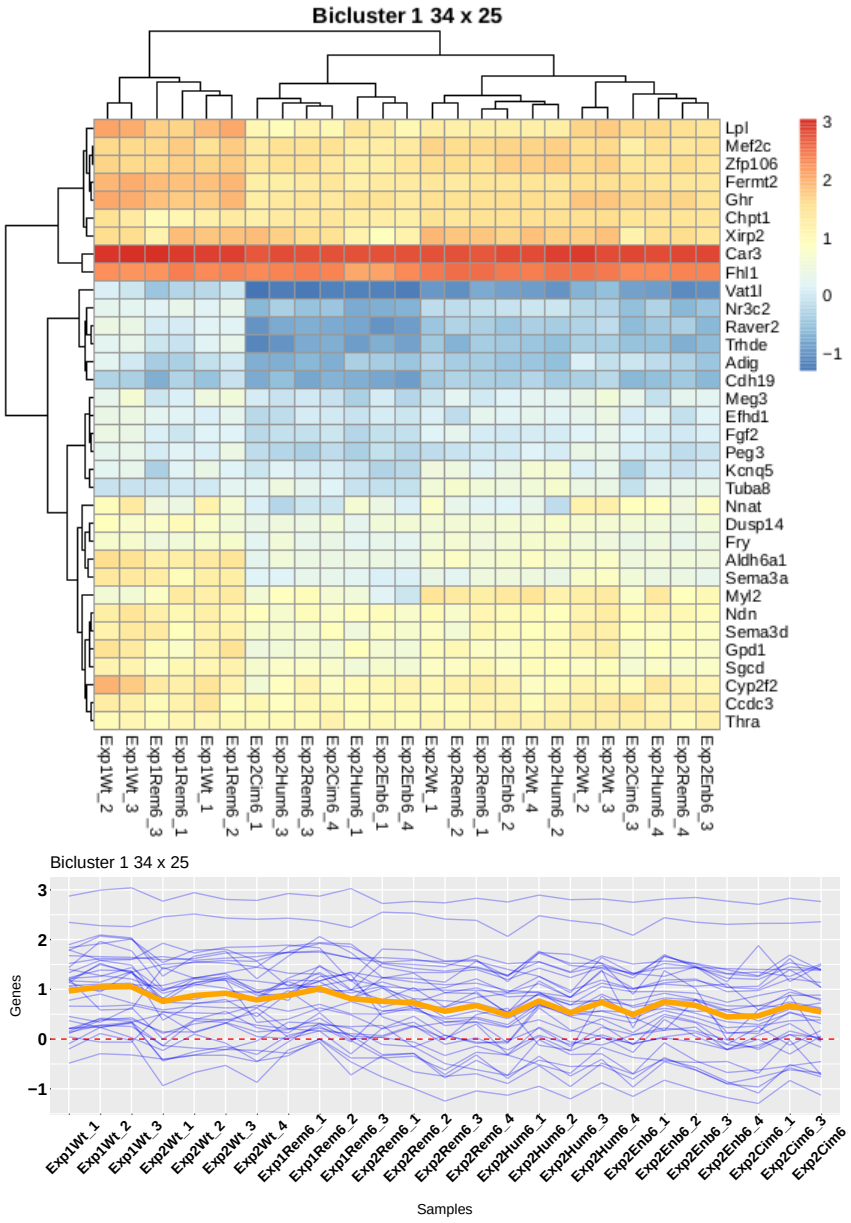


FIGURE 4.2: Example of a ISA bicluster with adequate sample representation but mild pattern

This observation leads us to another conclusion concerning the patterns located by the current algorithms. A pattern may not be clear in terms of differences. For example, in Plaid Bic#4 we observe two main hierarchical groups: one with Tg samples and one with 3 out of the 4 anti-TNFα factors. However, the two groups are brought together because of their differences.

(iii) Depending on further processing that we apply on the biclusters that an algorithm returns, we can influence the sample and gene representation thus directly influenvcing the results in an arbitrary way. In our case a strict analysis that was applied as the intersection of grouped biclusters leads in many samples being left out from the final bicluster even if their appearance was often in the group. Potentially, this could be resolved by the use of other measures such as choosing samples and genes that appear with a frequency higher than a threshold. (Kasim

et al., 2016) The right choice of filters as well as the right choice of b/c algorithm is an matter of investigation that is really dataset specific and requires multiple attempt in order to distinguish the right one within a multitude of filtering methods and algorithms.

(iv) We can gain knowledge about secondary actions of the compounds that we suspected they possess but haven't yet established their mechanism: We observed that many muscle-related functions emerged from the functional analysis of the biclusters. Reduced muscle power and muscle loss are often found as a response to chronic inflammation in multiple rheumatic conditions. It is believed, though is not clear yet, that TNFα is linked with causing muscle loss by activation of the destruction of muscle protein and impedance of myogenic differentiation. (Demirkapi et al., 2017) Anti-TNFα treatments have been linked with improvement of muscle function but also with induced myositis in some cases of side effects. (Zengin et al., 2017) This analysis could add to our knowledge about the way different treatments affect different functions involved in normal or defected muscle activity by linking the compounds with different groups of genes.

4.2 Can biologists draw dependable conclusions about gene expression patterns from b/c?

- Type of dataset and choice of algorithm

One downside of biclustering is that depending on the mathematical used by the algorithm to identify biclusters, different modules might arise. In our work, ISA, which is a popular algorithm and was chosen as the only method of biclustering in many studies, in our case returned biclusters with very mild differences between the samples and smooth patterns. Therefore, the researcher might need either to test a number of algorithms or follow guidelines according to which a biclustering analysis appropriate for the data could be designed. Until now, there are no well-defined guidelines to choose the right algorithm according to the type of data in possession. However, the 'biclust' and 'superbiclust' R packages also used in the current work are able to implement a number of algorithms and validation measures. It should be noted though that such an extended analysis would be time consuming and would require great computational resources. In our case, observation of the results leads to the conclusion that both methods did not return significant observations, patterns or functions. This could be either due to wrong choice of algorithm or due to incorrect further analysis of the initially acquired modules. However, both these scenarios would require further trials and efforts to investigate.

- A very well-defined biological question: what do you expect to find from the analysis.

The milestone of a biclustering analysis, is to pose the right questions and build the right dataset in order to answer them. The work of Xie et al. (2018) may be a great example on

posing and answering a biological questions with biclustering. In our case, we expressed the need to "mine" differences and similarities between treatments and conditions with the purpose of revealing transcriptional modules described by correlated subsets of genes in subsets of samples. However, the analysis with both algorithms did not succeed to highlight important patterns or functions in the found modules.

4.3 Discussion on GA application

Since a genetic algorithm approach is based on the fitness values produced by a function that entails great flexibility of choice, the main issue when building an approach is the design and implementation of the function. In our case, the choice of the Dunn index was chosen after a series of implementations of simple distance ratios among the samples and the healthy and diseased samples that represented the centroids. These approaches seemed appropriate in an intuitive way but the results showed that the optimization of these ratios to was far from feasible. In consequence, we arrived at the solution of the Dunn index as a metric popular in literature used for the evaluation of the quality of a clustering solution. Also, the preliminary analysis showed that there was a possibility for optimization (*Appendix B. Course of fitness during an execution with all selection processes*)

However, further analysis showed that the initial results are not reproducible as different gene sets would arise from independent executions of GA with the same parameters. This observation lead to the hypothesis that multiple sub-optimal solutions exist from combinations of different genes. This could either be a result of the fitness function choice or of an unsuitability of the current dataset for feature selection process like the one proposed.

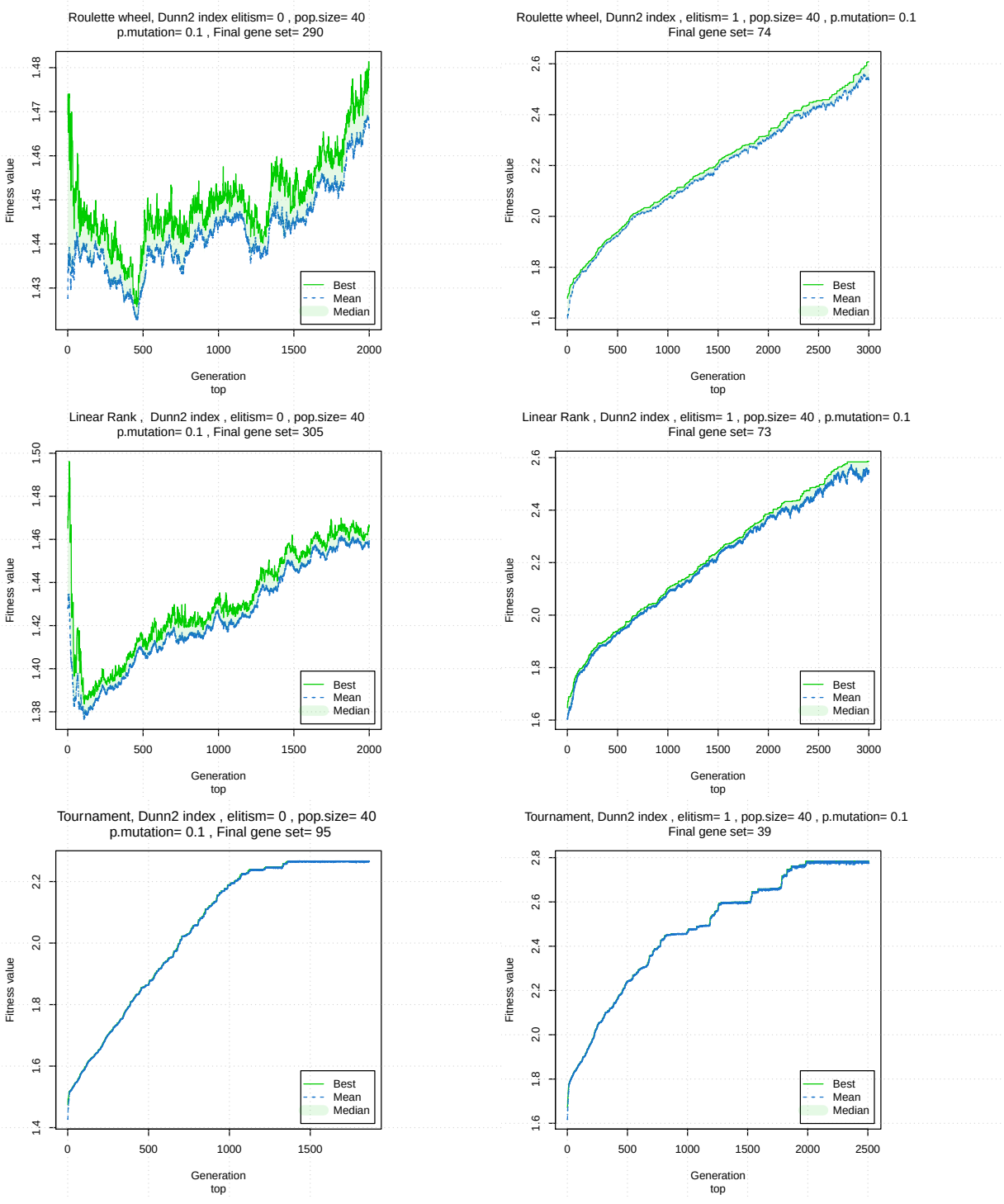
In conclusion, the idea of extracting informative feature with the use of genetic algorithms seems very appealing because of the flexibility and freedom that the fitness function can provide. Using or designing an appropriate fitness function can lead to the acquisition of very specialized results for a certain condition. Thus, the most significant characteristic is posing the right question about what you expect from the algorithm and using the right data that can lead you to them. The idea proposed in this part of the work seems appealing and it would be interesting to further test it on larger datasets or other types of datasets that are characterized from wider differences between the features (in our cases, genes).

Appendix A

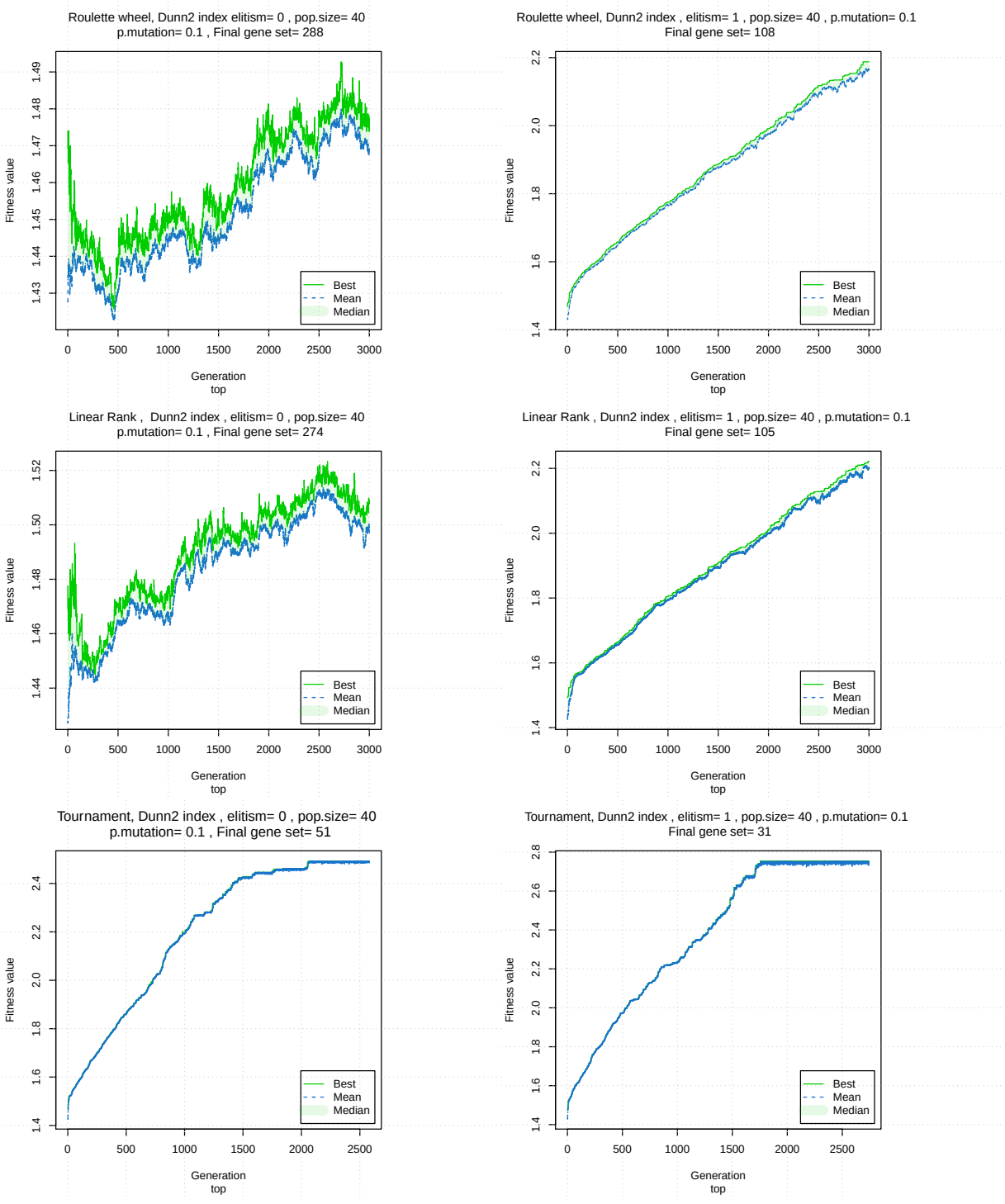
Genetic algorithms

A.1 Course of fitness during an execution with all selection modes

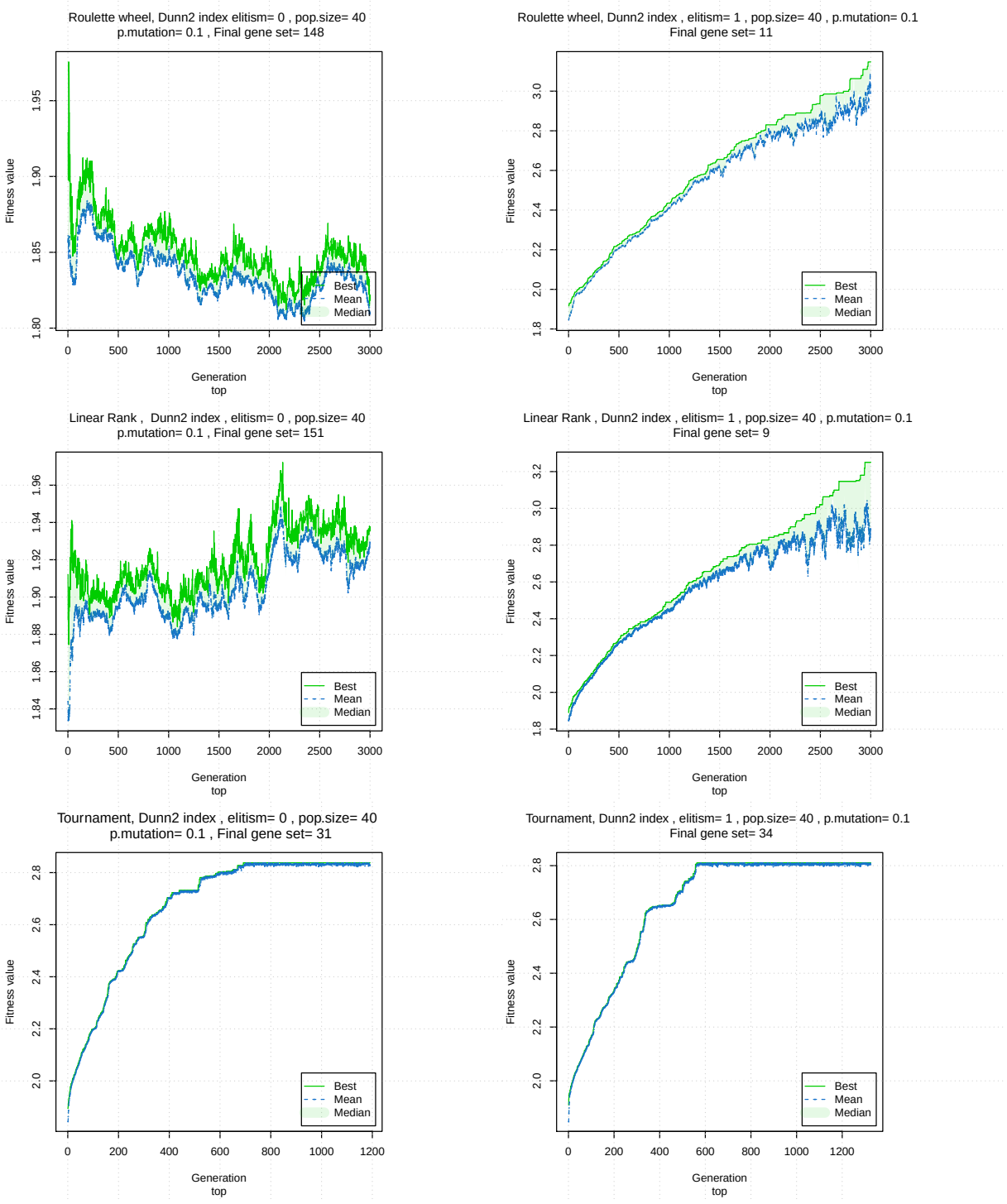
A.1.1 Humira


Roulette wheel, Dunn2 index elitism= 0 , pop.size= 40
p.mutation= 0.1 , Final gene set= 290
Roulette wheel, Dunn2 index , elitism= 1 , pop.size= 40 , p.mutation= 0.1
Final gene set= 74
Linear Rank , Dunn2 index , elitism= 0 , pop.size= 40
p.mutation= 0.1 , Final gene set= 305
Linear Rank , Dunn2 index , elitism= 1 , pop.size= 40 , p.mutation= 0.1
Final gene set= 73
Tournament, Dunn2 index , elitism= 0 , pop.size= 40
p.mutation= 0.1 , Final gene set= 95
Tournament, Dunn2 index , elitism= 1 , pop.size= 40 , p.mutation= 0.1
Final gene set= 39
Fitness value
Generation
top
Best
Mean
Median

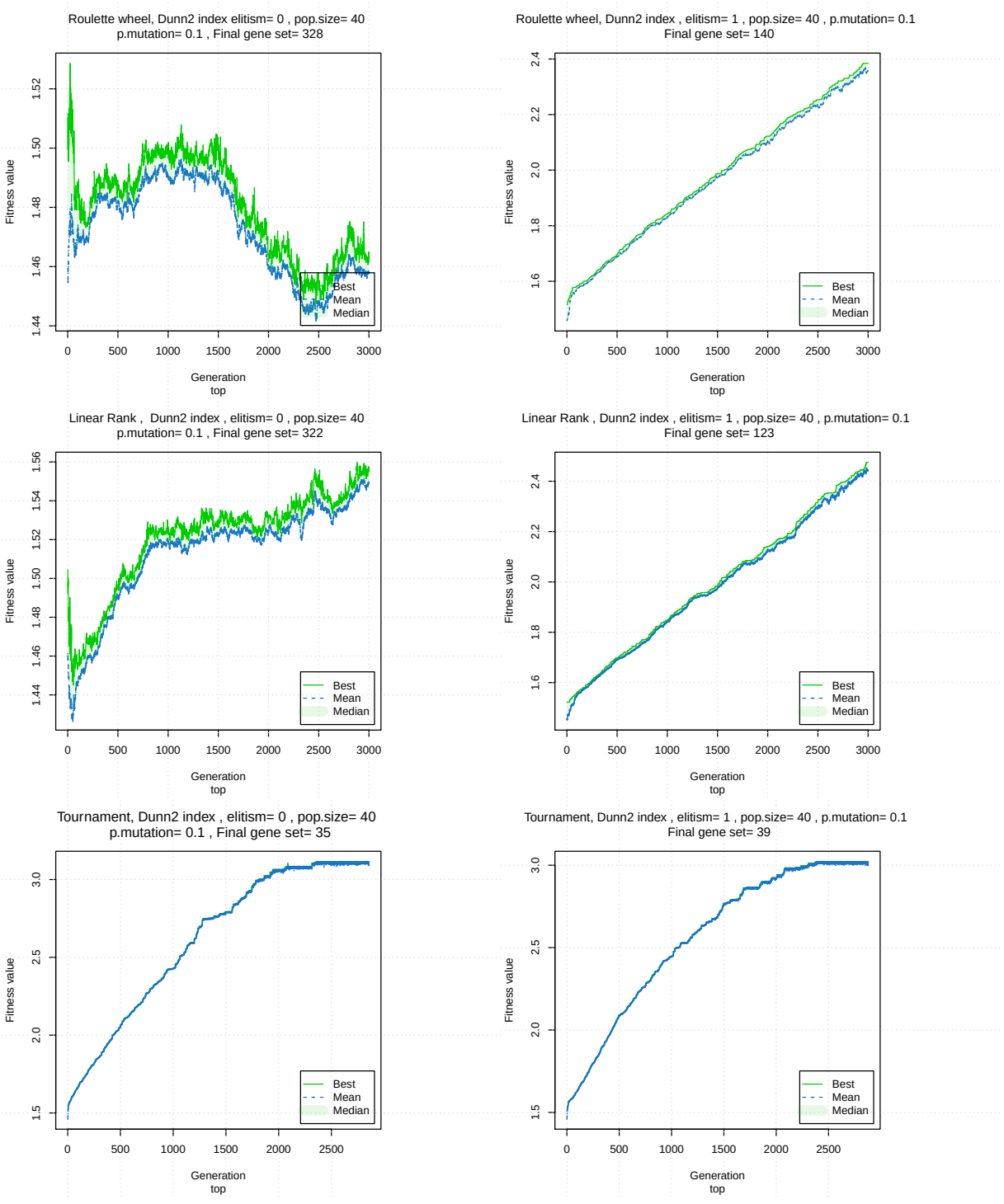
A.1.2 Cimzia


Roulette wheel, Dunn2 index elitism= 0 , pop.size= 40
p.mutation= 0.1 , Final gene set= 288
Roulette wheel, Dunn2 index , elitism= 1 , pop.size= 40 , p.mutation= 0.1
Final gene set= 108
Linear Rank , Dunn2 index , elitism= 0 , pop.size= 40
p.mutation= 0.1 , Final gene set= 274
Linear Rank , Dunn2 index , elitism= 1 , pop.size= 40 , p.mutation= 0.1
Final gene set= 105
Tournament, Dunn2 index , elitism= 0 , pop.size= 40
p.mutation= 0.1 , Final gene set= 51
Tournament, Dunn2 index , elitism= 1 , pop.size= 40 , p.mutation= 0.1
Final gene set= 31
Fitness value
Generation
top
Best
Mean
Median

A.1.3 Remicade



A.1.4 Enbrel


Roulette wheel, Dunn2 index elitism= 0 , pop.size= 40
p.mutation= 0.1 , Final gene set= 328
Roulette wheel, Dunn2 index , elitism= 1 , pop.size= 40 , p.mutation= 0.1
Final gene set= 140
Linear Rank , Dunn2 index , elitism= 0 , pop.size= 40
p.mutation= 0.1 , Final gene set= 322
Linear Rank , Dunn2 index , elitism= 1 , pop.size= 40 , p.mutation= 0.1
Final gene set= 123
Tournament, Dunn2 index , elitism= 0 , pop.size= 40
p.mutation= 0.1 , Final gene set= 35
Tournament, Dunn2 index , elitism= 1 , pop.size= 40 , p.mutation= 0.1
Final gene set= 39
Fitness value
Generation
top
Best
Mean
Median

Bibliography

Asquith, D. L., Miller, A. M., McInnes, I. B., and Liew, F. Y. (2009). Animal models of rheumatoid arthritis. *European Journal of Immunology*, 39(8):2040–2044.

Bergmann, S., Ihmels, J., and Barkai, N. (2003). Iterative signature algorithm for the analysis of large-scale gene expression data. *Physical Review E*, 67(3):031902.

Bryan, K., Terrile, M., Bray, I. M., Domingo-Fernandéz, R., Watters, K. M., Koster, J., Versteeg, R., Stallings, R. L., Domingo-Fernandé Z, R., Watters, K. M., Koster, J., Versteeg, R., and Stallings, R. L. (2014). No Title. *Nucleic acids research*, 42(3):e17.

Caplazi, P., Baca, M., Barck, K., Carano, R. A. D., Devoss, J., Lee, W. P., Bolon, B., and Diehl, L. (2015). Mouse Models of Rheumatoid Arthritis. *Veterinary Pathology*, 52(5):819–826.

Csárdi, G., Kutalik, Z., and Bergmann, S. (2010). Modular analysis of gene expression data with R. *Bioinformatics*, 26(10):1376–1377.

Demirkapi, M., Yildizgören, M. T., Gület, H., and Turhanoglu, A. D. (2017). The effect of anti-tumor necrosis factor-alpha treatment on muscle performance and endurance in patients with ankylosing spondylitis: A prospective follow-up study. *Arch Rheumatol*, 32(4):309–314. 29901011[pmid].

Eren, K., Deveci, M., Küçüktunç, O., and Çatalyürek, V. (2013). A ccomparative analysis of biclustering algorithms for gene expression dat. *Briefings in Bioinformatics*, 14(3):279–292.

Firestein, G. S. and McInnes, I. B. (2017). Immunopathogenesis of Rheumatoid Arthritis. *Immunity*, 46(2):183–196.

Furst, D. E., Schiff, M. H., Fleischmann, R. M., Strand, V., Birbara, C. A., Compagnone, D., Fischkoff, S. A., and Chartash, E. K. (2003). Adalimumab, a fully human anti tumor necrosis factor-alpha monoclonal antibody, and concomitant standard antirheumatic therapy for the treatment of rheumatoid arthritis: results of star (safety trial of adalimumab in rheumatoid arthritis). *The Journal of Rheumatology*, 30(12):2563–2571.

Goel, N. and Stephens, S. (2010). Certolizumab pegol. *MAbs*, 2(2):137–147.

Goh, F. G. and Midwood, K. S. (2011). Intrinsic danger: activation of toll-like receptors in rheumatoid arthritis. *Rheumatology*, 51(1):7–23.

Hartigan, J. A. (1972). Direct Clustering of a Data Matrix. *Journal of the American Statistical Association*, 67(337):123–129.

Hayden, M. S. and Ghosh, S. (2014). Regulation of nf-kb by tnf family cytokines. *Semin Immunol*, 26(3):253–266. 24958609[pmid].

Hochreiter, S., Bodenhofer, U., Heusel, M., Mayr, A., Mitterecker, A., Kasim, A., Khamiakova, T., Van Sanden, S., Lin, D., Talloen, W., Bijnens, L., Göhlmann, H. W. H., Shkedy, Z., and Clevert, D.-A. (2010). Fabia: factor analysis for bicluster acquisition. *Bioinformatics*, 26(12):1520–1527. btq227[PII].

Holland, J. (1975). *Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence*. University of Michigan Press.

Holmdahl, R., Malmström, V., and Burkhardt, H. (2014). Autoimmune priming, tissue attack and chronic inflammation - The three stages of rheumatoid arthritis. *European Journal of Immunology*, 44(6):1593–1599.

Kaiser, S. and Leisch, F. (2008). A Toolbox for Bicluster Analysis in R. *Compstat*.

Kannan, K., Ortmann, R. A., and Kimpel, D. (2005). Animal models of rheumatoid arthritis and their relevance to human disease. *Pathophysiology : the official journal of the International Society for Pathophysiology*, 12(3):167–81.

Kasim, A., Shkedy, Z., Kaiser, S., Hochreiter, S., and Talloen, W. (2016). *Applied Biclustering Methods for Big and High-Dimensional Data Using R*.

Kim, T.-H., Choi, S. J., Lee, Y. H., Song, G. G., and Ji, J. D. (2014). Gene expression profile predicting the response to anti-tnf treatment in patients with rheumatoid arthritis; analysis of geo datasets. *Joint Bone Spine*, 81(4):325 – 330.

Kluger, Y., Basri, R., Chang, J. T., and Gerstein, M. (2003). Spectral biclustering of microarray data: Coclustering genes and conditions. *Genome Res*, 13(4):703–716. 19x[PII].

Koczan, D., Drynda, S., Hecker, M., Drynda, A., Guthke, R., Kekow, J., and Thiesen, H.-J. (2008). Molecular discrimination of responders and nonresponders to anti-tnfalpha therapy in rheumatoid arthritis by etanercept. *Arthritis Res Ther*, 10(3):R50–R50. ar2419[PII].

Kollias, G., Papadaki, P., Apparailly, F., Vervoordeldonk, M. J., Holmdahl, R., Baumans, V., Desaintes, C., Di Santo, J., Distler, J., Garside, P., Hegen, M., Huizinga, T. W. J., Jüngel, A., Klareskog, L., McInnes, I., Ragoussis, I., Schett, G., Hart, B. '., Tak, P. P., Toes, R., van den Berg, W., Wurst, W., and Gay, S. (2011). Animal models for arthritis: innovative tools for prevention and treatment. *Annals of the Rheumatic Diseases*, 70(8):1357–1362.

Lazzeroni, L. and Owen, A. (2002). PLAID MODELS FOR GENE EXPRESSION DATA. *Statistica Sinica*, 12:61–86.

Li, G., Ma, Q., Tang, H., Paterson, A. H., and Xu, Y. (2009). QUBIC: a qualitative biclustering algorithm for analyses of gene expression data. *Nucleic Acids Research*, 37(15):e101–e101.

Lindberg, J., af Klint, E., Catrina, A. I., Nilsson, P., Klareskog, L., Ulfgren, A.-K., and Lundeberg, J. (2006). Effect of infliximab on mrna expression profiles in synovial tissue of rheumatoid arthritis patients. *Arthritis Res Ther*, 8(6):R179–R179. ar2090[PII].

Lindberg, J., Wijbrandts, C. A., van Baarsen, L. G., Nader, G., Klareskog, L., Catrina, A., Thurlings, R., Vervoordeldonk, M., Lundeberg, J., and Tak, P. P. (2010). The gene expression profile in the synovium as a predictor of the clinical response to infliximab treatment in rheumatoid arthritis. *PLoS One*, 5(6):e11310. 10-PONE-RA-16245R1[PII].

Lindqvist, A.-K. B., Bockermann, R., Johansson, C., Nandakumar, K. S., Johannesson, M., and Holmdahl, R. (2002). Mouse models for rheumatoid arthritis. *Trends in Genetics*, 18(6):S7–S13.

Ma, X. and Xu, S. (2013). Tnf inhibitor therapy for rheumatoid arthritis. *Biomed Rep*, 1(2):177–184. br-01-02-0177[PII].

Madeira, S. and Oliveira, A. (2004). Biclustering algorithms for biological data analysis: a survey. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 1(1):24–45.

Madeira, S. C., Teixeira, M. C., Sa-Correia, I., and Oliveira, A. L. (2010). Identification of regulatory modules in time series gene expression data using a linear time biclustering algorithm. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 7(1):153–165.

McCall, J. (2005). Genetic algorithms for modelling and optimisation. *Journal of Computational and Applied Mathematics*, 184(1):205 – 222. Special Issue on Mathematics Applied to Immunology.

McInnes, I. B. and Schett, G. (2011). The Pathogenesis of Rheumatoid Arthritis. *New England Journal of Medicine*, 365(23):2205–2219.

McInnes, I. B. and Schett, G. (2017). Pathogenetic insights from the treatment of rheumatoid arthritis. *The Lancet*, 389(10086):2328–2337.

Mitchell, M. (1998). *An Introduction to Genetic Algorithms*. MIT Press, Cambridge, MA, USA.

Monach, P. A., Mathis, D., and Benoist, C. The k/bxn arthritis model. *Current Protocols in Immunology*, 81(1):15.22.1–15.22.12.

O'Dell, J. R. (2004). Therapeutic Strategies for Rheumatoid Arthritis. *New England Journal of Medicine*, 350(25):2591–2602.

Padilha, V. A. and Campello, R. J. (2017). A systematic comparative evaluation of biclustering techniques. *BMC Bioinformatics*, 18(1):1–25.

Prelić, A., Bleuler, S., Zimmermann, P., Wille, A., Bühlmann, P., Gruissem, W., Hennig, L., Thiele, L., and Zitzler, E. (2006). A systematic comparison and evaluation of biclustering methods for gene expression data. *Bioinformatics*, 22(9):1122–1129.

Quan, L., Thiele, G., Tian, J., and Wang, D. (2008). The development of novel therapies for rheumatoid arthritis. *Expert Opinion on Therapeutic Patents*, 18(7):723–738.

Roy, S., Bhattacharyya, D. K., and Kalita, J. K. Analysis of Gene Expression Patterns Using Biclustering. *Methods in Molecular Biology*.

Szekanecz, Z., Meskó, B., Poliska, S., Váncsa, A., Szamosi, S., Végh, E., Simkovics, E., Laki, J., Kurkó, J., Besenyei, T., Mikecz, K., Glant, T. T., and Nagy, L. (2013). Pharmacogenetics and pharmacogenomics in rheumatology. *Immunol Res*, 56(0):325–333. 23564183[pmid].

Tanay, A., Sharan, R., and Shamir, R. (2002). Discovering statistically significant biclusters in gene expression data. *Bioinformatics*, 18(suppl$_1$) : $S136 - -S144$.

Thwaites, R., Chamberlain, G., and Sacre, S. (2014). Emerging role of endosomal toll-like receptors in rheumatoid arthritis. *Frontiers in immunology*, 5:1–1. 24474949[pmid].

Tomancak, P., Berman, B. P., Beaton, A., Weiszmann, R., Kwan, E., Hartenstein, V., Celniker, S. E., and Rubin, G. M. (2007). Global analysis of patterns of gene expression during drosophila embryogenesis. *Genome Biology*, 8(7):R145.

Turner, H., Bailey, T., and Krzanowski, W. (2005). Improved biclustering of microarray data demonstrated through systematic performance tests. 48:235–254.

van Baarsen, L. G. M., Wijbrandts, C. A., Gerlag, D. M., Rustenburg, F., van der Pouw Kraan, T. C. T. M., Dijkmans, B. A. C., Tak, P. P., and Verweij, C. L. (2010). Pharmacogenomics of infliximab treatment using peripheral blood cells of patients with rheumatoid arthritis. *Genes And Immunity*, 11:622 EP –. Original Article.

Wegman, E. J. (1990). Hyperdimensional data analysis using parallel coordinates. 85(411):664–675.

Wijbrandts, C. and Tak, P. (2017). Prediction of response to targeted treatment in rheumatoid arthritis. *Mayo Clinic Proceedings*, 92(7):1129 – 1143.

Xie, J., Ma, A., Fennell, A., Ma, Q., and Zhao, J. (2018). It is time to apply biclustering: a comprehensive review of biclustering applications in biological and biomedical data. *Briefings in Bioinformatics*, (March):1–16.

Zampeli, E., Vlachoyiannopoulos, P. G., and Tzioufas, A. G. (2015). Treatment of rheumatoid arthritis: Unraveling the conundrum. *Journal of Autoimmunity*, 65:1–18.

Zeisel, A., Muñoz-Manchado, A. B., Codeluppi, S., Lönnerberg, P., La Manno, G., Juréus, A., Marques, S., Munguba, H., He, L., Betsholtz, C., Rolny, C., Castelo-Branco, G., Hjerling-Leffler, J., and Linnarsson, S. (2015). Cell types in the mouse cortex and hippocampus revealed by single-cell rna-seq. *Science*, 347(6226):1138–1142.

Zengin, O., Onder, M. E., Alkan, S., Kimyon, G., Hüseynova, N., Demir, Z. H., Kısacık, B., and Onat, A. M. (2017). Three cases of anti-tnf induced myositis and literature review. *Revista Brasileira de Reumatologia (English Edition)*, 57(6):590 – 595.