Dissertation

submitted to the

Combined Faculty of Natural Sciences and Mathematics

of the Ruperto Carola University Heidelberg, Germany

for the degree of

Doctor of Natural Sciences

Presented by

M.A. Holly Amelia Rebecca Giles

born in Portsmouth

Oral examination: Dec 15, 2021?

Drug - microenvironment - gene interplay in chronic lymphocytic leukemia

Referees: Dr. Judith Zaugg

Prof. Dr. Michael Boutros

*To ...*

# Acknowledgements

I want to thank a few people.

# List of publications

### Thesis related

- Peter-Martin Bruch∗, **Holly A. R. Giles**∗, Carolin Kolb, Sophie Herbst, Tina Becirovic, Tobias Roider, Junyan Lu, Sebastian Scheinost, Lena Wagner, Jennifer Huellein, Ivan Berest, Mark Kriegsmann, Katharina Kriegsmann, Christiane Zgorzelski, Peter Dreger, Judith Zaugg, Carsten Mueller-Tidow, Thorsten Zenz, Wolfgang Huber, Sascha Dietrich et al. *in preparation.* "Mapping drug-microenvironment-genetic interplay in CLL reveals trisomy 12 as a modulator of microenvironment."***Journal bioRxiv***. doi:

### Other contributions

The author of this thesis also contributed to a number of other projects throughout the PhD. The following have been published:

- Berest, Ivan∗, Christian Arnold∗, Armando Reyes-Palomares, Giovanni Palla, Kasper Dindler Rasmussen, **Holly Giles**, and Peter-Martin Bruch et al. 2019. "Quantification Of Differential Transcription Factor Activity And Multiomics-Based Classification Into Activators And Repressors: DiffTF". *Cell Reports* 29 (10): 3147-3159.e12. doi:10.1016/j.celrep.2019.10.106.
- Lu, Junyan∗, Ester Cannizzaro∗, Fabienne Meier-Abt, Sebastian Scheinost, Peter-Martin Bruch, **Holly A. R. Giles**, and Almut Lütge et al. 2021. "Multi-Omics Reveals Clinically Relevant Proliferative Drive Associated With Mtor-MYC-OXPHOS

Activity In Chronic Lymphocytic Leukemia". ***Nature Cancer***. doi:10.1038/s43018-021-00216-6.

# Zusammenfassung

Abstract in German Aktuelle Entwicklungen im Bereich der “omik”-Technologien tragen wesentlich zurBeschleunigung des Fortschritts in der Krebsforschung bei etc

# Abstract

The tumour microenvironment and genetic alterations collectively influence disease biology and drug resistance in Chronic Lymphocytic Leukaemia (CLL). To establish an integrative understanding of these factors in CLL biology, we performed a combinatorial assay using 12 drugs individually co-applied with each of 17 microenvironmental stimuli on 192 samples of CLL peripheral blood mononucleated cells. We examined microenvironment response across a heterogeneous patient cohort and identified four distinct CLL subgroups that differed in their response landscapes and in patient outcomes. By combining our data with whole-exome sequencing, DNA-methylation, RNA-sequencing and copy number variant data of the same tumours, we systematically searched for molecular determinants of stimulus response and found trisomy 12 as a key modulator. Our data suggest that the amplifying effect of trisomy 12 on the response to environmental signals is mediated by the transcription factors Spi-B and PU.1. We generated a comprehensive map of drug-microenvironment interactions in CLL, and profiled the modulating impact of genetic features on these antagonistic and synergistic effects. Interleukin (IL) 4 and Toll-Like Receptor (TLR) 7/8/9 stimuli showed the most interactions. Both pathways were more active in CLL-infiltrated lymph nodes than in healthy samples ($p<0.001$), and high IL4 activity in lymph nodes correlated with shorter survival ($p=0.038$). We provide a multi-layered resource to investigate microenvironmental and drug interplay in CLL (Repository & Shiny). Our results highlight the importance of cell-extrinsic influences on drug response and disease progression, and how these further depend on molecular features.

# Table of Contents

**Acknowledgements** **i**

**List of publications** **iii**

**Zusammenfassung** **vii**

**Abstract** **ix**

**List of Abbreviations** **xiv**

List of Figures . . . . . xv

List of Tables . . . . . xvii

**1 Introduction** **1**

1.0.1 Disease Characteristcs . . . . . 1

1.0.2 IGHV status . . . . . 1

1.0.3 The role of BCR signalling . . . . . 1

1.0.4 Recurrent genetic features in CLL . . . . . 1

1.0.5 The incompletely understood role of trisomy12 . . . . . 1

1.0.6 Stratification of CLL patients . . . . . 1

1.1 The tumour microenvironment in CLL . . . . . 1

1.1.1 The concept of the protective niche . . . . . 2

1.1.2 Microenvironmeental pathways and cross-talk . . . . . 2

1.2 Therapies in CLL . . . 2
1.2.1 Clinical regimes . . . 2
1.2.2 New Targeted therapies . . . 2
1.2.3 Drug resistances in CLL . . . 2
1.2.4 BET inhibitors, and their potential for the treatment of leukemia . 2
1.3 Background to the approache used in this thesis . . . 2
1.3.1 Ex-vivo drug sensitivity screens . . . 2
1.3.2 Studies of the microenivornment, and attempts to replicate the impact of the microenvironment . . . 3
1.3.3 Measurement of outcomes? . . . 3
1.3.4 Multiomics datasets to study CLL . . . 3
1.3.5 Mathematical modelling . . . 3
1.3.6 CLL as a model for studying tumour biology . . . 3
1.4 Aims of this thesis . . . 3
1.5 Overview of the approach . . . 3

**2 Methods** **5**
2.1 Experimental methods: . . . 5
2.1.1 Drug - Stimulus Combinatorial Pertubation Assay . . . 5
2.1.2 Preparation of screening plates . . . 6
2.2 Drug-Stimulation assay . . . 7
2.3 WES, RNA Sequencing, Targeted Sequencing and DNA Copy number variants . . . 7
2.3.1 Follow - up investigations . . . 7
2.3.2 Investigation of Ibrutinib + IL4 + IBET-762 . . . 9
2.4 Data availability . . . 11
2.5 Statistical Analysis . . . 11
2.5.1 Screening Data . . . 11
2.5.2 Follow up . . . 13

2.5.3 Ibrutinib - IBET-762 - IL-4 analysis . . . . . 17

**3 Data** **19**

3.1 Drug screens and experiments . . . . . 19

3.1.1 High-throughput combinatorial pertubation assay . . . . . 19

3.1.2 Validation experiments . . . . . 19

3.2 Characteristics of drugs used in the screen . . . . . 19

3.2.1 Drug pathways . . . . . 20

3.2.2 Drug responses . . . . . 20

3.2.3 Genetic predictors of drug responses . . . . . 20

3.2.4 Drug - Drug Correlations . . . . . 20

3.3 Characteristics of stimuli used in the screen . . . . . 20

3.3.1 Stimulus responses . . . . . 20

3.3.2 Stimulus - Stimulus Correlations . . . . . 20

3.4 Characteristics of patient samples used in the screen . . . . . 20

3.4.1 Genetic Data available for each patient . . . . . 21

3.5 Processing of raw values obtained from cell viability assay . . . . . 21

3.5.1 Data normalization and quality control . . . . . 21

3.6 Generation of public resource . . . . . 21

3.6.1 Shiny app . . . . . 21

3.6.2 Package . . . . . 21

**4 Ex-vivo sensitivity to microenvironmental stimulation in primary CLL cells** **23**

4.1 Prolifing responses to the panel of stimuli . . . . . 24

4.1.1 *ex vivo* assay demonstrated functional diversity of cytokines and microenvironmental stimuli . . . . . 24

4.1.2 Microenvironmental response profiling identifies discrete patient subgroups . . . . . 26

4.2 Functional characterisation of patient clusters . . . . . 29

4.2.1 C1 - C4 showed distinct response profiles with the panel of stimuli 29
4.2.2 The clusters show differences in disease dynamics . . . . . 30
4.2.3 The clusters showed differential responses to drugs *in vitro* . . . . 31
4.2.4 The clusters are enriched for different genetic features . . . . . 32
4.2.5 GSEA of DE genes between subgroups . . . . . 33
4.3 Additional Analysis . . . . . 34
4.4 Summary . . . . . 34
4.5 Discussion . . . . . 35

**5 Genetic modulators of responses to microenvironmental stimulation 43**
5.1 Systematic analysis of the effect of genetic features on responses to stimuli: 43
5.1.1 Univariate analysis identifies IGHV status and tirsomy 12 as key modulators . . . . . 43
5.1.2 Multivariate analysis demosntrates multiple layers of biology involved . . . . . 44
5.1.3 Individual cytokine – gene interactions of note . . . . . 44
5.2 Profiling the effects of trisomy 12 on cytokine response, and how this may be possible . . . . . 46
5.2.1 Trisomy 12 modulates responses to IL4, TGFbeta and TLR stimuli 46
5.2.2 Gene dosage effects . . . . . 46
5.2.3 Investigation of non-trisomy12 samples that are phenocopies of trisomy12 CLL+ (based on their responses to stimuli) . . . . . 47
5.2.4 Investigation of TF activity in trisomy12+ CLL and finding that SPIB and PU1 TFs are upregulated in trisomy12+ CLL . . . . . 48
5.3 Profiling the downstream effects of SPIB and PU1: . . . . . 49
5.3.1 Over-representation of BCR, TLR, interleukin and TGFbeta response genes amongst SPIB and PU1 targets in ChIPseq data . . . 49
5.3.2 Visualisation of SPIB and PU1 binding sites from ChIPseq data in BCR, TLR, interleukin and TGFbeta signalling genes . . . . . 49

5.3.3 Investigation of DE genes within IL4 and TGFbeta pathways in SPIB/PU1 DKO mice . . . . . 49
5.3.4 SPIb and PU1 linked genes . . . . . 50
5.3.5 SPIB/PU1 DKOs in trisomy 12+ cell lines . . . . . 50
5.4 Conclusions . . . . . 51

**6 Results** **59**
6.1 Review of Results . . . . . 60

**7 Results** **61**
7.1 Review of Results . . . . . 62

**8 Discussion** **63**

**References** **65**

**Appendix** **67**
Figures . . . . . 67
Tables . . . . . 68

# List of Abbreviations

| | |
|---|---|
| AKT | Protein kinase B |
| BCR | B-cell receptor |
| BTK | Brutons tyrosine kinase |
| CDF | Cumulative Distribution Function |
| CLL | Chronic lymphocytic leukemia |
| CNV | Copy number variation |
| CpG ODN | CpG oligodeoxynucleotides |
| ERK/MAPK signalling | Mitogen-activated protein kinase signalling |
| FBS | Fetal bovine serum |
| FDR | False discovery rate |
| GSEA | Gene set enrichment analysis |
| Ig | Immunoglobulin |
| IGHV | Immunoglobulin heavy chain variable region |
| IL | Interleukin |
| JAK | Janus kinase |
| M-CLL | CLL with somatic hypermutations in IGHV loci |
| NF$\kappa$B | Nuclear factor kappa-light-chain-enhancer of activated B cells |
| NOTCH | Neurogenic locus notch |
| OS | Overall survival |
| PBMC | Peripheral blood mononuclear cells |
| PCA | Principal component analysis |
| PI3K | Phosphoinositide 3-kinase |
| U-CLL | CLL without somatic hypermutations in IGHV loci |
| TGF$\beta$ | Transforming growth factor $\beta$ |
| TLR | Toll-like receptor |
| TTT | Time to next treatment |
| SF3B1 | Splicing factor 3B subunit 1 |
| STAT | Signal transducer and activator of transcription |
| SYK | Spleen tyrosine kinase |

# List of Figures

3.1 Schematic of experimental protocol. By combining 12 drugs and 17 stimuli, we systematically queried the effects of simultaneous stimulation and inhibition of critical pathways in CLL (n=192). Integrating functional drug-stimulus response profiling with four additional omics layers, we identified pro-survival pathways, underlying molecular modulators of drug and microenvironment reponses, and drug-stimulus interactions in CLL. . 20

3.2 Schematic of validatory data. . . . . 20

3.3 Bar plot of the drugs used in screen. . . . . 20

3.4 Log transformed viability values for all drugs that were included in the screen after qulaity control. p values from student's t test. . . . . 21

4.1 Heatmap of Pearson correlation coefficients. Coefficients for each pair of stimuli were calculated using log transformed viability values normalised to untreated control, and ordered according to hierarchical clustering. Figure adapted from Bruch & Giles et al. 2021. . . . . 24

4.2 Scatter plot of log-transformed viability values, normalised to DMSO controls, for (A) treatment with JAK-STAT agonists IL4 and IL6 and (B) treatment with TLR agonists CpG ODN and Resiquimod. . . . . 25

4.3 Log transformed viabilities after treatment with each stimulus. Where stimuli decreased viability relative to control, points are shown in blue, whilst increased viability is shown in red. Figure adapted from Bruch & Giles et al. 2021. . . . . 26

4.4 The heatmap matrix shows the viability measurements for 192 samples (columns) and 17 stimuli (rows). Viability was measured via ATP-based assay after 48h stimulation. The data are shown normalised to DMSO-treated controls, and scaled within each row according to the Median Absolute Deviation (MAD). Limits were applied to scaling factor for optimal visualisation. The colour bars to the right show sample annotations. Consensus Clustering was used to define column tree layout, using hierarchical clustering with the Euclidean metric. Figure from Bruch & Giles et al. 2021. . . . . 27

4.5 Summaries of the CDFs of the consensus matrices. Consensus CDF graphic showing the CDFs of the consensus matrix for k = 2 - 7, as indicated in the legend, estimated using 100 bin histogram (left). Relative change in area under the CDF curve, for k = 2 - 7, to compare k with k - 1. In the case of k = 2, there is no k - 1, so the total area is plotted. Line shows relative increase in consensus between each value of k (right). 28

4.6 Assignment of patient samples (columns), to each cluster, for k = 1 - 7 (rows) to demonstrate stability of cluster membership. Cluster colour for k = 4 match those in heatmap in 4.4. . . . . 29

4.7 (A) Lymphocyte doubling time (LDT) stratified by cluster, p-values from Student's t-test. (B) Kaplan-Meier curves to show TTT for each cluster. p-values from univariate Cox proportional hazard models comparing IGHV-U enriched C1 with C2, and IGHV-M enriched C3 with C4. Figure from Bruch & Giles et al. 2021. . . . . 31

4.8 Log-transformed normalised viability values, stratified by cluster, for each drug. Drugs targeting the same pathway are grouped together. P-values from Student't t-test. . . . . 36

4.9 Distribution of selected genetic features (rows) within each cluster for all patient samples (columns). Where a patient sample is not annotated for a feature, this is marked in white. . . . . 37

4.10 Multinomial regression with lasso penalisation to identify enrichment or depletion of genetic features within each cluster. Matrix of genetic features (p=39), and IGHV status (encoded as M = 1 and U = 0) were used to identify multivariate predictors of cluster assignment. x axis shows genetic predictors, y axis indicates value and sign of coefficient assigned to feature, for each cluster (positive coefficients are enriched in the cluster, negative coefficients are depleted). Coefficients shown are mean coefficients from 50 bootstrapped repeats and error bars represent the mean ś standard deviation. Genetic features with >20% missing values were excluded, and only patients with complete annotation were included in the model (n=137). Figure from Bruch & Giles et al 2021. . . . . 38

4.11 Volcano plot of differentially expressed genes between C3 and C4. X axis indicates log2 fold change values, calculated using the `DESeq2` package (**DESeq2?**), y axis gives corresponding -log10(adjusted p value). P values adjusted using BH method. Genes are labeled where ajdusted $p < 0.05$. Figure from Bruch & Giles et al. 2021. . . . . 39

4.12 Gene set enrichment analysis (GSEA) to compare expression of genes in samples from C3 and C4 reveals upregulation of Hallmark pathways involved in microenvironmental signalling, stress response, metabolism and proliferation in C3. Normalised enrichment scores (NES) are shown for top 10 most significant pathways upregulated in C3 versus C4. Bars coloured according to adjusted p-value. Genes are ranked based on Wald statistic, calculated using the Deseq2 package and GSEA performed using the fgsea algorithm. Figure from Bruch & Giles et al. 2021. . . . . 40

4.13 Enrichment plots of selected pathways. Gene set enrichment analysis (GSEA) was performed with the Hallmark gene sets from the GSEA Molecular Signatures Database. Wald statistic was used to rank the genes. The green curve corresponds to the Enrichment Score curve, which is the running sum of the weighted enrichment score obtained from GSEA software. Figure from Bruch & Giles et al. 2021 . . . . . 41

5.1 (ref:stimuliGeneAssosciations) . . . . . 44

5.2 (ref:stimuliGeneAssosciationsMulti) . . . . . 45

5.3 (ref:IHGVBCR) . . . . . 46

5.4 (ref:TLRIHGVtri12) . . . . . 47

5.5 (ref:TLRGenes) . . . . . 48

5.6 (ref:IL4KRAS) . . . . . 49

5.7 (ref:tri12cytResponse) . . . . . 50

5.8 (ref:SMAD3geneDosage) . . . . . 51

5.9 (ref:STATgeneDosage) . . . . . 52

5.10 (ref:IRAK4geneDosage) . . . . . 53

5.11 (ref:tri12Classfier) . . . . . 53

5.12 (ref:tri12PhenocopiesCytResponse) . . . . . 54

5.13 (ref:tri12PhenocopiesRNA) . . . . . 55

5.14 (ref:diffTFexplainer) . . . . . 56

5.15 (ref:tri12diffTF) . . . . . 56
5.16 (ref:SpiBshRNAKD) . . . . . 57

6.1 Estimated residuals from model XXX3. ... . . . . 60

7.1 Estimated residuals from model XXX4. ... . . . . 62

8.1 Analysis of ATACseq dataset of two trisomy 12 and two non-trisomy 12 untreated CLL PMBC samples. The volcano plot depicts change in TF activity (x axis) versus BH-adjusted p-values (y axis) for 636 TFs, comparing trisomy 12 and non-trisomy 12 samples. The (**diffTF?**)[REFERENCE] package was ran in analytical mode to calculate TF activity, measured as weighted mean difference. TFs are labeled if adjusted p-value $< 0.01$ and absolute weighted mean difference $> 0.15$. . . . . 67

# List of Tables

4.1 Table depicting results of Multivariate Cox Proportional Hazard Model to test prognostic value of key genetic features and clusters using Time to Next Treatment and C3 as reference. . . . . 32

8.1 Detailed descriptive statistics of location and dispersion for 2100 observed swap rates for the period from February 15, 1999 to March 2, 2007. Swap rates measured as 3.12 (instead of 0.0312). . . . . 68

# Chapter 1

# Introduction

(theoretical background) ## CLL Disease biology

### 1.0.1 Disease Characteristcs

Dietrich et al. (2017) shows that... (Dietrich et al. 2017).

### 1.0.2 IGHV status

### 1.0.3 The role of BCR signalling

### 1.0.4 Recurrent genetic features in CLL

-mutations

-copy number variants

### 1.0.5 The incompletely understood role of trisomy12

### 1.0.6 Stratification of CLL patients

## 1.1 The tumour microenvironment in CLL

### 1.1.1 The concept of the protective niche

Lymph nodes etc

### 1.1.2 Microenvironmeental pathways and cross-talk

pathways esp IL4 and TLR
The microenvironment in CLL constitutes cross talk via soluble factors and cell-cell contacts, in the blood, bone marrow and lymph nodes.

**??**

*Figure adapted from ...*

**??**

*Figure adapted from ...*

## 1.2 Therapies in CLL

### 1.2.1 Clinical regimes

### 1.2.2 New Targeted therapies

in particular Ibrutinib and idelalisib

### 1.2.3 Drug resistances in CLL

incl ME -mediated and ibrutinib - IL4

### 1.2.4 BET inhibitors, and their potential for the treatment of leukemia

## 1.3 Background to the approache used in this thesis

### 1.3.1 Ex-vivo drug sensitivity screens

### 1.3.2 Studies of the microenivornment, and attempts to replicate the impact of the microenvironment

### 1.3.3 Measurement of outcomes?

LDT, TTT, TFT, OS

### 1.3.4 Multiomics datasets to study CLL

The power of multi-omics approaches especially the use of ATAC sequencing and combining ATAC and RNA

### 1.3.5 Mathematical modelling

Mathematical background in particular, lasso-regularised regression, and the caveats associated with feature selection wrt cluster assignment

### 1.3.6 CLL as a model for studying tumour biology

## 1.4 Aims of this thesis

## 1.5 Overview of the approach

Including diagrams

# Chapter 2

# Methods

Methods in quotation marks are taken from Bruch & Giles et al. 2021 and I have authored the original text, unless stated otherwise. Paragraphs without quotation marks were rewritten for the thesis.

Routine: all headings should be here at the end of each chapter, make sure all methods are included Copy things in and put in quotation marks if from Bruch & Giles et al. 2021, mark if they were not written by me, or if they were written jointly Add REFERENCE and TABLE for now Return to this later and neaten / expand on methods, and provide all referencing and tables

## 2.1 Experimental methods:

### 2.1.1 Drug - Stimulus Combinatorial Pertubation Assay

#### Patient Sample Preparation

*with Peter-Martin Bruch* Original Peripheral Blood was taken from 192 patients for the initial drug-stimulation assay. Blood was separated by Ficoll gradient (GE Healthcare) and mononuclear cells were cryopreserved.

### 2.1.2 Preparation of screening plates

*with Peter-Martin Bruch* Original "Sample preparation, cell-culture, drug-stimulation profiling and genomic annotation was performed on 192 CLL patient samples as previously described[REFERENCE] with the following adjustments. Stimuli and drugs were mixed and preplated in the culture plates directly before adding the cell suspensions. RPMI-1640 and supplements were acquired from Gibco by Life Technologies, human serum was acquired from PAN Biotech (Cat.No. P40-2701, Lot.No:P-020317). Luminescence was measured after 48h on a Perkin Elmer EnVision."

*with Peter-Martin Bruch* Original Compounds were obtained from Selleckchem, MedChemExpress and Sigma-Aldrich, dissolved in DMSO and stored at -20°C. 12 drugs were used in two concentrations [TABLE]. Final DMSO concentration did not exceed 0.3% in all experiments. Carfilzomib, Panobinostat and Venetoclax were removed from the analysis as they showed inconsistent toxicity depending on used media. Insert table

Recombinant cytokines and stimulatory agents were dissolved according to manufacturer's protocol. 17 stimuli were selected [TABLE]. HS-5 conditioned medium was produced by incubating HS-5 stroma cell line (gifted by Martina Seiffert, DKFZ, Heidelberg) for 4 days at 37°C and 5% CO2, after which the supernatant was centrifuged and stored at -20°C, the final concentration of HS-5 CM was 20%. Bead immobilised anti-IgM was removed from the analysis due to storage instability. "

Paper: "Compounds and stimulatory agents were dissolved, stored and diluted according to manufacturer's protocol. HS-5 conditioned medium was produced by incubating HS-5 stromal cell line to >80% confluency and cell removal by centrifugation. For a detailed list of stimuli and drugs and associated concentrations see Supp. Tables 1 and 2. Final DMSO concentration did not exceed 0.3%."

Original Compounds were preplated in 384-well polypropylene storage plates (Greiner Bio-One Cat.No.:781271), which were stored at -20°C. For each batch of samples tested on the same day a new storage plate was thawed and diluted in serum free RPMI, with or

without stimuli. 5$\mu$L of drug-stimulation dilution were added into each well of the assay plates, 20 $\mu$L of cell suspension were added. The final cell concentration was 8*10ˆ5 cells/ml. Cells were thawed as previously described[REFERENCE].

## 2.2 Drug-Stimulation assay

*with Peter-Martin Bruch* Original Drug-stimulation assays were performed with RPMI-1640 (Gibco by Life Technologies) supplemented with Penicillin Streptomycin (Gibco) final concentration of 100 Units/ml and 100 $\mu$g/ml respectively, L-Glutamine (Gibco) final concentration 2mM, and 10% pooled, heat-inactivated and sterile filtered human type AB male off the clot serum (PAN Biotech, Cat.No. P40-2701, Lot.No:P-020317).

Drugs and stimuli were added to the cells simultaneously. Culture was performed in 384-well plates (Greiner Bio-One Cat.No.: 781904) and cells were incubated at 37°C and 5% CO2 for 48h.

Cell Viability was determined using the ATP-based CellTiter-Glo assay (Promega, Cat.No.:G7573). Luminescence was measured for the drug-stimulation assays using a Perkin Elmer EnVision and for the drug-drug interaction experiments using a Perkin Elmer EnSight, with a measurement time of 100ms per well.

## 2.3 WES, RNA Sequencing, Targeted Sequencing and DNA Copy number variants

*Myself* Original Sequencing data generated by Dietrich et al[REFERENCE], and all sequencing and data processing were performed as described there. For

### 2.3.1 Follow - up investigations

**Lymphocyte doubling time**

**Survival data**

**ATACsequencing of 4 CLL samples**

*with Peter-Martin Bruch (mostly edited by me)* Paper "Peripheral blood was taken from 4 CLL patients and separated by Ficoll gradient (GE Healthcare), mononuclear cells were cryopreserved on liquid nitrogen. Samples were later thawed from frozen as previously described[REFERENCE] and MACS sorted for CD19 positive cells (Milteny autoMACS). The cells were resuspended in RPMI (GIBCO, Cat.No. 21875-034), with the addition of 2mM glutamine (GIBCO, Cat.No. 25030-24), 1% Pen/Strep (GIBCO, Cat.No. 15140-122) and 10% pooled, heat-inactivated and sterile filtered human type AB male off the clot serum (PAN Biotech, Cat.No. P40-2701, Lot.No:P-020317). 5ml of cell suspension was cultured in 6-well plates (Greiner Bio-One Cat.No. 657160). After thawing, cells were incubated at 37°C and 5% CO2 for 6 hours in 0.2% DMSO. The final cell concentration was 2x106 cells/ml. Cell viability and purity was assessed using FACS. All samples had a viability over 90% and over 95% of CD19+/CD5+/CD3- cells."

*with Nayara* Paper "ATAC-seq libraries were generated as described previously[REFERENCE]. Cell preparation and transposition was performed according to the protocol, starting with 5x104 cells per sample. Purified DNA was stored at 20°C until library preparation was performed. To generate multiplexed libraries, the transposed DNA was initially amplified for 5x PCR cycles using 2.5 $\mu$L each of 25 M PCR Primer 1 and 2.5 $\mu$L of 25 $\mu$M Barcoded PCR Primer 2 (included in the Nextera index kit, Illumina, San Diego, CA, USA), 25 $\mu$L of NEBNext High-Fidelity 2x PCR Master Mix (New England Biolabs, Boston, Massachusetts) in a total volume of 50 $\mu$L. 5 $\mu$L of the amplified DNA was used to determine the appropriate number of additional PCR cycles using qPCR. Additional number of cycles was calculated through the plotting of the linear Rn versus cycle, and corresponds to one-third of the maximum fluorescent intensity. Finally, amplification was performed on the remaining 45 $\mu$L of the PCR reaction using the optimal number of cycles determined for each library by qPCR (max. 13 cycles in total). The amplified fragments were purified with two rounds of SPRI bead clean-up (1.4x). The size distribution of the libraries was assessed on Bioanalyzer with a DNA High Sensitiv-

ity kit (Agilent Technologies, Santa Clara, CA), concentration was measured with Qubit DNA High Sensitivity kit in Qubit 2.0 Flurometer (Life Technologies, Carlsbad, CA). Sequencing was performed on NextSeq 500 (Illumina, San Diego, CA, USA) using 75bp paired-end sequencing, generating 450 million paired-reads per run, with an average of 55 million reads per sample."

**Spi-B and PU.1 shRNA Knockdowns**

**Immunohistochemistry of patient lymph nodes**

*Peter-Martin Bruch* Paper "LN biopsies of CLL-infiltrated and non-neoplastic lymph nodes were Paraffin fixed and arranged in Tissue Microarrays. Consecutively they were stained for PAX-5 (790-4420, Roche), CD3 (790-4341, Roche), pSTAT6 (ab28829, abcam), STAT6 (519-4290, Zytomed Systems) and pIRAK4 (ab216513, abcam). The slides were analysed using Qupath 50 and the recommended protocol. Cell based data on mean staining intensity was exported and further analysed using R."

**Ibrutinib - IL4 - STAT6i interaction assay**

### 2.3.2 Investigation of Ibrutinib + IL4 + IBET-762

**Ibrutinib - IL4 - IBET-762 interaction assay**

**Ibrutinib - IL4 - IBET -762 treatment**

**ATACsequencing**

*with Peter Martin Bruch* get from diffTF Peripheral blood was taken from 4 CLL patients and separated by Ficoll gradient (GE Healthcare), mononuclear cells were cryopreserved on liquid nitrogen. Samples were later thawed from frozen as previously described[REFERENCE] and MACS sorted for CD19 positive cells (Milteny autoMACS). The cells were resuspended in RPMI (GIBCO, Cat.No. 21875-034), with the addition of 2mM glutamine (GIBCO, Cat.No. 25030-24), 1% Pen/Strep (GIBCO, Cat.No. 15140-122) and 10% pooled, heat-inactivated and sterile filtered human type AB male off the clot

serum (PAN Biotech, Cat.No. P40-2701, Lot.No:P-020317). 5ml of cell suspension was cultured in 6-well plates (Greiner Bio-One Cat.No. 657160). After thawing, cells were incubated at 37°C and 5% CO2 for 6 hours in 0.2% DMSO. The final cell concentration was 2x106 cells/ml. Cell viability and purity was assessed using FACS. All samples had a viability over 90% and over 95% of CD19+/CD5+/CD3- cells.

*with Nayara* get from diffTF ATAC-seq libraries were generated as described previously[REFERENCE]. Cell preparation and transposition was performed according to the protocol, starting with 5x104 cells per sample. Purified DNA was stored at 20°C until library preparation was performed. To generate multiplexed libraries, the transposed DNA was initially amplified for 5x PCR cycles using 2.5 $\mu$L each of 25 $\mu$M PCR Primer 1 and 2.5 $\mu$L of 25 $\mu$M Barcoded PCR Primer 2 (included in the Nextera index kit, Illumina, San Diego, CA, USA), 25 $\mu$L of NEBNext High-Fidelity 2x PCR Master Mix (New England Biolabs, Boston, Massachusetts) in a total volume of 50$\mu$L. 5 $\mu$ L of the amplified DNA was used to determine the appropriate number of additional PCR cycles using qPCR. Additional number of cycles was calculated through the plotting of the linear Rn versus cycle, and corresponds to one-third of the maximum fluorescent intensity. Finally, amplification was performed on the remaining 45 $\mu$L of the PCR reaction using the optimal number of cycles determined for each library by qPCR (max. 13 cycles in total). The amplified fragments were purified with two rounds of SPRI bead clean-up (1.4x). The size distribution of the libraries was assessed on Bioanalyzer with a DNA High Sensitivity kit (Agilent Technologies, Santa Clara, CA), concentration was measured with Qubit DNA High Sensitivity kit in Qubit 2.0 Flurometer (Life Technologies, Carlsbad, CA). Sequencing was performed on NextSeq 500 (Illumina, San Diego, CA, USA) using 75bp paired-end sequencing, generating 450 million paired-reads per run, with an average of 55 million reads per sample.

**RNAsequencing**

**Proteomics**

## 2.4 Data availability

European Genome-Phenome Archive (EGA) accession .... The data for .... and computational analysis code used in this study are available from ....

Additionally, we obtained CLL ATACseq data[REFERENCE] from the European Genome-phenome Archive (EGA: EGAD00001002110).

For the ChIPseq analysis of SPIB binding sites, we accessed the data 37 from the NCBI GEO database 51, accession GEO: GSE56857. We made use of the data for SPIB in the OCILY3 DLBCL cell line (GSM1370276).

## 2.5 Statistical Analysis

Mixutre of paper and original Integrative data analysis of screening data, DNA and RNA sequencing, CNV and methylation profiles and follow up experiments was performed using R version 4 (REFERENCE R Core Team, 2018) with the RStudio interface (REFERENCE RStudio Team, 2016) and using packages that included (**DESeq2?**)[REFERENCE], (**survival?**)[REFERENCE], (**Glmnet?**)[REFERENCE], (Wilkerson and Hayes 2010)[REFERENCE], (**clusterProfiler?**)[REFERENCE], (**ChIPseeker?**)[REFERENCE (**genomation?**)[REFERENCE] and (**BloodCancerMultiOmics2017?**)[REFERENCE] to perform univariate association tests, multivariate regression with and without lasso penalization, Cox regression, generalised linear modelling and clustering.

### 2.5.1 Screening Data

#### Processing of screening data

*with Peter-Martin Bruch* Paper To quantify the response of a patient sample to each treatment condition, we used viability relative to the control, i.e., the CellTiter Glo luminescence readout of the respective well divided by the median of luminescence readouts of the DMSO control wells on the same plate.

### Quality Assessment and Control

### Drug -drug and stimulus - stimulus correlations

*myself* original complete Pearson correlation coefficients were calculated for each drug - drug and stimulus - stimulus pair, using the `cor` functions of the (**stats?**)[REFERENCE] package with log transformed viability values which were normalised to untreated controls.

### Characterising stimulus responses across all samples

*myself* original complete (just check) For the heatmap in figure 4.4, the viability data is represented as z scores. Log transformed viability values, normalised to DMSO controls, are row-scaled according to the Median Absolute Deviance. Limits were then applied to this row scaling factor for optimal visualisation. The matrix of z scores was plotted using the (**pheatmap?**)[REFERENCE] package. The ordering of the columns (patient samples) was obtained from the dendrogram that resulted from running `ConsensusClusterPlus`, from the (Wilkerson and Hayes 2010) [REFERENCE] package. The rows were ordered using the dendrogram order produced by `hclust` with default branch arrangement.

### Univariate gene - repsonse testing

### Penalised multivariate regression of gene - response

*myself* original For the plots in Figure 3B, we used a Gaussian linear model with L1-penalty. As the dependent variable, the normalised viability value was used for all stimuli. We used the same features as above, plus Methylation Cluster(coded as 0, 0.5, 1). As a measure of explained variance the reduction in cross-validated mean squared error relative to the null model was calculated.

### Correlation of RNA-Receptor Expression with viability

### Linear modelling of drug - stimulus interactions

#### Modelling of drug - stimulus - gene interactions

*myself* original To identify drug - stimulus interactions that were dependent on genetic features, we first fitted linear models to the viability matrix, for each drug - stimulus combination, using the formula in Eqn2. We extracted drug - stimulus interaction coefficients for each patient to generate a response matrix for n = 137 patients. Next, we performed multivariate regression using a Gaussian linear model with L1-penalty (i.e., lasso regression) as implemented in the R package glmnet[REFERENCE]. As the dependent variable, the matrix of drug - stimulus interaction coefficients for each patient sample was used. As input to the model, genetic features with more than 20% missing values were excluded, and only patients with complete annotation were included in the model ( n= 137 ). As predictors, the genetic mutations and CNVs (p= 39), IGHV status (coded as 0-1) and Methylation CLuster (coded as 0, 0.5, 1) were used, using 3-fold cross-validation. Misclassification error was used as loss for cross- validation.

### 2.5.2 Follow up

#### Lympocyte doubling times

*Herbst et al., 2020c* - update "Patients which had lymphocyte counts available for less than 4 timepoints between the sample collection date and the time of the next treatment and patients currently in treatment were ex-cluded. Thus, ... patients with enough data remained. Lymphocyte growth rates were calculated by fitting a linear model to the log10 transformed lymphocyte counts of all timepoints between the sample collection date and the time of the next treatment versus the period of time."

#### Survival analyses

*myself* original Survival analyses were performed using TTT, TFT and OS as metrics. Follow-up information to calculate OS was available for all 192 CLL patients. For 188 of 192 CLL patients treatment information after sample collection was available.

*Herbst et al* "Time to next treatment (TTT) was calculated from the date of sample col-

lection to subsequent treatment initiation. Patients without treatment initiation during the observation time and patients who died before treatment initiation were censored at the latest follow-up contact."

*myself* original For for the purpose of visualising Kaplan-Meier plots , optimal cut-points of staining levels were calculated using maximally selected rank statistics as computed by the package (**maxstat?**) REFERENCE. Based on these cut points, patients were split into two groups, and their survival data were plotted using the Kaplan-Meier method, using the R package (**survminer?**) REFERENCE .

For Cox proportional hazards regression, the `coxph` function of the R package survival [REFERENCE] (Therneau, 2020) was used. Expand

### Penalised multivariate regression genetic predictors of cluster membership

*Myself* original (edit so not similar to JCI) I used a multinomial linear model with L1-penalty, implemented in the (**cvglmnet?**) package[REFERENCE]. As the dependent variable, the cluster assignment for each patient was used. As input to the model ,genetic features with more than 20% missing values were excluded, and only patients with complete annotation were included in the model ( n= 137 ). As predictors, the genetic mutations and CNVs (p= 39) and IGHV status (coded as 0-1) were used, using 3-fold cross-validation. Misclassification error was used as loss for cross- validation. The resulting coefficients indicated associations between genetic features and each cluster.

### Gene expression and gene set enrichment analysis between clusters

*myself* original For the n=49 patient samples for which viability data and RNA–Seq data for matching samples was available, the R package (**DESeq2?**)[REFERENCE] was used to search for associations of these two data types.RNA-Seq read count data was regressed on to the patient clusters C3, C4 (design formula ~ IGHV.status + Cluster). Genes were ranked by their test statistics and Gene Set Enrichment Analysis (GSEA) (implementing the fgsea algorithm with the (**clusterProfiler?**) package[REF-

ERENCE]) was applied to the ranked lists with the KEGG pathway gene set selections from the MSigDBdatabase[REFERENCE].

### Associations of ex-vivo stimulus responses with genomic features

We tested for associations between stimulus viability assay results and genomic features by Student's t-tests (two-sided, with equal variance). We tested somatic mutations (aggregated at the gene level), copy number aberrations and IGHV status. We restricted the analysis to features that were present in at least 3 patient samples (63 features). p-values were adjusted for multiple testing by applying the Benjamini-Hochberg procedure.

### ATACseq processing

*with Ivan Berest* original (update to paper) We downloaded CLL data[REFERENCE] from the European Genome-phenome Archive (EGA: EGAD00001002110). Original dataset had 88 ATAC-seq samples from 55 patients, however we used for the analysis only one sample per patient passing quality checks, resulting in 52 samples. We used an in-house constructed ATAC-seq processing pipeline[REFERENCE] to obtain final bam files mapped to the hg19 annotation genome that were corrected for CG bias.

### Annotation of trisomy 12 status

*with Ivan Berest* original (update to paper) As trisomy 12 status was not included in the original metadata, we used a mean amount of reads in the chromatin accessible peaks for each sample to distinguish trisomy 12 patients. All samples containing 1.4 times more reads in the peaks located on chromosome 12, compared to the peaks on all other chromosomes, were classified as trisomy 12 patients.

### diffTF analysis of TF activity in trisomy 12 CLL

*with Ivan Berest* original (update to paper) FOr the larger ATACseq dataset, we ran analytical mode of diffTF[REFERENCE] pipeline to investigate differential TF activity between trisomy 12 and non-trisomy 12 patients using the HOCOMOCO v10 database

[REFERENCE] with the following design formula: "~ Batch + Gender + IGHV status + trisomy12 status."

For the smaller ATACseq dataset, ATAC-seq data generated from our CLL samples were processed similarly with the in-house ATAC-seq pipeline, as described previously[REFERENCE], with the only exception that we didn't use CG bias correction step. We used a similarly analytical mode of diffTF with HOCOMOCO v10 database[REFERENCE] using the following parameters: minOverlap = 1; design formula = "~ Patient + Trisomy 12 status."

### Functional enrichment analysis of SPIB ChIPseq data

*myeslf* original (update to paper) We downloaded SPIB ChIPseq data[REFERENCE] from the NCBI GEO database[REFERENCE], accession `GEO:GSE56857`. We made use of the data for SPIB in the OCILY3 DLBCL cell line (GSM1370276). SPIB ChIP peaks were filtered for significance (q value<0.05). We used the annotatePeaks function from the package clusterProfiler[REFERENCE] to annotate the nearest gene for each ChIP-peak. We filtered peaks within ś1kb of a TSS of a gene and performed over-representation of KEGG and Reactome pathways (using the clusterProfiler package[REFERENCE]) amongst the resulting list of genes.

### Linear modelling of drug - stimulus interactions

*myself* original To identify drug-stimulus interactions, we fitted linear models to the viability matrix, for each drug - stimulus combination, using the formula in Eqn1. Drug - stimulus interaction coefficients and associated p values were extracted and used to define significant interactions.

### Drug-Drug interaction assay

An independent patient cohort of 16 patients was used in the drug-drug interaction experiments.

**Survival analysis of immunohistory chemistry data**

### 2.5.3 Ibrutinib - IBET-762 - IL-4 analysis

**RNAseq processing**

**Deseq2 analysis**

**ATACseq processing**

**diffTF analysis**

**Generation of Gene Regulatory Network**

**Proteomics Processing**

**Proteomics Analysis**

# Chapter 3

# Data

Characterisation of primary CLL samples by high-throughput combinatorial screening and multi-omic profiling. An introduction to the dataset that my PhD is based on.

## 3.1 Drug screens and experiments

### 3.1.1 High-throughput combinatorial pertubation assay

We measured the effects of 17 cytokines and microenvironmental stimuli on cell viability in 192 primary CLL samples and combined each one with 12 drugs to investigate the influence on spontaneous and drug-induced apoptosis (Figure 3.1). Viability was assessed by ATP measurement via CellTiterGlo after 48 h of culture and normalised to untreated controls8.

### 3.1.2 Validation experiments

In addition to the screen, we acquired the following validatory data (Figure 3.2).

## 3.2 Characteristics of drugs used in the screen

.

**Figure 3.1:** Schematic of experimental protocol. By combining 12 drugs and 17 stimuli, we systematically queried the effects of simultaneous stimulation and inhibition of critical pathways in CLL (n=192). Integrating functional drug-stimulus response profiling with four additional omics layers, we identified pro-survival pathways, underlying molecular modulators of drug and microenvironment reponses, and drug-stimulus interactions in CLL.

.

**Figure 3.2:** Schematic of validatory data.

### 3.2.1 Drug pathways

We screened 17 different drugs (Figure 3.3).

.

**Figure 3.3:** Bar plot of the drugs used in screen.

### 3.2.2 Drug responses

The drug responses were as follows (Figure 3.4).

### 3.2.3 Genetic predictors of drug responses

### 3.2.4 Drug - Drug Correlations

## 3.3 Characteristics of stimuli used in the screen

### 3.3.1 Stimulus responses

### 3.3.2 Stimulus - Stimulus Correlations

## 3.4 Characteristics of patient samples used in the screen

**Figure 3.4:** Log transformed viability values for all drugs that were included in the screen after qulaity control. p values from student's t test.

### 3.4.1 Genetic Data available for each patient

WES, CNVs, Methylation, Transcriptomic, ATACseq, LDT, survT, IHC PLus Ibrutinib + IBET + IL4 treated samples - ATACseq, RNAseq, proteomics

## 3.5 Processing of raw values obtained from cell viability assay

### 3.5.1 Data normalization and quality control

## 3.6 Generation of public resource

### 3.6.1 Shiny app

### 3.6.2 Package

- Present the underlying economic model/theory and give reasons why it is suitable to answer the given problem[1].

[1] Here is an example of a footnote.

# Chapter 4

# Ex-vivo sensitivity to microenvironmental stimulation in primary CLL cells

The screen included 17 cytokines and microenvironmental stimuli, which were selected based on evidence in the literature that each stimulus had been shown to impact on CLL viability *in vitro*. We attempted to put together the largest panel of stimuli for an assay of this kind in CLL, aiming to minimise redundancy among the compounds (Bruch & Giles et al. 2021).

CLL cells do not proliferate *in vitro*, but rather undergo spontaneous apoptosis in the absence of stimulation (Collins et al. 1989). The effect of each stimuli on CLL viability was thus quantified by comparing ATP counts in treated primary samples, compared with those in DMSO wells where a positive value indicates that the sample viability was increased relative to control.

The assay represents a reductionist model of microenvironmental signalling, making it possible to dissect the effect of individual stimuli on baseline viability. The large

patient cohort also enabled us to study the differential impact of these stimuli across heterogeneous samples, and identify distinct subgroups of patients whose samples show similar response profiles to the panel of stimuli.

## 4.1 Prolifing responses to the panel of stimuli

### 4.1.1 *ex vivo* assay demonstrated functional diversity of cytokines and microenvironmental stimuli

To investigate heterogeneity amongst responses to the stimuli, I calculated Pearson correlation coefficients for each stimulus pair, using the log-transformed normalised viabilities (Bruch & Giles et al. 2021). The resulting coefficients were ordered using hierarchical clustering and visualised in a symmetrical heatmap (Figure 4.1). The hierarchical clustering distinguished clusters of stimuli, including a larger group corresponding to agonists of TLR and Nf$\kappa$B pathways and a smaller group encompassing IL4 and TLR stimuli.

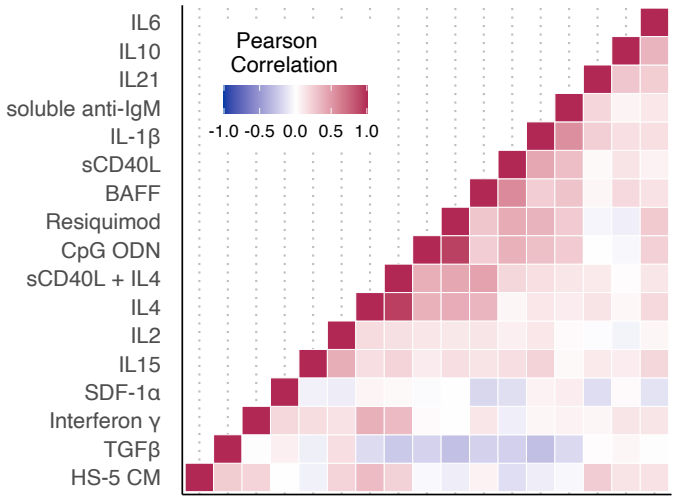


**Figure 4.1:** Heatmap of Pearson correlation coefficients. Coefficients for each pair of stimuli were calculated using log transformed viability values normalised to untreated control, and ordered according to hierarchical clustering. Figure adapted from Bruch & Giles et al. 2021.

98.5% of stimulus pairs showed little correlation (R < 0.6), including those that targeted similar downstream pathways, indicating a high degree of functional diversity amongst microenvironmental signals. For example, JAK-STAT agonists such as IL4 and IL6 showed little correlation (Figure 4.2A).

Only two stimulus pairs showed correlations where R > 0.6, and in both cases these targeted near identical receptors or downstream pathways. These included CpG ODN (TLR 9) and Resiquimod (TLR 7 and 8) (Figure 4.2B), and IL4 and IL4 + CD40L which primarily target JAK3 - STAT6.

Repeating the analysis to correlate drug - drug pairs demonstrated that drugs targeting components of the same pathway were highly correlated. For quality control purposes, this indicated that our data sensitively and specifically reflect inter-individual differences in pathway dependencies (Dietrich et al. 2017).

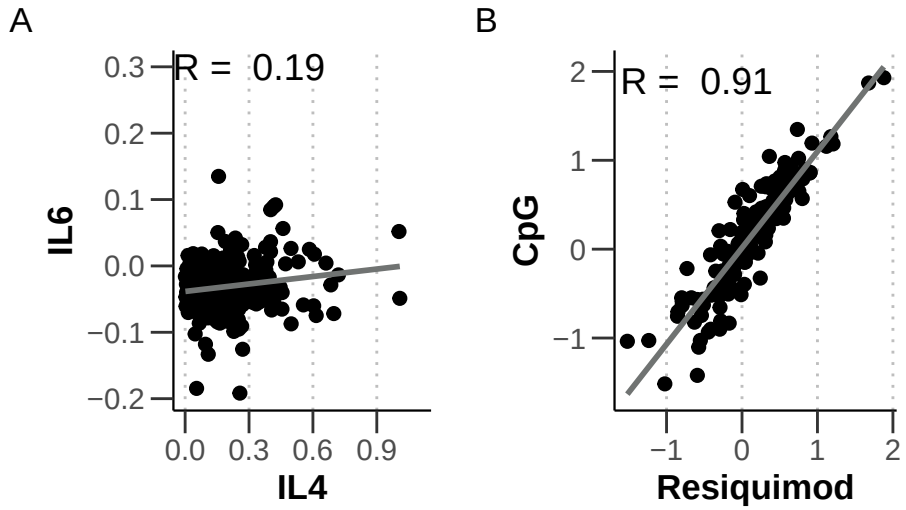


**Figure 4.2:** Scatter plot of log-transformed viability values, normalised to DMSO controls, for (A) treatment with JAK-STAT agonists IL4 and IL6 and (B) treatment with TLR agonists CpG ODN and Resiquimod.

Microenvironmental stimulation induced diverse phenotypes between patient samples, and across different stimuli. To gain a global overview of these phenotypes, I visualised log-transformed viability values normalised to DMSO controls for all patient samples

and all stimuli (Figure 4.3 (Bruch & Giles et al. 2021).

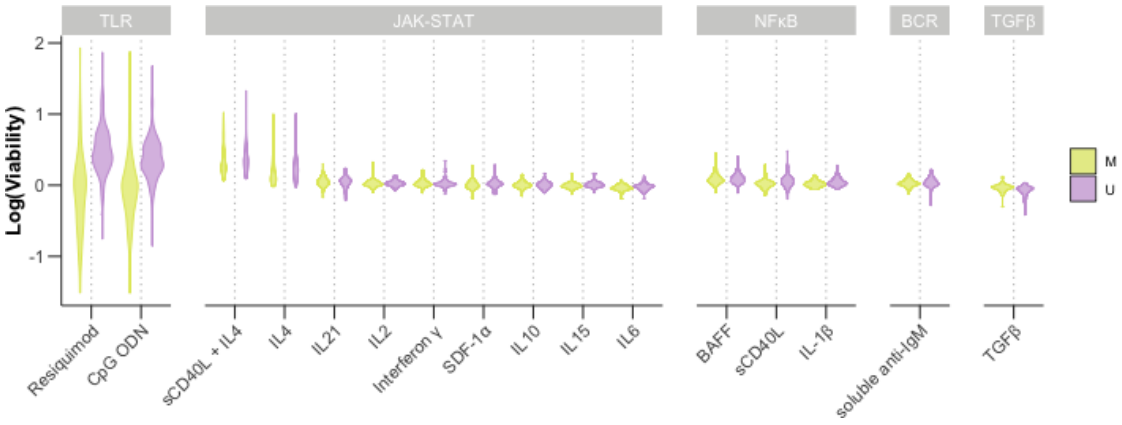


**Figure 4.3:** Log transformed viabilities after treatment with each stimulus. Where stimuli decreased viability relative to control, points are shown in blue, whilst increased viability is shown in red. Figure adapted from Bruch & Giles et al. 2021.

The majority of the stimuli increased viability, underlining the supportive nature of the microenvironment in CLL. However, three out of 17 reduced CLL viability relative to control, namely IL6, TGF$\beta$ and TLR 7/8/9 agonists in IGHV-mutated (IGHV-M) samples.

IL4 and TLR7/8/9 agonists Resiquimod and CpG ODN induced the strongest responses, an indication of their potency in modulating CLL cell survival. Notably, TLR agonists increased viability in certain samples, in most cases IGHV-U, and decreased viability in others, mostly IGHV-M. Both IL4 and TLR7/8/9 emerged as the key pathways in the screen, and play an important role throughout the results of this thesis.

### 4.1.2 Microenvironmental response profiling identifies discrete patient subgroups

To further investigate the variability in responses across the cohort, we visualised z scores of the log-transformed viabilities and performed consensus clustering on the resulting

matrix to generate a heatmap of all stimuli responses across all samples (Figure 4.4.

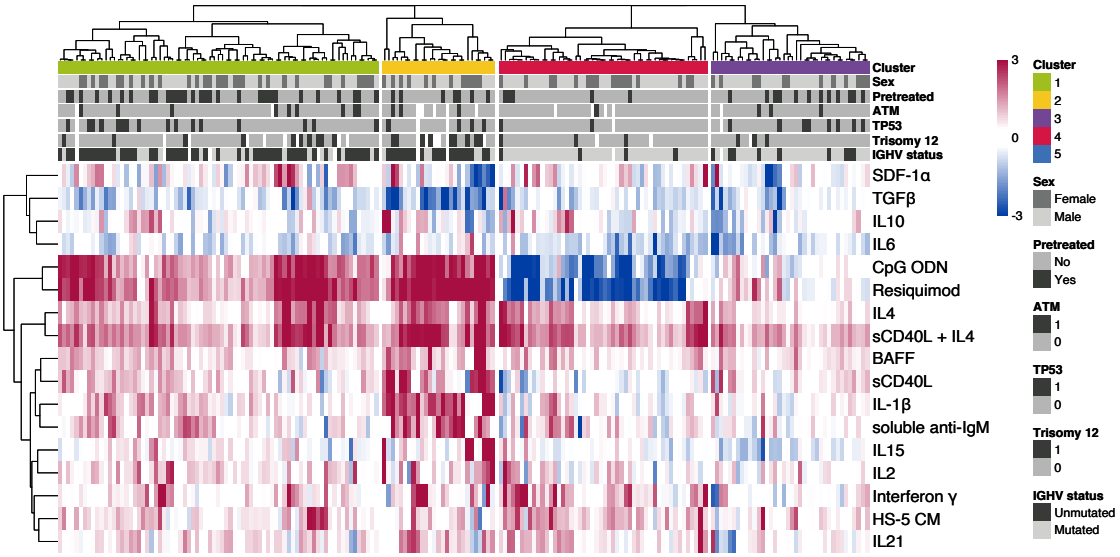


**Figure 4.4:** The heatmap matrix shows the viability measurements for 192 samples (columns) and 17 stimuli (rows). Viability was measured via ATP-based assay after 48h stimulation. The data are shown normalised to DMSO-treated controls, and scaled within each row according to the Median Absolute Deviation (MAD). Limits were applied to scaling factor for optimal visualisation. The colour bars to the right show sample annotations. Consensus Clustering was used to define column tree layout, using hierarchical clustering with the Euclidean metric. Figure from Bruch & Giles et al. 2021.

The heatmap reflects heterogeneity in responses across samples and stimuli, once again underlining the potency of IL4 in increasing sample viability across diverse genetic backgrounds and the diversity in responses to TLR stimulation by Resiquimod and CpG ODN.

To generate the column-wise clustering, consensus clustering was performed on the matrix of z scores. Consensus clustering allows the user to subsample from the matrix of values, to generate hierarchical clustering for a given number of clusters, k. From this, it is possible to calculate a consensus matrix for each value of k, indicating for each

pair of values the proportion of time they occupy the same cluster when subsampled together.

I visualised the clustered heatmap in Figure 4.4 for different values of k, and concluded on the existence of four robust clusters within the cohort (add these to appendix?). Each cluster shows a unique response profile to the panel of stimuli. We termed the clusters C1 to C4: C1 and C2 were enriched for IGHV-U whilst the samples in C3 and C4 were mostly IGHV-M.

To validate the choice of four clusters, I visualised the summaries of the consensus matrix, using the `ConsensusClusterPlus` package (Wilkerson and Hayes 2010), to quantify the degree of confidence in the clusters for different values of k.

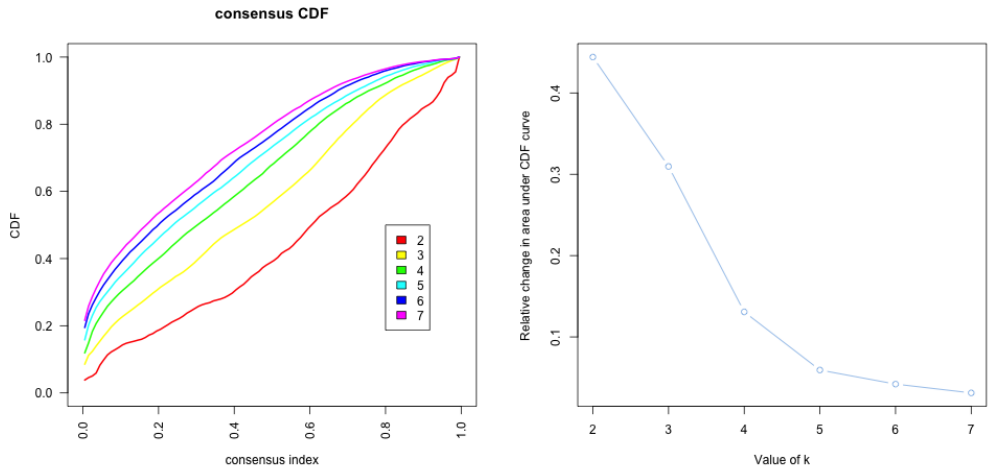


**Figure 4.5:** Summaries of the CDFs of the consensus matrices. Consensus CDF graphic showing the CDFs of the consensus matrix for k = 2 - 7, as indicated in the legend, estimated using 100 bin histogram (left). Relative change in area under the CDF curve, for k = 2 - 7, to compare k with k - 1. In the case of k = 2, there is no k - 1, so the total area is plotted. Line shows relative increase in consensus between each value of k (right).

The graph of the CDFs of the consensus matrix for each k indicated that the CDF reaches a maximum and cluster confidence is maximised at k = 7 , though above k = 4

there is little appreciable increase. Figure 4.5. This is confirmed in the graph showing relative change in the area under the CDF curve, showing there is only a small increase in consensus between k = 4 and k =5, supporting the choice of k = 4. The cluster tracking plot depicts how each patient sample is assigned for each value of k. For k = 4, the plot indicates that C3 and 4 in particular are highly stable (Figure 4.6).

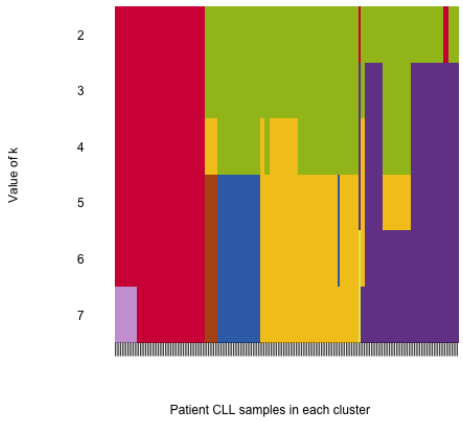


**Figure 4.6:** Assignment of patient samples (columns), to each cluster, for k = 1 - 7 (rows) to demonstrate stability of cluster membership. Cluster colour for k = 4 match those in heatmap in 4.4.

## 4.2 Functional characterisation of patient clusters

### 4.2.1 C1 - C4 showed distinct response profiles with the panel of stimuli

The heatmap in Figure 4.4 demonstrated that each cluster responded differently to the panel of stimuli.

Amongst the IGHV-U enriched C1 and C2, both showed strong, positive responses to IL4 and TLR7/8/9 stimulation. C2 could be distinguished by stronger responses to the stimuli overall, in particular to NF$\kappa$B agonists IL1$\beta$, anti-IgM, BAFF and sCD40L.

Amongst the IGHV-M enriched clusters, C3 showed weaker responses to the majority of stimuli, and C4 was defined by a negative response to TLR7/8/9 stimulation (Bruch & Giles et al. 2021). Figure **??** summarises these findings in more detail, showing responses stratified by cluster for a subset of the stimuli.

```
cairo_pdf
        2
```

### 4.2.2 The clusters show differences in disease dynamics

To validate the potential biological significance of these four clusters, we investigated whether the groups showed differential *in vivo* disease progression (Bruch & Giles et al. 2021). The study design was such that not all patients in the cohort were treatment - free, which confounded the analysis. For that reason, lymphocyte doubling time (LDT) and time to next treatment (TTT) were used to quantify CLL proliferative capacity, independently of treatment.

C1 and C2 showed a shorter LDT than C3 and C4, which is expected due to the differential proportions of IGHV-U and M patient samples in these groups (Figure 4.7A). Notably, within the IGHV-M enriched clusters C3 and C4, samples in C3 showed a significantly shorter LDT (Student's t-test, p-value = 0.025).

To further validate this, we observed that TTT in the IGHV-M enriched C3 was significantly shorter than C4 and comparable to the progression dynamics of IGHV-U enriched C1 and 2 (Figure 4.7B).

The difference in disease progression between the clusters indicated that microenvironmental response represents an additional biological layer, holding information relevant to disease dynamics. To validate that these clusters were not simply an indication of any underlying genetic features, we checked whether the observed differences in progression dynamics could be explained by other prognostic markers (Bruch & Giles et al. 2021).

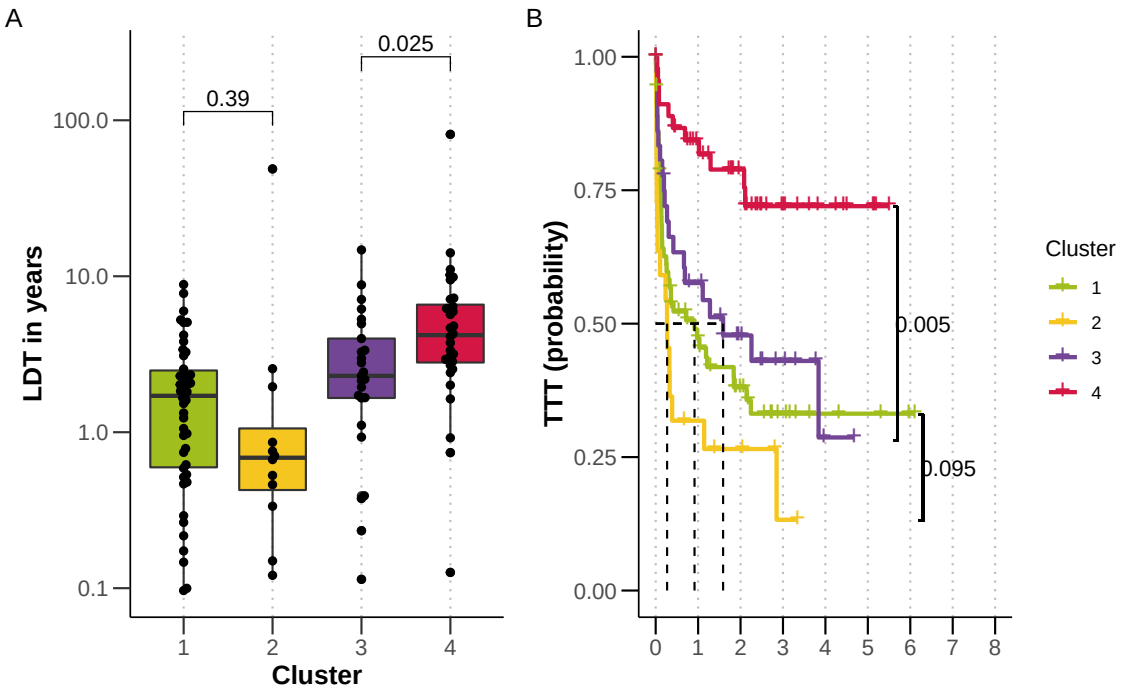


**Figure 4.7:** (A) Lymphocyte doubling time (LDT) stratified by cluster, p-values from Student's t-test. (B) Kaplan-Meier curves to show TTT for each cluster. p-values from univariate Cox proportional hazard models comparing IGHV-U enriched C1 with C2, and IGHV-M enriched C3 with C4. Figure from Bruch & Giles et al. 2021.

A multivariate Cox proportional hazard model accounting for IGHV status, trisomy 12 and TP53 in addition to the cluster assignment indicated an independent prognostic value for cluster assignment between C3 and C4 (p= 0.039, Table 4.1).

### 4.2.3 The clusters showed differential responses to drugs *in vitro*

The potential clinical relevance of the clusters was underlined by my observation that the samples within each group showed differential responses to drugs *in vitro* (Figure 4.8).

As expected, the IGHV-U enriched clusters C1 and 2 were more sensitive to BCR inhibition by ibrutinib, idelalisib and PRT062607, than C3 and 4. Between C1 and C2,

**Table 4.1:** Table depicting results of Multivariate Cox Proportional Hazard Model to test prognostic value of key genetic features and clusters using Time to Next Treatment and C3 as reference.

| Factor | coef | exp(coef) | se(coef) | z | p.value |
|---|---|---|---|---|---|
| Cluster 3 vs Cluster 1 | -0.03979 | 0.96099 | 0.29813 | -0.13347 | 0.89382 |
| Cluster 3 vs Cluster 2 | 0.51595 | 1.67522 | 0.37741 | 1.36708 | 0.1716 |
| Cluster 3 vs Cluster 4 | -0.82011 | 0.44038 | 0.39760 | -2.06267 | 0.03914 |
| IGHV.status | 0.55192 | 1.73658 | 0.27253 | 2.02513 | 0.04285 |
| trisomy 12 | -0.13357 | 0.87496 | 0.35617 | -0.37503 | 0.70764 |
| TP53 | 1.38977 | 4.01395 | 0.26072 | 5.33058 | <0.0001 |

C2 was more sensitive to a number of the drugs, including idelalisib (SYK) (p-value = 0.012), everolimus (mTOR) (p-value = 0.02) and the chemotherapeutics fludarabine (p-value = 0.031) and nutlin-3a (p-value = 0.042). Amongst C3 and C4, C3 showed lower sensitivity to everolimus (p-value = 0.051) and to fludarabine (p-value < 0.001) and nutlin-3a (p-value = 0.01). This aligns with the observation that patients in C3 have a poorer prognosis, despite most of these samples annotated as IGHV-M. C4 also showed a positive repsonse to Nf$\kappa$B inhibition by BAY-11-7085, and p38 MAPK inhibition by Ralimetinib.

Such differential drug response patterns suggests that microenvironmental response may reflect disease-specific CLL biology, in the same way as molecular profiling, and thus may have the potential to guide therapy decisions in future.

### 4.2.4 The clusters are enriched for different genetic features

Next we assessed differences in the molecular profiles of samples within each cluster. Visually, it appeared that certain clusters were enriched or depleted for various genetic features recurrent in CLL (Figure 4.9.

To quantify this, I ran a multinomial model, with lasso regularisation, to predict cluster

membership (C1 -4) based on the matrix of genetic features for all the patient samples (Bruch & Giles et al 2021). The model assigned coefficients to genetic features, where a positive coefficient indicated that this feature was enriched in the cluster, and a negative coefficient indicated it was depleted.

The approach used 3-fold cross validation, selecting the optimal model using lamda min. To ensure that the resulting coefficients were robust, we selected coefficients that satisfied certain cut-offs. Coefficients were selected if they were assigned in $> 60\%$ of 50 bootstrapped repeats, and were larger than 0.35. Figure 4.10 shows the mean coefficients and associated standard deviation, for each genetic feature that met these criteria in each cluster.

As we expected, IGHV status was the main feature that predicted cluster membership. Beyond IGHV status, trisomy 12 and *SF3B1* mutations were enriched in C2, which showed enhanced responses to many stimuli. C4, which was associated with slow in-vivo progression, showed depletion of *TP53*, *ATM*, RAS/RAF mutations and gain8q.

### 4.2.5 GSEA of DE genes between subgroups

In addition to genetic features, I investigated differential expression of genes within each cluster. For n = 49 samples, RNAseq data was available for matched PBMC samples. I focused on the difference between clusters 3 and 4, for which 21 RNAseq samples were available (Bruch & Giles et al. 2021).

To quantify differential gene expression, I began by filtering out immunoglobulin genes, including genes at the heavy, light and kappa loci that encode the antigen receptor of B cells. The clusters each show differential enrichment of IGHV-M and U samples, and thus the differential expression analysis would otherwise be dominated by immunoglobulin genes that are well known to affected by this biomarker.

I followed the Deseq2 protocol using a design formula to quantify the difference between clusters, and accounting for the confounding effect of IGHV status. 87 genes were differentially expressed (adjusted $p < 0.05$) between C3 and 4 (Figure 4.11).

To assign biological meaning to the differentially expressed genes, I quantified the enrichment of Hallmark pathways amongst the genes. I ranked the genes based on the Wald statistic, and then ran GSEA was using the fgsea algorithm (Figure 4.12) (Bruch & Giles et al. 2021).

Several pathways were upregulated amongst samples in C3, compared to C4, indicating that these pathways may relate in some way to the shorter TTT and LDT of patients within this cluster. Pathways associated with higher disease aggression were regulated in C3 including genesets relating to stress response (Unfolded Protein Response, UV Response Up, P53 Pathway), metabolism (Oxidative Phosphorylation) and proliferation (G2M Checkpoint, MYC Targets V1, MTORC1 Signaling, E2F Targets) (Figure 4.13).

In addition, C3 showed upregulation of microenvironmental signalling pathways relative to C4, including TNFa Signalling via NFKB and Interferon Gamma Response (Figure 4.13) (Bruch & Giles et al. 2021). This finding underlines our hypothesis that differential activity of microenvironmental signalling, both *in vivo* and *ex vivo* may be relevant to disease prognosis.

## 4.3 Additional Analysis

Can microenvironmental response aid prognostic models? Run a Multivariate model to predict survival/drug response with genetic matrix and stimuli response matrix?

## 4.4 Summary

The screen represents an attempt to comprehensively dissect the impact of individual microenvironmental pathways on CLL viability. Our assay has enabled us to highlight key, broad spectrum signals such as IL4. Our heterogeneous cohort also reveals pathways that operate in subsets of patients, such as TLR. IL4 and TLR represent the key pathways that we focus on throughout the rest of this thesis.

In addition, this approach enables us to use microenvironmental response profiles, as an alternative to molecular profiling, to look for the existence of subgroups.

We uncovered four such subgroup with distinct response profiles and molecular properties and clinical outcomes, suggesting that microenvironmental response holds biologically significant information that may be relevant to prognosis and treatment decision making.

## 4.5 Discussion

Gosia suggested discussing each results section individually?

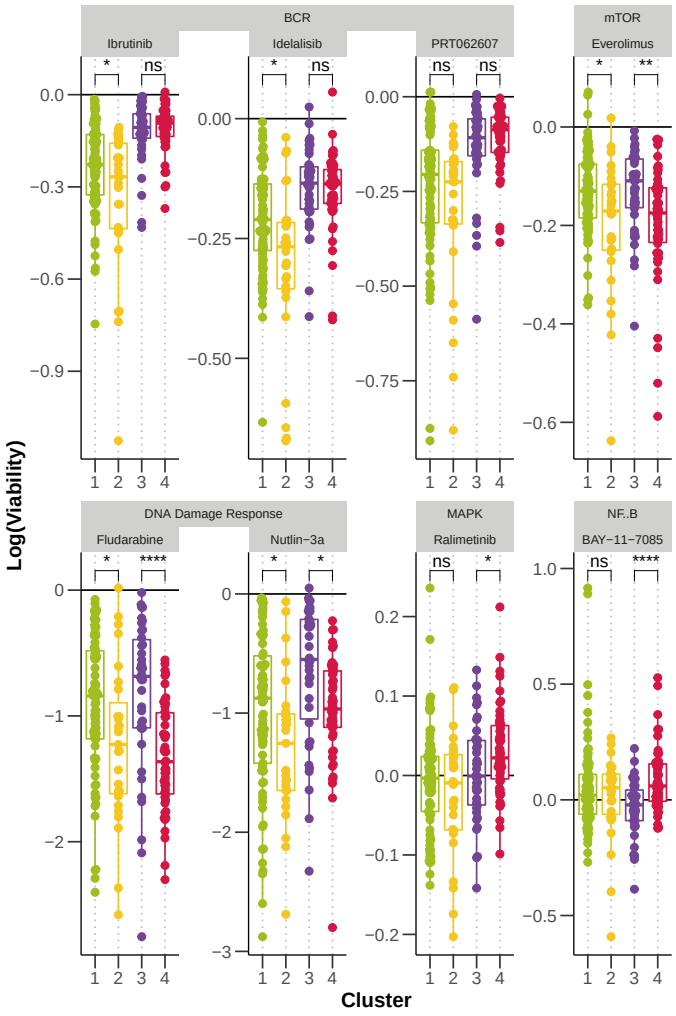


**Figure 4.8:** Log-transformed normalised viability values, stratified by cluster, for each drug. Drugs targeting the same pathway are grouped together. P-values from Student't t-test.

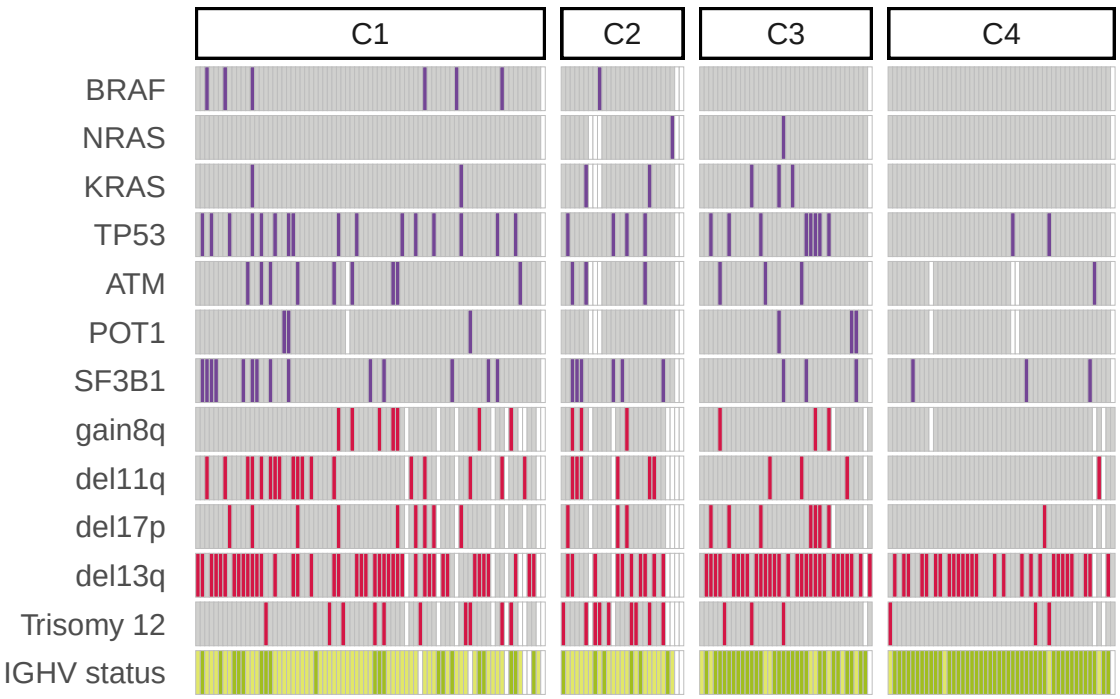


**Figure 4.9:** Distribution of selected genetic features (rows) within each cluster for all patient samples (columns). Where a patient sample is not annotated for a feature, this is marked in white.

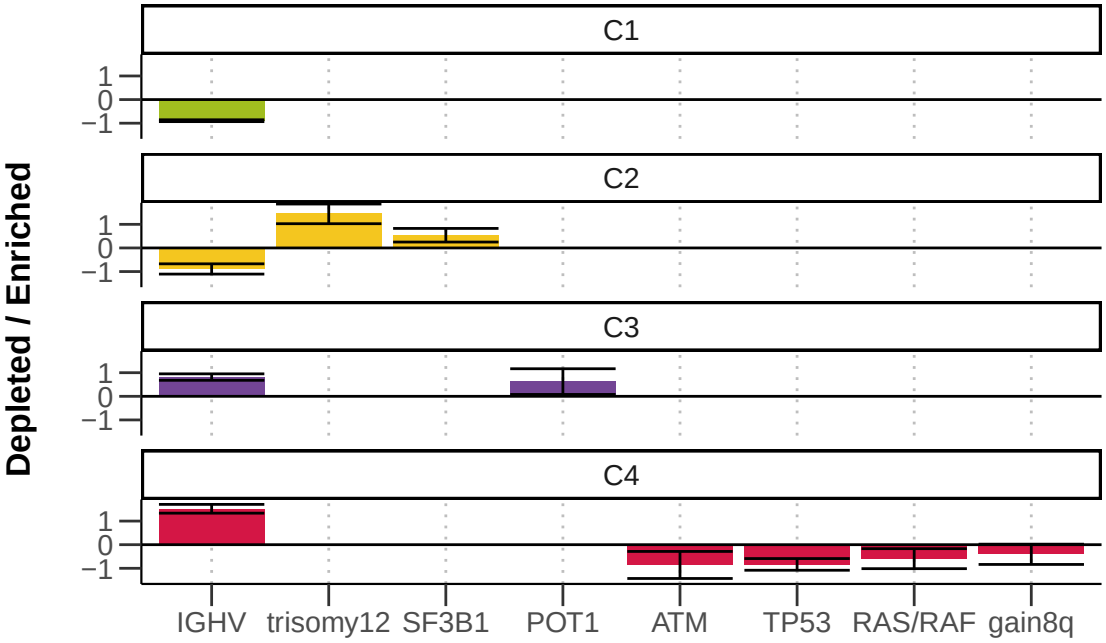


**Figure 4.10:** Multinomial regression with lasso penalisation to identify enrichment or depletion of genetic features within each cluster. Matrix of genetic features (p=39), and IGHV status (encoded as M = 1 and U = 0) were used to identify multivariate predictors of cluster assignment. x axis shows genetic predictors, y axis indicates value and sign of coefficient assigned to feature, for each cluster (positive coefficients are enriched in the cluster, negative coefficients are depleted). Coefficients shown are mean coefficients from 50 bootstrapped repeats and error bars represent the mean ś standard deviation. Genetic features with >20% missing values were excluded, and only patients with complete annotation were included in the model (n=137). Figure from Bruch & Giles et al 2021.

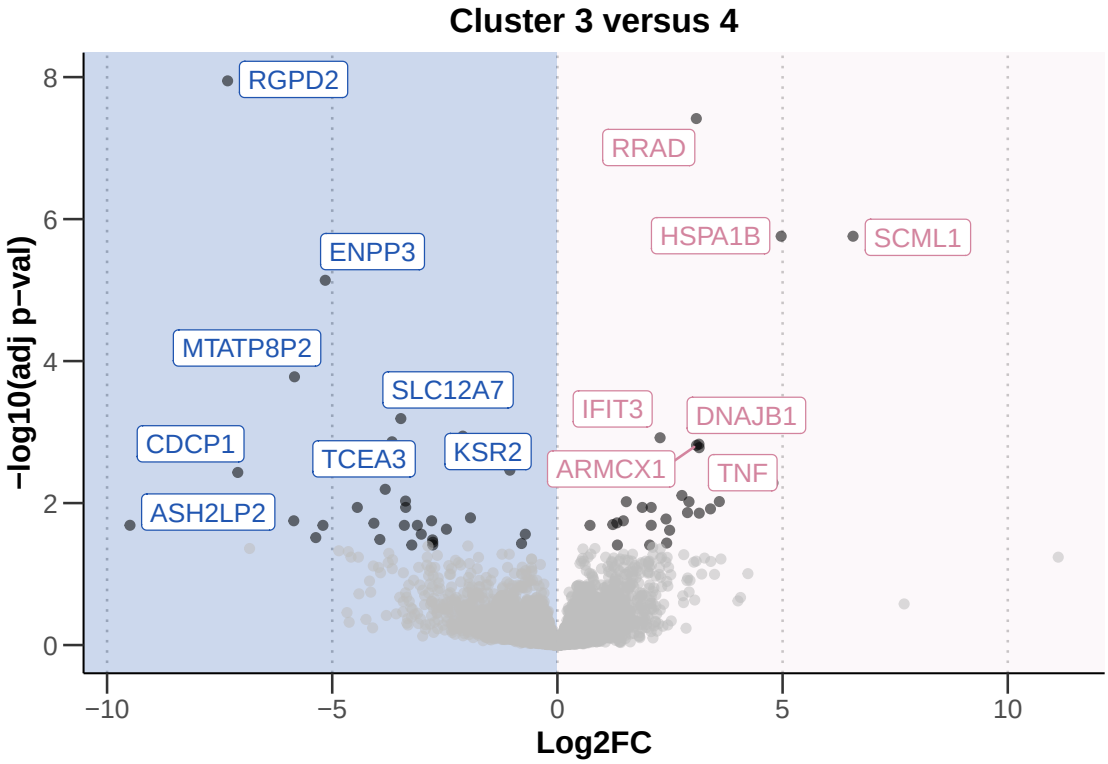


**Figure 4.11:** Volcano plot of differentially expressed genes between C3 and C4. X axis indicates log2 fold change values, calculated using the `DESeq2` package (**DESeq2?**), y axis gives corresponding -log10(adjusted p value). P values adjusted using BH method. Genes are labeled where ajdusted $p < 0.05$. Figure from Bruch & Giles et al. 2021.

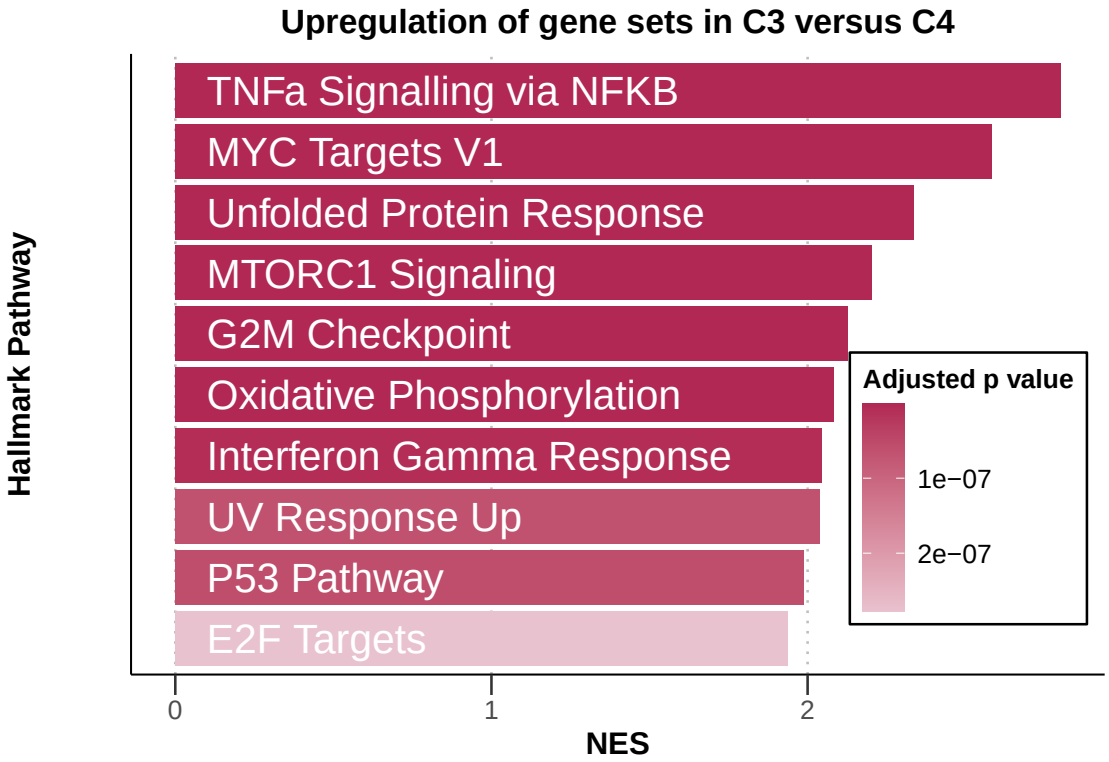


**Figure 4.12:** Gene set enrichment analysis (GSEA) to compare expression of genes in samples from C3 and C4 reveals upregulation of Hallmark pathways involved in microenvironmental signalling, stress response, metabolism and proliferation in C3. Normalised enrichment scores (NES) are shown for top 10 most significant pathways upregulated in C3 versus C4. Bars coloured according to adjusted p-value. Genes are ranked based on Wald statistic, calculated using the Deseq2 package and GSEA performed using the fgsea algorithm. Figure from Bruch & Giles et al. 2021.

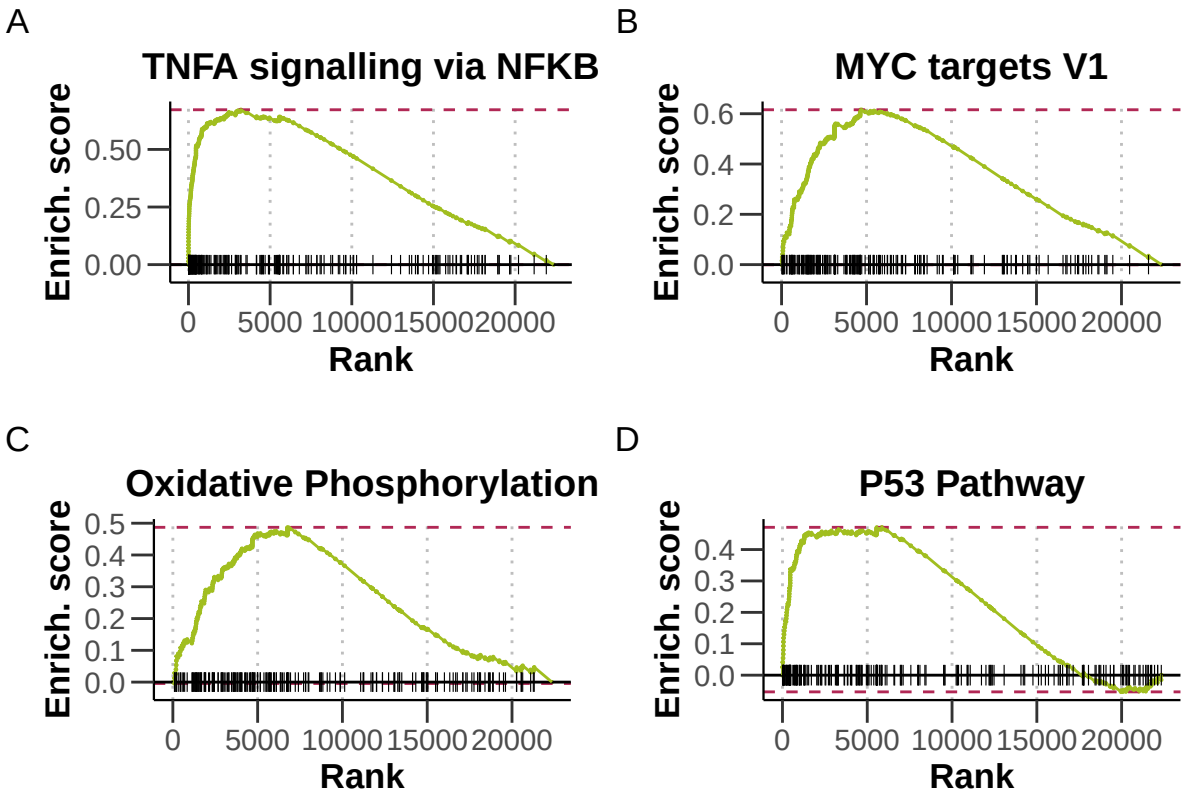


**Figure 4.13:** Enrichment plots of selected pathways. Gene set enrichment analysis (GSEA) was performed with the Hallmark gene sets from the GSEA Molecular Signatures Database. Wald statistic was used to rank the genes. The green curve corresponds to the Enrichment Score curve, which is the running sum of the weighted enrichment score obtained from GSEA software. Figure from Bruch & Giles et al. 2021

# Chapter 5

# Genetic modulators of responses to microenvironmental stimulation

To continue: add the additional plots text and plot arrangement also update the cytokine labelling and gene labelling any additional plots text dump

Continute to figure 6 (see at clsoe of tomorrow if you are better of leaveing this till i get back, then only one week behind)

## 5.1 Systematic analysis of the effect of genetic features on responses to stimuli:

### 5.1.1 Univariate analysis identifies IGHV status and tirsomy 12 as key modulators

(Figure 5.1).

(ref:stimuliGeneAssosciations)

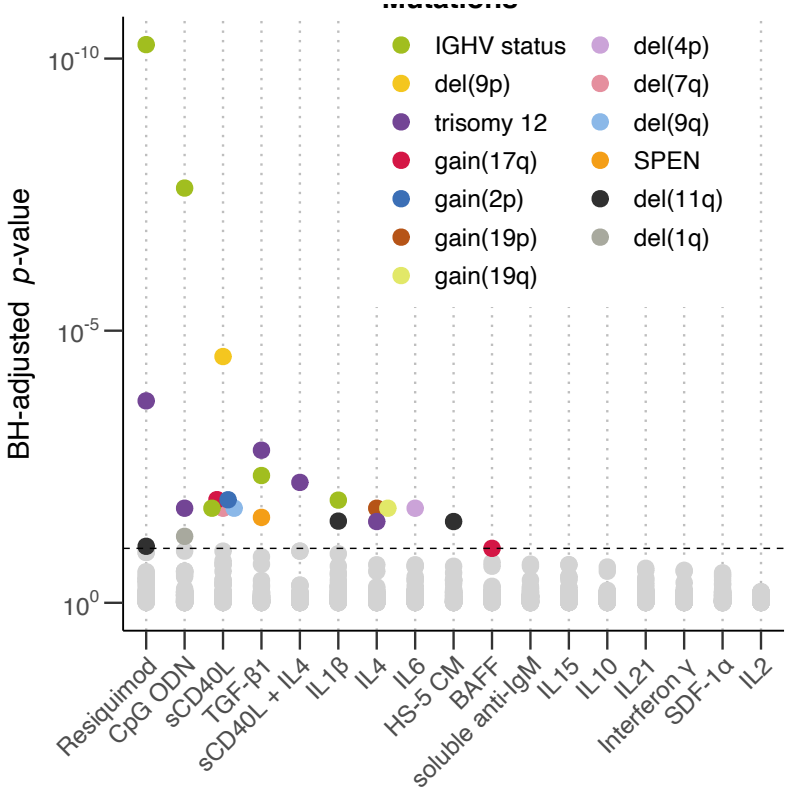


**Figure 5.1:** (ref:stimuliGeneAssosciations)

### 5.1.2 Multivariate analysis demosntrates multiple layers of biology involved

(Figure 5.2).

(ref:stimuliGeneAssosciationsMulti)

### 5.1.3 Individual cytokine – gene interactions of note

Role of IGHV status and BCR signalling in Cytokine response (Figure 5.3).

(ref:IHGVBCR) TLR response stratified by IGHV status (and trisomy 12) and discussion of interactions between BCR and TLR pathways (Figure 5.4).

(ref:TLRIHGVtri12) -Modulators of TLR signalling incl. effect of del11q and ATM mutations on TLR response (Figure 5.5).

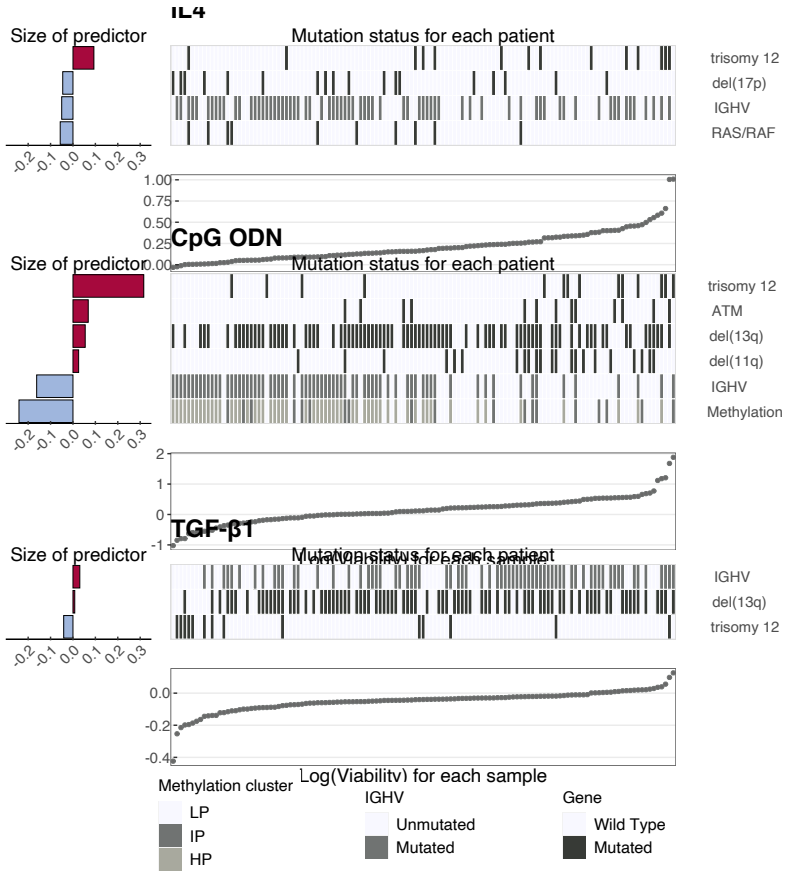


**Figure 5.2:** (ref:stimuliGeneAssosciationsMulti)

(ref:TLRGenes)

Potentially show effect of IGHV, then ivnestgiation of TLR combined boave and belwo, then look intro tri 12

These plots are not working right now - can work on them if i haev time -samples with KRAS mutations benefit less from IL4 stimulation

(Figure 5.6).

(ref:IL4KRAS)

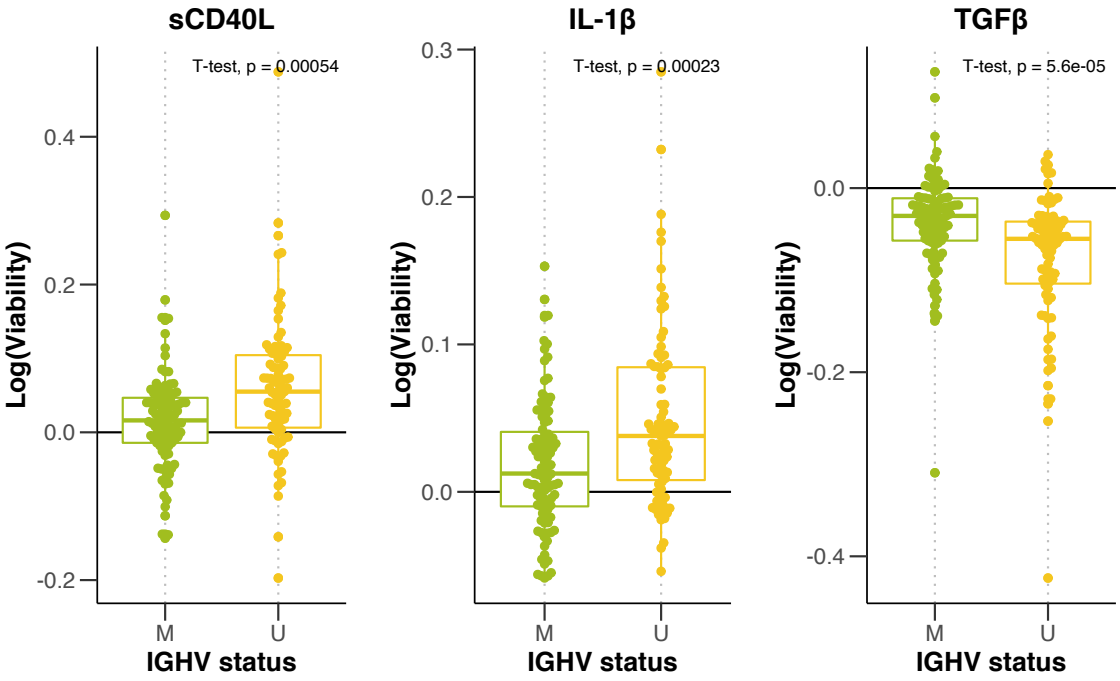


**Figure 5.3:** (ref:IHGVBCR)

## 5.2 Profiling the effects of trisomy 12 on cytokine response, and how this may be possible

### 5.2.1 Trisomy 12 modulates responses to IL4, TGFbeta and TLR stimuli

(Figure 5.8).

(ref:SMAD3geneDosage)

### 5.2.2 Gene dosage effects

(Figure ??.

(ref:tri12geneDosage)

*The Proteomics dataset was kindly shared by Sophie Herbst, from her paper...* (Figure 5.9).

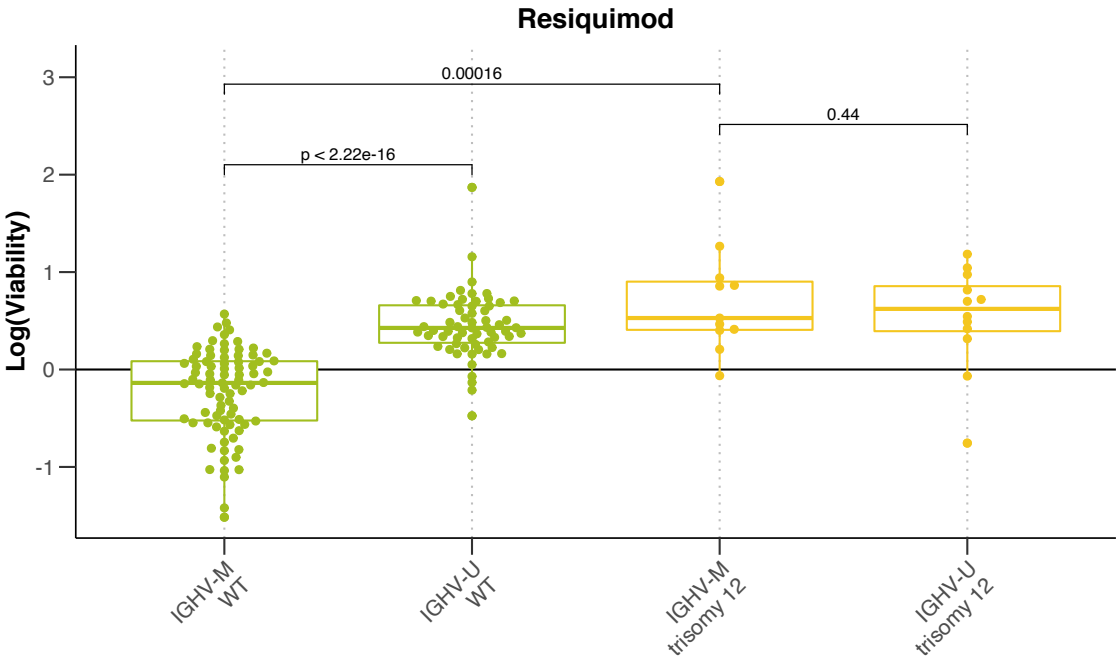


**Figure 5.4:** (ref:TLRIHGVtri12)

(ref:STATgeneDosage) (Figure 5.10).

(ref:IRAK4geneDosage)

### 5.2.3 Investigation of non-trisomy12 samples that are phenocopies of trisomy12 CLL+ (based on their responses to stimuli)

(Figure 5.11.

(ref:tri12Classfier) (Figure 5.12.

(ref:tri12PhenocopiesCytResponse) (Figure 5.13.

(ref:tri12PhenocopiesRNA) Pat033 has duplicated regions: -p13.31 - 42 copies -q24.13 - 10 copies -q24.33 - 21 copies See what Junyan says about this and potentially add a figure

(Figure **??**.

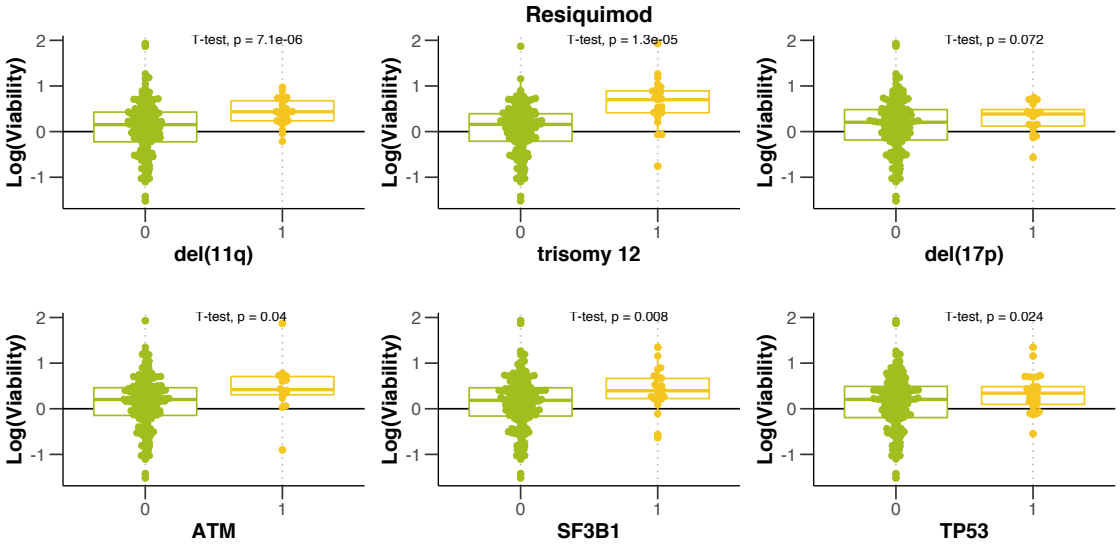


**Figure 5.5:** (ref:TLRGenes)

(ref:tri12PhenocopiesCNV)

### 5.2.4 Investigation of TF activity in trisomy12+ CLL and finding that SPIB and PU1 TFs are upregulated in trisomy12+ CLL

tri 12 annotation

running diffTF, add a diagram to explain?

(Figure 5.14.

(ref:diffTFexplainer) visualisation of results

(Figure **??**.

(ref:tri12annotation)

(Figure 5.15.

(ref:tri12diffTF)

Also refer to : (Figure 8.1

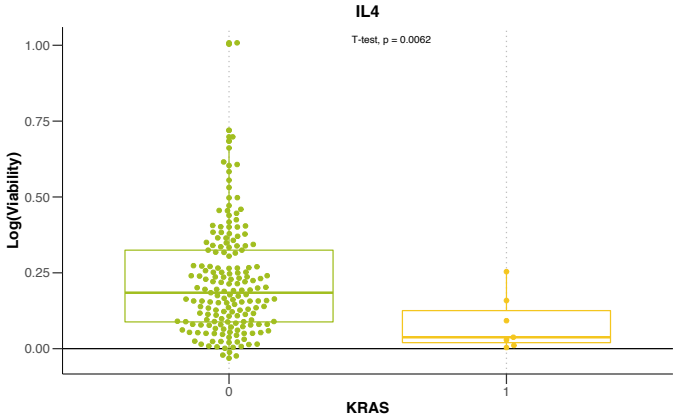


**Figure 5.6:** (ref:IL4KRAS)

## 5.3 Profiling the downstream effects of SPIB and PU1:

### 5.3.1 Over-representation of BCR, TLR, interleukin and TGFbeta response genes amongst SPIB and PU1 targets in ChIPseq data

(Figure **??**.

(ref:SpiBChipSeq)

### 5.3.2 Visualisation of SPIB and PU1 binding sites from ChIPseq data in BCR, TLR, interleukin and TGFbeta signalling genes

(Figure **??**.

(ref:SpiBBindingSites)

### 5.3.3 Investigation of DE genes within IL4 and TGFbeta pathways in SPIB/PU1 DKO mice

(Figure **??**.

(ref:SpiBDEgenes)

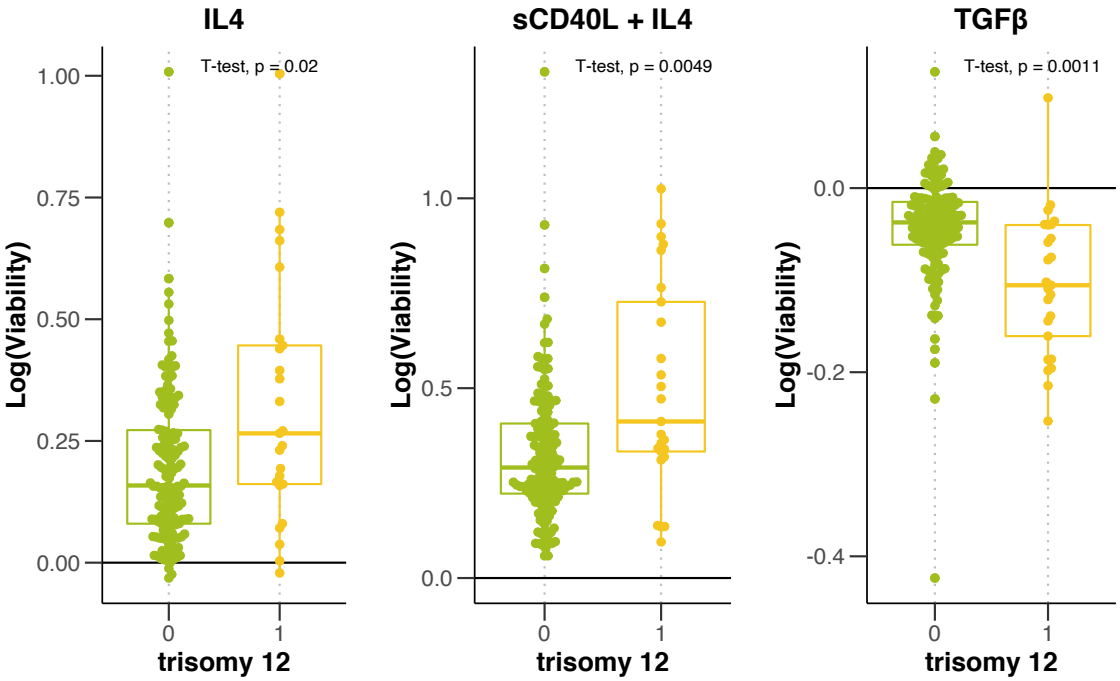


**Figure 5.7:** (ref:tri12cytResponse)

### 5.3.4 SPIb and PU1 linked genes

Key result: generation of GRN to identify SPIB and PU1 regulated genes, and investigation of associated log2FCs of these genes in trisomy 12 vs non-trisomy12 samples (including CD79A and B)

(Figure **??**.

(ref:SpiBGRN)

### 5.3.5 SPIB/PU1 DKOs in trisomy 12+ cell lines

(Figure 5.16.

(ref:SpiBshRNAKD)

*The data here was generated by Tina Bericoviv and was included in Bruch & Giles et al. 2021. It is included in this thesis for completeness.*

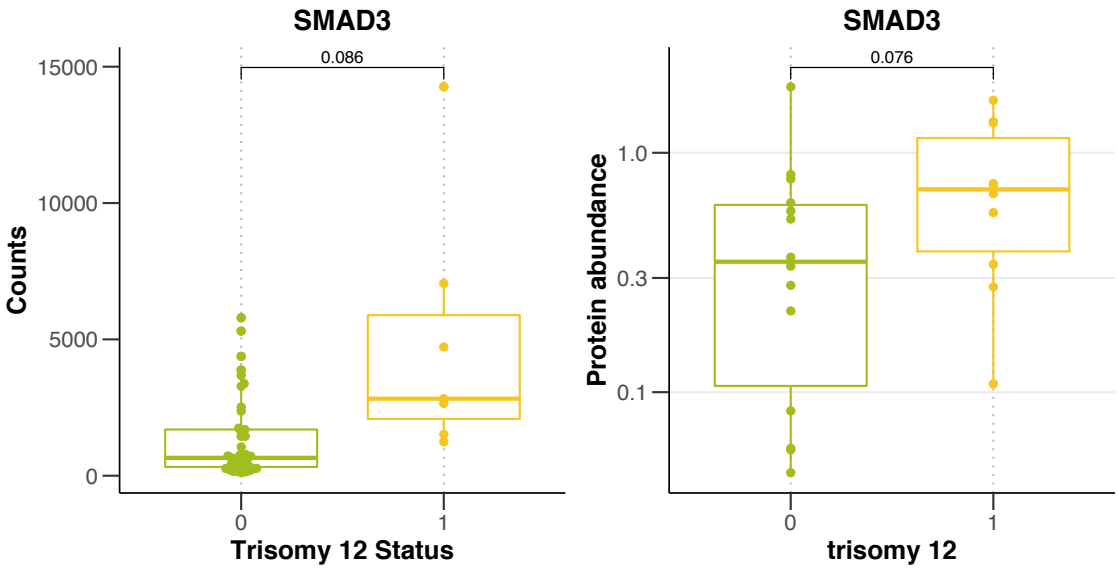


**Figure 5.8:** (ref:SMAD3geneDosage)

## 5.4 Conclusions

so everything is here focus on the content you have and finihsh - add additioanl if time

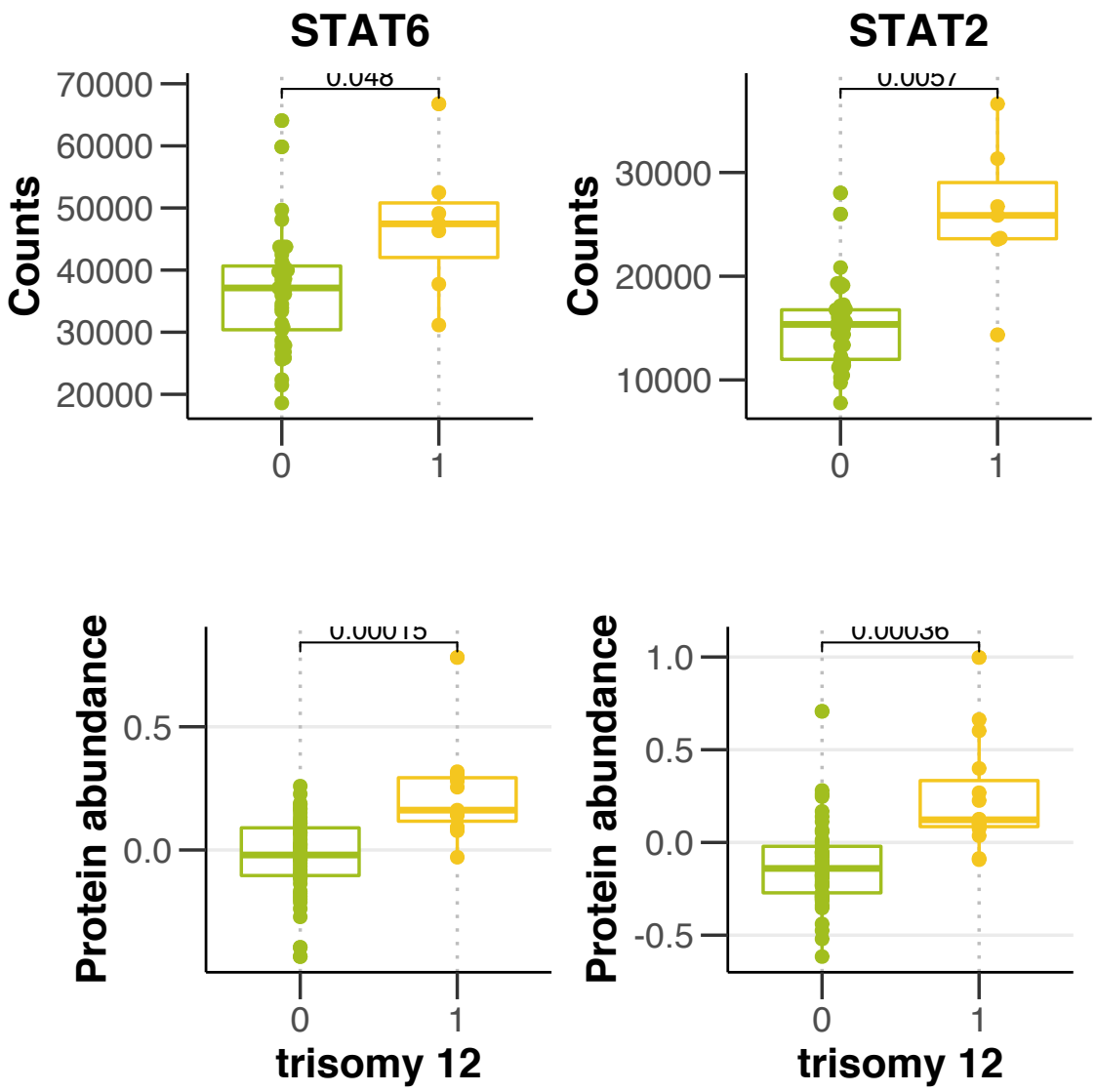


**Figure 5.9:** (ref:STATgeneDosage)

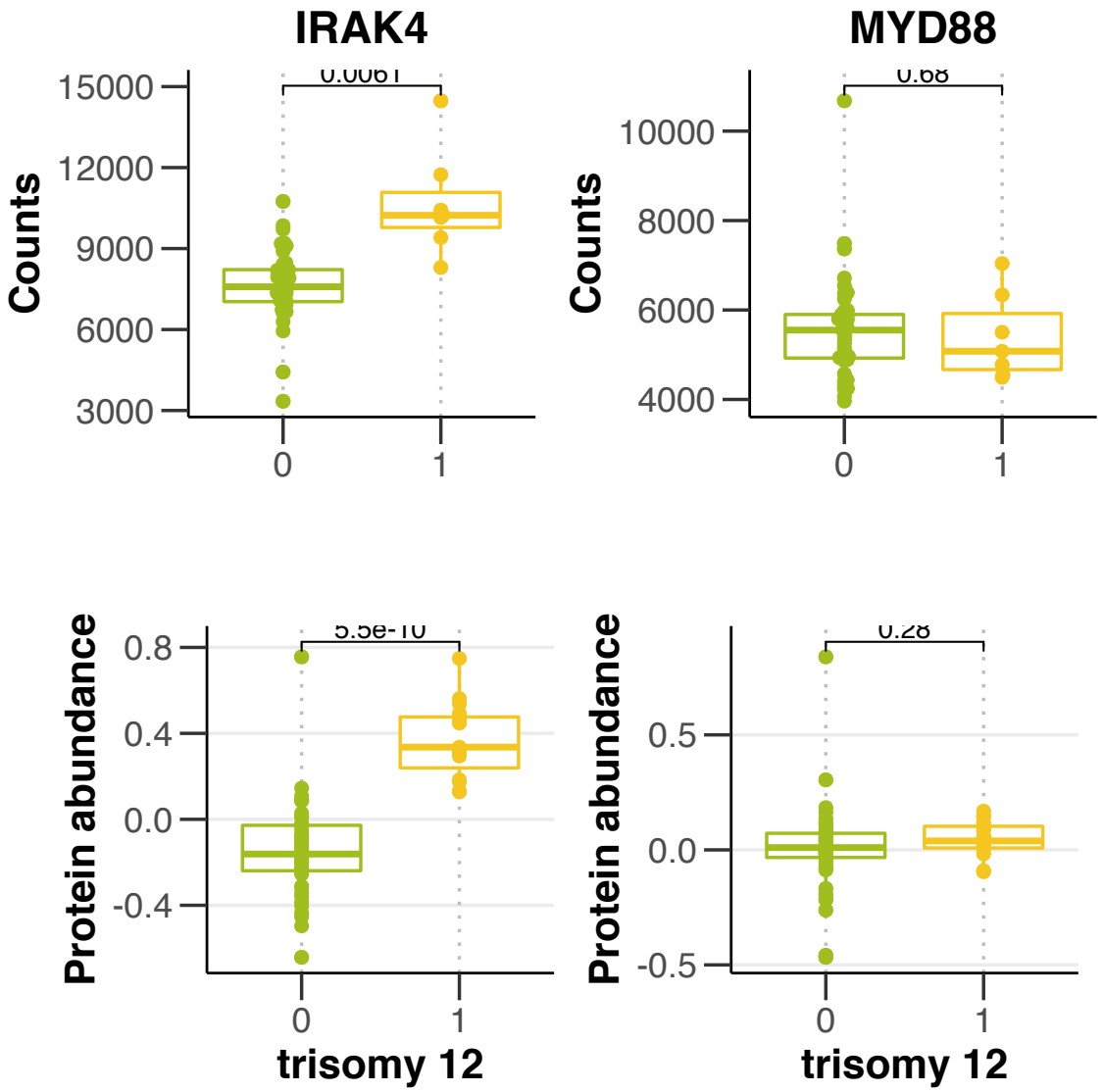


**Figure 5.10:** (ref:IRAK4geneDosage)

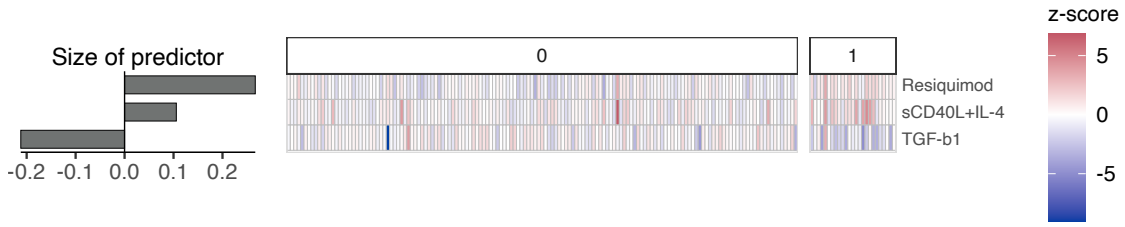


**Figure 5.11:** (ref:tri12Classfier)

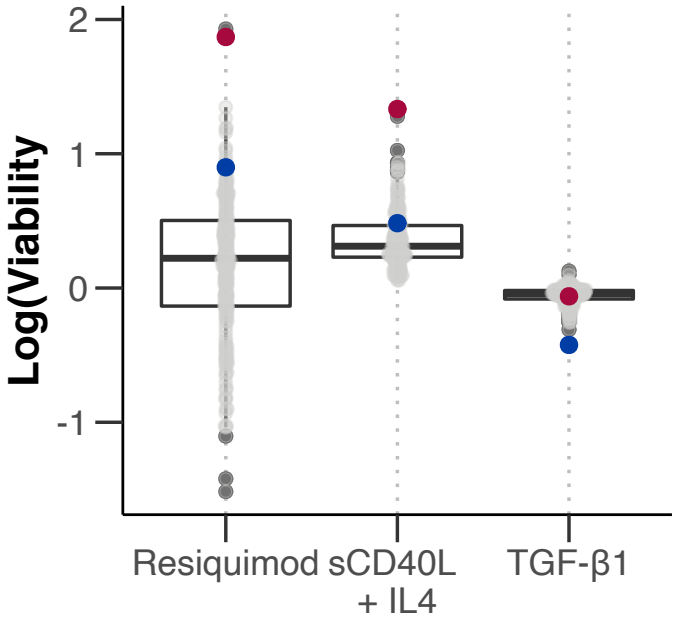


**Figure 5.12:** (ref:tri12PhenocopiesCytResponse)

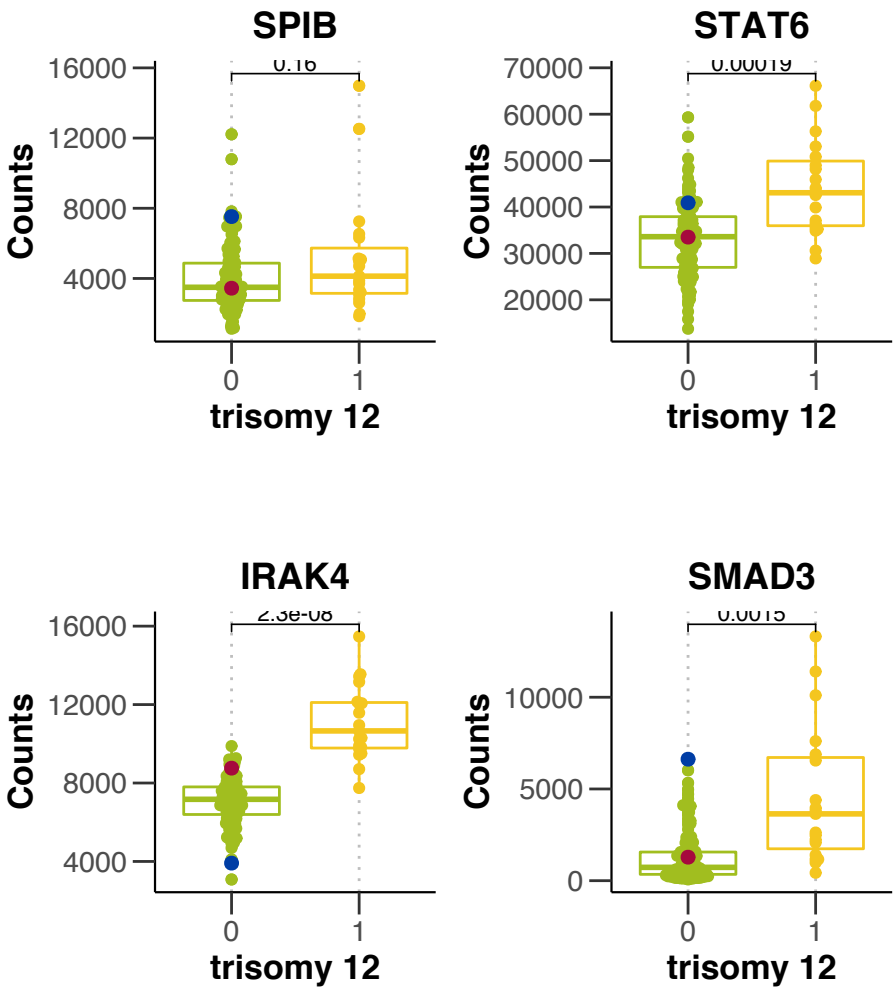


**Figure 5.13:** (ref:tri12PhenocopiesRNA)

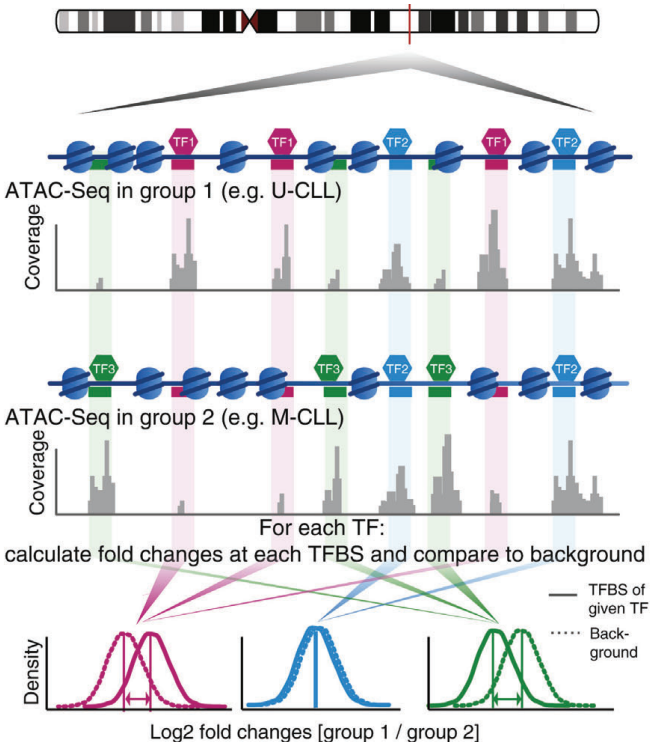


**Figure 5.14:** (ref:diffTFexplainer)

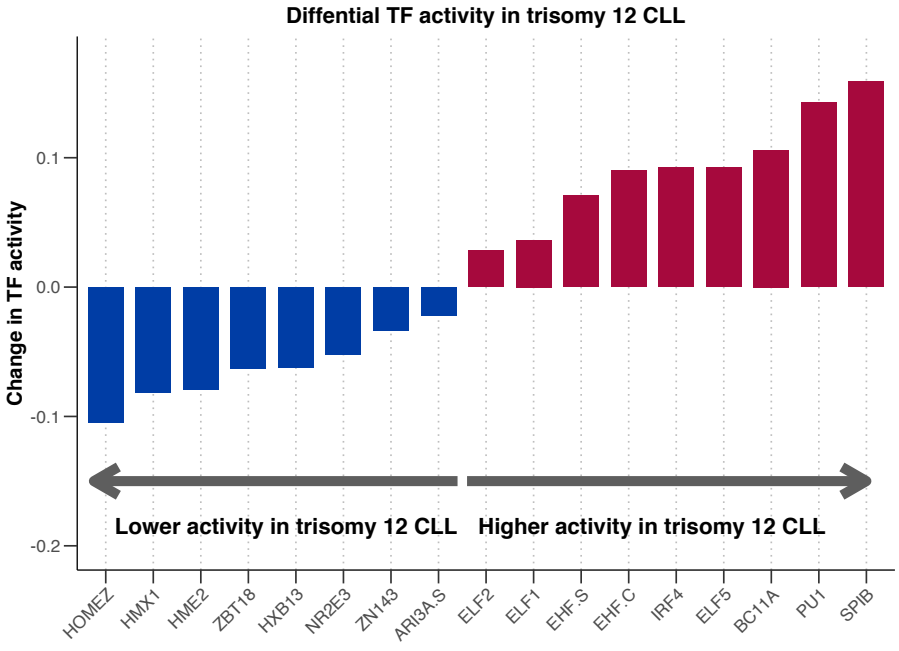


**Figure 5.15:** (ref:tri12diffTF)

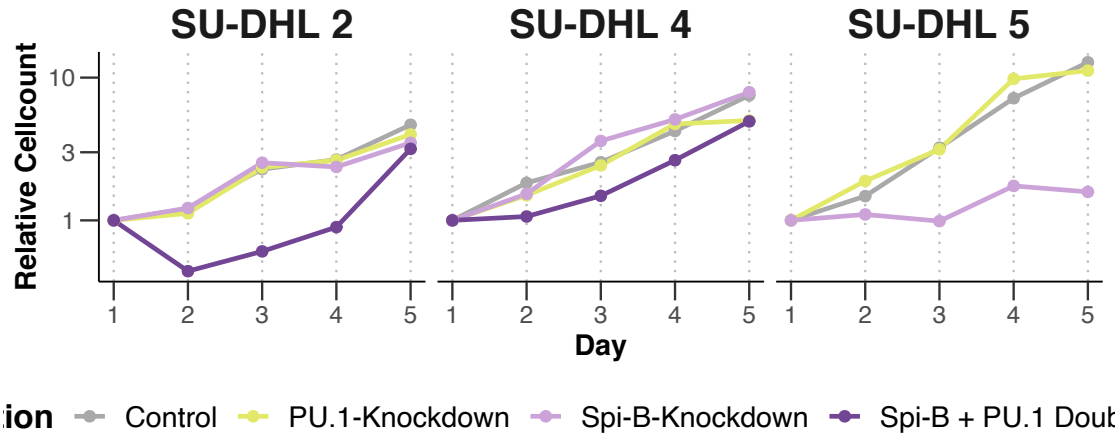


**Figure 5.16:** (ref:SpiBshRNAKD)

# Chapter 6

# Results

- Organize material and present results.
- Use tables, figures (but prefer visual presentation):
  - Tables and figures should supplement (and not duplicate) the text.
  - Tables and figures should be provided with legends.
  - *Figure* **??** *shows how to include and reference graphics. The graphic must be labelled before. Files must be in* ***.eps*** *format. You can do this really easily in R Markdown with* ***`knitr::include_graphics()`***!
  - Figures can be referenced with `\@ref(fig:<name>)`, where `<name>` is the name of the code chunk.
- Tables and graphics may appear in the text or in the appendix, especially if there are many simulation results tabulated, but is also depends on the study and number of tables resp. figures. The key graphs and tables must appear in the text!
- R Markdown can also supports math equations just like *LaTeX*!

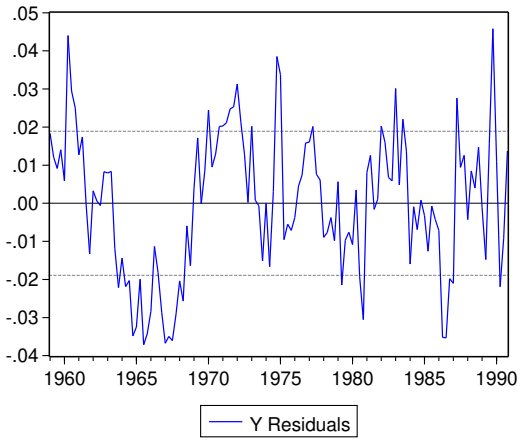


**Figure 6.1:** Estimated residuals from model XXX3. ...

– *Equation* (7.1) *represents the ACs of a stationary stochastic process:*

$$f_y(\lambda) = (2\pi)^{-1} \sum_{j=-\infty}^{\infty} \gamma_j e^{-i\lambda j} = (2\pi)^{-1} \left( \gamma_0 + 2 \sum_{j=1}^{\infty} \gamma_j \cos(\lambda j) \right) \tag{6.1}$$

*where* $i = \sqrt{-1}$ *is the imaginary unit,* $\lambda \in [-\pi, \pi]$ *is the frequency and the* $\gamma_j$ *are the autocovariances of* $y_t$*.*

– Equations can be referenced with `\@ref(eq:<name>)`, where name is defined by adding `(\#eq:<name>)` in the line immediately before `\end{equation}`.

## 6.1 Review of Results

- Do the results support or do they contradict economic theory ?
- What does the reader learn from the results?
- Try to give an intuition for your results.
- Provide robustness checks.
- Compare to previous research.

# Chapter 7

# Results

- Organize material and present results.
- Use tables, figures (but prefer visual presentation):
  - Tables and figures should supplement (and not duplicate) the text.
  - Tables and figures should be provided with legends.
  - *Figure* **??** *shows how to include and reference graphics. The graphic must be labelled before. Files must be in* ***.eps*** *format. You can do this really easily in R Markdown with* ***`knitr::include_graphics()`***!
  - Figures can be referenced with `\@ref(fig:<name>)`, where `<name>` is the name of the code chunk.
- Tables and graphics may appear in the text or in the appendix, especially if there are many simulation results tabulated, but is also depends on the study and number of tables resp. figures. The key graphs and tables must appear in the text!
- R Markdown can also supports math equations just like *LaTeX*!

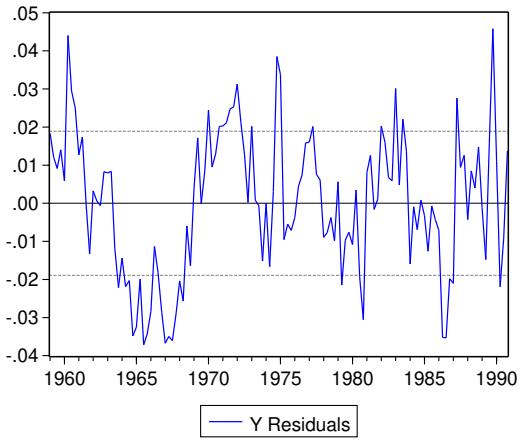


**Figure 7.1:** Estimated residuals from model XXX4. ...

– *Equation* (7.1) *represents the ACs of a stationary stochastic process:*

$$f_y(\lambda) = (2\pi)^{-1} \sum_{j=-\infty}^{\infty} \gamma_j e^{-i\lambda j} = (2\pi)^{-1} \left( \gamma_0 + 2 \sum_{j=1}^{\infty} \gamma_j \cos(\lambda j) \right) \tag{7.1}$$

*where* $i = \sqrt{-1}$ *is the imaginary unit,* $\lambda \in [-\pi, \pi]$ *is the frequency and the* $\gamma_j$ *are the autocovariances of* $y_t$*.*

– Equations can be referenced with `\@ref(eq:<name>)`, where name is defined by adding `(\#eq:<name>)` in the line immediately before `\end{equation}`.

## 7.1 Review of Results

- Do the results support or do they contradict economic theory ?
- What does the reader learn from the results?
- Try to give an intuition for your results.
- Provide robustness checks.
- Compare to previous research.

# Chapter 8

# Discussion

# References

10 Collins, Russell J., Louise A. Verschuer, Brian V. Harmon, Roger L. Prentice, John H. Pope, and John F. R. Kerr. 1989. "Spontaneous programmed death (apoptosis) of Bchronic lymphocytic leukaemia cells following their culture in vitro." *British Journal of Haematology* 71 (3): 343–50. https://doi.org/10.1111/j.1365-2141.1989.tb04290.x.

Dietrich, Sascha, Magorzata Ole, Junyan Lu, Leopold Sellner, Simon Anders, Britta Velten, Bian Wu, et al. 2017. "Drug-perturbation-based stratification of blood cancer." *Journal of Clinical Investigation* 128 (1): 427–45. https://doi.org/10.1172/JCI93801.

Wilkerson, Matthew D, and D Neil Hayes. 2010. "ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking." *Bioinformatics (Oxford, England)* 26 (12): 1572–73. https://doi.org/10.1093/bioinformatics/btq170.

# Appendix

## Figures

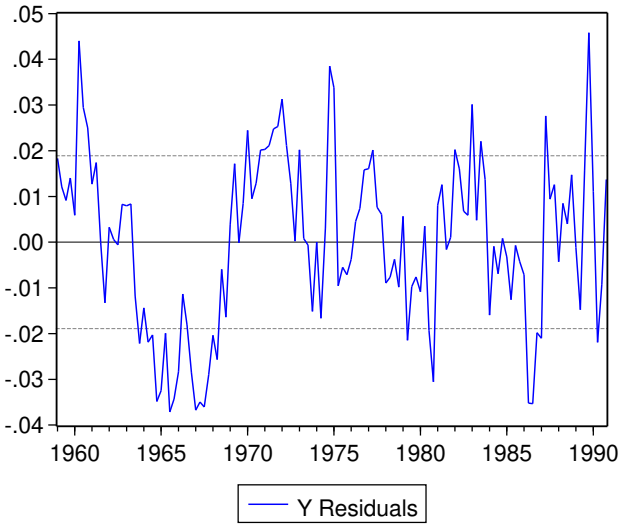


**Figure 8.1:** Analysis of ATACseq dataset of two trisomy 12 and two non-trisomy 12 untreated CLL PMBC samples. The volcano plot depicts change in TF activity (x axis) versus BH-adjusted p-values (y axis) for 636 TFs, comparing trisomy 12 and non-trisomy 12 samples. The (**diffTF?**)[REFERENCE] package was ran in analytical mode to calculate TF activity, measured as weighted mean difference. TFs are labeled if adjusted p-value $< 0.01$ and absolute weighted mean difference $> 0.15$.

# Tables

**Table 8.1:** Detailed descriptive statistics of location and dispersion for 2100 observed swap rates for the period from February 15, 1999 to March 2, 2007. Swap rates measured as 3.12 (instead of 0.0312).

| | 3m | 6m | 1yr | 2yr | 3yr | 5yr | 7yr | 10yr | 12yr | 15yr |
|---|---|---|---|---|---|---|---|---|---|---|
| Mean | 3.138 | 3.191 | 3.307 | 3.544 | 3.756 | 4.093 | 4.354 | 4.621 | 4.741 | 4.878 |
| Median | 3.013 | 3.109 | 3.228 | 3.490 | 3.680 | 3.906 | 4.117 | 4.420 | 4.575 | 4.759 |
| Min | 1.984 | 1.950 | 1.956 | 2.010 | 2.240 | 2.615 | 2.850 | 3.120 | 3.250 | 3.395 |
| Max | 5.211 | 5.274 | 5.415 | 5.583 | 5.698 | 5.805 | 5.900 | 6.031 | 6.150 | 6.295 |
| StD | 0.915 | 0.919 | 0.935 | 0.910 | 0.876 | 0.825 | 0.803 | 0.776 | 0.768 | 0.762 |