INSTITUT FÜR KARDIOGENETIK

Identification of genetic risk variants for atherosclerosis using oxidative stress assays in vascular smooth muscle cells and bioinformatic approaches

Identification genetischer Risikovarianten für Artheriosklerose via oxidativem Stress Assay in glatten Muskulaturzellen und bioinformatische Ansätze

Masterarbeit

verfasst am
Institut für Kardiogenetik

im Rahmen des Studiengangs
Molecular Life Science
der Universität zu Lübeck

vorgelegt von
Torben Falk

ausgegeben und betreut von
Prof. Dr. Jeanette Erdmann

mit Unterstützung von
Dr. Tobias Reinberger

Lübeck, den 21. Juli 2022

Eidesstattliche Erklärung

Ich erkläre hiermit an Eides statt, dass ich diese Arbeit selbständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe.

_____

Torben Falk

Zusammenfassung

Ich muss das Ding wohl irgendwann auch noch in Deutsch schreiben...

Abstract

Placeholder, this is more or less what I am doing:

I am currently writing my Master thesis at the university of Lübeck at the Institute for Cardiogenetics on the topic of "Identification of genetic risk variants for atherosclerosis using oxidative stress assays in vascular smooth muscle cells and bioinformatic approaches":

Coronary artery disease (CAD) describes the arterial build-up of fatty deposits to a point where the blood supply to the heart gets interrupted. It is one of the major causes of death worldwide. Risk factors for CAD are typical lifestyle factors like smoking or physical inactivity, but also include hereditary factors (**cdcCoronaryArteryDisease2021**; **CoronaryHeartDisease2018**). These can provide access to the molecular pathology of the disease. One amazing resource for studying these interactions are genome wide association studies (GWAS). Unfortunately, GWAS are just the first step in a longer journey of establishing causal loci to gene links, uncovering the molecular basis of disease, and implementing tools for clinical risk prediction. A plethora of follow-up analyses (postGWAS) can and need to be performed (**lichouFunctionalStudiesGWAS2020a**).

We hypothesize that oxidative stress in smooth muscle cells plays a role in stability of atherosclerotic plaques. For this reason, I am cultivating and differentiating primary human smooth muscle cells and characterizing them using oxidative stress assay, qPCR, seahorse assay & immunofluorescence (IF).

Additionally, I am working with GWAS data on CAD (**aragamDiscoverySystematicCharacterization2021a**). Curating further publicly available data that can be used for bioinformatic follow-up analyses like the enrichment for involved tissues. Further, I am using the data to build a web application that allows co-visualization and visual exploration.

Acknowledgements

Daaanke an alle!

Contents

Abkürzungsverzeichnis

A adenin
AF488 Alexa Fluor® 488
bp basepairs
C cytosin
DNA deoxyribonucleic acid
dsDNA doublestranded DNA
E. coli *Eschericha coli*
EDTA Ethylenediaminetetraacetic acid
G guanine
HCl hydrogen chloride
qPCR quantative polymerase chain reaction
T thymin
HAoSMC human arortic smooth muscle cell
IF immunoflourescence

1

Introduction

1.1 Coronary artery disease

- CAD is serious. - Describe the pheontype - Describe prevalence - Check intro of Anja, papers for more info - Go into treatment and Risk - herditary parts in disease

1.2 GWAS

GWAS

What are GWAS. Describe how it works, why we do it.

Post GWAS

And their limitations. Possible follow up studies. Focus on computational and cell based assays.

1.3 Muscle Cells in CAD

- We now that smooth muscle cells play a key role - It is widely accepted that there is not only one type of smooth muscle cell - Go into the contractile phenotype and synthetic phenotype - TGFb or PDGF used to induce them - contractile phenotype is thought to be protective

1.4 PDGF Signaling and Oxidative Stress

PDGF Signaling

ROS

- ROS in general and the role of ROS in disease

ROS in PDGF Signaling

- ROS as a second messenger in PDGF signaling

1.5 Complementary High Through Put Methods

Linkage Disequilibrium

Linkage disequilibrium is a parameter from populations genetics that describes the non-random association of two or more allels. The LD is often quantified using the correlation coefficient r^2 (**slatkinLinkageDisequilibriumUnderstanding2008**).

$$D_{AB} = p_{AB} - p_A p_B$$

$$r^2 = \frac{D_{AB}^2}{p_A(1-p_A) \times p_B(1-p_B)}$$

Where pA and pB is the frequency of the alleles A and B respectively. pAB is the frequency of the AB haplotype.

Locus To Gene Scores

Regulatory Build

The ensembl regulatory build compiles a summary of regulatory regions found in the human genome. It is build on the basis publically available epigenetic marks and transcription factor binding and contains Promotors, Proximal enhancers, distal enhancers and CTCF binding sites (**zerbinoEnsemblRegulatoryBuild2015**).

ENCODE cCREs

scATAC Seq

ABC Model

TADs

1.6 Aim of the thesis

- Build tool for visual exploration of the CAD GWAS data. - Establish a system to test the role of ROS in CAD.

2

Material

2.1 Manufactors

Manufacturer	Seat
Agilent Technologies, Inc.	Santa Clara, CA, USA
BRAND GMBH & Co. KG	Wertheim, DE
Eppendorf AG	Hamburg, DE
Merck KGaA	Darmstadt, DE
Keyence Corporation	Osaka, JP
Sarstedt AG & Co.	Nürnberg, DE
Sigma-Aldrich Co. LLC.	St. Louis, MO, USA
Thermo Fisher Scientific Inc.	Waltham, MA, USA

2.2 Antibodies

Name	Species	Manufacturer
8oxoG	?	?
other oxStress	?	?
Fibronectin	?	?
secondary ones	?	?

2.3 Celllines

Name	Celltype	Manufacturer
Human Aortic Smooth Muscle Cell (HAoSMC)	prim. human cell	

2.4 Primer

Target	Name	Sequence
CNN1	Fw	5'-seq-3'
	Rv	5'-seq-3'
GAPDH	Fw	5'-seq-3'
	Rv	5'-seq-3'
MMP9	Fw	5'-seq-3'
	Rv	5'-seq-3'

2.5 Chemicals

Name	Manufacturer
5x First STardn buffer	
Antimycin	
BSA	
CellROX	
Collagen I	
dNTP	
DTT	
Ethanol	
FCCP	
Hoechst	
IL1	
M-MLVRT	
NAC	
NaHCO3	
NaOH	
Oligomycin	
Oligos	

Continued on next page

(Continued)

Name	Manufacturer
PBS	
PDGF	
RiboLick	
Seahorse XF calibrant	
SYBR GREEN Master Mix	
TGF	

2.6 Media, Supplements

Name	Manufacturer
Cryomedium	even necessary?
Fetal Bovine Serum (FBS)	?
Medium 231 (M231)	LIFE TECHNOLOGIES EUROPE BV
Smooth Muscle Cell Growth Supplement (SMGS)	?
XF Base Medium	Agilent Technologies, Inc

2.7 Solutions

Name	Manufacturer
Interleukin 1 beta (IL-1)	IL-1 BSA in PBS
N-Acetylcystein (NAC)	0.25 M NAC in water, ~pH 7
Platelet-derived growth factor-BB (PDGF-BB)	PDGF-BB BSA in PBS
Transforming Growth Factor beta (TGF)	TGF BSA in PBS

2.8 Kits

Kit	Manufacturer
Total RNA Purification Kit	Jena Bioscience GmbH

2.9 Consumables

Name	Manufacturer
0.2 mL tubes	
0.5 mL tubes	
1.5 mL tubes	
2 mL tubes	
5 mL tubes	
24-well plate	
Seahorse cultivation plate	
Seahorse sensor plate	
384 Well Multiply PCR plates	
pasteur pipettes	
tips gray	
tips yellow	
tips blue	
tips gray, filter	
tips yellow, filter	
tips blue, filter	
syringe	
filter	
Wiege Schälchen	
Cell counter thinys	
falcon 15 ml	
falcon 50 ml, blue cap	
falcon 50 ml	
5 ml pipette	
10 ml pipette	
20 ml pipette	
50 ml pipette	
parafilm	

Continued on next page

(Continued)

Name	Manufacturer
T75 cellculture flask	
cover foil for qPCR	
pH Papier	

2.10 Devices

Name	Manufacturer
Cell culture	
Bench 1	
Bench 2	
Pipette very small	
Pipette small	
Pipette medium	
Pipette large	
Incubator 1	
Incubator 2	
Cell Counter	
Water Bath	
Centrifuge cell culture	
microscope	
pipette boy	
big lab	
Centrifuge RNA place	
large centrifuge	
Seahorse	
qPCR maschine	
scale	
vortex	
magnetic stirer	
keyence	
Thermo Cycler	
pH electrode	

Continued on next page

(Continued)

Name	Manufacturer
other	
nanodrop	
Ice maschine	

2.11 Programs & Modules

Programs

Program	Version	Manufacturer
Affinity Designer	1.10	Serif (Europe) Ltd.
Excel	Version 2205	Microsoft Corporation
GitHub		GitHub, Inc
keyence sftware?!		
MiKTeX	2.9	Christian Schenk
python	3.9	Python Software Foundation
PyCharm (Community edition)	2021.2.2	JetBrains s.r.o.
SchemaSpy	5.0.0	John Currier
SDS	2.2.2	Thermo Fisher Scientific GmbH Im
sqlite3_analyzer	3.38.5.	The SQLite Consortium
Wave Controller	2.6.3	Agilent Technologies

Python Modules

Module	Version	Info
beautifulsoup4	4.11.1	crummy.com/software/BeautifulSoup
bokeh	2.4.1	bokeh.org
numpy	1.21.4	numpy.org
pandas	1.3.4	pandas.pydata.org
Pillow	8.4.0	python-pillow.org
pyliftover	0.4	github.com/konstantint/pyliftover
python standard library	3.9	docs.python.org
matplotlib	3.4.3	matplotlib.org

Continued on next page

(Continued)

Module	Version	Info
requests	2.26.0	requests.readthedocs.io
scipy	1.7.3	scipy.org
seaborn	0.11.2	seaborn.pydata.org
urllib3	1.26.7	urllib3.readthedocs.io
wget	3.2	bitbucket.org/licface/pywget

Frameworks

- This thesis was generated with the uzl-thesis class have been written and kindly provided by Till Tantau.
- Styling of the GWAS Visualizer was done with the CSS Framework Bootstrap.

2.12 Public Data

3

Methods

3.1 Cultivation and differentiation of HAoSMCs

For the following experiments human arortic smooth muscle cells were used. A cell type commonly used for the study of cardiovascular function and disease [Reference for this claim]. Cells were kept at 37°C and 5% CO2 when ever possibile.
Cells were differentiated treated first with TGFb and then with IL-1 & PDGF-BB to induce a synthetic phenotype. For more imformation please check the section on smooth muscle cells in CAD.

Thawing & Cultivation

Cells were cultivated to a maximum passage of 10, after that new passage cells were thawed. For long time storage cells were kept in liquid nitrogen in [cryo medium]. When required new cells (6th passage) were need cells were thawed at 37°C in the water bath and transfered to a falcon. After centrifugation for 2 min at 300xg the supernatant was remoced and the cell pellet taken up in 14 mL of M231 + SMSG and cultivated in a T75 flask. Every other day 2/3 of the medium were removed and replaced by fresh.

Passaging

When reaching a maximum of 80% confluency (approx. once a week) the medium was removed completely and cells were washed once with 5 mL of PBS. Then the cells were incubation with 3 mL trypsin for 4 min at 37°C. After 7 mL M231 were added to the deattched cells and the cells were transfered to a falcon and pelleted for 4 min at 300xg. The supernatant was removed and the pellet resuspenden in M231 + SMGS, seeding 500×10^3 cells per T75 flask.

Preparation of Collagen I matrix

For preparation of the collagen matrix (1.8 mg/mL) all the components were mixed, adding the collagen last. All components were stored at 4°C and all pipetting steps were carried out on ice:

Table 3.1: Col I Matrix composition The stupid matrix

component	concentration	volume (µL)
H20	-	38.9
M231	-	53.3
SMGS	20x	5,3
NaOH	1 M	2,7
NaHCO3	7.5 %	2.1
Col I	5 mg/mL	57.6
total	-	160

160 µL of matrix mix were transfered in each used well of a 24-well plate, fully coating the bottom of the well. For polymerization the matrix was incubated at 37°C for at least 60 min.

Differentitation of HAoSMCs

Differentiation was carried out in 24 wells plates with 1 mL M231 supplemented with 1 % FBS and different cytokines:

- **Day 0:** Matrix and cells were prepareed as described in the sections Preparation of Col I matrix and Passaging. Seeding of 40×10^3 in M231 + SMGS on plastic or on 160 µL collagen 1 matrix.
- **Day 1:** After 24 h the medium was replaced with 1 mL M231 + 1% FBS + 5 ng/mL TGFb (or just 1 mL M231 + 1% FBS).
- **Day 5:** The medium was replaced with 1 mL M231 + 1% FBS + 10 ng/mL IL-1 + 10 ng/mL PDGF-BB (or just 1 mL M231 + 1% FBS).
- **Day 7:** Potentially further stimulation described in the section of the used assay.

3.2 mRNA Quantification

qPCR was utilized to assess the mRNA concentration of the two reporter genes CNN1 and MMP9 in HAoSMC differentiated as described in section on differentiation. Using the house keeping gene GAPDH for reference.
SYBR® Green is an intercalating DNA dye that allows for the monitoring of DNA amplification. Flourescence is measured after every amplification cycle of the PCR yielding a crossing point when signal reaches a certain threshhold. A lower quantification cycle (Cq) corresponds to an higher inital DNA concentration (**huggettStandardisationReportingNucleic2011**).

RNA Isolation

RNA was isolated using the kit and extraction was performed according to the corresponding protocol, using an extra washing step with 700 µL 80 % ethanol and eluting with 30 µL of RNase-free water. Determination of nucleic acid concentration was carried out with the NanoDrop.

Reverse Transcription

For reverse transcription RNA samples were diluted to yield 10 µL of 10 ng/µL RNA. The samples were heated for 5 min at 68°C before adding 10 µL of the RT reaction mix described in the following table:

Table 3.2: RT Master Mix

component	concentration	volume (µL)
First Strand Buffer	5x	4
DTT		2
dNTP		1
Oligos		1
RiboLock		1
M-MLVRT		1

The reaction was carried out for 60 min at 37°C before inactivating the enzyme for 5 min at 95°C. cDNA was used for qPCR or stored at -20°C.

qPCR

The samples were prepared in a 384-well plate using SYBR® Green Master Mix:

Table 3.3: qPCR samples The samples for the qPCR

component	conentration	volume (µL)
SYBR GREEN Master Mix	1:2	3.75
Primer (forward + reverse)	5 pM (each)	1.125
H20	-	1.125
cDNA	-	1.5

Wells were sealed, thoroughly mixed by invertation of the plate and the assay performed using follwing programme on the TaqMan:

Table 3.4: qPCR programme The programme for the qPCR

step	time (s)	temperature (°C)	loop to	passes
1	120	50		1
2	600	95		1
3	15	60		40
4	60	60	3	40
5	600	95		1
6	-	16		1

Processing of Data

The Cq was automatically calculated by the software SDS2.2.2 and exported for further analysis. The arithmetic mean of three 3 technical was calculated for each sample, disregarding values that are obvious outliers. For normalization the mean ct of the reference gene GAPDH was substracted from the mean ct of the gene of interest:

$$\Delta ct = ct(\text{geneofinterest}) - ct(\text{GAPDH})$$

Taking into account the exponential amplification of DNA in PCR, the Δct can then be transformed into an relative expression level. Where 10×10^6 is just a constant to yield values that are easier to work with:

$$\text{rel.expr.} = 2^{-\Delta ct 10 \times 10^6}$$

In total four biological replicates were done. Data visualization and statistical analysis was done in python using the modules: pandas, numpy, scipy as well as pyplot and seaborn. Assuming a normal distribution, student's t-test was used, a p-value of 0.05 is considered as significant. For detailed information please check the script.

3.3 Energy Profiling

Seahorse Assay was utilized to assess the energy profile of HAoSMC differentiated as described in section on differentiation. For this assay cells were not differentiated in a 24 well plate but the seahorse plate, using 5 technical repeats and on control well for 4 tested conditions. Since the plate would not fit the matrix cells were cultivated in plastic!
The Seahorse XF Analyzer allows real time measurement of dissolved oxygen and protons in a confined small volume by using solid state sensor probes. These are used to calculate the oxygen consumption rate (OCR) and extracellular acidification rate (ECAR) of a cell monolayer. The OCR and ECAR are indicators for mitochondrial respiration and glycolisis respectively and can be used to assess the metabolic function of cells (**HowAgilentSeahorse**).

Seahorse Assay

On the day before the assay the Seahorse XF Analyzer was turned on to calibrate and the sensor cartrigae was left to equilibrate in Seahorse XF calibrant over night at 37 °C (in non-CO2 environment).
On the day of the assay, cells were washed with 500 µL PBS each and afterwards 500 µL supplemented XF BASE medium were added, cels were left to incubate for 1 h at 37°C in non-CO2 environment. During this time toxins for disruption of the respiratory chain were prepared and added to the catridge:

Table 3.5: toxins for seahorsetoxins for seahorse. :)

component	concentration in cartridge(µM)	volume in cartridge(µL)	concentration in well (µM)
Oligomycin	14	55	1.4
FCCP	10	60	2.0
Antimycin	50	65	5.0

The compound catridge was loaded into the XF Analyser for calibration, after successful calibration the hydration catridge was replaced with the cell plate. Measurement was carried out as following:

– Calibration of the probes.

- Equilibration
- 3 Repeats of:
  - Mixing (1 min)
  - Pause (2 min)
  - Detection of OCR and EACR (4 min)
- Pause (2 min)
- Injection of 55 µL Oligomycin
- 3 Repeats of:
  - Mixing (1 min)
  - Pause (2 min)
  - Detection of OCR and EACR (4 min)
- Pause (2 min)
- Injection of 60 µL FCCP
  - Mixing (1 min)
  - Pause (2 min)
  - Detection of OCR and EACR (4 min)
- Pause (2 min)
- Injection of 55 µL Antimycin
- 3 Repeats of:
  - Mixing (1 min)
  - Pause (2 min)
  - Detection of OCR and EACR (4 min)

Finally the medium was removed and cells were stained for 15 min with Hoechst (concentration in PBS) and photographed in the keyence to determine cell count for normalization.

Processing of Data

Cells were quantified using (what exactly does Tobias script do) with a python script provided by my supervisor Dr. Tobias Reinberger, cell count and the signal of the control wells were used to normalize the OCR and EACR calculated by the XF Analyzer with the accompanying software.
In total three biological were recorded of which on was excluded because no changes in OCR and EACR could be detected and cells deacttched from the bottom of the wells during Hoechst staining. For the remaining two replicates the least fitting of the 5 technical repeats for each condition was manually excluded.
Further , again using a modified python script provided by Dr. Tobias Reinberger. Assuming a normal distribution, student's t-test was used, a p-value of 0.05 is considered as significant. For detailed information please check the script.

3.4 Oxidative Stress Assay

CellROX Green assay was used to assess generation of reactive oxygen species in HAoSMC differentiated as described in section on differentiation, after further stimulation (from here

on refered to as 'boost') with PDGF. Additionally a recovery experiment was performed using NAC, a potent anioxidant, to quench generation of ROS.
CellROX Green is a flourescent dye that gets oxidized by ROS and then binds to DNA, showing bright green flourescence (**CellROXGreenReagent**).

CellROX Assay

For the assay cells were washed with PBS then the boost was performed using variable concentrations of PDGF in 300 µL HBSS. For ROS quenching with NAC, 0.25 M NAC solution was added to the wells 2 h prior to the experient and also added to HBSS during the experiment.

Table 3.6: Seahorse Assay

component	concentration	final concentration	volume (µL)
HBSS	-	-	300
PDGF	?	variable (0 - 400 ng/mL)	variable
Hoechst	?	?	0.3
CellROX (1:500)	?	?	0.6
NAC	0.25 M	variable (0 - 8 mM)	variable
total	-	-	∼300

Cells further kept at 37°C in 5 % CO2 environment, incubation time is indicated with the results of the respective experiment. Imaging was done at the keyence microscope using standard sensitivity and a exposure time of s in the green channel and s for the blue channel.

Processing of Data

For time resolved PDGF-BB boost titration 7 biological repeats were performed of which one was excluded because of high signal in the negative control. For NAC quench 4 biological repeats were performed of which one was excluded because no signal in the positive control. For quantification of signal itensity, pixels with a green value higher than 90 were counted. Differents in cell count were adjusted by divison by the number of pixels with a blue value bigger than 80 (CHECK TRESHHOLD VALUES!!). Z-STACK FOR THE NAC quench images. To adjust for large variance in total signal intensity between biological repeats, values were adjusted by divison through the total signal of all recorded conditions.
Mann Whitney U Test was used, a p-value of 0.05 is considered as significant. For detailed information please check the scripts.

3.5 Immunoflourescence

Fibronektion as marker of matrix. Used cells. Maybe also the anti-8-oxoG AB?

Protocol

First cells were washed with PBS and fixated for 40 min with 200 µL -20°C methanol-aceton (1:1), after the removal of methanol-aceton cells were left to dry for 20 min. Then cells were

treated for 30 min with 250 µL permeabilization buffer, folllowed by 30 min with 250 µL blocking buffer and incubated with 300 µL of the primary AB over night at 4°C:

The next day the primary AB was removed and cells wre incubated for 60 min at RT with the secondary AB:

After removal of the secondary AB cells were washed 3 times with PBS and stained for 15 min with DAPI (1:5000 in water).

Processing of Data

Me counting pixels.

3.6 Curation of Data for postGWAS Analyses

GWAS data and data for postGWAS analyses and co-visualization was downloaded from public resources. Processing of the data and further annotation is briefly described in the follwing listing. For a complete overview of all the data collected please refer to table or the download scripts. For explainaition of the data please refer to the introdcution or the sources themselves.

- **GWAS Summary Statistics:** The CAD GWAS summary statistics as well as a list of indetified proxy SNPs from the study was annotated with [what did Tobias do] via the ensembl REST API by Dr. Tobias Reinberger.
- **HGNC Gene List** The newest quarterly update to the complete HGNC dataset was downloaded via the FTP server. The dataset was used to generate a list of all approved symbols, mapping to their HGNC id. Further a list of all symbols (approved, alias and previous) was generated, mapping to their HGNC id.
- **Linked SNPs** LD r^2 values for variants in a 500 kb window around all variants in the list of CAD GWAS proxy variants, were computed and downloaded via the ensembl REST API. For humans ensembl calculates the LD with data from the 1000 Genomes project (see table). In the same process linked SNPs were annotated with their most severe consequence (by VEP) via the ensembl REST API.

Table 3.7: 1000 Genomes Populations

Name	Size (individuals)	Description
1000GENOMES:phase3:ALL	2504	All phase 3 individuals
1000GENOMES:phase3:AMR	347	Americans
1000GENOMES:phase3:EAS	504	East Asians
1000GENOMES:phase3:EUR	503	European
1000GENOMES:phase3:SAS	489	South Asian

- **Ensembl Genome Annotatation** The newest ensembl build (ensembl release 106) was downloaded via the FTP server. Features annotated as genes of the type protein coding, lncRNA or miRNA were extracted, further gene symbols were mapped to their HGNC id if possible.
- **Open Target Genetics l2g Scores** The lattest list of Open Target Genetics l2g Scores was downloaded via the FTP server. Entrie were annotated with their HGNC

ID when ever possible, entries that do not map to a gene that is approved by the HGNC were dropped.
- **Ensembl Regulatory Build** The newest ensembl regulatory build (ensembl release 106) was downloaded via the FTP server.
- **TSS** Transcription start sides for protein coding genes were extracted from a USCS Brower dump.
- **Associated traits from GWAS catalog** The SNP trait associations from the lattest release of the GWAS catalog as well as the accompanying list of studies was downloaded via the FTP server. SNP-trait correlations missing a the position or a p-value for the association were dropped from the data set. Further the column for Odds Ration or beta was seperated in to two columns.
- **TADs** The data on TAD was downloaded via the download link on the 3D genome browser.
- **scATAC-seq from Miller et al.** The processed scATAC seq data was scraped from the Miller Lab GitHub repository.
- **scATAC-seq from CATlas** The processed scATAC seq data was scraped from the Ren Lab website.
- **ABC model** The ABC model data was downloaded from the Engreitz Lab FTP server. The data was translated from Hg19 to Hg38 with pyliftover.
- **ENCODE cCREs** Done by Dr. Tobias Reinberger.

3.7 Visualization of GWAS data

For visualization of the data, an bokeh application was build, that fetches the data from the database and renders it to a webbrowser.
Bokeh is a python module that allow easy and interactive visualization of data. It combines the powerful data processing tools of python with the interactivity of JavaScript running in the browser. The python side of bokeh creates python objects which are serialized into JSON data and handed over to bokehJS which deserializes them into JavaScript objects that are rendered to the browser. The intergrated bokeh server additionally offers the possiblity to synchronize data between the underlying python environment and brower side JavaScript library, allowing real time updates to the displayed data.

According to good design principles the concerns of the application are split into two sections. Reading of data from the database and further processing steps are managed by a data provide and enclosed in one class. In contrast to the model-controler-view architecture, a popular architectural pattern for the design of user interfaces, there is no partition between a view and a controler. Since data visualization as well as the control widgets are created by bokeh, it is convienient to use the build in event listeners of the library to handle the required callbacks. Therefore the main file is responsilbe for the creation of all plots and widgets as well as listening for inputs.

STYLING USING A bootstrap thingy

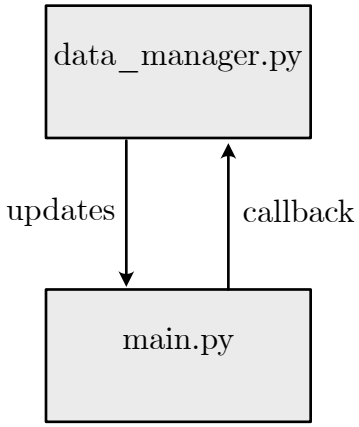


Figure 3.8: Architecture of the vis tool

3.8 Enrichment analysis

Based in the data in the database initial postGWAS studies were run. Annotation enrichment analyses are a popular tool for the identification of terms that are over-represent in a list of interest. The most prominent application probably being their application as gene set enrichment analyses (GESA). GESA are used to check for the over-representtion of a candidate gene list in a predefined set of genes (**tipneyIntroductionEffectiveUse2010**). In this case the method is used to determine if regulatory elements which overlap with SNPs associated with CAD are enriched in a biosample, using fisher's exact test.
As a list of regulatory elements the cell type specific cCREs that are part of the ENCODE project were used (excluding cCREs marked as unclassified). As a list of SNPs the list of CAD associated proxy SNPS and linked variants (european population, r2 >= 0.6) was used. The four values required for calculation of the enrichment factor () and p-value require are shown in the Contingency of figure:

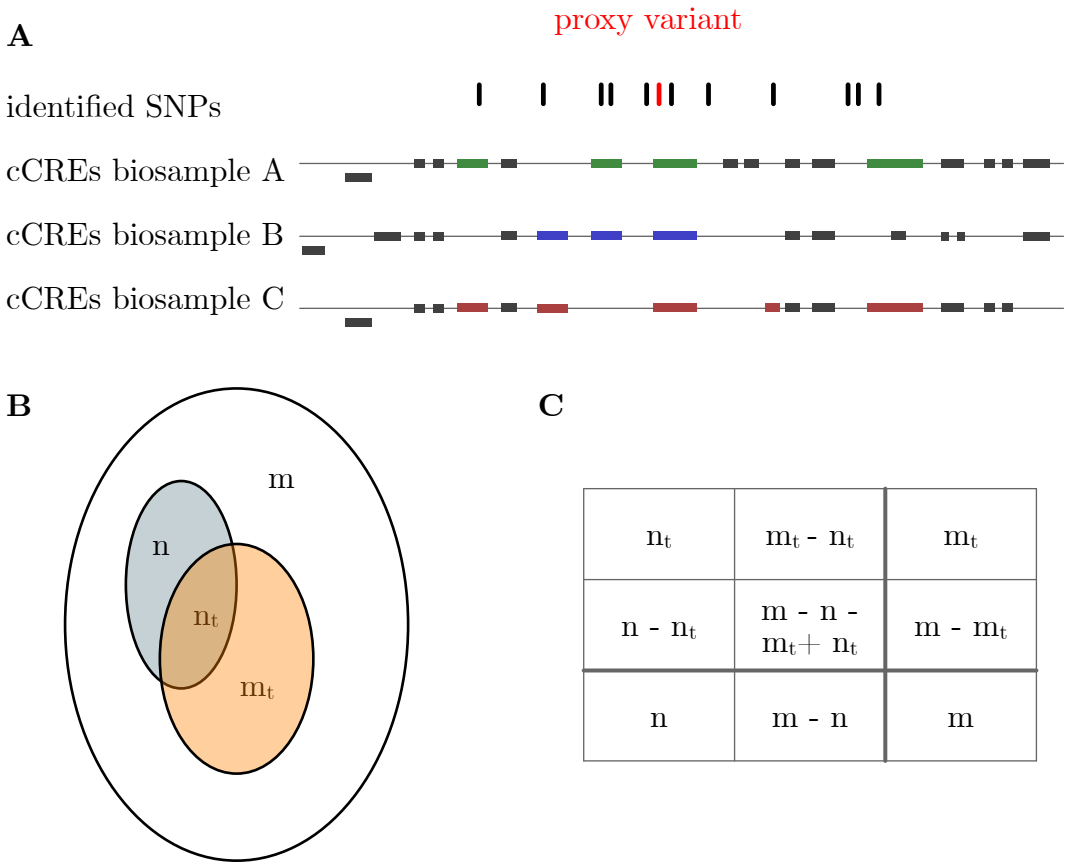
- Number of distinct cCREs among all biosamples (m)
- Number of distinct cCREs that are annotated in the biosample of interest (mt)
- Number of distinct cCREs that overlap with a SNP in the SNP list in any biosample (n)
- Number of distinct cCREs that overlap with a SNP in the SNP list in the biosample of interest (nt)

The p-value for the number of overlaps to be greater than or equal to the observation can be calculated as the cumulative distribution function of the hypergeometric distribution.

$$P(\sigma_t \geq n_t) = \sum_{k=n_t}^{min(m_t,n)} \frac{\binom{n}{k}\binom{m-n}{m_t-k}}{\binom{m}{m_t}}$$

To account for the multiple comparisons problem, p-values were adjusted with Bonferroni correction where n is the number of tests (≡ number of biosamples):

$$p_{ajd.} = p * n$$



n_t	m_t - n_t	m_t
n - n_t	m - n - m_t+ n_t	m - m_t
n	m - n	m

Figure 3.9: enrichment

The analysis and visualization were done in python. A p-value of 0.05 is considered as significant. For detailed information please check the analysis script and the visualization script.

4

Results

4.1 Differentitaion

Characterization of the phenotype we are inducing.

Expression of CNN1 & MMP9

Markers CNN1 and MMP9.

Further getting a quick glimps into the energy profile of the cells.

Energy profile

Looking at the extracellulat matrix.

4.2 Evaluation of oxidative Stress

PDGF boost of out cells indcues oxidative stress

Characterization of the CellROX Assay

Rescue of ROS production using NAC

4.3 Visualization of GWAS data

After tinkering around for quite a bit we decided to build a webapplication for a multitude of reasons. After tinkering around for even more time we decided to build a backend in the for of sqlite database because it is the most sustainable. Relational databases are used so much for a reason.

Curation of Data for postGWAS analyses

Describe all the data that is in the database and how we processed it.

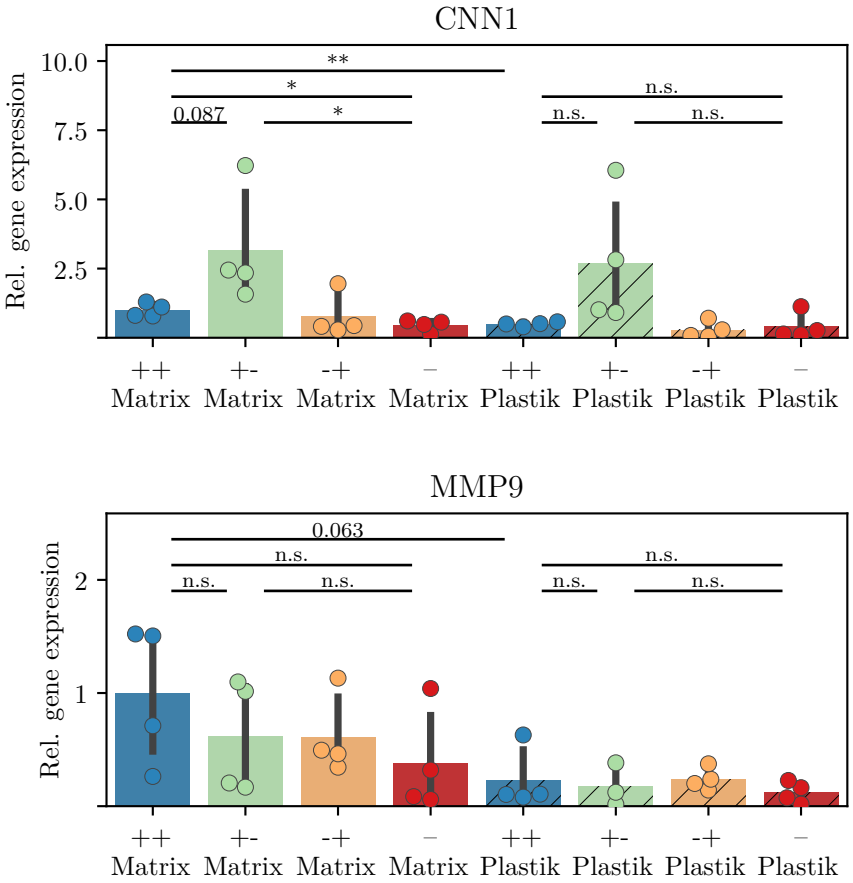


Figure 4.1: Expression of CNN1 & MMP9 in HAoSMCs
WUHU!. Meine erste Abbildung!

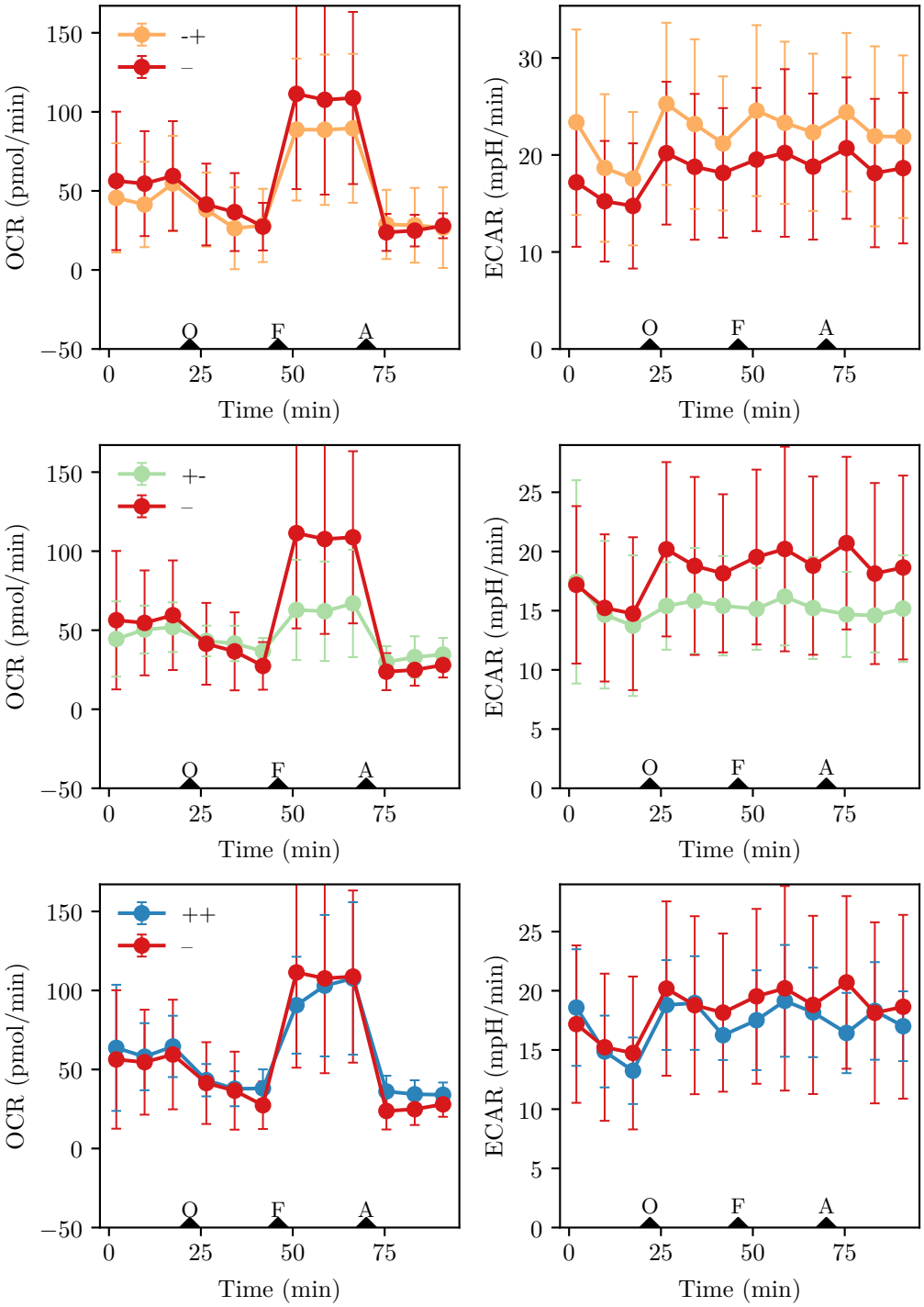


Figure 4.2: Seahorse tracks The seahorse tracks I did record.

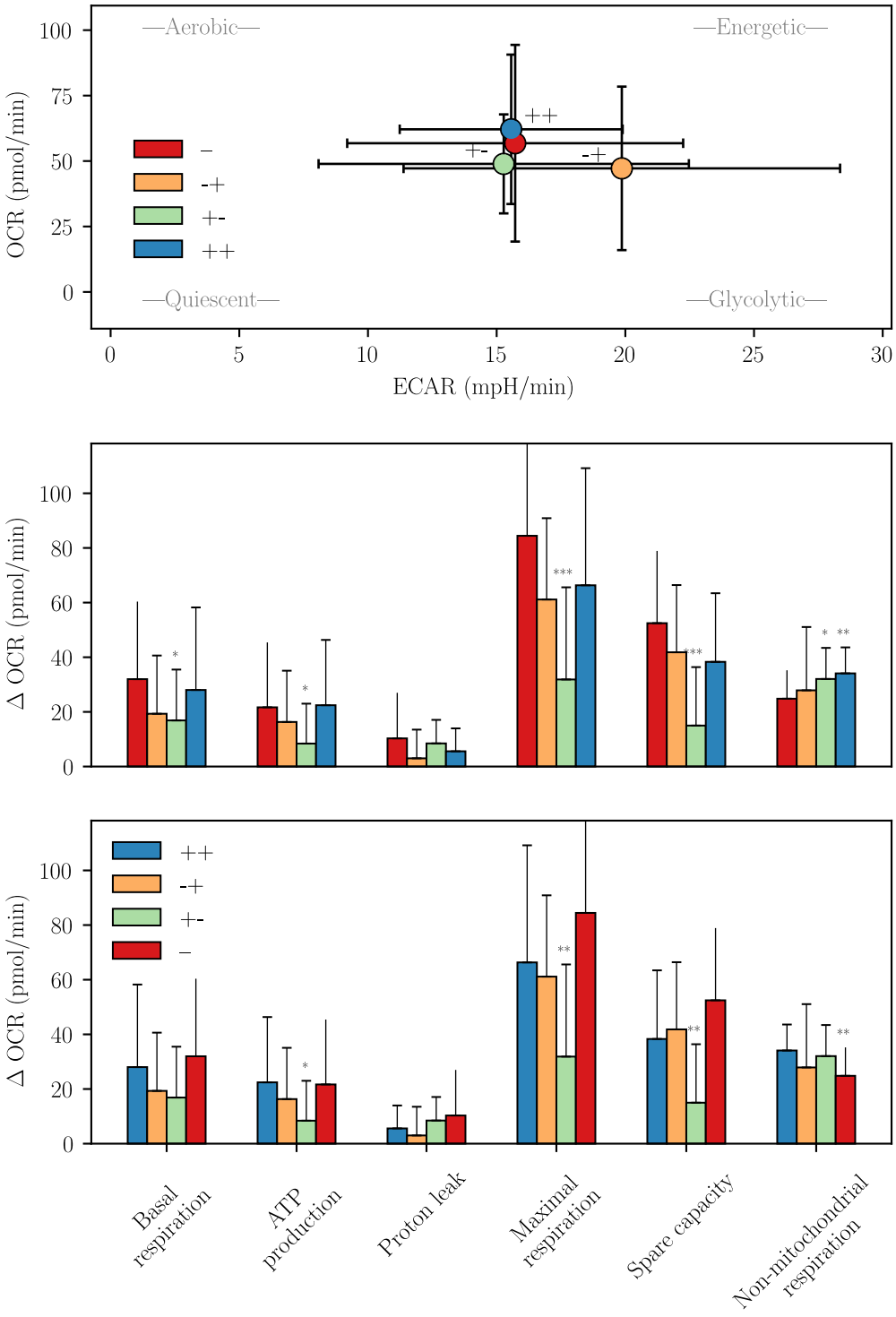


Figure 4.3: Energy profile Evaluation of those seahorse tracks.

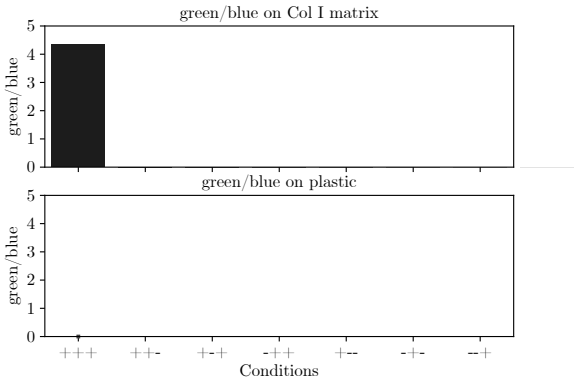


Figure 4.4: Stimulation with PDGF induces oxidative stress.
Repeat of the result already shown by Tobi.

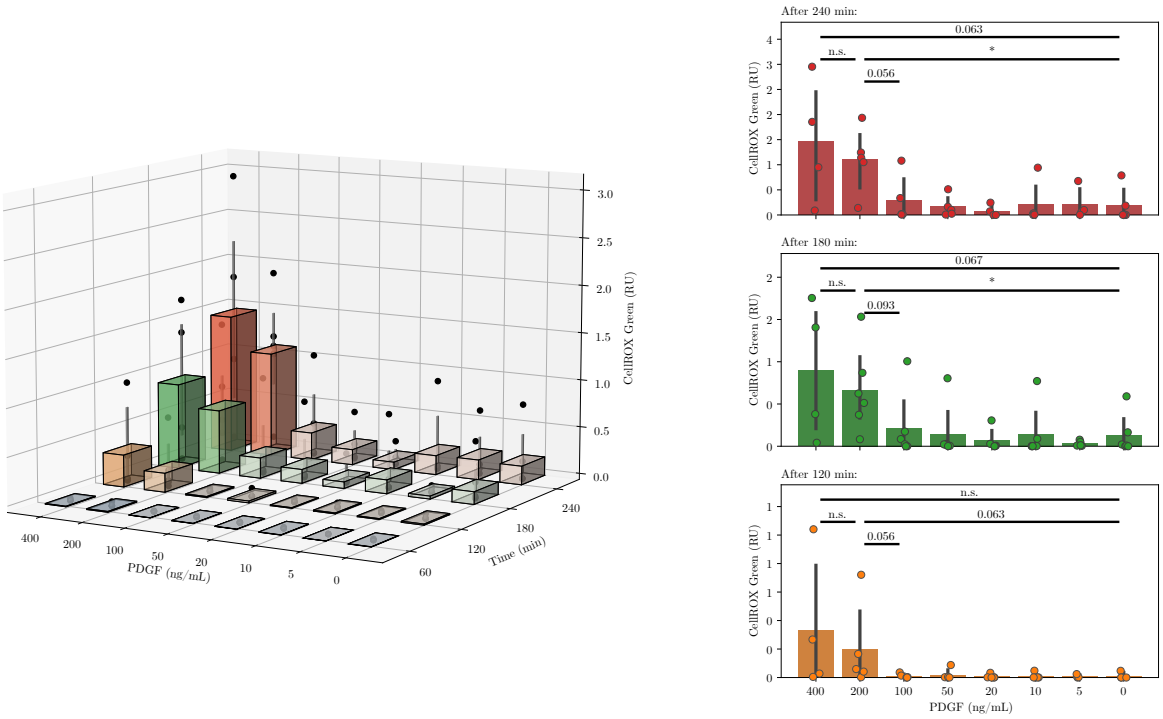


Figure 4.5: CellROX titration Titration

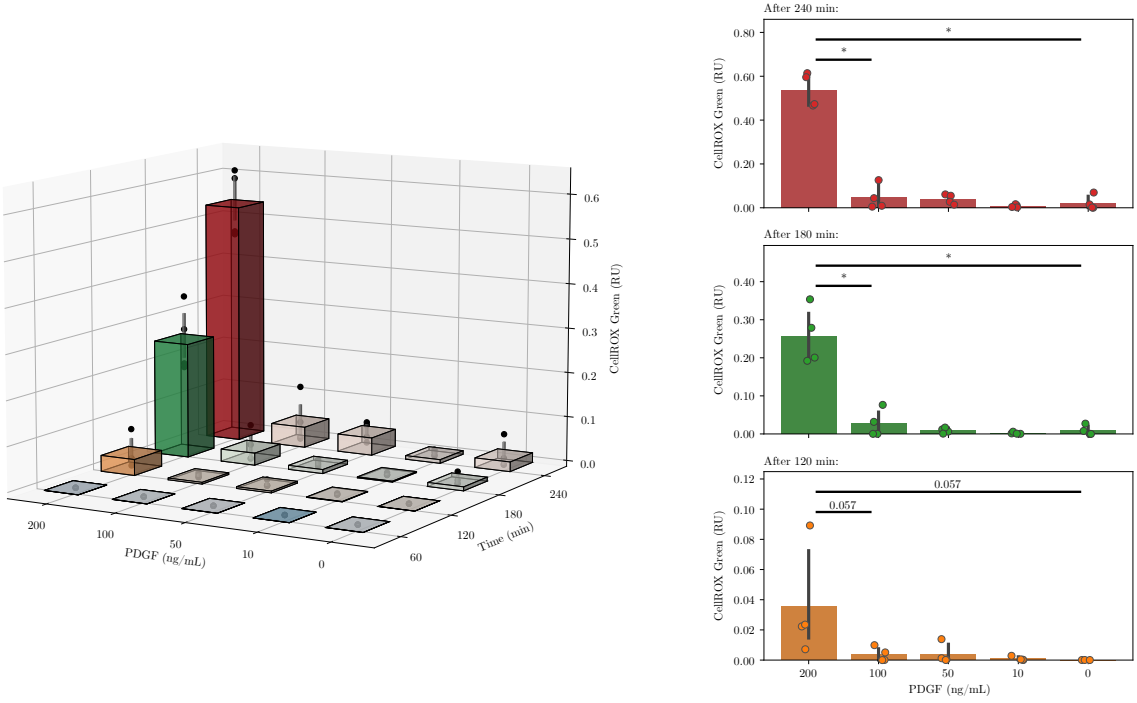


Figure 4.6: Stimulation with PDGF induces oxidative stress - normalized.
My attempt at normalization.

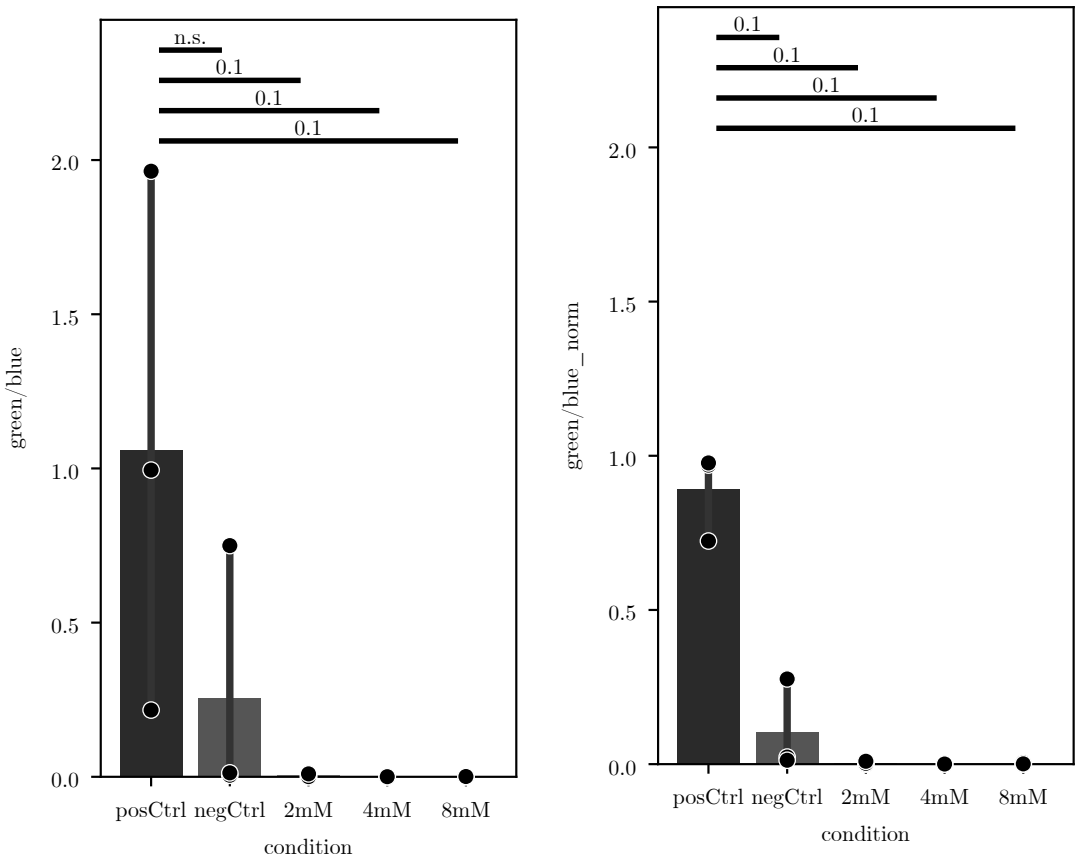


Figure 4.7: NAC quench The NAC quench. :).

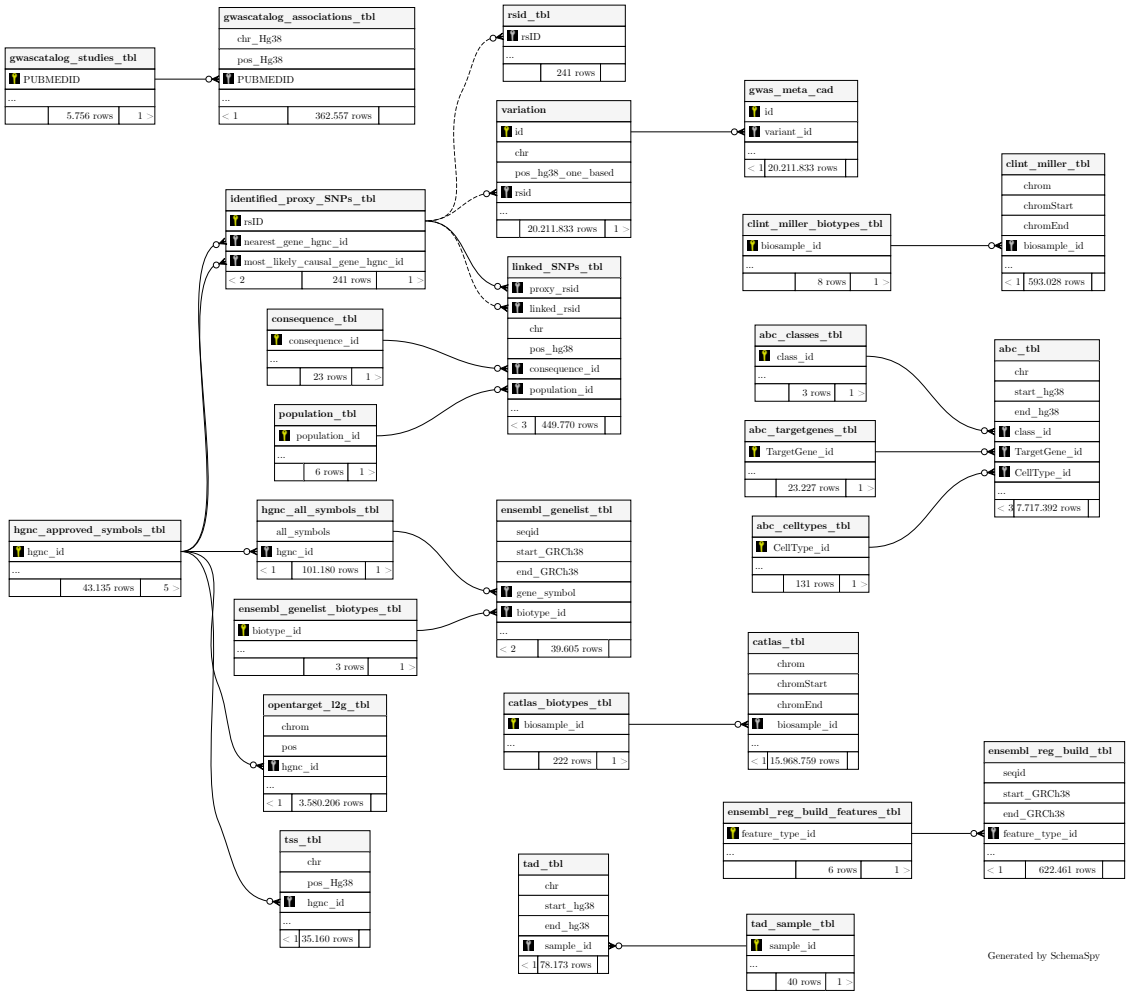


Figure 4.8: My super duper databaseDATABASE!!

Visualization of GWAS data

The first use case of this database and our original goal is the bokeh app for visualization. Wuhu! Show some screenshots but easiest would just be to try it out.

Enrichment analysis

Second the data that we are curating can als be put to use for other scenarios. Running a first test analysis as a proof of concept.

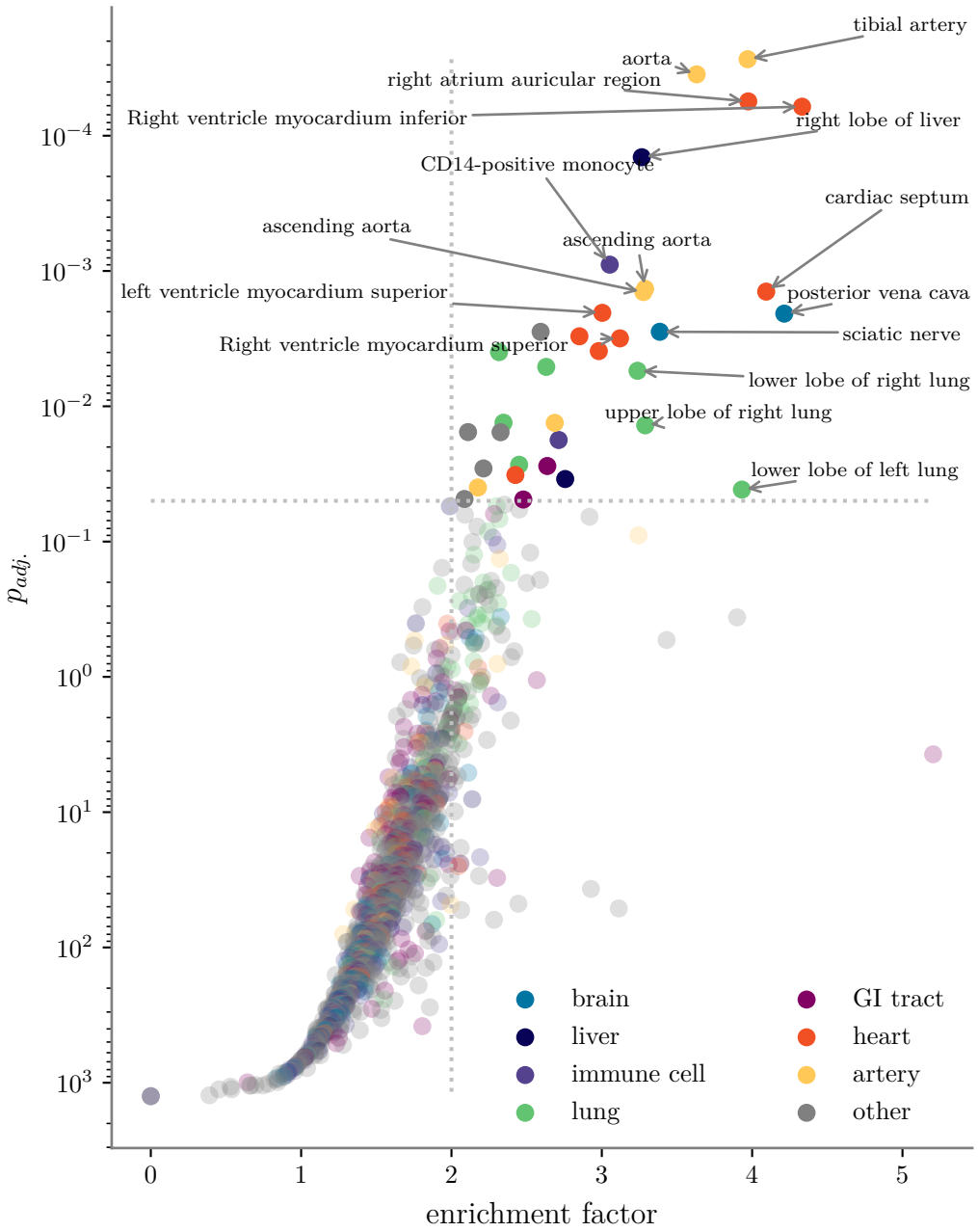


Figure 4.9: Enrichment Analysis This stuff actually seems to be working!!

5

Discussion

Here I'll discuss my results should I ever finish the rest of my thesis.

6

Conclusion & Outlook

We are closer to doing postGWAS analyses, we really hope that the database makes everything smoother. And we have a system where we can functionally access these identified SNPs. We are close to a point were we can combine both part of the project.