# Deciphering mechanisms underlying tumor heterogeneity using Multi-Omics approaches

**Joonas Avik**

**Supervisors:** Doctor Daniel Sobral at FCT, and Professor Joakim Lundeberg and PhD student Alma Andersson at KTH, Stockholm.

Computational Multi-Omics lab, FCT/UNL, Departamento de Ciências da Vida, Caparica, Portugal

**Abstract:**
Cancer is a complex disease and presents one of the greatest challenges in modern medicine. Despite remarkable advances in treatment of several cancer types, relapse and resistance to therapy remain recurring outcomes in patients, which underscores a need for personalized treatment approaches. These complications have been related to the high genetic diversity observed within tumors, termed intratumor heterogeneity (ITH). While specific mutational profiles have been associated with the development of heterogeneous tumors, the relationship between ITH and phenotype could unveil features that undergo selection and convey fitness. Features presented in the transcriptome, as markers of heterogeneity, might therefore be valuable biomarkers. In this project, these features are explored by assuming a linear relationship between genetic ITH measures and gene expression data from The Cancer Genome Atlas samples. By first reducing the number of variables among the transcriptome to the differentially expressed genes between low and high ITH samples, the association between specific gene expression profiles and ITH is sought with a linear model. By using two different methods for estimating ITH, called *Expands* and *PhyloWGS*, the association was modeled with each method. Interestingly, the model based on *Expands* captured the elevated expression of a chaperone gene *DNAJC18* as being consistently associated with lower ITH in four cancer types. On the other hand, models based on *PhyloWGS* presented lower predictive power. These results demonstrate that the transcriptome can be used to predict genetic ITH, although this depends on the method used for characterizing ITH.

Key words: Cancer, intratumor heterogeneity, gene expression, linear model, lasso regularization, TCGA

## Introduction:

Cancers are characterized by unstable genomes, which leads to diversity within tumors that is manifested as cellular subpopulations with distinct genotypes. This diversity, termed intra-tumor heterogeneity (ITH) develops through clonal expansions caused by the initial accumulation of drivers that undergo selection in the tumor microenvironment accompanied by neutral passenger mutations (Greenman *et al.*, 2007). Exploring the mutational landscape of a tumor provides a time frame in which the early, clonal mutations are present in all tumor cells while subclones are characterized by an additional, less prevalent set of mutations (Carter *et al.*, 2012). Deciphering this information from sequencing data has identified drivers underlying clonal expansions as well as alleles responsible for therapeutic resistance (Landau *et al.*, 2013). Furthermore, a heterogeneous cancer in theory can provide an assemblage of subclones resistant to any therapeutic agent, which means that heterogeneity might be a potential, quantifiable biomarker for cancer prognosis (Merlo and Maley, 2010). This has led to the development of a variety of algorithmic methods for measuring ITH from genomic data. These have for example been applied to investigate how ITH may be related to clinical outcome on data generated in connection to large scale genomic initiatives such as *The Cancer Genome Atlas* (TCGA) project. These studies generally correlate increased ITH measures to poor clinical outcome. For example, estimating ITH with the *Expands* method (Andor *et al.*, 2014) has been used to correlate adverse patient outcome to moderate ITH (Andor *et al.*, 2016) while estimates made with *PyClone* (Roth *et al.*, 2014) has associated tumors with higher ITH to poorer survival rates (Morris *et al.*, 2016). Both of these methods quantify ITH as a number of cellular subpopulations by grouping single nucleotide variants (SNVs) according to their estimated cellular prevalence while accounting for copy number variations (CNVs).

This clinical significance of ITH has led to the search for mechanisms underlying it. On the genetic level, the association between genetic ITH estimates and the variants from which they originate have been studied. As an example, the *PhyloWGS* method, which considers both SNVs and CNVs in its measure of heterogeneity (Deshwar *et al.*, 2015), has been used to study the association between ITH and genomic instability as expressed with both SNVs and CNVs (Raynaud *et al.*, 2018). Similarly, the association between SNV load and ITH estimates made with the *MATH* approach (Mroz and Rocco, 2013) have been made (De Matos *et al.*, 2019). Moreover, in search for specific causal variants, mutations in epigenetic modifier genes have been noted in tumors of higher heterogeneity (De Matos *et al.*, 2019). These drivers of ITH were found by applying linear models on the association between ITH and sets of mutated genes with a shrinkage method called *lasso* or L1-shrinkage (Friedman, Hastie and Tibshirani, 2010). Linear modeling with *lasso* has also been used to study the effect of specific variants on the transcriptome (Gerstung *et al.*, 2015). In this case, by using principal components analysis on the transcriptome, the association between general expression profiles and specific genetic variants was modeled. Through this multi-omics approach, the authors linked tumor genotypes to their phenotypic profiles which undergo selection.

In this study, the genotype-phenotype relation is further explored by modeling the association between the transcriptome and the genome data sets available in TCGA. Here, this is done by assuming a linear relationship between the expression of specific genes and the genetic background in the form of ITH measures. This approach promises to pinpoint potential biomarkers that could be further tested in *e.g.* proteomic assays.

# Results:

We performed the analysis with a pan-cancer approach, the goal of which was to identify ITH features in common among cancers. To achieve this, a strategy that relies on differential expression analysis (DEA) for variable selection among expression data ahead of modeling was used (see methods for details).

## Searching for Markers of ITH that are Common Among Cancers

The ITH measures used as response variables in the modeling were defined as the number of subclonal populations (SPs). Here, SPs were calculated for the samples of 33 cancer types available in TCGA by using the *Expands* method, which was run successfully for 8274 out of 9850 samples.

To find genes with possible association with ITH the search space among expression data was first reduced to potentially meaningful genes. To do this, differential expression analysis (DEA) between groups of low- and high ITH, based on *Expands* ITH estimates was performed (see methods). Since cancers with less than 6 samples per group were excluded, DEA was performed for 23 out of the 33 cancers. For these, DEA yielded significant (p-adjusted < 0.05) differentially expressed (DE) genes for 17 cancers. Then, the significant DE genes were filtered for intersecting genes among cancers, which yielded 38 genes among 6 cancer types to be used for model fitting: LUAD, STAD, BRCA, HNSC, UCEC, BLCA.

Using normalized expression of the filtered 38 genes as predictors and *Expands* SPs as response variables, the linear model fitting was performed with the shrinkage method *lasso,* which during the calculation of coefficients served to extract a subset of genes showing an association with ITH (see methods for details). As a result, we obtained significant coefficients for LUAD, STAD, BRCA, HNSC, UCEC and these are compiled in a heatmap (**Figure 1**). Interestingly, the elevated expression of *DNAJC18* is consistently shown to be significantly associated with lower genetic ITH in four cancer types: STAD, BRCA, HNSC and UCEC.

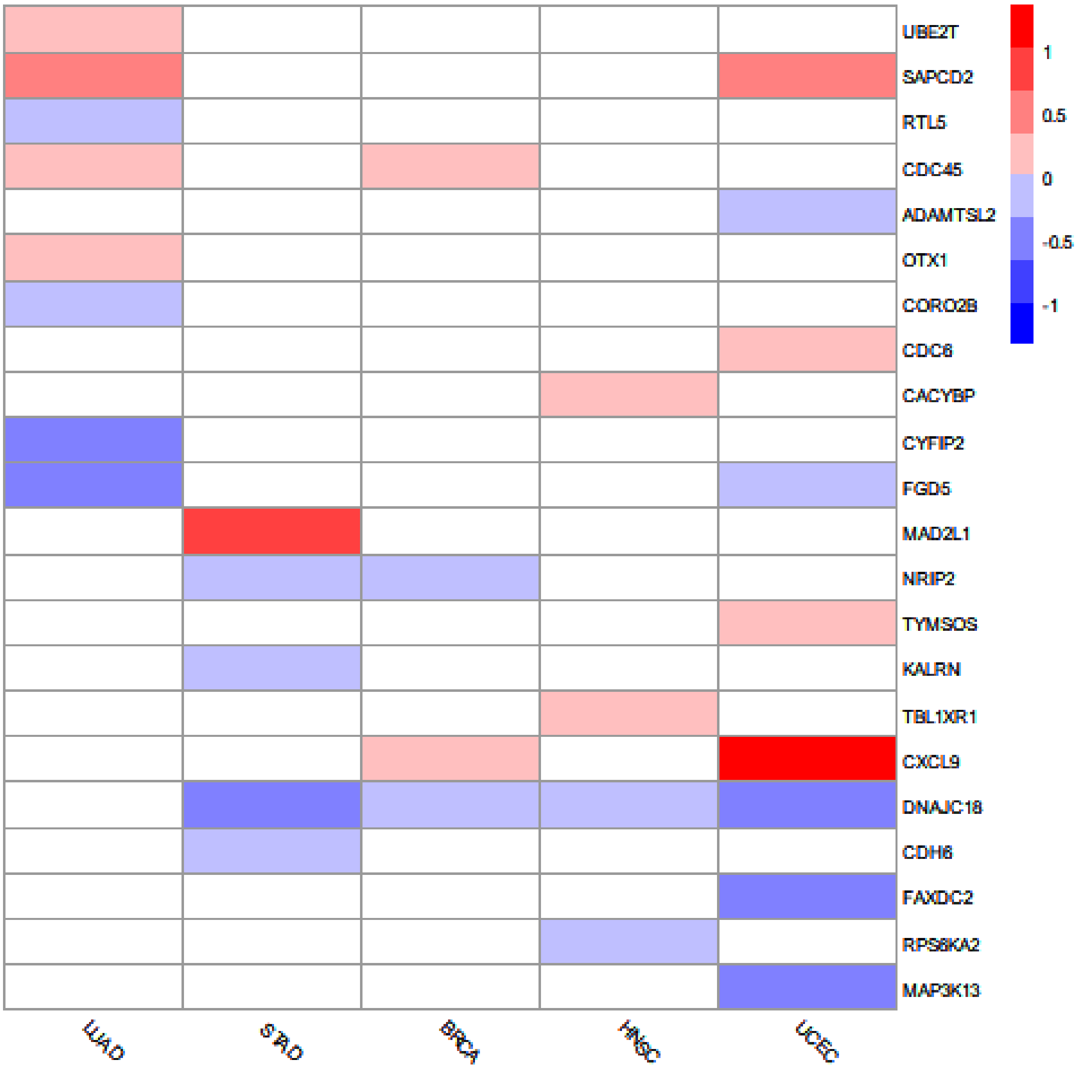

**Figure 1:** *Heatmap of coefficients calculated for each gene (y-axis) in each model based on Expands ITH estimates for the cancer types (x-axis). A negative (blue) and positive (red) coefficient of a gene indicates that its elevated expression is associated with a decrease, and increase in ITH, respectively. The coefficients are calculated for a set of 38 genes that were differentially expressed in 6 cancer types: LUAD, STAD, BRCA, HNSC, UCEC and BLCA. Since only zero-coefficients were calculated for BLCA, it is therefore excluded from the plot. Notably, an elevated expression of DNAJC18 is consistently associated with low ITH in four cancer types: STAD, BRCA, HNSC and UCEC.*

## Testing the models

To test the models, the root mean square errors (RMSE) were calculated between model predictions and *Expands* ITH estimates for LUAD, STAD, BRCA, HNSC and UCEC (**Table 1**). The relatively large RMSE's of these models indicate that the predictions are not precise (**Table 1**). However, Pearson correlation calculated between model predictions and the *Expands* ITH estimates was significant (p-value < 0.05) for all cancer types mentioned above, which indicates that the calculated coefficients have some predictive power for the change in ITH (**Table 1**). This being said, the relatively low correlation estimates for BRCA, HNSC and UCEC (0.20, 0.33 and 0.21) mean that the associations of these models might be less meaningful than those of LUAD and STAD, for which higher Pearson correlation was estimated (0.51 and 0.55 respectively).

**Table 1:** *Root mean-square errors (RMSE) and Pearson's correlation estimates along with p-values for models based on Expands. RMSE's are shown to be high, indicating large variance among predictions. However, the estimated Pearson's correlations between model predictions and Expands ITH estimates are significant (p-value < 0.05), further indicating that the calculated coefficients succeed in displaying an association between the expression of the captured subset of genes and genetic ITH.*

| *Cancer type* | *LUAD* | *STAD* | *BRCA* | *HNSC* | *UCEC* |
|---|---|---|---|---|---|
| *Pearson's cor.* | *0.51* | *0.55* | *0.20* | *0.33* | *0.21* |
| *p-value* | *2.58e-9* | *2.16e-8* | *3.15e-3* | *2.93e-4* | *0.033* |
| *RMSE* | *3.47* | *3.77* | *2.90* | *2.77* | *5.11* |

## Searching for Markers Cancer by Cancer

Since ITH estimates of different algorithmic methods developed for data obtained from single tumor biopsy samples are known to vary (Abécassis *et al.*, 2019), it is necessary to test whether linear models could be obtained with another ITH method. Here, the *PhyloWGS* method was tested with estimates brought from literature (Raynaud *et al.*, 2018). To provide a comparison between models based on *Expands* and *PhyloWGS*, the DEA output was filtered based on log fold change and average expression (among samples) to extract genes in a cancer-by-cancer search (see methods for details). This resulted in four cancer types for which models based on *PhyloWGS* and *Expands* ITH estimates could be tested and compared: LUAD*,* HNSC*,* BRCA and STAD (**Table 2**). Although models were obtained with both ITH methods, models based on *Expands* seem to be more relevant in their capacity to predict ITH. Overall, *Expands* yielded models for STAD, BRCA, COAD, HNSC, LUAD, all of which gave non-zero coefficients and significant Pearson's correlations. I.e. all of the tested models based on *Expands* displayed some association between the filtered expression data and genetic ITH, most significantly for STAD and COAD with Pearson's correlation estimates of $cor_P = 0.693$ and $cor_P = 0.633$, respectively and p-values of 3.32e-8 and 6.69e-6, respectively. Overall, the Pearson's correlation test showed significant correlations (p-value < 0.05) for all models based on *Expands*, while for *PhyloWGS*, the only model with any predictive power was obtained for STAD ($cor_P = 0.310$). The model test errors as RMSE and Pearson's correlation measured between predictions and original ITH estimates, along with p-values are summarized in **Table 2**. While the RMSE values of the predictions would indicate worse precision of models based on *Expands*, it must be noted that these values are calculated on different scales and are therefore not comparable.

**Table 2:** *Pearson's correlation coefficient and root mean-square error (RMSE) measured on independent test sets between model predictions and ITH estimates for PhyloWGS and Expands, respectively. Expands yielded models showing significant association for all 4 cancer types that could be compared, while PhyloWGS models did not yield coefficients for BRCA, or LUAD, and for HNSC no correlation was measured despite calculated coefficients. The RMSE values measured for either method is based on different scales and are therefore not directly comparable. P-values < 0.05 are highlighted.*

| | Method: | **PhyloWGS** | ***Expands*** |
|---|---|---|---|
| STAD | Pearson correlation: | 0.310 | 0.693 |
| | p-value: | 0.0301 | 3.32e-08 |
| | RMSE: | 1.86 | 2.67 |
| BRCA | Pearson correlation: | NA | 0.278 |
| | p-value: | NA | 2.43e-04 |
| | RMSE: | 1.13 | 2.55 |
| HNSC | Pearson correlation: | 0.189 | 0.484 |
| | p-value: | 0.0518 | 1.49e-07 |
| | RMSE: | 1.49 | 2.59 |
| LUAD | Pearson correlation: | NA | 0.282 |
| | p-value: | NA | 0.045 |
| | RMSE: | 3.06 | 3.90 |

Here, by using genes filtered by highest log fold change, we get significant coefficients with models based on *Expands*, as well as *PhyloWGS*. However, with *Expands* these models seem to be more relevant in their capacity to predict ITH. Overall, *Expands* yielded models for STAD, BRCA, COAD, HNSC, LUAD, all of which gave non-zero coefficients and significant Pearson's correlations. I.e. all of the tested models based on *Expands* displayed some association between the expression data and genetic ITH, most significantly for *STAD* and *COAD* with Pearson's correlation estimates of $\text{cor}_P = 0.693$ and $\text{cor}_P = 0.633$, respectively and p-values of 3.32e-8 and 6.69e-6, respectively.

Overall, the Pearson's correlation test showed significant correlations (p-value < 0.05) for all models based on *Expands*, while for *PhyloWGS*, the only model with any predictive power was obtained for *STAD* ($\text{cor}_P = 0.310$). The model test errors as RMSE and Pearson's correlation measured between predictions and original ITH estimates, along with p-values are summarized in **Table 2**. While the RMSE values of the predictions would indicate worse precision of models based on *Expands*, it must be noted that these values are calculated on different scales and are therefore not comparable.

# Methods:

Linear models were used to study the association between the transcriptomic data and genetic intratumor heterogeneity (ITH) in single tumor biopsy samples. The model building comprised three main parts: **1**) Generating response variables in form of genetic ITH estimates for each tumor sample; **2**) Selecting predictor variables based on the results of differential expression analysis (DEA) between groups of *high-* and *low ITH*; **3**) Linear model fit with variable subset selection and model testing.

All analyses were performed in *R* programming language https://www.r-project.org/ version 3.6.2.

## The code

The complete code for performing all computational methods is available at: https://github.com/joonasavik/ITH-code**.**

## Data sets

Publicly available data generated in the context of The Cancer Genome Atlas (TCGA) program was downloaded from the Genomic Data Commons (GDC) Data Portal in November 2019 https://portal.gdc.cancer.gov/repository

### *Genetic data: SNV and CNV data sets*

The TCGA data that were used for ITH estimation with *Expands* were simple nucleotide variant (SNV) and copy number variation (CNV) data sets. The downloaded SNV data were MAF files (Mutation Annotation Format) produced with the mutation calling algorithm *Mutect2* (Cibulskis *et al.*, 2013) on whole exome sequencing data (*File Format: MAF - GDC Docs*, no date). As such, the data comprises the coordinates of the variants on the GRCh38 reference genome and the allelic frequency of each variant as the ratio of sequencing reads with the mutation to total reads across the locus. Masked SNV data sets were used, i.e. predicted germline variants have been filtered in the TCGA workflow to protect privacy. The downloaded CNV data was copy number segments generated from Affymetrix SNP 6.0 array data through the TCGA CNV pipeline (*Bioinformatics Pipeline: Copy Number Variation Analysis - GDC Docs*, no date). As such, the CNV data comprises coordinates of genomic regions and the copy number for these regions estimated from microarray intensities.

### *Transcriptomic data: RNA sequencing reads*

The transcriptome data used were HTSeq counts as the number of mRNA sequencing reads for each gene, produced through the TCGA workflow (*Bioinformatics Pipeline: mRNA Analysis - GDC Docs*, no date). Ensembl Gene ID's present in the TCGA data were converted to HGNC gene names using a match table downloaded from Ensembl Biomart in April 2020 https://www.ensembl.org/biomart/martview.

## Estimating genetic ITH with *Expands*

ITH was defined as the number of clonal subpopulations (SPs) and was estimated for each sample with the Expanding Ploidy and Allele Frequency on Nested Subpopulations (*Expands*) method (Andor *et al.*, 2014) using its R package (version 2.1.2).

*Expands* takes simple nucleotide variant (SNV) as well as copy number variation (CNV) data sets as input. TCGA provides copy numbers in form of segment mean values which were converted back into copy numbers by 2*2^(segment mean). *Expands* also requires a binary value indicating if a variant is germline. Here, since masked data is used, all variants are treated as being of tumor origin. Then, to assign the average copy number (among all cells) estimated for regions provided in the CNV data set to the overlapping variants in the SNV set, the `assignQuantityToMutation` function was used. The cellular frequencies of each mutation were then calculated with the `computeCellFrequencyDistributions` function which also calculates the density distributions for the probabilities of each mutation existing in a fraction of the cells. Finally, mutations with similar cellular frequencies were grouped with the `clusterCellFrequencies` function which applies hierarchical clustering on the probability distributions of the cellular frequencies. *Expands* was run on default parameters as is done in the demonstration of *Expands* with TCGA data (R-package vignette): maximum ploidy of mutated cells is set to 6, the upper threshold for the noise score of subpopulation detection is 0.7, the precision with which SPs are measured was set to 0.018.

This operation was performed for 9850 TCGA tumor samples on a computer cluster.

## Differential Expression Analysis

To reduce the parameter search space in the modeling step to genes with potential association to ITH, differential expression analysis (DEA) was performed between sample groups of *high-* and *low ITH* within each cancer type and then filtered DEA results for modeling.

### *Grouping samples according to low- and high ITH*

A sample was assigned to the *high ITH* group if the number of SPs for that sample was above the value defined for the upper quartile (*high ITH*: #SPs > 3Q.), and to the *low ITH* group if the number of SPs was below the value of the lower quartile (*low ITH*: #SPs < 1Q.), the quartiles being defined by the distribution of the number of SPs within a cancer type. Samples in the interquartile range were assigned to the *moderate ITH* group (*moderate ITH*: 1Q. ≤ SPs ≥ 3Q.) and were excluded from DEA.

### *Normalization of count data with edgeR*

The *edgeR* package version 3.28.0 (Robinson, Mccarthy and Smyth, 2010) was used to prepare the input data used for DEA from the downloaded count data. Firstly, to deal with variance among genes with low expression the `filterByExpr` function was used to remove genes with low counts. Next, the raw library size of each sample was scaled by their relative library sizes to make samples comparable. To do this, the `calcNormFactors` function was used, which implements the TMM (trimmed mean of M values) method for scaling. Finally, the counts within samples by counts per million mapped reads (CPM) were normalized using the `cpm` function.

### *DEA performed with limma package*

To perform DEA, the *limma* package (Ritchie *et al.*, 2015) version 3.42.2 was used. As *limma* was developed for microarray data, we used the `voom` function (Law *et al.*, 2014) to transform RNA-Seq counts so they can be used for *limma*. Additionally, `voom` estimates the mean-variance relationship in normalized count data and assigns weights to the counts of each gene according to its variance. A linear model is then fit for each gene with `lmFit` and an empirical Bayes method with `eBayes` function is applied to test whether the difference between groups is significant based on the model fit. DEA was performed for cancer types with >5 samples per each group of *low*- and *high ITH*, using the *low* group as reference. Finally, the output of the DEA was filtered to provide specific predictor variables for the subsequent modeling (covered below).

## Linear model fitting and variable subset selection with *lasso*

To model the association between gene expression and genetic ITH, generalized linear models were fit on gene expression data and genetic ITH estimates. To infer common markers between cancer types with modeling, significant differentially expressed genes (adjusted p-value < 0.05) in common between cancers were selected by filtering the DEA output for each cancer.

### *Constructing model training set and test set*

The downloaded count data for the *filtered* genes was normalized with the *edgeR* package (using TMM and CPM) as described above. For the model fitting process, a data matrix comprising samples as rows and the normalized counts of the *filtered* genes (predictor variables) as columns, plus an additional column for the ITH estimates (response variable) was created. For cancers with more than 150 samples, the rows of each data matrix were split into training set (75% of samples) and test set (25%) by sampling without replacement.

### *Fitting the generalized linear model with the glmnet package*

Coefficients of each predictor variable were calculated by fitting a generalized linear model with the shrinkage method *lasso* as applied in the *glmnet* package (Friedman, Hastie and Tibshirani, 2010) version 3.2-0. To fit the models, the tuning parameter for the *lasso* penalty term was selected by using the `cv.glmnet` function with 10-fold cross validation. The linear models were then fit with the `glmnet` function on the training data for each cancer type. To illustrate similarities between cancers, the obtained model coefficients were summarized in a heatmap (**Figure 1**) using the *pheatmap* package (Raivo Kolde, 2019) version 1.0.12.

### Testing the models

For each cancer with a test set, the selected models were first applied to obtain ITH predictions. Next, the root mean-square error (RMSE) was calculated between model predictions and the observed ITH estimates for the test set (**Table 1**, **Table 2**). Additionally, Pearson's correlation was estimated between model predictions and the original estimates with the `cor.test` function (base *R*).

### Searching for Markers Cancer by Cancer

Since repeating the above strategy with *PhyloWGS* estimates brought from literature (Raynaud *et al.*, 2018) yielded no models, models were built to be more comparable between *Expands* and *PhyloWGS* methods. For a fair comparison, only samples for which both methods had obtained estimates were used. Furthermore, the predictor variables for the models based on each ITH method from the DEA output were selected according to highest fold change and highest average expression (among samples). This was done by filtering 500 genes with highest absolute log fold change among the statistically significant (adjusted p-value < 0.05) and of these, 100 genes with highest mean expression (among samples) were used for the modeling.

## Discussion:

Performing the search for ITH markers with a pan-cancer approach revealed increased expression of DNAJC18 to be associated with lower ITH in *STAD, BRCA, HNSC* and *UCEC,* when estimating ITH with *Expands*. The biological meaning of DNAJC18 (and the other genes') expression in the context of genetic heterogeneity should be investigated further. An overview of the gene provided by UniProtKB, describes it as a putative member of the DnaJ family of chaperone proteins. As a homologue of known DnaJ proteins, it has been identified in the human genome through alignment and its annotation has been reviewed (Swiss-Prot). As chaperones, the DnaJ family proteins are associated with protein folding and have been functionally and structurally characterized elsewhere (Qiu *et al.*, 2006). Although the expression of this gene is clearly observed on transcript level, the existence of a protein has not been confirmed according to UniProtKB (https://www.uniprot.org/uniprot/Q9H819). This is important, since one of the advantages of identifying a specific gene among expression data, as I do in this project, is the potential use of its corresponding protein as a biomarker of ITH.

A cancer by cancer search was conducted to compare models based on *Expands* and *PhyloWGS* methods. The varying results obtained here with models based on *PhyloWGS* and *Expands* can be explained with how the models depend on the DEA. The fact that less predictive models are obtained with *PhyloWGS* indicate that the expression of the genes selected via DEA, although significantly differing between groups of *low*- and *high ITH*, in fact vary among the rest of the samples to an extent that no linear association can be modeled. To clarify on this, the *low*- and *high ITH* groups comprise samples with SPs estimated below and above the interquartile range of SPs in the sample cohort, respectively. This means that ca. half of the samples are excluded from DEA, while all samples are used in modeling. This would mean that groups formed according to *PhyloWGS* estimates might not be meaningful in the context of the differential expression. This can be tested by forming groups randomly and comparing the DEA output. To increase interpretability of the DEA results, a gene set enrichment analysis could be conducted. However, while the modeling results are related to the DEA output, it remains to be explained why the DEA results vary between groups defined with either ITH method. This demands a more detailed understanding of the algorithms used for estimating ITH with each method. How the estimations have accounted for the effect of copy number variations is of particular interest, as this can affect gene expression significantly.

## Acknowledgements:

First, I would like to thank Daniel Sobral, my supervisor at the Multi-Omics lab at FCT in Portugal for helping and supporting me throughout the project and throughout the quarantine, and for granting me the opportunity. I also want to thank Alma Andersson, my co-supervisor, for supporting me from Sweden, and Joakim Lundeberg for taking on the project at KTH. I also want to extend my gratitude to Arsénio Fialho at IST for making it possible for me to work on this project during my exchange at IST, and I want to thank Patrik Ståhl for examining the work at KTH. Finally, I want to thank Ana Rita Grosso, Rui Pinto, and João Neto at the Multi-Omics lab for the good times!

The computational analysis with Expands was performed using resources funded by the Lisboa2020 Operational Program through the INCD project (LISBOA-01-0145-FEDER-022153).

## References:

Abécassis, J. *et al.* (2019) 'Assessing reliability of intra-tumor heterogeneity estimates from single sample whole exome sequencing data', *PLOS ONE*. Edited by K. Ellrott, 14(11), p. e0224143. doi: 10.1371/journal.pone.0224143.

Andor, N. *et al.* (2014) 'Genome analysis EXPANDS: expanding ploidy and allele frequency on nested subpopulations', 30(1), pp. 50–60. doi: 10.1093/bioinformatics/btt622.

Andor, N. *et al.* (2016) 'Pan-cancer analysis of the extent and consequences of intratumor heterogeneity', *Nature Medicine*. Nature Publishing Group, 22(1), pp. 105–113. doi: 10.1038/nm.3984.

*Bioinformatics Pipeline: Copy Number Variation Analysis - GDC Docs* (no date). Available at: https://docs.gdc.cancer.gov/Data/Bioinformatics_Pipelines/CNV_Pipeline/ (Accessed: 28 April 2020).

*Bioinformatics Pipeline: mRNA Analysis - GDC Docs* (no date). Available at: https://docs.gdc.cancer.gov/Data/Bioinformatics_Pipelines/Expression_mRNA_Pipeline/#mrna-expression-workflow (Accessed: 28 April 2020).

Carter, S. L. *et al.* (2012) 'Absolute quantification of somatic DNA alterations in human cancer', *Nature Biotechnology*, 30(5), pp. 413–421. doi: 10.1038/nbt.2203.

Cibulskis, K. *et al.* (2013) 'Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples'. doi: 10.1038/nbt.2514.

*CRAN - Package pheatmap* (no date). Available at: https://cran.r-project.org/web/packages/pheatmap/index.html (Accessed: 30 April 2020).

Deshwar, A. G. *et al.* (2015) 'PhyloWGS: Reconstructing subclonal composition and evolution from whole-genome sequencing of tumors', *Genome Biology*. BioMed Central Ltd., 16(1), p. 35. doi: 10.1186/s13059-015-0602-8.

*File Format: MAF - GDC Docs* (no date). Available at: https://docs.gdc.cancer.gov/Data/File_Formats/MAF_Format/ (Accessed: 28 April 2020).

Friedman, J., Hastie, T. and Tibshirani, R. (2010) 'Regularization paths for generalized linear models via coordinate descent', *Journal of Statistical Software*. University of California at Los Angeles, 33(1), pp. 1–22. doi: 10.18637/jss.v033.i01.

Gerstung, M. *et al.* (2015) 'Combining gene mutation with gene expression data improves outcome prediction in myelodysplastic syndromes', *Nature Communications*. Nature Publishing Group, 6(1), pp. 1–11. doi: 10.1038/ncomms6901.

Greenman, C. *et al.* (2007) 'Patterns of somatic mutation in human cancer genomes', *Nature*, 446(7132), pp. 153–158. doi: 10.1038/nature05610.

Landau, D. A. *et al.* (2013) 'Evolution and impact of subclonal mutations in chronic lymphocytic leukemia’, *Cell*, 152(4), pp. 714–726. doi: 10.1016/j.cell.2013.01.019.

Law, C. W. *et al.* (2014) ‘Voom: Precision weights unlock linear model analysis tools for RNA-seq read counts’, *Genome Biology*. BioMed Central, 15(2), p. R29. doi: 10.1186/gb-2014-15-2-r29.

De Matos, M. R. *et al.* (2019) ‘A systematic pan-cancer analysis of genetic heterogeneity reveals associations with epigenetic modifiers’, *Cancers*. MDPI AG, 11(3). doi: 10.3390/cancers11030391.

Merlo, L. M. F. and Maley, C. C. (2010) ‘The role of genetic diversity in cancer’, *Journal of Clinical Investigation*, pp. 401–403. doi: 10.1172/JCI42088.

Morris, L. G. T. *et al.* (2016) ‘Pan-cancer analysis of intratumor heterogeneity as a prognostic determinant of survival’, *Oncotarget*. Impact Journals LLC, 7(9), pp. 10051–10063. doi: 10.18632/oncotarget.7067.

Mroz, E. A. and Rocco, J. W. (2013) ‘MATH, a novel measure of intratumor genetic heterogeneity, is high in poor-outcome classes of head and neck squamous cell carcinoma’, *Oral Oncology*. NIH Public Access, 49(3), pp. 211–215. doi: 10.1016/j.oraloncology.2012.09.007.

Qiu, X. B. *et al.* (2006) ‘The diversity of the DnaJ/Hsp40 family, the crucial partners for Hsp70 chaperones’, *Cellular and Molecular Life Sciences*. Springer, pp. 2560–2570. doi: 10.1007/s00018-006-6192-6.

Raynaud, F. *et al.* (2018) ‘Pan-cancer inference of intra-tumor heterogeneity reveals associations with different forms of genomic instability’, *PLOS Genetics*. Edited by M. Kool. Public Library of Science, 14(9), p. e1007669. doi: 10.1371/journal.pgen.1007669.

Ritchie, M. E. *et al.* (2015) ‘Limma powers differential expression analyses for RNA-sequencing and microarray studies’, *Nucleic Acids Research*. Oxford University Press, 43(7), p. e47. doi: 10.1093/nar/gkv007.

Robinson, M. D., Mccarthy, D. J. and Smyth, G. K. (2010) ‘edgeR: a Bioconductor package for differential expression analysis of digital gene expression data’, *BIOINFORMATICS APPLICATIONS NOTE*, 26(1), pp. 139–140. doi: 10.1093/bioinformatics/btp616.

Roth, A. *et al.* (2014) ‘PyClone: Statistical inference of clonal population structure in cancer’, *Nature Methods*. Nature Publishing Group, 11(4), pp. 396–398. doi: 10.1038/nmeth.2883.