TÉCNICO LISBOA

Deciphering mechanisms underlying tumor heterogeneity using Multi-Omics approaches

Joonas Avik

Thesis to obtain the Master of Science Degree in

Biotechnology

Supervisors: Dr. Daniel Vieira Noro e Silva Sobral,

Prof. Arsénio do Carmo Sales Mendes Fialho

Examination Committee

Chairperson: Prof. Leonilde de Fátima Morais Moreira

Supervisor: Dr. Daniel Vieira Noro e Silva Sobral

Members of the Committee: Dr. Marta Belchior Lopes

June 2020

The work presented in this thesis was performed at Departamento de Ciências da Vida of FCT/UNL (Caparica, Portugal), during the period December-May 2020, under the supervision of Daniel Sobral, using resources funded by the Lisboa2020 Operational Program. The thesis was co-supervised at Royal Institute of Technology (KTH) by Joakim Lundeberg and Alma Andersson. The work is presented as the master's thesis in biotechnology at Instituto Superior Técnico, under the examination of Arsénio Fialho, and master's thesis at KTH in medicinal biotechnology, under the examination of Patrik Ståhl.

Summary:

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

Table of Contents

List of abbreviations, figures, and tables: 6
Abbreviations 6
Figures 7
Tables 8
Introduction: 8
How Heterogeneity Arises 8
Measuring Heterogeneity 9
Heterogeneity as a Biomarker of Clinical Outcome 10
Searching for Markers of Heterogeneity 10
Project Goals 10
Results: 11
Searching for Markers of ITH that are Common Among Cancers 11
Testing the models 12
Performing the Same Search with ITH estimates of *PhyloWGS* yielded no models 13
Searching for Markers Cancer by Cancer 14
Methods: 20
The code 20
Data sets 20
Estimating genetic ITH with *Expands* 20
Differential Expression Analysis 21
Linear model fitting and variable subset selection with lasso 22
Testing the models 22
Comparing models based on *Expands* and *PhyloWGS* 22
Discussion: 23
Future Perspectives: 25
Acknowledgements: 26
Bibliography: 26

List of abbreviations, figures, and tables:

Abbreviations

ITH	intratumor heterogeneity
TCGA	The Cancer Genome Atlas
GDC	Genomic Data Commons
SNV	single nucleotide variation
CNV	copy number variation
DEA	differential expression analysis
lasso	least absolute shrinkage and selection operator
SPs	subpopulations
RMSE	root mean-square error
SDEGs	significantly differentially expressed genes
cor_P	Pearson's correlation coefficient

Cancer types (TCGA study abbreviations):

LUAD	lung adenocarcinoma
STAD	stomach adenocarcinoma
BRCA	breast invasive carcinoma
HNSC	head and neck squamous cell carcinoma
UCEC	uterine corpus endometrial carcinoma
BLCA	bladder urothelial carcinoma
THYM	thymoma
KIRC	kidney renal clear cell carcinoma

Figures

Figure 1: *Heatmap of coefficients calculated with each model (cancer type, x-axis) for each predictor variable (gene, y-axis). The coefficients are those of linear models based on gene expression data and genetic ITH as defined with Expands. 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 (most of which yield zero-coefficients (white) due to the lasso penalty) that were differentially expressed in 6 cancer types: LUAD, STAD, BRCA, HNSC, UCEC, BLCA. No non-zero coefficients were calculated for BLCA, which 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, UCEC.*

Figure 2: *Plot of model predictions (y-axis) as a number of subpopulations (SPs) against observed SPs provided with Expands (x-axis) for each sample in the test set of STAD. This model predicts ITH in STAD with a linear combination of the expressions of a small subset of 5 genes (**Figure 1**). The plot displays predictions with relatively high variance – a root mean-square error (RMSE) of 3.77 (**Table 1**) because the Expands ITH estimates have a high range between 1 and 20 SPs. Nonetheless, correlation is clearly visible, and this is also portrayed by a Pearson's correlation estimate of 0.55 with a very small p-value of 2.16e-8.*

Figure 3: *Boxplots displaying the distribution of SPs (x-axis) with the line represents median SPs, the length of the box representing the interquartile range, and the dots are outliers. Boxplots are displayed for all cancers (y-axis) for different methods: PhyloWGS (left) and Expands (right). The. Cancer types are ordered according to reducing cohort size from the top. The ordering of the cancers according to cohort size reveals no distinct pattern of similarity between methods. Instead, the most noticeable difference is that PhyloWGS output (left) are continuous values and Expands (right) outputs integer values for SPs.*

Figure 4: *Stacked barplots displaying the sample groups of STAD with either method: PhyloWGS (left) and Expands (right). The bars represent the sizes of the groups formed with one method and the bars are coloured according to the groups to which the samples would befall with the other method. As expected, this shows little similarity between the composition of the groups. Despite classifying samples according to quartiles, the size difference in Expands groups was expected due to the methods discrete output values, as opposed to continuous PhyloWGS output. Groups sizes demand careful consideration as they are likely to affect DEA outcome. This difference in group composition and group sizes was noted for all cancer types.*

Figure 5: *ITH estimates (SPs) from PhyloWGS (left) and Expands methods (right) plotted against model ITH predictions for STAD, BRCA and HNSC. The genes used for modeling were selected separately for each method and cancer type from their respective DEA output. Expands yielded models with significant Pearson correlation for all cancer types displayed here. Models for LUAD were also compared but are not displayed in plots. Both PhyloWGS and Expands methods gave significant correlation between predictions and estimates for STAD, yet with varying results (*cor_P *= 0.31 and 0.69).*

Tables

Table 1: *Test errors (RMSE) calculated for models based on Expands 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, further indicating that the calculated coefficients succeed in displaying an association between the expression of the captured subset of genes and genetic ITH.*

Table 2: *Adjusted Rand index was calculated to measure the similarity between the groups of samples based on the ITH estimates of PhyloWGS and Expands methods. Adjusted Rand index of 1 would indicate identical labels of low, moderate, and high, as applied here according to the ITH estimates of either method. The near-zero adjusted Rand indices displayed here suggest very low similarity between the ITH estimates of PhyloWGS and Expands. In the Table are given 8 cancer types. Overall, adjusted Rand index below 0.1 was measured for 29 out of 32 cancer types, and below 0.01 for 19 cancer types.*

Table 3: *This Table summarizes the DEA results as the number of significantly differentially expressed genes (SDEGs) obtained from DEA runs with groups of low- and high ITH based on either ITH method, with varying group sizes depending on which ITH methods is used (Expands samples, and PhyloWGS samples) are displayed. The groups are formed with samples for which both ITH methods had estimated ITH (Merged samples). The varying composition of the DEA outputs obtained with the two ITH methods is displayed as the number of SDEGs in common among DEA outputs (Intersecting SDEGs). Finally, after filtering top 100 genes with highest fold change and highest average expression among the SDEGs of each method, the similarity between these groups of 100 are displayed as the number of intersecting genes (Intersecting top 100).*

Table 4: *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, as can be seen from the plots in Figure 5 and are therefore not directly comparable.*

Introduction:

Cancer is a complex disease and presents one of the greatest challenges in modern medicine. Despite remarkable advances in treatment of several cancer types, cancer relapse and resistance to therapy remain recurring outcomes in patients. Complications in cancer treatment and prognosis owe in part to the vast variety of cases within cancer types (termed inter-tumor heterogeneity) as well as the diverse cellular architecture of individual tumors, known as intratumor heterogeneity (ITH).

How Heterogeneity Arises

Cancers are characterized by unstable genomes, accumulating mutation faster than can be explained by increased cell division rates (Loeb, 2010). As variants accumulate individually in tumor cells, genetic intratumor heterogeneity (ITH) is manifested in cellular subpopulations which possess distinct genotypes. This diversity 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). While all variants of the initial clone are inherited in the expanding population, subclonal mutations continue to arise and cause the branching of the tumor phylogenetic tree (Gerlinger et al., 2012). Exploring the mutational landscape of a tumor then 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).

Measuring Heterogeneity

A heterogeneous cancer in theory can provide an assemblage of subclones resistant to any therapeutic agent, which means heterogeneity might be a potential, quantifiable biomarker for cancer prognosis (Merlo & Maley, 2010). This has led to the development of a variety of algorithmic methods for measuring ITH (further denoted ITH methods) from genomic data. In brief, inferring heterogeneity from bulk data can be done by assuming the heritability of genotypes. This would mean that single nucleotide variants (SNVs) of similar observed frequencies might be representative of a subpopulation of closely related cells. The allelic frequency of a SNV is observed as sequencing reads carrying the variant over total reads mapped to the mutated locus. A method called *Mutant Allele Tumor Heterogeneity* (MATH) quantifies heterogeneity based on the distribution of these frequencies, with a wider distribution representing higher genetic diversity among the sampled cells (Mroz & Rocco, 2013). However, observed variant allele frequencies are not always representative of the fraction of cells carrying the variant. Copy number variations (CNVs), which are frequently observed in cancer genomes affect the observed frequency of variants embedded in the altered region (Carter et al., 2012). Accounting for this effect is especially important for cancers with highly aberrant genomes where CNVs affect large portions of the genomes (Noorbakhsh et al., 2018).

To improve heterogeneity estimation, methods such as *Expands* (short for *Expanding Ploidy and Allele Frequencies on Nested Subpopulations*) account for CNVs while providing a reconstruction of the tumor subpopulations. In brief, to estimate the number of subpopulations in the tumor sample, *Expands* corrects the observed allelic frequencies of SNVs for copy number alterations by overlaying CNV and SNV data sets. Then, for each SNV, the method calculates the fraction of cells carrying the SNV and the probability that the SNV existists in that fraction of the cells. It then uses hierarchical clustering on the distribution of those probabilities, grouping SNVs with similar cell fractions into subpopulations. The clustering is based on the assumption that passenger mutations accumulate in the cell before a driver event leads to clonal expansion and that these mutations consequently have the same frequencies in the population (Andor et al., 2014). Another popular method called PyClone provides a similar reconstruction of subpopulations while using different mathematical models (Roth et al., 2014).

Furthermore, a method called *PhyloWGS* was developed for the automated reconstruction of phylogenetic relationships between tumor subpopulations. It works by fitting variant allele frequencies into rooted tree structures defined by a set of rules that comply with evolutionary modes observed in earlier studies (Nik-Zainal et al., 2012). It thereby attempts to model linear and branched tumor phylogenies and relies on a Bayesian method to find the tree that best fits the data. As a result of this methodology, *PhyloWGS* outputs the number of subpopulations as the number of leaf nodes in the best tree.

Heterogeneity as a Biomarker of Clinical Outcome

Large scale genomic initiatives such as *The Cancer Genome Atlas* (TCGA) project, has generated plentiful genetic data which has been systematically processed and characterized in centralized TCGA workflows for tumor biopsies. Among other data types, TCGA has profiled SNVs as well CNVs for many cancer types, as well as clinical data of cases. Given the plenitude of data, several ITH methods have been applied for investigating how ITH may be related to clinical outcome. These studies generally correlate increased ITH measures to poor clinical outcome. For example, estimation using 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 mutations according to their estimated prevalence while accounting for copy number alterations.

Searching for Markers of Heterogeneity

The apparent clinical significance of the estimations made with a variety of ITH methods has spurred the search for underlying mechanisms of ITH. On the genetic level, the association between genetic ITH estimates and the variants from which they originate have been studied. 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). Following the same logic, the association between SNV load and ITH estimates made with the *MATH* approach (Mroz & Rocco, 2013), which considers only SNVs, have been made (De Matos et al., 2019).

In search for specific causal variants, mutations in epigenetic modifier genes have been noted in tumors of higher heterogeneity (De Matos et al., 2019). The authors modeled ITH (quantified with the *MATH* approach) on sets of mutated genes by applying a linear models with a shrinkage method called *lasso* or L1-shrinkage (Friedman et al., 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 expression data of ca. 20 000 genes, the association between general expression profiles and specific genetic variants was modeled. Through this integrative approach, the authors linked tumor genotypes to their phenotypic profiles which undergo selection.

Lasso (short for *Least Absolute Shrinkage and Selection Operator*) is a useful method when working with large data sets because it yields sparse models. Similar to ridge regression, it fits the model by calculating coefficients that minimize the residual sum of squares of the fit while penalizing large coefficients. However, the difference between the regularization techniques is that *lasso* shrinks the coefficients of less important features to zero, and thereby provides a model of lower complexity.

Project Goals

In this project, the genotype-phenotype relation is further explored by modeling the association between the transcriptome and the genome. 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, if associations with

specific genes can be found, promises to pinpoint potential biomarkers that could be further tested in proteomic assays. The project is carried out as a pan-cancer study aimed at identifying ITH markers that are common among cancers, making use of all available SNV and CNV data gathered in TCGA projects.

Results:

The goal of the project was to find specific gene expression profiles that are associated with genetic intratumor heterogeneity (ITH) estimates for TCGA samples by using a linear model. I performed the analysis with a pan-cancer approach, the goal of which was to identify ITH features in common among cancers. To achieve this, I chose a strategy that relies on differential expression analysis (DEA) for variable selection among expression data ahead of modeling (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, I calculated SPs 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 I first reduced the search space among expression data to potentially meaningful genes. To do this, I performed differential expression analysis (DEA) between groups of low- and high ITH, based on *Expands* ITH estimates (see methods). Since I excluded cancers with less than 6 samples per group, 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, I performed the linear model fitting with a shrinkage method called *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, I 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.

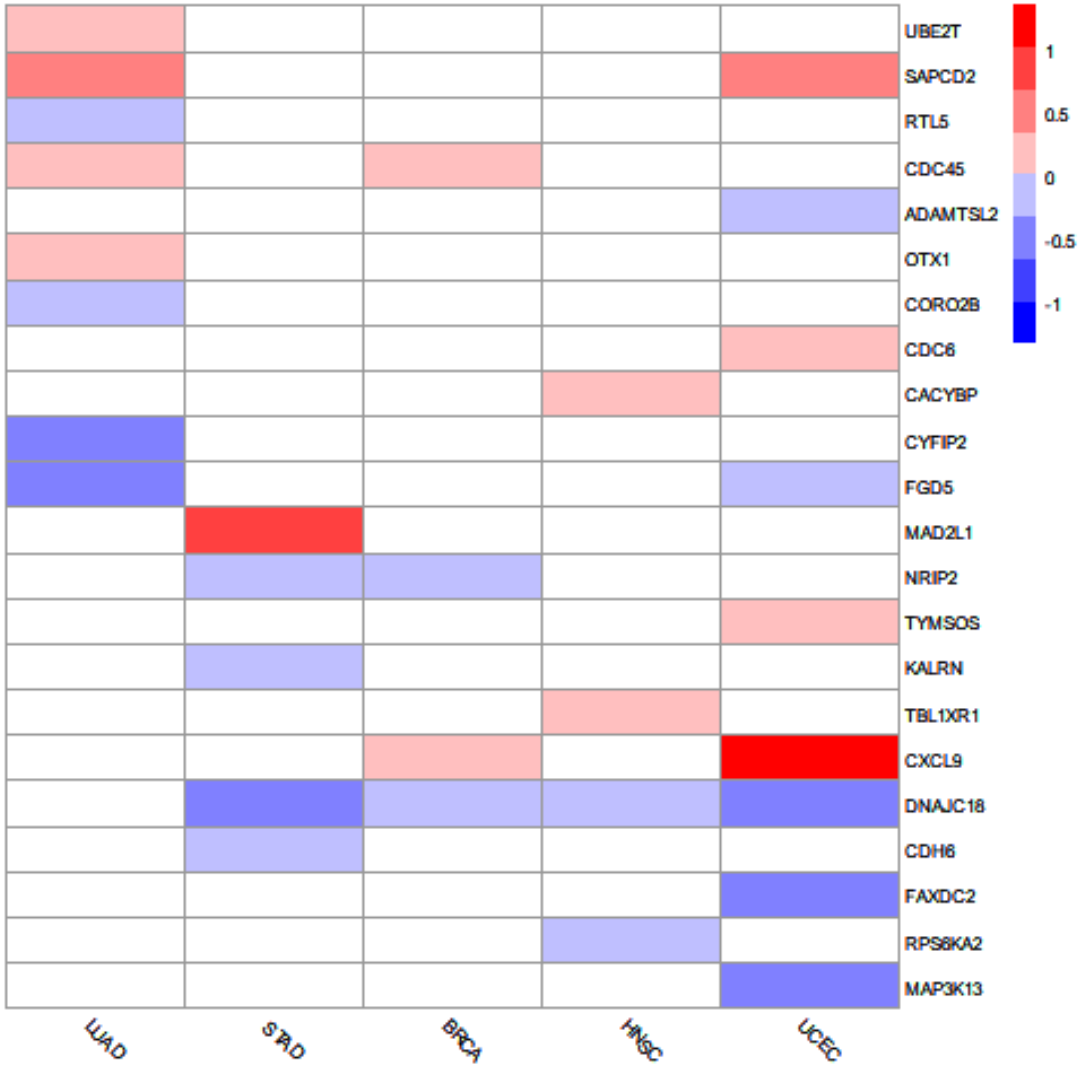

Figure 1: *Heatmap of coefficients calculated with each model (cancer type, x-axis) for each predictor variable (gene, y-axis). The coefficients are those of linear models based on gene expression data and genetic ITH as defined with Expands. A negative (blue) and positive (red) coefficient of a gene 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 (most of which yield zero-coefficients (white) due to the lasso penalty) that were differentially expressed in 6 cancer types: LUAD, STAD, BRCA, HNSC, UCEC, BLCA. No non-zero coefficients were calculated for BLCA, which 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, UCEC.*

Testing the models

To test the models, I calculated the root mean square errors (RMSE) between model predictions and *Expands* ITH estimates for LUAD, STAD, BRCA, HNSC, UCEC (**Table 1**). The relatively large RMSE's of models built for LUAD, STAD, BRCA, HNSC and UCEC indicate that the predictions are not precise (**Table 1**). However, Pearson correlation calculated between model predictions and the *Expands* ITH estimates (**Figure 2**) 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**). However, 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: *Test errors (RMSE) calculated for models based on Expands 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, 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*

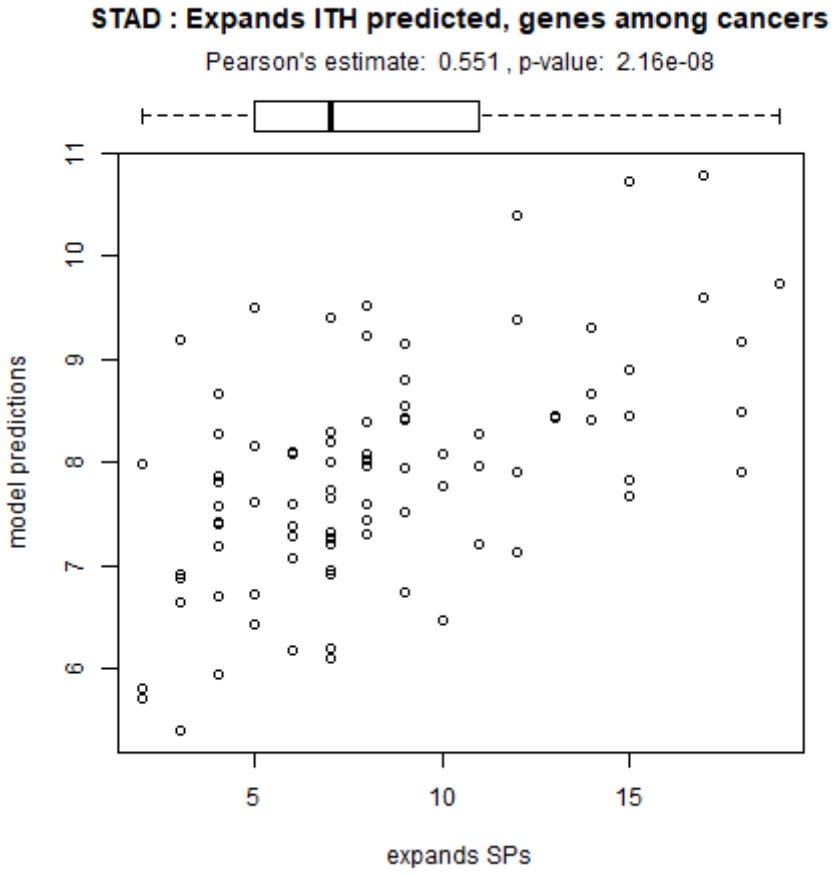

Figure 2: *Plot of model predictions (y-axis) as a number of subpopulations (SPs) against observed SPs provided with Expands (x-axis) for each sample in the test set of STAD. This model predicts ITH in STAD with a linear combination of the expressions of a small subset of 5 genes (**Figure 1**). The plot displays predictions with relatively high variance – a root mean-square error (RMSE) of 3.77 (**Table 1**) because the Expands ITH estimates have a high range between 1 and 20 SPs. Nonetheless, correlation is clearly visible, and this is also portrayed by a Pearson's correlation estimate of 0.55 with a very small p-value of 2.16e-8.*

Performing the Same Search with ITH estimates of *PhyloWGS* yielded no models

The same strategy for finding features in common among cancers was tested with *PhyloWGS* estimates taken from literature (Raynaud et al., 2018). Firstly, performing DEA and filtering genes with significant (p-adjusted < 0.05) differential expression across cancers yielded 35 genes which were in common among 6 cancer types: THYM, BRCA, PRAD, LUAD, STAD, KIRC. Of this set of 35 genes, only 2 intersected with the set of 38 genes filtered

for *Expands* (*UBE2T* and *MAD2L1*). Further, the modeling process was performed exactly as before, but yielded only one coefficient with the model based on THYM samples, and no coefficients for the rest of the cancers.

Searching for Markers Cancer by Cancer

The failure of obtaining any coefficients for the linear models based on ITH estimated with *PhyloWGS* could have been due to the DEA results that were filtered ahead of modeling. To find features in common among cancers, I had only used intersecting DE genes for modeling. This might have provided genes with low log fold changes between groups of *low*- and *high ITH*, as well as high variability between samples excluded from DEA but included in modeling (see methods).

Here, to see whether linear models could be obtained with *PhyloWGS*, I instead filtered the DEA output based on log fold change and average expression (among samples) to extract potentially more meaningful genes in a cancer-by-cancer search (see methods section). To provide a comparison between *Expands* and *PhyloWGS* methods, I only used samples for which both methods had obtained ITH estimates.

Firstly, to compare the output of each method as the number of subpopulations (SPs), the ranges of SPs were displayed in a boxplot (**Figure 3**). As expected, both *PhyloWGS* and *Expands* display varying ranges for SPs between cancers due to the variability between cancers known as inter tumor heterogeneity. However, when ordering cancers according to cohort size, decreasing from the top, no striking pattern of similarity can be noticed between methods despite the same samples being used. Furthermore, Pearson's correlation (cor_p) test between SPs of either method revealed very low, yet significant overall correlation of $cor_p = 0.2$ with all samples. An important distinction between the methods' output is that *PhyloWGS* estimates are on a continuous range, while *Expands* provides discrete values for the number of SPs (**Figure 3**).

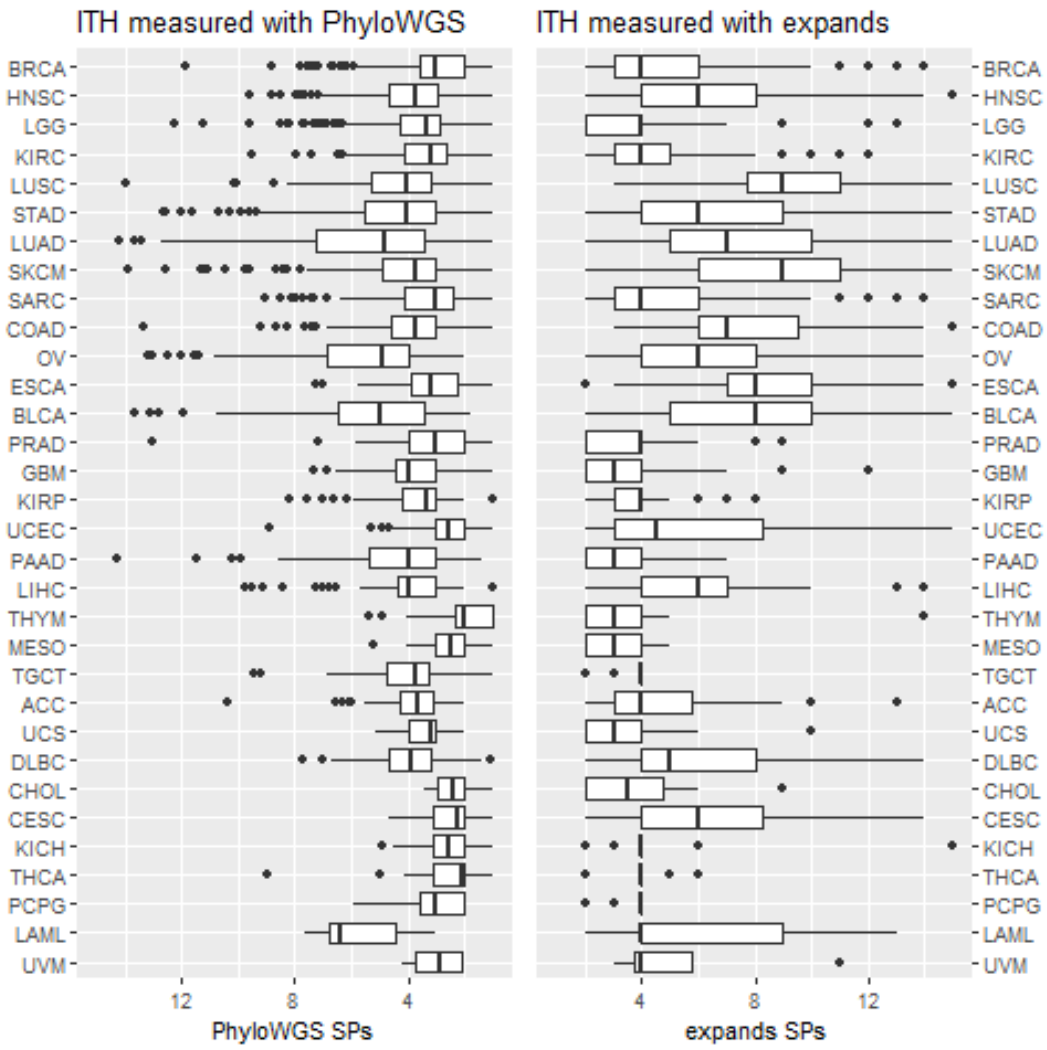

Figure 3: *Boxplots displaying the distribution of SPs (x-axis) with the line represents median SPs, the length of the box representing the interquartile range, and the dots are outliers. Boxplots are displayed for all cancers (y-axis) for different methods: PhyloWGS (left) and Expands (right). The. Cancer types are ordered according to reducing cohort size from the top. The ordering of the cancers according to cohort size reveals no distinct pattern of similarity between methods. Instead, the most noticeable difference is that PhyloWGS output (left) are continuous values and Expands (right) outputs integer values for SPs.*

As before, the selection of genes for the modeling was based on the results of DEA conducted between groups of samples of *low*- and *high ITH*. These groups were based on each ITH method separately. As the groups are formed for each cancer according to the distributions displayed in the boxplot (**Figure 3**), the noted distinction between output data types (i.e. continuous for *PhyloWGS*, and discrete for *Expands*) was expected to have consequences for the sizes of the groups based on either ITH method (see methods). Furthermore, a comparison of the composition of the groups based either ITH method showed very low similarity (**Figure 4**). As a measure of similarity, the Adjusted Rand index was calculated between samples' group labels as assigned with each ITH method (**Table 2**). Adjusted Rand index below 0.1 was calculated for 29 out of 32 cancer types and 19 of these were below 0.01. This means that for each ITH method, the DEA was conducted between groups composed of very different samples and that the cases of intersecting samples were likely random (probability was not estimated).

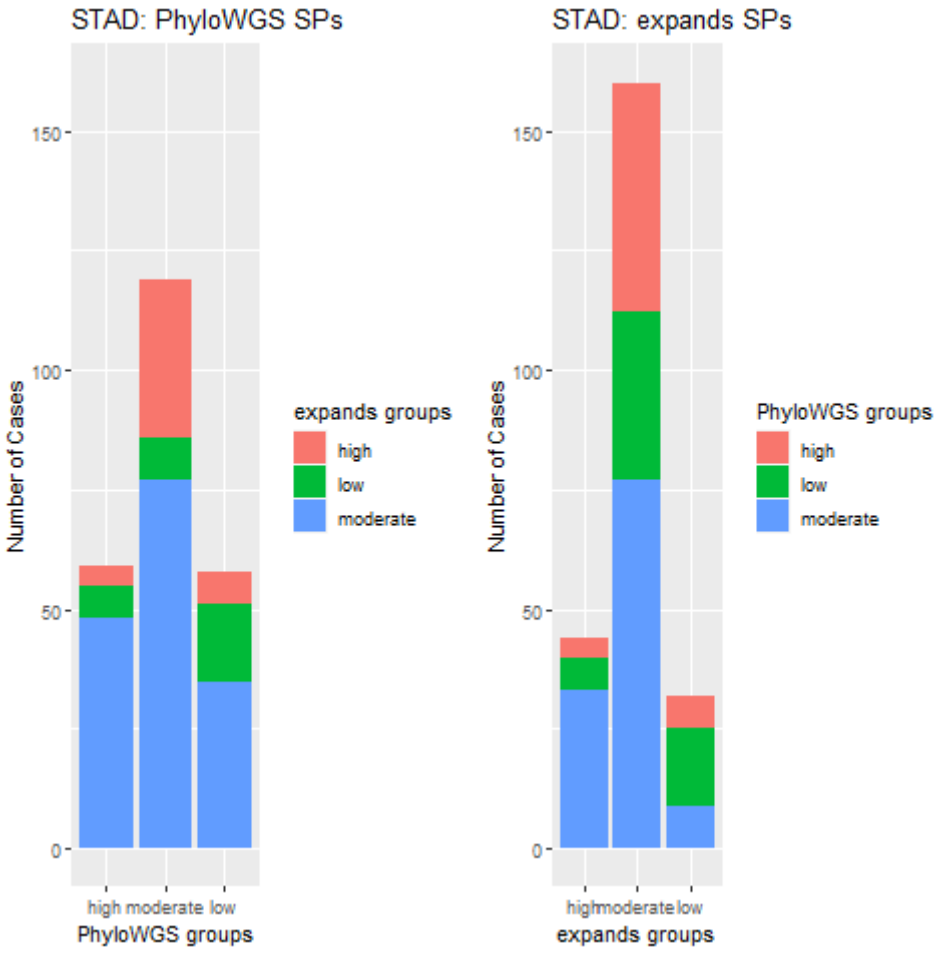

Figure 4: *Stacked barplots displaying the sample groups of STAD with either method: PhyloWGS (left) and Expands (right). The bars represent the sizes of the groups formed with one method and the bars are coloured according to the groups to which the samples would befall with the other method. As expected, this shows little similarity between the composition of the groups. Despite classifying samples according to quartiles, the size difference in Expands groups was expected due to the methods discrete output values, as opposed to continuous PhyloWGS output. Groups sizes demand careful consideration as they are likely to affect DEA outcome. This difference in group composition and group sizes was noted for all cancer types.*

Cancer type	Adjusted Rand Index
BRCA	-0.0064
HNSC	0.0015
LGG	0.0045
KIRC	0.0017
LUSC	0.019
STAD	0.011
LUAD	-0.0013
UCEC	0.11

Table 2 (left): *Adjusted Rand index was calculated to measure the similarity between the groups of samples based on the ITH estimates of PhyloWGS and Expands methods. Adjusted Rand index of 1 would indicate identical labels of low, moderate, and high, as applied here according to the ITH estimates of either method. The near-zero adjusted Rand indices displayed here suggest very low similarity between the ITH estimates of PhyloWGS and Expands. In the Table are given 8 cancer types. Overall, adjusted Rand index below 0.1 was measured for 29 out of 32 cancer types, and below 0.01 for 19 cancer types.*

The DEA yielded significant differentially expressed genes (SDEGs, p-adjusted < 0.05) with both ITH methods for 5 cancers: LUAD, LUSC, HNSC, BRCA, AND STAD (**Table 3**). The DEA results varied largely in the number of SDEGs and number of intersecting genes (**Table 3**) between ITH methods which was expected given the noted difference in the composition of the DEA groups based on either ITH method displayed earlier (**Figure 4**).

Table 3: *This Table summarizes the DEA results as the number of significantly differentially expressed genes (SDEGs) obtained from DEA runs with groups of low- and high ITH based on either ITH method, with varying group sizes depending on which ITH methods is used (Expands samples, and PhyloWGS samples) are displayed. The groups are formed with samples for which both ITH methods had estimated ITH (Merged samples). The varying composition of the DEA outputs obtained with the two ITH methods is displayed as the number of SDEGs in common among DEA outputs (Intersecting SDEGs). Finally, after filtering top 100 genes with highest fold change and highest average expression among the SDEGs of each method, the similarity between these groups of 100 are displayed as the number of intersecting genes (Intersecting top 100).*

Cancer types	Merged samples	DEA group	*Expands* samples	PhyloWGS samples	*Expands* SDEGs	PhyloWGS SDEGs	Inter-secting SDEGs	Inter-secting top 100
LUAD	212	low:	50	62	2457	2113	930	33
		high:	45	45				
LUSC	265	low:	64	68	21	153	3	-
		high:	66	66				
HNSC	426	low:	60	118	3658	617	339	4
		high:	94	94				
BRCA	686	low:	136	210	7025	3440	1995	29
		high:	134	134				
STAD	197	low:	28	56	7262	1763	631	1
		high:	35	44				

To perform modeling, I filtered the DE genes by highest fold change and highest average expression (among samples) to obtain sets of 100 genes to be used as predictor variables for each cancer and each ITH method separately. 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 4**).

Table 4: *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, as can be seen from the plots in Figure 5 and are therefore not directly comparable.*

	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

In contrast to the first experiment, in which using intersecting DE genes did not yield models for *PhyloWGS*, here by using genes filtered by highest log fold change, I get significant coefficients with models based on *Expands*, as well as *PhyloWGS*. This is unsurprising because I select genes with more varied expression between discrete groups of *low-* and *high ITH* before modeling their expression on a scale from low to high ITH.

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 $cor_P = 0.693$ 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 4**. While the RMSE values of the predictions would indicate worse precision of models based on *Expands*, in must be noted that these values are calculated on different scales (**Figure 5**) and are therefore not comparable.

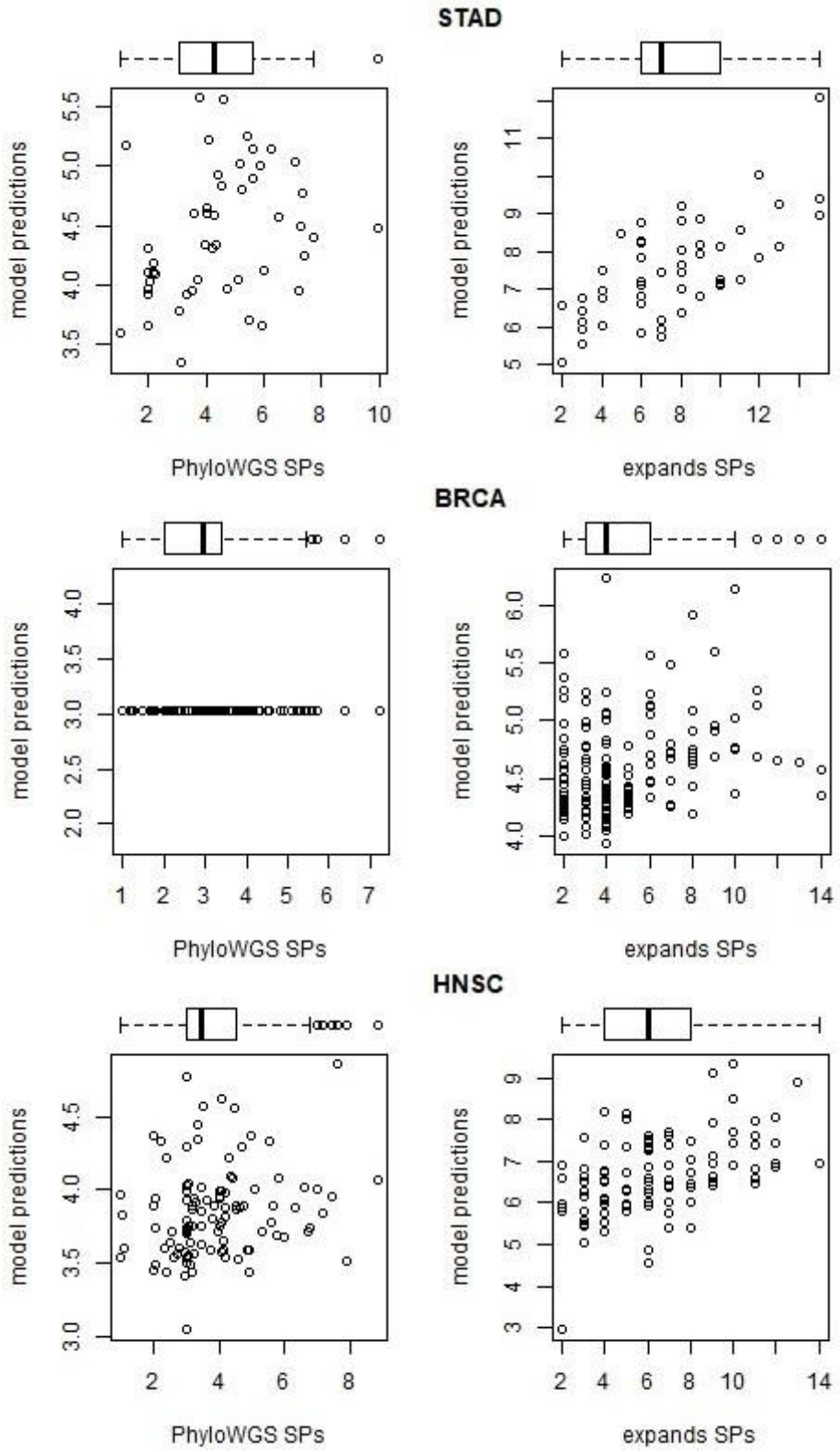

Figure 5: *ITH estimates (SPs) from PhyloWGS (left) and Expands methods (right) plotted against model ITH predictions for STAD, BRCA and HNSC. The genes used for modeling were selected separately for each method and cancer type from their respective DEA output. Expands yielded models with significant Pearson correlation for all cancer types displayed here. Models for LUAD were also compared but are not displayed in plots. Both PhyloWGS and Expands methods gave significant correlation between predictions and estimates for STAD, yet with varying results (*cor_P *= 0.31 and 0.69).*

Methods:

I used a linear model 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.

The genetic data: SNV and CNV sets

The TCGA data that I 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., 2013a) on whole exome sequencing data (*File Format: MAF - GDC Docs*, n.d.). 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. I used masked SNV data sets, 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*, n.d.). As such, the CNV data comprises coordinates of genomic regions and the copy number for these regions estimated from microarray intensities.

The 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*, n.d.). 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 is 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 I converted back into copy numbers by 2*2^(segment mean). *Expands* also requires a binary value indicating if a variant is germline. Here, since I am using masked data, 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 set to the overlapping variants in the SNV set, I used the `assignQuantityToMutation` function. I then computed cellular frequencies of each mutation with the `computeCellFrequencyDistributions` function which also calculates the density distributions for the probabilities of each mutation existing in a fraction of the cells. Finally, I grouped mutations with similar cellular frequencies with the `clusterCellFrequencies` function which applies hierarchical clustering on the probability distributions of the cellular frequencies. I run *Expands* 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 is 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, I performed differential expression analysis (DEA) 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

I assigned a sample 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

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

DEA performed with limma

To perform DEA, I used the *limma* package (Ritchie et al., 2015) version 3.42.2. As *limma* was developed for microarray data, I used the `voom` function (Law et al., 2014) to transform RNA-Seq counts so they can be used for *limma*. Additionally, I `voom` to estimates the mean-variance relationship in normalized count data and

assigns weights to the counts of each gene according to its variance. I then fit a linear model for each gene with `lmFit` and apply an empirical Bayes method with `eBayes` function to test whether the difference between groups is significant based on the model fit. I performed DEA for cancer types with >5 samples per each group of *low*- and *high ITH*, using the *low* group as reference. Finally, I filter the output of the DEA 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, I fit generalized linear models upon gene expression data and genetic ITH estimates. To infer common markers between cancer types with modeling, I selected significant differentially expressed genes (adjusted p-value < 0.05) in common between cancers by filtering the DEA output for each cancer.

Constructing the model training set and test set

I normalized the downloaded count data for the *filtered* genes with the *edgeR* package (using TMM and CPM) as described above. For the model fitting process, I created 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). For cancers with more than 150 samples, I split the rows of each data matrix into training set (75% of samples) and test set (25%) by sampling without replacement.

Fitting the generalized linear model with glmnet

I calculated the coefficients of each predictor variable by fitting a generalized linear model with the shrinkage method *lasso* as applied in the *glmnet* package (Friedman et al., 2010) version 3.2-0. To fit the models, I first select the tuning parameter [lambda] for the *lasso* penalty term by using the `cv.glmnet` function with 10-fold cross validation. I then fit the linear model with the `glmnet` function on the training data for each cancer type. To illustrate similarities between cancers, I summarized the obtained model coefficients 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, I first applied the model to obtain ITH predictions. Next, I calculated the root mean square error (RMSE) between model predictions and the observed ITH estimates for the test set (**Table 3**). Additionally, I calculated the Pearson's correlation between model predictions and the original estimates with the `cor.test` function (base *R*).

Comparing models based on *Expands* and *PhyloWGS*

Since repeating the above strategy with *PhyloWGS* estimates brought from literature (Raynaud et al., 2018) yielded no results, I built models to be more comparable between *Expands* and *PhyloWGS* methods. For a fair comparison, I used only samples for which both methods had obtained estimates. Furthermore, I filtered the predictor variables for the models based on each ITH method from the DEA output 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.

Ahead of modeling, I first compare the outputs of *Expands* and *PhyloWGS*, by creating a boxplot (**Figure 3**) with the *ggplot2* package (https://ggplot2.tidyverse.org/) version 3.3.0. Then, I compared the DEA input groups of *low*- and *high ITH* based on each ITH methods by calculating the adjusted Rand index between samples' ITH group labels (*low, moderate* and *high*) with the *mclust* package (Scrucca et al., 2016) version 5.4.5. Furthermore, to visualize the differences between groups based on each ITH method, I created stacked barplots (**Figure 5**) with the *ggplot2* package (https://ggplot2.tidyverse.org/) version 3.3.0. Finally, I compared the models that could be obtained with both *PhyloWGS* and *Expands*. To do this, I summarized the test errors and Pearson correlations for models based on both ITH methods in **Table 6** and created side-by-side scatterplots of model predictions and ITH estimates (**Figure 5**) by using the `par` function in base R.

Discussion:

In this project, I conducted a search for molecular markers of intratumor heterogeneity (ITH) by using a linear model for the connection between estimated genetic ITH, and normalized gene expression data of all available tumor samples collected in association to The Cancer Genome Atlas (TCGA) project. A linear model was used because the ITH measures used (as number of subpopulations, SPs) displayed a continuous distribution for among samples in each cancer. I performed the analysis with a pan-cancer approach, the goal of which was to identify ITH markers in common among cancers. To achieve this, I chose a strategy that relies on differential expression analysis (DEA) for variable selection among expression data ahead of modeling. Next, I also did a cancer by cancer search while comparing the DEA results based on *Expands* and *PhyloWGS* and the models obtained for the two ITH methods.

Firstly, I will focus on the matter of how I use DEA to prior to modeling, and how this affects my results. I have chosen to use DEA to reduce the number of dependent variables and thereby deal with potential model overfitting. By filtering differentially expressed (DE) genes common among cancers, I narrow down the search space for gene expression profiles that could be associated with ITH from ca. 20 thousand genes in the whole transcriptome, to 38 and 36 genes for *Expands* and *PhyloWGS* models, respectively. While this might help with overfitting, I do ditch nearly the whole transcriptomic data set, and reduce the number of cancers used for modeling from 33 to 6. My justification for this is that I am looking for biomarkers. Thereby the strict filters do serve my purpose, and the way I use modeling for extracting signals does pinpoint individual genes as I intended. However, as I am modeling the ITH association to individual genes as their linear combinations, my models are not representative of the transcriptomic data; they represent the DEA results. Moreover, the dependency of my models on DEA is further highlighted by the fact that I filter genes based on differential expression in the context of heterogeneity (groups of *low*- and *high ITH*). Therefore, the expression of the filtered genes is inherently more likely to display an association with a change in ITH in the modeling. Once again, this serves my purpose for biomarker discovery, but it is not optimal for modeling a meaningful genotype-phenotype relationship. Also, I miss out on most of the potential associations that the linear model could portray. In the cancer-by-cancer

search, I select more genes to be used in modeling based on log fold changes in each cancer separately, but for the sake of comparing models based on *PhyloWGS* and *Expands* the number of cancers observed is reduced from 33 to 4.

There are a number of points that can be addressed to improve upon these issues. Firstly, the choice of using intersecting genes among 6 cancer types models based on both *Expands* and *PhyloWGS* is arbitrary and could easily be extended to include more cancer types and far more genes. In fact, modeling could simply be tested with the whole transcriptome, rendering DEA obsolete. This still leaves the issue of potential overfitting, which needs to be tested. There are also a number of arbitrary decisions made in the modeling process, such as sampling 25% of the samples test sets only for cancers with at least 150 samples. Alternatives for both the relative and total size of the test sets can be tested. Also, the number of folds in cross validation performed for model selection needs to be reconsidered to obtain more representative models. The same issue applies for the cancer-by-cancer search, which could be conducted separately for the ITH methods and comparing the results more generally by including all cancers for which results are obtained with either ITH method. It may be that ITH as portrayed by either ITH method might require different models to represent the relationship between gene expression and the genetic background, as it is portrayed by an ITH method.

The varying results obtained here with models based on *PhyloWGS* and *Expands* can also 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 also be conducted.

This brings me to the second point that must be discussed, which is the matter of ITH estimates being unreliable. This is why I tested my modeling strategy with two different ITH methods: *Expands* and *PhyloWGS.* As the correlation between the estimates of these methods is very weak ($cor_P = 0.2$) they cannot be used to validate the results obtained with either method. This is especially true since the models are based on DEA results, yielding almost entirely different sets of genes for models based on either ITH methods. I have demonstrated this in the results as differing numbers of significantly differentially expressed genes obtained for either method (**Table 3**). This is especially notable for BRCA and STAD, for which DEA based on *Expands* yielded 7025 and 7262 significant genes respectively, while for *PhyloWGS*, 3440 and 1763 significant genes were obtained. Since genes were filtered by fold change for the comparison, it is likely that *Expands* yielded models with higher predictive power (**Table 4**) due to the filtered genes having higher differential expression between groups defined with *Expands*. This information was available as log fold changes among the DEA output, but is not included in the results (although, I have demonstrated varying DEA group compositions in **Figure 3**).

While my modeling results are related to the DEA output (as discussed above), 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. Important to note here is that the authors of the study from which *PhyloWGS* estimates have been taken (Raynaud et al., 2018) used ABSOLUTE (Carter et al., 2012) to calculate the effect of copy number alterations on SNV frequencies ahead of applying *PhyloWGS* for heterogeneity estimation. To note one issue, ABSOLUTE and *Expands* have been shown to estimate tumor purity differently (Andor et al., 2014), but more information about the performance of all computational methods involved is needed to explain the differences that I observe. For example, the mutation calling algorithms used might differ. While *my* SNV data is based on *Mutect2* (Cibulskis et al., 2013b), the method used to obtain the SNV upon which the *PhyloWGS* estimates are based is not reported (Raynaud et al., 2018). Furthermore, while the TCGA raw data used is the same, the data could be processed differently given any updated routines in GDC repository – I downloaded all data in late 2019, while for (Raynaud et al., 2018) the data was downloaded in 2015 and 2018. The effects of all these things should simply be avoided by running the ITH methods on exact same data sets.

To address the specific associations that I obtained, I showed increased expression of DNAJC18 to be associated with lower ITH in *STAD, BRCA, HNSC* and *UCEC,* when defining 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.

Most of the genes for which a linear association could be found between expression and ITH are not displayed because their relevance is highly questionable given the highly varying ITH estimates between methods. The focus has thus been on whether the relationship between ITH, as estimated with *Expands* or *PhyloWGS*, and gene expression can be portrayed with a linear model.

Future Perspectives:

The search for underlying mechanisms of tumor heterogeneity is an effort towards personalized treatment. As the heterogeneity of cancer implies the existence of highly variable cases among patients, the successful treatment and survivability among them is likely to increase given the development of personalized-medicine strategies that this study aims to contribute to. Specifically, improved diagnosis and patient outcome could be possible with the discovery of quantifiable biomarkers of ITH, as these might serve to aid risk stratification and

guide medical decision making. The link between the genetic and the phenotypic heterogeneity remains to be elucidated and will demand accurate estimates of ITH. Single sample biopsies are and will continue to be for at least a few years the most available source of data. Therefore, development of novel algorithmic methods to reconstruct cellular populations from bulk data will continue. The resolution by which heterogeneity can be portrayed in single biopsy samples will increase with increasing sequencing coverage, and new data will be generated in large scale collaborative efforts analogous to the concluded National Institute of Health's TCGA project, which mainly stores whole exome sequencing- and transcriptomic data. The International Cancer Genome Consortium (ICGC) has gathered whole genome sequencing data, aimed at covering the rest of the 99% of the cancer genome, and recently published results of the Pan Cancer Analysis of Whole Genomes (PCAWG) (Campbell et al., 2020). Similar data collection efforts will be carried out in pursuit of higher genomic resolution and more detailed and broader comparison between cases. Long read sequencing technologies and linked read methods will increasingly provide phasing capability, further resolving tumor evolution from bulk data. Meanwhile, development of single cell sequencing methods will demand new bioinformatics tools while overcoming the need for reconstructing tumor heterogeneity with algorithmic methods as they are applied to bulk data (Lawson et al., 2018).

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).

Bibliography:

Andor, N., Graham, T. A., Jansen, M., Xia, L. C., Aktipis, C. A., Petritsch, C., Ji, H. P., & Maley, C. C. (2016). Pan-cancer analysis of the extent and consequences of intratumor heterogeneity. *Nature Medicine*, *22*(1), 105–113. https://doi.org/10.1038/nm.3984

Andor, N., Harness, J. V, Mü Ller, S., Mewes, H. W., & Petritsch, C. (2014). *Genome analysis EXPANDS: expanding ploidy and allele frequency on nested subpopulations*. *30*(1), 50–60. https://doi.org/10.1093/bioinformatics/btt622

Bioinformatics Pipeline: Copy Number Variation Analysis - GDC Docs. (n.d.). Retrieved April 28, 2020, from https://docs.gdc.cancer.gov/Data/Bioinformatics_Pipelines/CNV_Pipeline/

Bioinformatics Pipeline: mRNA Analysis - GDC Docs. (n.d.). Retrieved April 28, 2020, from https://docs.gdc.cancer.gov/Data/Bioinformatics_Pipelines/Expression_mRNA_Pipeline/#mrna-expression-workflow

Campbell, P. J., Getz, G., Korbel, J. O., Stuart, J. M., Jennings, J. L., Stein, L. D., Perry, M. D., Nahal-Bose, H. K., Ouellette, B. F. F., Li, C. H., Rheinbay, E., Nielsen, G. P., Sgroi, D. C., Wu, C. L., Faquin, W. C., Deshpande, V., Boutros, P. C., Lazar, A. J., Hoadley, K. A., … Zhang, J. (2020). Pan-cancer analysis of whole genomes. *Nature*, *578*(7793), 82–93. https://doi.org/10.1038/s41586-020-1969-6

Carter, S. L., Cibulskis, K., Helman, E., McKenna, A., Shen, H., Zack, T., Laird, P. W., Onofrio, R. C., Winckler, W., Weir, B. A., Beroukhim, R., Pellman, D., Levine, D. A., Lander, E. S., Meyerson, M., & Getz, G. (2012). Absolute quantification of somatic DNA alterations in human cancer. *Nature Biotechnology*, *30*(5), 413–421. https://doi.org/10.1038/nbt.2203

Cibulskis, K., Lawrence, M. S., Carter, S. L., Sivachenko, A., Jaffe, D., Sougnez, C., Gabriel, S., Meyerson, M., Lander, E. S., & Getz, G. (2013a). *Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples*. https://doi.org/10.1038/nbt.2514

Cibulskis, K., Lawrence, M. S., Carter, S. L., Sivachenko, A., Jaffe, D., Sougnez, C., Gabriel, S., Meyerson, M., Lander, E. S., & Getz, G. (2013b). Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nature Biotechnology*, *31*(3), 213–219. https://doi.org/10.1038/nbt.2514

CRAN - Package pheatmap. (n.d.). Retrieved April 30, 2020, from https://cran.r-project.org/web/packages/pheatmap/index.html

De Matos, M. R., Posaioana, I., Carvalho, F. S., Morais, V. A., Grosso, A. R., & De Almeida, S. F. (2019). A systematic pan-cancer analysis of genetic heterogeneity reveals associations with epigenetic modifiers. *Cancers*, *11*(3). https://doi.org/10.3390/cancers11030391

Deshwar, A. G., Vembu, S., Yung, C. K., Jang, G. H., Stein, L., & Morris, Q. (2015). PhyloWGS: Reconstructing subclonal composition and evolution from whole-genome sequencing of tumors. *Genome Biology*, *16*(1), 35. https://doi.org/10.1186/s13059-015-0602-8

File Format: MAF - GDC Docs. (n.d.). Retrieved April 28, 2020, from https://docs.gdc.cancer.gov/Data/File_Formats/MAF_Format/

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, *33*(1), 1–22. https://doi.org/10.18637/jss.v033.i01

Gerlinger, M., Rowan, A. J., Horswell, S., Larkin, J., Endesfelder, D., Gronroos, E., Martinez, P., Matthews, N., Stewart, A., Tarpey, P., Varela, I., Phillimore, B., Begum, S., McDonald, N. Q., Butler, A., Jones, D., Raine, K., Latimer, C., Santos, C. R., … Swanton, C. (2012). Intratumor Heterogeneity and Branched Evolution Revealed by Multiregion Sequencing. *New England Journal of Medicine*, *366*(10), 883–892. https://doi.org/10.1056/NEJMoa1113205

Gerstung, M., Pellagatti, A., Malcovati, L., Giagounidis, A., Della Porta, M. G., Jädersten, M., Dolatshad, H., Verma, A., Cross, N. C. P., Vyas, P., Killick, S., Hellström-Lindberg, E., Cazzola, M., Papaemmanuil, E., Campbell, P. J., & Boultwood, J. (2015). Combining gene mutation with gene expression data improves outcome prediction in myelodysplastic syndromes. *Nature Communications*, *6*(1), 1–11. https://doi.org/10.1038/ncomms6901

Greenman, C., Stephens, P., Smith, R., Dalgliesh, G. L., Hunter, C., Bignell, G., Davies, H., Teague, J., Butler, A., Stevens, C., Edkins, S., O'Meara, S., Vastrik, I., Schmidt, E. E., Avis, T., Barthorpe, S., Bhamra, G., Buck, G., Choudhury, B., … Stratton, M. R. (2007). Patterns of somatic mutation in human cancer genomes. *Nature*, *446*(7132), 153–158. https://doi.org/10.1038/nature05610

Landau, D. A., Carter, S. L., Stojanov, P., McKenna, A., Stevenson, K., Lawrence, M. S., Sougnez, C., Stewart, C., Sivachenko, A., Wang, L., Wan, Y., Zhang, W., Shukla, S. A., Vartanov, A., Fernandes, S. M., Saksena, G., Cibulskis, K., Tesar, B., Gabriel, S., … Wu, C. J. (2013). Evolution and impact of subclonal mutations in chronic lymphocytic leukemia. *Cell*, *152*(4), 714–726. https://doi.org/10.1016/j.cell.2013.01.019

Law, C. W., Chen, Y., Shi, W., & Smyth, G. K. (2014). Voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biology*, *15*(2), R29. https://doi.org/10.1186/gb-2014-15-2-r29

Lawson, D. A., Kessenbrock, K., Davis, R. T., Pervolarakis, N., & Werb, Z. (2018). Tumour heterogeneity and metastasis at single-cell resolution. In *Nature Cell Biology* (Vol. 20, Issue 12, pp. 1349–1360). Nature Publishing Group. https://doi.org/10.1038/s41556-018-0236-7

Loeb, L. A. (2010). Mutator phenotype in cancer: Origin and consequences. In *Seminars in Cancer Biology* (Vol. 20, Issue 5, pp. 279–280). https://doi.org/10.1016/j.semcancer.2010.10.006

Merlo, L. M. F., & Maley, C. C. (2010). The role of genetic diversity in cancer. In *Journal of Clinical Investigation* (Vol. 120, Issue 2, pp. 401–403). https://doi.org/10.1172/JCI42088

Morris, L. G. T., Riaz, N., Desrichard, A., Senbabaoglu, Y., Ari Hakimi, A., Makarov, V., Reis-Filho, J. S., & Chan, T. A. (2016). Pan-cancer analysis of intratumor heterogeneity as a prognostic determinant of survival. *Oncotarget*, *7*(9), 10051–10063. https://doi.org/10.18632/oncotarget.7067

Mroz, E. A., & 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*, *49*(3), 211–215. https://doi.org/10.1016/j.oraloncology.2012.09.007

Nik-Zainal, S., Van Loo, P., Wedge, D. C., Alexandrov, L. B., Greenman, C. D., Lau, K. W., Raine, K., Jones, D., Marshall, J., Ramakrishna, M., Shlien, A., Cooke, S. L., Hinton, J., Menzies, A., Stebbings, L. A., Leroy, C., Jia, M., Rance, R., Mudie, L. J., … Campbell, P. J. (2012). The life history of 21 breast cancers. *Cell*, *149*(5), 994–1007. https://doi.org/10.1016/j.cell.2012.04.023

Noorbakhsh, J., Kim, H., Namburi, S., & Chuang, J. H. (2018). Distribution-based measures of tumor heterogeneity are sensitive to mutation calling and lack strong clinical predictive power. *Scientific Reports*, *8*(1), 1–12. https://doi.org/10.1038/s41598-018-29154-7

Qiu, X. B., Shao, Y. M., Miao, S., & Wang, L. (2006). The diversity of the DnaJ/Hsp40 family, the crucial partners for Hsp70 chaperones. In *Cellular and Molecular Life Sciences* (Vol. 63, Issue 22, pp. 2560–2570). Springer. https://doi.org/10.1007/s00018-006-6192-6

Raynaud, F., Mina, M., Tavernari, D., & Ciriello, G. (2018). Pan-cancer inference of intra-tumor heterogeneity reveals associations with different forms of genomic instability. *PLOS Genetics*, *14*(9), e1007669. https://doi.org/10.1371/journal.pgen.1007669

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., & Smyth, G. K. (2015). Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, *43*(7), e47. https://doi.org/10.1093/nar/gkv007

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

Roth, A., Khattra, J., Yap, D., Wan, A., Laks, E., Biele, J., Ha, G., Aparicio, S., Bouchard-Côté, A., & Shah, S. P. (2014). PyClone: Statistical inference of clonal population structure in cancer. *Nature Methods*, *11*(4), 396–398. https://doi.org/10.1038/nmeth.2883

Scrucca, L., Fop, M., Murphy, T. B., & Raftery, A. E. (2016). Mclust 5: Clustering, classification and density estimation using Gaussian finite mixture models. *R Journal*, *8*(1), 289–317. https://doi.org/10.32614/rj-2016-021