BINF-F401 — *Analysis of functional and comparative genomic data*
Study of tumorous purity in breast cancer

Robin Petit[†]

[†]robpetit@ulb.ac.be

Introduction

Table 1 contains all the abbreviations used in this paper. Please refer to this table in case of incomprehension.

Abbreviation	Meaning
BRCA	Breast Cancer
CDF	Cumulated Density Function
ER	Estrogen Receptor
GSE	Gene Set Enrichment
IHC	Immunohistochemistry
LCIS	Lobular Breast Carcinoma *in situ*
NMF	Non-Negative Matrix Factorization
PDF	Probability Density Function
PR	Progesterone Receptor
TCGA	The Cancer Genome Atlas

Table 1: Meaning of common abbreviations used in this paper.

The purity of a tumor is the proportion of cancerous cells that is present in the tumor. It is pathologists' responsibility to determine this purity by analysing histopathological slides of the tumor and counting all the different types of cells. Yet, a pathologist is unable to analyse the whole tumor which would take an enormous amount of time. Therefore, depending on the objective of the slide analysis, several computer-aided methods have been developed to either segment the slides (Komura and Ishikawa, 2018; Sirinukunwattana et al., 2016; Xing and Yang, 2016), or analyse the tumor genetically to find the somatic DNA alterations (Carter et al., 2012) or even use whole-genome and whole-exome sequencing (Oesper et al., 2014).

The Cancer Genome Atlas (TCGA) (Weinstein et al., 2013) has become an important database for cancer-related data. 250 samples have been taken from there in order to analyze tumorous purity (both ABSOLUTE (Carter et al., 2012) and IHC) through mutations in genes (see Table 2), NMF of mRNA-seq and GSE (Subramanian et al., 2005).

Results

Comparison between IHC and ABSOLUTE purities

When comparing IHC purity versus ABSOLUTE purity, we observed that these measures only slightly correlate: Spearman's $r_S = 0.355$ and Pearson's $r_P = 0.36$. Yet, they do correlate significantly with $p < 10^{-8}$.

More specifically, IHC measures tend to be greater than ABSOLUTE evaluations (Figure 1) with a mean of IHC $-$ ABSOLUTE of 0.202 and a standard deviation of 0.175. This indicates that either pathologists overestimate homogeneity in tumors or ABSOLUTE underestimates it. It must be kept in mind that ABSOLUTE has been known to underestimate purity in certain cases (Oesper et al., 2014). Still, ABSOLUTE estimations are known to be highly accurate (Carter et al., 2012).

Influence of mutations on ABSOLUTE purity

The distribution of the number of mutations of each gene from Table 2 is shown on Figure 2. We can observe that for each gene, more than 50% of samples are not mutated, and that CDH1, GATA3, MAP3K1 are highly non-mutated (more than 80% of samples).

Though, when summing the mutations of every gene, the mode becomes 1 mutation with only around 25% of samples having no mutations at all, therefore around 75% of samples have at least one of these five genes which is mutated.

Moreover, the only pairs of genes showing significant correlation ($p < 0.05$) in the number of mutations are (CDH1, GATA3), (CDH1, TP53), (MAP3K1, TP53) and (PIK3CA, TP53). All the other pairs cannot be considered correlated (see Table 3a). Also, these correlations are negative and light ($-0.21 \leq r \leq -0.127$) meaning that a bigger number of mutations in TP53 is correlated with a lower number of mutations of CDH1, MAP3K1 and PIK3CA, and a bigger number of mutations of GATA3 is correlated with a lower number of mutations of CDH1.

When aggregating all the mutations together, the correlation coefficient of these four gene pairs stays roughly constant except for (PIK3CA, TP53) for which the correlation

HGNC Approved Symbol	Entrez ID	Related Cancer Type
CDH1	999	LCIS (Berx et al., 1998)
GATA3	2625	ER-α related (Ciocca et al., 2009)
MAP3K1	4214	Invasive BRCA (Easton et al., 2007)
PIK3CA	5290	Invasive BRCA and ER/PR-related (Saal et al., 2005)
TP53	7157	BRCA (Gasco et al., 2002)

Table 2: Genes considered in this study with both their HGNC symbol which (used further down) and their Entrez ID.

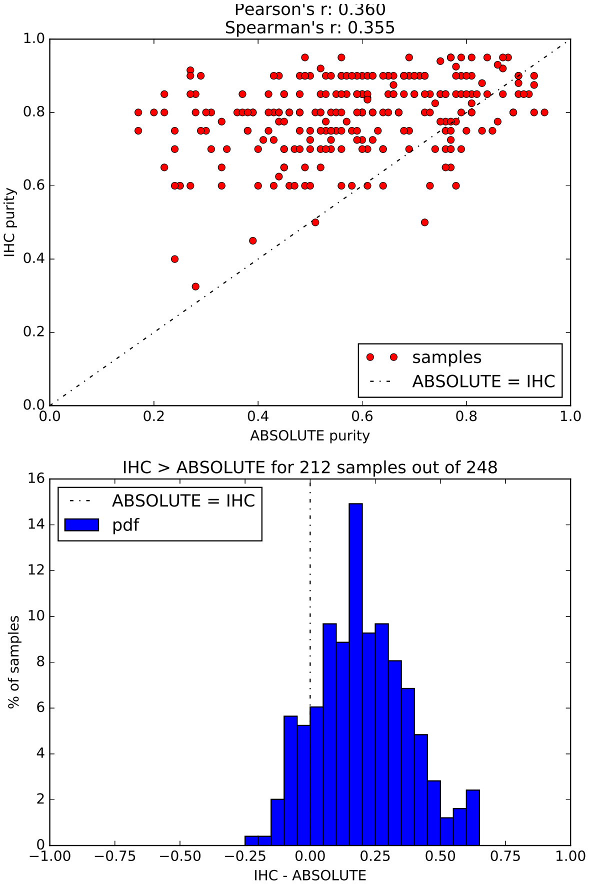


Figure 1: **(Up)** Couples (IHC, ABSOLUTE) purities. The dot-dashed line is the theoretical IHC = ABSOLUTE curve. It is known that pathologists are biased on their work (Fandel et al., 2008), in this case, IHC purities seem to overestimate the purity of tumors. **(Down)** Frequency of difference between IHC purity and ABSOLUTE purity.

increased from -0.127 to -0.161. Also a new pair of genes has become significantly correlated: (CDH1, PIK3CA); and except that one pair, no pair has changed significance threshold. We also noticed that this last pair has positive correlation.

No dependence between the number of mutations of any gene and the ABSOLUTE purity (nor from the sum of all mutations and the ABSOLUTE purity) was detectable (Figure 3a): these quantities are independent to one another. Even when aggregating all the mutations such that each gene

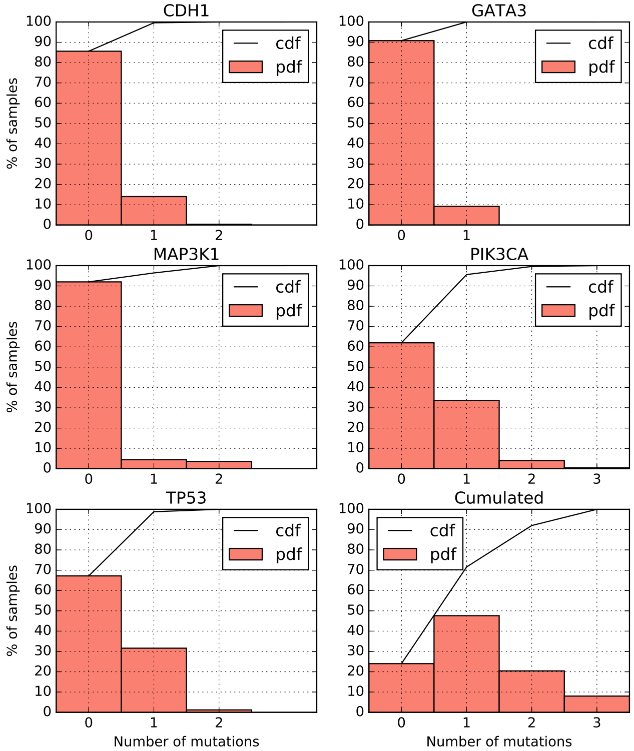


Figure 2: Distribution of the number of mutations of each gene from Table 2 and of the sum on all genes.

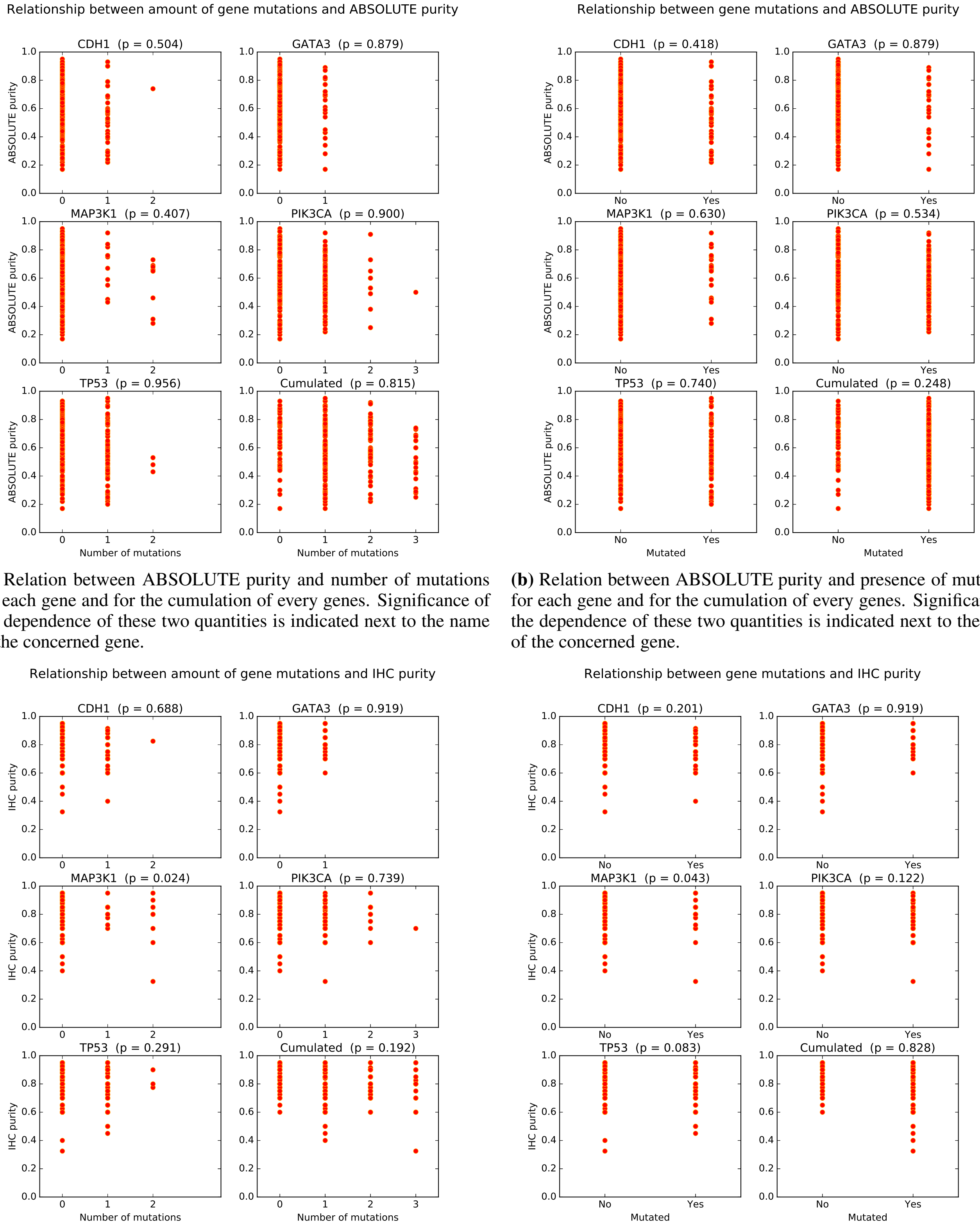
	GATA3	MAP3K1	PIK3CA	TP53
CDH1	-0.129^{*}	-0.034	$+0.114$	-0.210^{***}
GATA3		$+0.011$	-0.114	-0.078
MAP3K1			$+0.058$	-0.170^{**}
PIK3CA				-0.127^{*}

(a) Pearson's r coefficient of the number of mutations for each pair of genes.

	GATA3	MAP3K1	PIK3CA	TP53
CDH1	-0.131^{*}	-0.037	$+0.125^{*}$	-0.214^{***}
GATA3		$+0.008$	-0.107	-0.075
MAP3K1			$+0.073$	-0.175^{**}
PIK3CA				-0.161^{*}

(b) Pearson's r coefficient of the presence of mutations for each pair of genes.

Table 3: Correlation between the number of mutations or the presence of mutations for each pair of genes from Table 2.
Legend: *: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$.



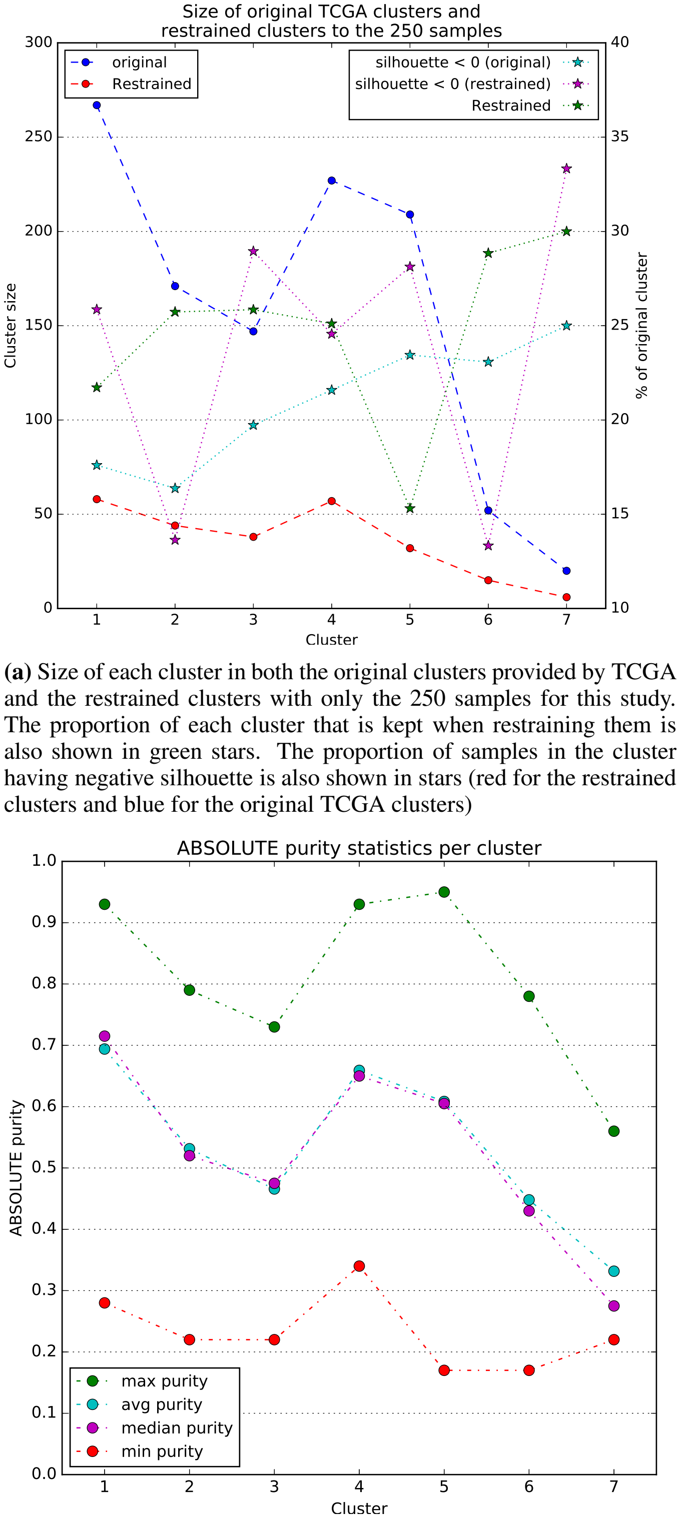
(a) Relation between ABSOLUTE purity and number of mutations for each gene and for the cumulation of every genes. Significance of the dependence of these two quantities is indicated next to the name of the concerned gene.

(b) Relation between ABSOLUTE purity and presence of mutations for each gene and for the cumulation of every genes. Significance of the dependence of these two quantities is indicated next to the name of the concerned gene.

(c) Adaptation of Figure 3a for IHC purity.

(d) Adaptation of Figure 3b for IHC purity.

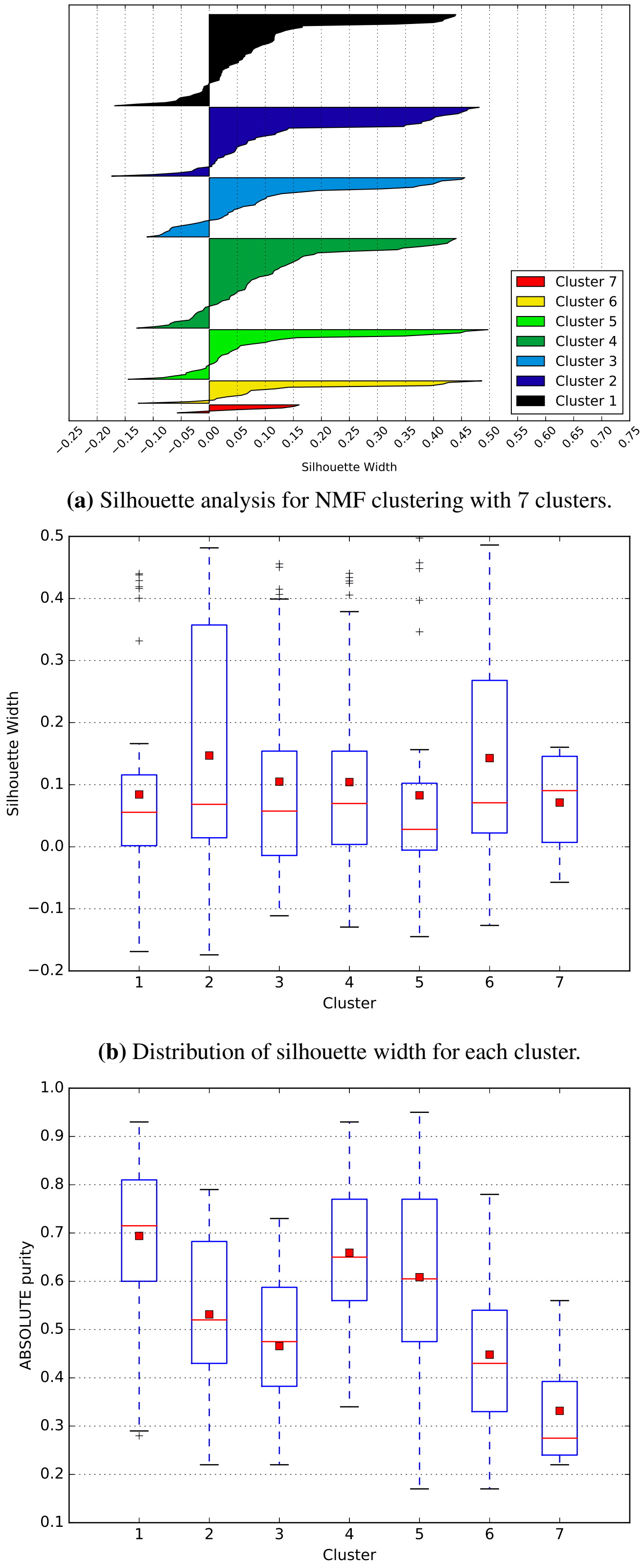
Figure 3: Relation between mutations and ABSOLUTE/IHC purity.



(a) Size of each cluster in both the original clusters provided by TCGA and the restrained clusters with only the 250 samples for this study. The proportion of each cluster that is kept when restraining them is also shown in green stars. The proportion of samples in the cluster having negative silhouette is also shown in stars (red for the restrained clusters and blue for the original TCGA clusters)

(b) ABSOLUTE purity minimum, mean, median and maximum per cluster.

Figure 4: Distribution of size, proportion and ABSOLUTE purity per cluster.



(a) Silhouette analysis for NMF clustering with 7 clusters.

(b) Distribution of silhouette width for each cluster.

(c) Distribution of ABSOLUTE purity for each cluster.

Figure 5: Clustering of the samples and silhouette/ABSOLUTE purity distribution.

is either mutated or non-mutated, no dependence is noticeable (Figure 3b).

The same result stands for the independence of the mutations (or number of mutations) and the IHC purity (see Figures 3c and 3d).

Clustering

Out of the 7 clusters provided by TCGA, clusters 6 and 7 are significantly smaller than the other ones (Figures 4a and 5a), and we observed that the distribution of the samples into the clusters is highly unequal (Figure 4a). Also, we noticed that beside not being equal in size in the original data, the clusters do not contain the same proportion of the studied samples: the kept proportion in the restrained clusters is not constant.

When looking at a cluster that would be linked to higher purity than the others, only clusters 1 and 4 showed significance (with respectively $p = 1.45 \times 10^{-7}$ and $p = 5.33 \times 10^{-4}$), and when looking at clusters that would be linked to higher purity than the whole set of samples, the same two were found (with respectively $p = 2.55 \times 10^{-5}$ and $p = 4.72 \times 10^{-3}$).

Significance of difference between the mean ABSOLUTE purity between each pair of cluster can be seen on Figure 6. We can indeed observe that cluster 1 shows to be statistically bigger than all the other clusters in terms of ABSOLUTE purity, except cluster 4.

Materials and Methods

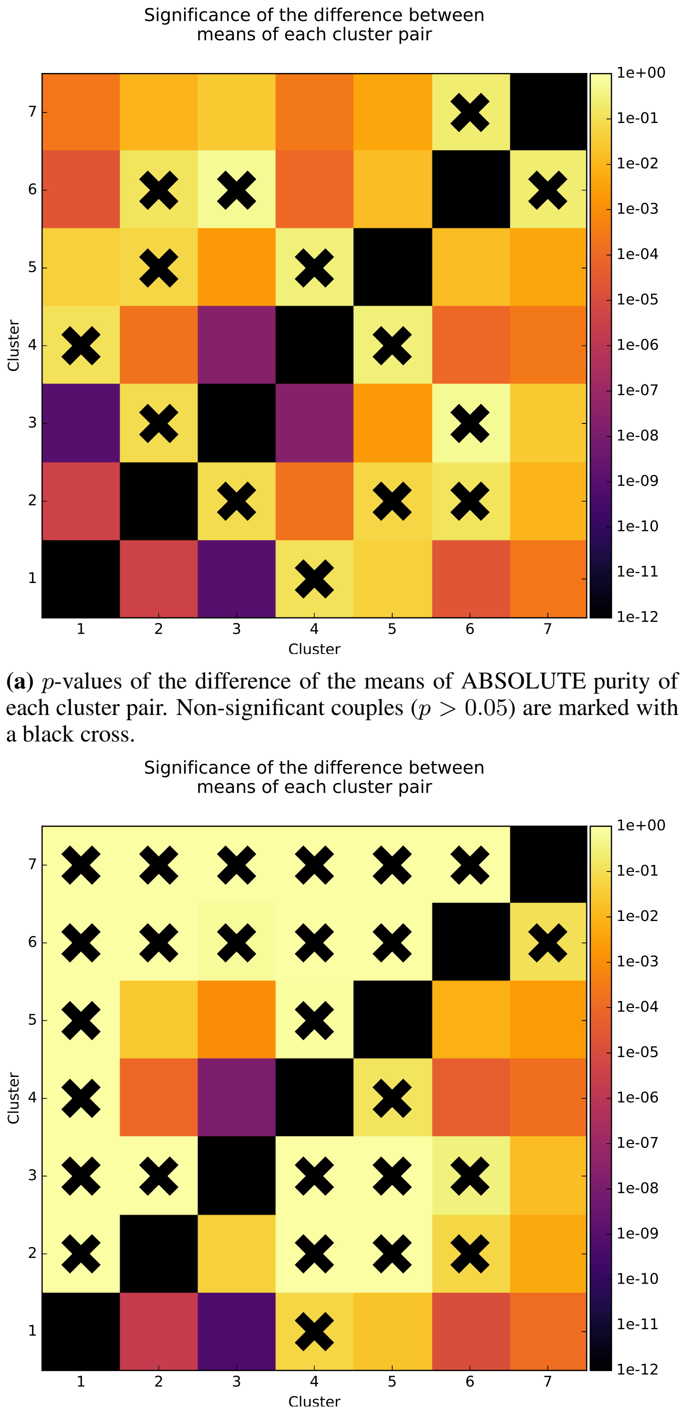
All of the (Python 3.5) source code used in this study as well as this very document are available at the following web page: https://github.com/RobinPetit/Breast-Cancer-Purity. The non-standard libraries used are `scipy` (1.1.0) and `numpy` (1.14.3).

The data used in this study are those provided for the project, comprising 250 samples with both IHC and ABSOLUTE purities and mutations of each gene of Table 2 as well as those required to be downloaded.

Significance of Pearson's r correlation coefficient

Significance indices (p-values) for Pearson's correlation coefficients are computed by a t-score: $t = r_P\sqrt{(N-2)/(1-r_P{}^2)}$ with r_P being Pearson's r and N the sample size since this statistic follows a Student's t with $N-2$ degrees of freedom (Lee Rodgers and Nicewander, 1988) (with bivariate normal distribution assumption which can be discarded provided the sample size is sufficiently big).

Figure 7 shows the associated p-value of every correlation coefficient (in absolute value) as well as significance threshold $p = 0.05$, $p = 0.01$, $p = 0.001$, and $p = 10^{-8}$ and their associated quantiles. The lower right corner of the figure is a bit messy due to numerical stability issues. Nonetheless, the p-values keep decreasing for higher values of $|r|$.



(a) p-values of the difference of the means of ABSOLUTE purity of each cluster pair. Non-significant couples ($p > 0.05$) are marked with a black cross.

(b) p-values of 1-tailed Mann-Whitney U-test on each pair of clusters testing if one is greater than the other.

Figure 6: p-values of Mann-Whitney U tests on clusters.

For Figure 1, $N = 248$ because samples TCGA-C8-A130-01 and TCGA-C8-A133-01 did not have provided IHC purity, so they were removed from the samples for this correlation analysis. For Table 3, $N = 250$, i.e. every sample is used to test correlation.

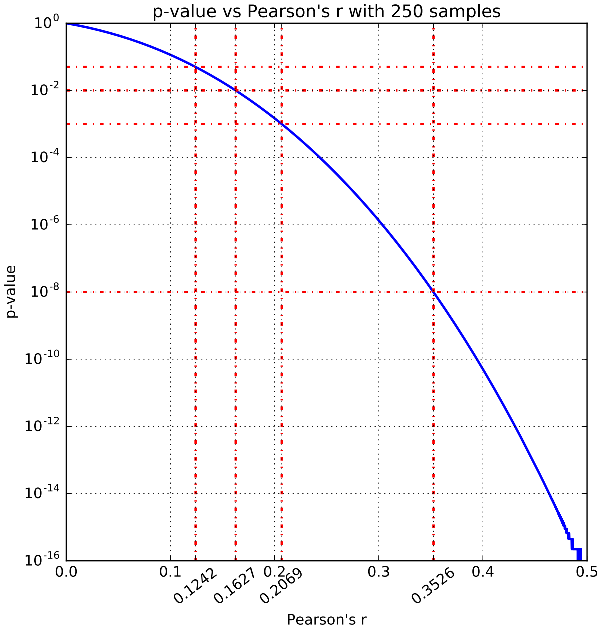


Figure 7: Associated p-value for $|r|$. Dotted red lines represent the significance thresholds and their associated quantiles. Quantiles are displayed in diagonal on the horizontal axis.

Independence tests

Independence tests between two variables were performed using a χ^2-test. For Figures 3a and 3c, the purity is a continuous variable, therefore it has been discretized into 20 subintervals of equal size in order to make a proper contingency table for the χ^2. Still, this test doesn't seem appropriate because it is way too sensitive to the discretization of the variable (number of intervals when splitting).

NMF Clustering

The NMF clustering of the samples comes from the one performed by TCGA which was performed on 1093 samples including the 250 samples used in this study. Only the samples of interest have been kept, the others have been discarded.

The difference of distribution between each pair of clusters has been tested by a two-sided Mann-Whitney U test (Figure 6) because the distribution of ABSOLUTE purity per cluster (Figure 8) is clearly not Gaussian-like, and samples are not numerous enough for asymptotic results to be of any help. Also, a one-sided Mann-Whitney U test has been performed in order to test the hypothesis that one cluster has higher mean ABSOLUTE purity than another.

Finally, two more hypotheses have been tested regarding the clustering: each cluster has been tested to show if its ABSOLUTE purity mean is higher than the union of either all of the other clusters or all the samples; i.e. is cluster i's mean purity higher than mean purity in the whole dataset or higher than mean purity in the whole dataset except samples of cluster i? Again, this has been performed with a one-sided Mann-Whitney U test.

Conclusion

To summarize, it has been found in this project that IHC purity and ABSOLUTE purity slightly correlate but are still very different, meaning that at least one of these is not a good measure. We followed by assuming that the ABSOLUTE purity was the proper way to measure purity, finding that the 70% purity criterion of TCGA is not well fulfilled by the provided samples. When looking for genes mutations associated to high ABSOLUTE purity, no convincing result were extracted, but this is probably also due to the wrong statistical framework applied to it. Also, among the subtypes of cancer that were pre-computed (clusters), two showed statistically significant greater ABSOLUTE purity than the rest of the samples.

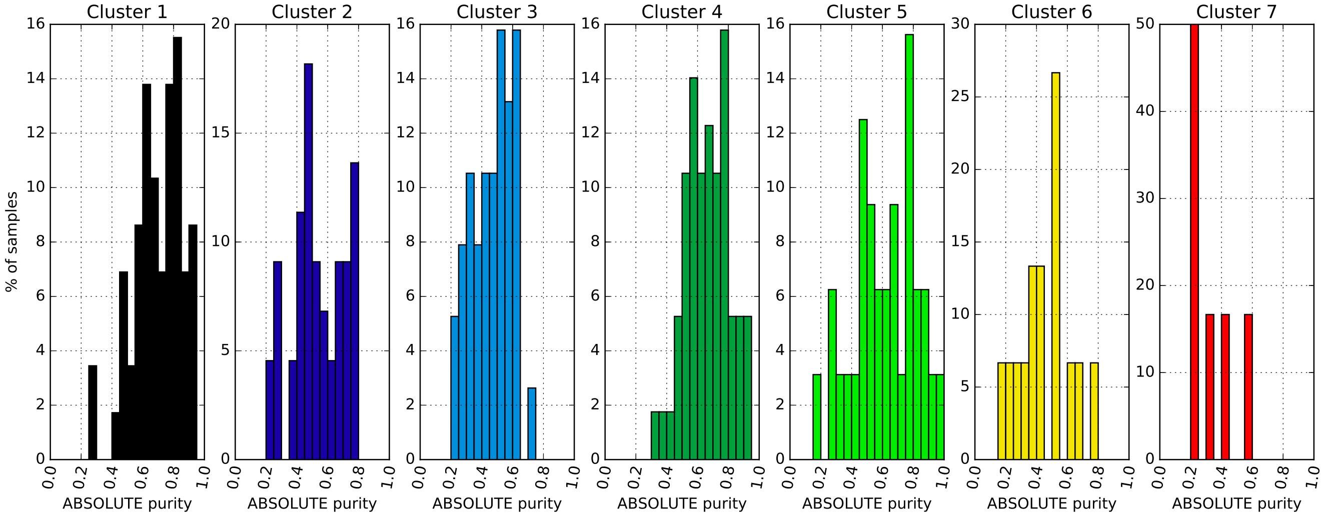


Figure 8: Distribution of the ABSOLUTE purity for each cluster.

References

Berx, G., Becker, K.-F., Höfler, H., and Van Roy, F. (1998). Mutations of the human e-cadherin (cdh1) gene. *Human mutation*, 12(4):226–237.

Carter, S. L., Cibulskis, K., Helman, E., McKenna, A., Shen, H., Zack, T., Laird, P. W., Onofrio, R. C., Winckler, W., Weir, B. A., et al. (2012). Absolute quantification of somatic dna alterations in human cancer. *Nature biotechnology*, 30(5):413.

Ciocca, V., Daskalakis, C., Ciocca, R. M., Ruiz-Orrico, A., and Palazzo, J. P. (2009). The significance of gata3 expression in breast cancer: a 10-year follow-up study. *Human pathology*, 40(4):489–495.

Easton, D. F., Pooley, K. A., Dunning, A. M., Pharoah, P. D., Thompson, D., Ballinger, D. G., Struewing, J. P., Morrison, J., Field, H., Luben, R., et al. (2007). Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature*, 447(7148):1087.

Fandel, T., Pfnür, M., Schäfer, S., Bacchetti, P., Mast, F., Corinth, C., Ansorge, M., Melchior, S., Thüroff, J., Kirkpatrick, C., et al. (2008). Do we truly see what we think we see? the role of cognitive bias in pathological interpretation. *The Journal of pathology*, 216(2):193–200.

Gasco, M., Shami, S., and Crook, T. (2002). The p53 pathway in breast cancer. *Breast Cancer Research*, 4(2):70.

Komura, D. and Ishikawa, S. (2018). Machine learning methods for histopathological image analysis. *Computational and Structural Biotechnology Journal*.

Lee Rodgers, J. and Nicewander, W. A. (1988). Thirteen ways to look at the correlation coefficient. *The American Statistician*, 42(1):59–66.

Oesper, L., Satas, G., and Raphael, B. J. (2014). Quantifying tumor heterogeneity in whole-genome and whole-exome sequencing data. *Bioinformatics*, 30(24):3532–3540.

Saal, L. H., Holm, K., Maurer, M., Memeo, L., Su, T., Wang, X., Jennifer, S. Y., Malmström, P.-O., Mansukhani, M., Enoksson, J., et al. (2005). Pik3ca mutations correlate with hormone receptors, node metastasis, and erbb2, and are mutually exclusive with pten loss in human breast carcinoma. *Cancer research*, 65(7):2554–2559.

Sirinukunwattana, K., Raza, S. E. A., Tsang, Y.-W., Snead, D. R., Cree, I. A., and Rajpoot, N. M. (2016). Locality sensitive deep learning for detection and classification of nuclei in routine colon cancer histology images. *IEEE transactions on medical imaging*, 35(5):1196–1206.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43):15545–15550.

Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R. M., Ozenberger, B. A., Ellrott, K., Shmulevich, I., Sander, C., Stuart, J. M., Network, C. G. A. R., et al. (2013). The cancer genome atlas pan-cancer analysis project. *Nature genetics*, 45(10):1113.

Xing, F. and Yang, L. (2016). Robust nucleus/cell detection and segmentation in digital pathology and microscopy images: a comprehensive review. *IEEE reviews in biomedical engineering*, 9:234–263.