

The Involvement of Hypoxia in APOBEC-Mediated Mutagenesis in Multiple Cancers



Jiachen Liang
St Edmund Hall
University of Oxford

A thesis submitted for the degree of
Master of Science by Research

Trinity 2018

Acknowledgements

I would like to thank my research supervisor, Dr Francesca Buffa, for her indispensable role in overseeing the project and providing me with thorough guidance throughout my degree.

I thank Dr Wei-Chen Cheng, Dr Alessandro Barberis, Ms Sakura Maezono, Dr Jagat Chauhan, and Dr Andrew Dhawan for their support and insight in data handling and analysis.

I thank my mother, father, brother, and family for their sacrifices and support in helping me achieve this degree and in all my endeavours.

Abstract

The Involvement of Hypoxia in APOBEC-Mediated Mutagenesis in Multiple Cancers

Jiachen Liang

Department of Oncology, St Edmund Hall
University of Oxford

Mutational signatures have been studied to associate mutations to their originating causes. One such signature is that of the APOBEC family of proteins. APOBEC proteins are normally involved in innate immunity, playing a role in the defence against foreign genetic material by inducing mutations in single-stranded DNA and RNA. APOBECs have recently been associated to mutations in the genomes of multiple human cancers. The regulation of APOBEC genes is currently poorly understood, especially in the context of cancer. Recent evidence suggests that certain APOBEC genes can be upregulated in immune cells exposed to hypoxia. I explore this path further by studying the association between hypoxia and APOBEC expression and activity in cancer using RNA-Seq, mutation, and amplification data from 6723 patients in 13 cancer types. My results show evidence of correlation between hypoxia and APOBEC gene expression as well as mutagenic activity, specifically APOBEC3A and 3B. I also show that APOBEC3B is involved in pathways that overlap with those associated with hypoxia, contrary to the other APOBEC family members. Finally, I provide a comparison of the methods used to evaluate the APOBEC mutational signature. Overall, this study provides preliminary evidence for the association between APOBEC and hypoxia. It also raises many questions, but it provides solid grounds for future studies to be conducted in more depth about this topic.

Word count: 23600

(excluding figures)

Table of Contents

Acknowledgements.....	3
Abstract.....	5
Table of Contents	7
List of Tables	10
List of Figures.....	11
Terminology.....	13
1. Introduction	15
1.1 Mutations and Mutational Signatures	15
1.2 Drawbacks of Using Reporter Genes	17
1.3 A Mathematical Approach to Mutational Signatures.....	18
1.4 APOBEC Proteins Are Enzymatic Mutators of the Genome.....	20
1.5 APOBEC Mutagenesis and Cancer.....	25
1.6 The APOBEC Mutational Signature.....	27
1.7 APOBEC and Hypoxia.....	31
1.8 Objectives of the Present Study	33
2. Methods.....	37
2.1 Data Source.....	37
2.2 Data Pre-Processing.....	41
2.3 Immune Infiltration	42
2.4 Correlation Between APOBEC and Hypoxia Signature Expression	43
2.5 APOBEC and Hypoxia Gene Co-Expression Analysis	45
2.6 Analysis of the APOBEC Mutational Signature	47
2.7 Software Deployment	51

3. Results	53
3.1 APOBEC is Expressed and Functional Across Cancers	53
3.1.1 <i>APOBEC3 Genes are Significantly and Differentially Expressed Across Cancers</i>	53
3.1.2 <i>Decomposing APOBEC Expression into its Components</i>	59
3.1.3 <i>The Expression of APOBEC3 Family Genes is Highly Self-Correlated</i>	64
3.1.4 <i>The APOBEC Mutational Signature is Present in Cancer</i>	69
3.2 APOBEC is not Significantly Mutated in Breast Cancer	70
3.3 APOBEC Amplification Does not Drive the Differences in Expression Between Breast Cancer and Normal Breast Samples.....	73
3.4 Hypoxia is Present in Tumours and Correlates with APOBEC3A and APOBEC3B Expression.....	76
3.4.1 <i>A Hypoxia Gene Signature is Highly and Preferentially Expressed in Tumours</i>	76
3.4.2 <i>APOBEC3A and APOBEC3B Expression Correlates with the Hypoxia Signature</i>	89
3.4.3 <i>A Principal Component Regression Model to Explain the Relationship between APOBEC and Hypoxia</i>	94
3.5 Immune Infiltration Has a Negligible Effect on APOBEC Presence.....	102
3.6 Subtypes of Breast Cancer Show Different Relationships between APOBEC and Hypoxia.....	119
3.7 The APOBEC Mutational Signature is Present in Multiple Cancers	130
3.7.1 <i>Count of Mutations at a Consensus Motif</i>	131
3.7.2 <i>Enrichment of Mutations at a Consensus Motif</i>	138
3.7.3 <i>Enrichment of Consensus Mutational Patterns in Samples</i>	144
3.7.4 <i>Enrichment of Curated Mutational Patterns in Samples</i>	160
3.7.5 <i>Comparing the Four Methods</i>	177
3.8 The APOBEC Mutational Signature Correlates with Hypoxia in Breast Cancer	185
3.9 APOBEC3B and Hypoxia are Associated with Common Pathways	200

4. Discussion	213
4.1 Expression of APOBEC Family Genes in Patient Samples across Cancer Types	213
4.2 The Relationship between APOBEC and Hypoxia on the Gene Expression Level	218
4.3 APOBEC and Hypoxia in Subtypes of Breast Cancer	225
4.4 Mutational Signatures to Study the Relationship Between APOBEC and Hypoxia	228
4.5 Pathway Analysis Shows a Functional Relationship between APOBEC and Hypoxia in Breast Cancer	234
4.6 Significance of the Work	236
4.7 General Remarks	238
4.8 Conclusion	240
Bibliography	241
Supplementary Tables	251
Supplementary Figures	261

List of Tables

Table 1: Characteristics of APOBEC Proteins	23
Table 2: Cancer Types Chosen for Study	40
Table 3: Principal Component Coefficients of the Expression of APOBEC3 Genes in Breast Cancer.....	61
Table 4: Mutations in APOBEC Genes	71
Table 5: Correlation Between Somatic Copy Number and Normalised mRNA Count of APOBEC Genes in Breast Cancer	75
Table 6: Principal Component Regression of the Expression of APOBEC3 Genes in Relation to the Buffa Hypoxia Signature.....	97
Table 7: COSMIC Mutational Signatures Selected for Analysis.....	164
Table 8: Summary of Correlation between the Expression of A3A and A3B and the Presence of the APOBEC Mutational Signature	184
Table S1: Genes in the Buffa Hypoxia Signature	251
Table S2: Genes Co-Expressed with APOBEC3 Genes in Breast Cancer.....	254
Table S3: Top Genes Co-Expressed with the Buffa Hypoxia Signature in Breast Cancer	258

List of Figures

Figure 1: APOBEC3A/3B Mutagenesis and Downstream Effects	30
Figure 2: Pipeline of Analysis Methods	52
Figure 3: Mean Expression of APOBEC Genes across 13 Cancer Types	56
Figure 4: Expression of APOBEC Genes in Breast Cancer	57
Figure 5: Expression of APOBEC3B across 13 Cancer Types	58
Figure 6: APOBEC3A and 3B Versus Principal Components 1 and 2 of the Expression of All APOBEC Genes	62
Figure 7: Self-Correlation of the Expression of APOBEC Genes	65
Figure 8: Overview of Self-Correlation of Expression in the Buffa Hypoxia Signature Across 13 Cancer Types	79
Figure 9: Correlation Between Mean, Median, and PC1 of the Expression of the Buffa Signature across 13 Cancer Types	81
Figure 10: Mean expression of the Buffa et al. hypoxia signature across 13 cancer types	88
Figure 11: Correlation between the Expression of APOBEC Genes and the Mean Buffa Hypoxia Signature	91
Figure 12: Relationship between the Expression of the Hypoxia Signature and APOBEC3A/3B in Breast Cancer	92
Figure 13: Correlation between Individual APOBEC Genes and Individual Genes in the Buffa Hypoxia Signature in Breast Cancer	93
Figure 14: xCell Scores for TCGA Samples of 13 Cancer Types	105
Figure 15: Correlation Between APOBEC Genes and the Mean Buffa Hypoxia Signature Partial to Selected xCell Signatures	118
Figure 16: Mean Expression of APOBEC Genes in Breast Cancer Subtypes	121
Figure 17: Expression of APOBEC3A and APOBEC3B in Breast Cancer Subtypes	122
Figure 18: Mean Expression of the Buffa Hypoxia Signature across Breast Cancer Subtypes	124
Figure 19: Correlation between APOBEC Genes and the Mean Buffa Hypoxia Signature in Subtypes of Breast Cancer	125
Figure 20: Relationship between the Expression of the Hypoxia Signature and APOBEC3A/3B in Subtypes of Breast Cancer	126
Figure 21: Relationship between the APOBEC3A and 3B and the Number of APOBEC-Induced Mutations	133
Figure 22: Relationship between the APOBEC3A and 3B and the Enrichment of APOBEC-Induced Mutations	140

Figure 23: deconstructSigs Scores for the 30 COSMIC Signatures.....	148
Figure 24: Percentage of Non-Null deconstructSigs Scores for the 30 COSMIC Signatures ..	151
Figure 25: Relationship between the APOBEC3A and 3B and deconstructSigs Scores	155
Figure 26: deconstructSigs Scores for COSMIC Signatures Curated for each CancerTtype ..	165
Figure 27: Percentage of Non-Null deconstructSigs Scores for the 30 COSMIC Signatures ..	169
Figure 28: Relationship between the APOBEC3A and 3B and deconstructSigs Scores Using only Curated Signatures.....	172
Figure 29: Correlation between the Expression of Random genes and the APOBEC Mutational Signature	181
Figure 30: Z Scores for A3A and A3B Relative to the Distribution of Random Genes in Terms of Correlation with the APOBEC Mutational Signature	183
Figure 31: Relationship between Hypoxia and the APOBEC Mutational Signature.....	187
Figure 32: Correlation between the APOBEC Mutational Signature and the Expression of Random Genes in Relation to the Correlation with the Mean Expression of the Buffa Hypoxia Signature	197
Figure 33: Z Scores for the Mean Buffa Signature Relative to the Distribution of Random Genes in Terms of Correlation with the APOBEC Mutational Signature	199
Figure 34: Top Enriched Pathways in Genes Co-Expressed with 5 or more APOBEC3 Family Genes in Breast Cancer	203
Figure 35: Top Enriched Pathways in Genes Co-Expressed with APOBEC3B in Breast Cancer	205
Figure 36: Top Enriched Pathways in Genes Co-Expressed with Hypoxia Genes in Breast Cancer	207
Figure 37: Overlap between Genes Correlated with any 5 APOBEC3 Genes, APOBEC3B, and Hypoxia in Breast Cancer in Terms of Genes	209
Figure 38: Overlap between Genes Correlated with any 5 APOBEC3 Genes, APOBEC3B, and Hypoxia in Breast Cancer in Terms of Pathways	211
Figure S1: Expression of APOBEC Genes in 12 Cancer Types	261
Figure S2: Self-Correlation of the Expression of the Buffa Hypoxia Signature in 13 Cancer Types.....	268
Figure S3: Residuals Histogram and Q-Q Plots for the Principal Component Regression of the Expression of APOBEC3 Genes in Relation to the Mean Hypoxia Signature.....	282
Figure S4: Difference in Spearman's Rho between Partial Correlation and Direct Correlation between APOBEC and Hypoxia.....	286

Terminology

APOBEC: Apolipoprotein B mRNA-editing catalytic subunit

COSMIC: Catalogue of Somatic Mutations in Cancer

DCS: deconstructSigs

FDR: False-discovery rate

NMF: Non-negative matrix factorisation

PCA: Principal component analysis

RSEM: RNA-Seq by expectation maximisation

ssDNA/RNA: Single-stranded DNA/RNA

TCGA: The Cancer Genome Atlas

1. Introduction

1.1 Mutations and Mutational Signatures

Cancer is caused by mutations, often a series of mutations, conferring a selective advantage to the tumour progenitor cell, enabling it to survive, escape destruction, and eventually proliferate into a full-blown cancer (Hanahan and Weinberg, 2011). Much of cancer research has been focused on identifying, classifying, and designing therapies around oncogenic mutations. With the advent of Sanger sequencing (Sanger et al., 1977), the study of mutations in their genomic context has accelerated rapidly. The first studies to examine mutations relative to the context in which they are generated in cancer have started with well-known reporter genes such as TP53. TP53 was sequenced in patients in several studies that associated distinct mutational patterns in the gene to specific mutational sources. This elucidates aetiologies for these patterns and defines a 'mutational signature' for a certain mutation-inducing phenomenon. For example, tumour samples of squamous cell skin carcinoma have been shown to be enriched in the UV light-induced CC to TT and C to T transitions at dipyrimidine sites in TP53 (Brash et al., 1991). Another notable case is G to T transversions occurring primarily at codon 249 of TP53 being identified in hepatocellular carcinoma patients exposed to aflatoxin (Bressac et al., 1991). A later study also linked tobacco smoke to mutations occurring in a context dependent on more than the mutated base itself. The group identified an increased prevalence of G to T

transversions at endogenously methylated CpG motifs in TP53 of lung cancer patients who have a history of smoking (Pfeifer et al., 2002).

1.2 Drawbacks of Using Reporter Genes

Although these previously mentioned studies proved to be insightful and paved the way for years of work to come, they had major drawbacks. First, in cancers that have multiple sources of mutation, it is impossible to disentangle the effects of each individual source. Only a mix of mutational signatures can be reported as a combined signature, thus reducing the ability to link a signature to an aetiology. Second, examining reporter genes such as TP53 for mutations includes the effects of selection and the specific genomic architecture of the gene, which can affect the rate of mutations observed. Since we are observing mutations in a limited region of the genome, we cannot assume or extrapolate with great accuracy that what we observe is the case for mutations elsewhere, including passenger mutations that do not lead to an apparent phenotype and mutations that occur in non-coding regions.

The advent of next-generation sequencing (NGS) and the completion of the Human Genome Project have made possible the analysis of entire genomes for mutational signatures. Catalogues of somatic mutations have been generated in various cancer types, starting with melanoma (Plesance et al., 2010a), lung cancer (Plesance et al., 2010b), and breast cancer (Nik-Zainal et al., 2012). Searching for mutational signatures in the whole genome eliminates the limitations and caveats of using a reporter gene, as outlined above. Passenger mutations and mutations in non-coding regions are regarded the same way as driver mutations in genes such as TP53 and the effects of selection are minimal, as the genes affected by selection only account for a small portion of the genome.

1.3 A Mathematical Approach to Mutational Signatures

The first studies on mutational signatures used a hypothesis-driven approach – the research started with empirical evidence of mutational signatures and went on to validate them at a large scale in the genome to formally assign the signature to a cause. For example, we know for a few decades now that UV light induces CC to TT and C to T mutations (Witkin, 1969), and this has been observed in genomes of melanoma using NGS (Pleasant et al., 2010a). Although the hypothesis-driven discovery of signatures yields highly accurate and clinically significant results, it does not have the power to fully independently discover new signatures based solely on the genetic sequence and mutations, i.e. without prior knowledge of potential signatures to drive the search.

One study focused on creating a catalogue of somatic mutations in 21 breast cancers and extracted five mutational signatures using a forward mathematical approach (Nik-Zainal et al., 2012). The study used no prior definition of mutational signatures to drive their search and used only the mutations sequenced to extract signatures. Essentially, the authors developed an *a priori* method for the discovery of mutational signatures, one without influence or bias from prior knowledge.

To bring about a consensus in the way of reporting mutational signatures, the authors classify each substitution by its class (C>A, C>G, C>T, T>A, T>C, and T>G) as well as the bases immediately 5' and 3' of the mutated base. This yields 96 possible trinucleotide contexts (6 substitutions multiplied by 4 possible adjacent bases, twice.) Each substitution-type mutation in the sequenced

genomes is classified into one of these 96 contexts, and the prevalence of each context is recorded for each genome.

With information pertaining to the prevalence of each substitution type in each sample, the authors use non-negative matrix factorisation (NMF) to reduce the dimensions in the data and extract common features from it. In this case, the features are mutational signatures that are present in the sequenced genomes and contribute with different weights to each of them.

Briefly, NMF works by decomposing a matrix V of size $a \times b$ into two smaller matrices called factors, W and H , the product of which approximate the original matrix (Berry et al., 2007). W is of size $a \times r$, and H is of size $r \times b$. Thus, we have r rows and columns in H and W . r defines the features that we wish to extract from V . r must be smaller than both a and b , or else the most prominent features extracted will simply correspond exactly to the input parameter that has the least dimensions. In other words, we cannot extract more features than there are dimensions in our data. It is to note that the $V \approx W \times H$ problem cannot be solved analytically in most applied cases, instead requiring a numerical approach to the solution. In the case of mutational signatures, Nik-Zainal et al. started with a matrix of 21 genomes by 96 mutational contexts. The end result is the extraction of 5 features from the data set that are the new mutational signatures. These features are profiles of the relative prevalence of the 96 trinucleotide mutations. Linear combinations of these five profiles describe the mutational landscape of each of the 21 genomes.

1.4 APOBEC Proteins Are Enzymatic Mutators of the Genome

While many mutations arise from sources exogenous to the organism, some mutations arise from endogenous mechanisms that serve an essential purpose. For example, class-switch recombination in the immune system requires the action of activation-induced deaminase (AID, or AICDA for 'activation-induced cytosine deaminase'), which mutates regions of the variable chain of immunoglobulins (Muramatsu et al., 1999, 2000). This process introduces variability in the antigen binding sites of immunoglobulins in order for them to recognise a wide pattern of antigens.

In the same family as AID are APOBEC proteins (Salter et al., 2016). APOBEC stands for 'apolipoprotein B mRNA-editing catalytic subunits.' The APOBEC family of genes and proteins consists of APOBECs 1, 2, 3A, 3B, 3C, 3D, 3F, 3G, 3H, and 4, as well as APOBEC1 complementation factor (A1CF). Although their names imply the editing of RNA, APOBECs mostly edit DNA, with RNA-editing limited to a few family members (Table 1). The different family members have different roles, are expressed in different circumstances, and are localised in different cellular compartments. Table 1, adapted from Salter et al. 2016, shows an overview of the various differences among APOBEC proteins. Common to all APOBEC deaminases is the zinc-dependent deaminase domain (Smith et al., 2012). Even APOBEC4, which has no known enzymatic activity, contains one such domain, although it slightly differs from the consensus domain found in other APOBECs.

APOBEC1 is the first APOBEC protein to be identified (Teng et al., 1993) and the one that gave the family its name (Davidson et al., 1995). APOBEC1 was found to deaminate C₆₆₆₆ to U in the mRNA of the apolipoprotein B gene to induce a premature stop codon, thereby creating a shorter variant of the apolipoprotein B polypeptide called apolipoprotein-B48 (Navaratnam et al., 1993). APOBEC1 does not edit any arbitrary cytosine; it recognises a specific motif and targets C₆₆₆₆ and C₆₈₉₃ in the apolipoprotein B mRNA (Navaratnam et al., 1991). It is to note that the C₆₈₉₃ mutation is of little clinical significance, as it only occurs when C₆₆₆₆ is mutated and a stop codon is introduced, and occurs at a much lower rate than mutations in C₆₆₆₆ (Navaratnam et al., 1991).

The APOBEC3 (A3) genes, A3A, A3B, A3C, A3D, A3F, A3G, and A3H, are arranged in a tandem cluster on chromosome 22 and closely share sequence homology (Jarmuz et al., 2002). A3 proteins differ in function, but are all involved in immunity and defence against viruses and retroelements by binding and mutating single-stranded DNA (Salter et al., 2016). A3G is the main A3 protein involved in the defence against HIV (Sheehy et al., 2002) and other retroviruses (Mangeat et al., 2003), while A3D, A3F, and A3H also have anti-HIV properties, but their activity is lesser than that of A3G (Chaipan et al., 2013). A3B, A3C, A3F, and A3G are expressed in hepatocytes as part of a mechanism to defend against the replication of the hepatitis B virus, and the expression of A3B, A3C, and A3F can be increased by IFN- α , further suggesting a direct role in innate immunity (Bonvin et al., 2006). A3A, on the other hand, has been shown to be expressed and active in monocytes and macrophages, inducing deamination upon activation by IFN- α (Thielen et al., 2010). Interestingly, A3A activity in monocytes and

macrophages is not limited to DNA; like A1, it can bind and deaminate RNA as well (Sharma et al., 2015). A3G has also been shown to deaminate RNA, but there is only cell culture evidence to support this claim, and its RNA deaminase activity has not yet been observed in a physiological context (Sharma et al., 2016).

Little is known about the function of APOBEC 2 and 4. Both A2 and A4 were discovered by searching for sequence homology to APOBEC1 (Harris et al., 2002; Rogozin et al., 2005). A2 has been shown to be expressed in cardiac and skeletal muscle tissue and is involved in the development of muscle tissue (Sato et al., 2010). Its expression is controlled by NF- κ B and increased by TNF- α and IL-1 β , elucidating a role in innate immunity and the inflammation response (Matsumoto et al., 2006). Increased A2 activity has also been linked to liver and lung cancers (Okuyama et al., 2012). A4 is shown to be expressed in the testis but has no known or observed deamination activity (Marino et al., 2016).

Table 1: Characteristics of APOBEC Proteins

	Genomic location	Tissue expression	Cellular localisation	Binding activity	Deaminase activity	Preferred sequence
AID	12p13	Activated B cells	n/C	ssDNA RNA	ssDNA RNA	W <u>R</u> C
APOBEC1	12p13.1	GI tract	n/C	ssDNA RNA	ssDNA RNA	A <u>C</u> N
APOBEC2	6p21	Heart, skeletal muscle, TNF- α activated liver cells	N/C	-	-	-
APOBEC3A	22q13.1	monocytes, macrophages, non-progenitor cells	N (C in monocytes)	ssDNA RNA	ssDNA RNA	T <u>C</u> W
APOBEC3B	22q13.1	PKC induced cancer cells, IFN- α activated liver cells	N	ssDNA RNA	ssDNA	T <u>C</u> W
APOBEC3C	22q13.1	immune centres, peripheral blood cells	N/C	ssDNA RNA	ssDNA	T <u>C</u> W
APOBEC3D	22q13.1	immune centres, peripheral blood cells	C	ssDNA RNA	ssDNA	T <u>C</u> W
APOBEC3F	22q13.1	immune centres, peripheral blood cells, IFN- α activated liver cells	C	ssDNA RNA	ssDNA	T <u>C</u> W
APOBEC3G	22q13.1	immune centres, peripheral blood cells, IFN- α activated liver cells	C	ssDNA RNA	ssDNA	<u>C</u> CW
APOBEC3H	22q13.1	immune centres, peripheral blood cells	N/C	ssDNA RNA	ssDNA	T <u>C</u> W
APOBEC4	1q25.3	-	-	ssDNA RNA	-	-
A1CF	10q11.23	varied	n/C	-	-	-

Characteristics of APOBEC cytidine deaminases, adapted from Salter et al. 2016. Members listed include all APOBEC genes/proteins, including A1CF, a complement factor to APOBEC1 with no deaminase domain. A dash denotes the absence of evidence in literature. In cellular localisation, 'N' denotes nuclear, 'C' denotes cytoplasmic, 'N/C' denotes a mix of nuclear and cytoplasmic, and 'n/C' denotes a mostly cytoplasmic localisation with presence in the nucleus. For the preferred sequence, the underlined cytidine is deaminated to uracil. W denotes adenine or thymine. N denotes adenine or uracil. R denotes adenine or guanine.

1.5 APOBEC Mutagenesis and Cancer

It is not unexpected that an enzymatic mutator of DNA would be linked to cancer. Particularly, A3B is constitutively localised in the nucleus, unlike any other APOBEC protein (Burns et al., 2013a; Lackey et al., 2012). The strict nuclear localisation of A3B implies that functional A3B has the possibility of acting on genomic DNA and that an increase in A3B expression is likely to lead to a higher mutation load in the genome. On the contrary, A3G, the most active deaminase of the A3 family in anti-viral defence, is strictly cytoplasmic, greatly reducing its potential for genomic mutagenesis (Bennett et al., 2008).

A3B is first linked to breast cancer by a study showing that the overexpression of A3B causes an increase in mutations and genomic uracil (Burns et al., 2013a). In a series of knockdown experiments, the authors confirm that A3B is responsible for nuclear cytosine deamination in breast cancer cell lines, accounting up to half of all mutations when A3B is expressed. The study also shows correlation between A3B expression and C to T mutation load, a marker of cytosine deamination after DNA repair, as well as evidence that TP53 is selectively mutated by A3B. The association between A3B and cancer is then extended to multiple cancer types by an analysis of genomic data from The Cancer Genome Atlas (Burns et al., 2013b). The study shows that A3B is upregulated in bladder urothelial carcinoma, uterine corpus endometrioid carcinoma, head and neck squamous cell carcinoma, lung squamous cell carcinoma, and lung adenocarcinoma. The same cancers have an enrichment in mutations in a 5'-TC context similar to those generated by recombinant A3B

expressed in an in vitro model (Burns et al., 2013a). Overall, these two studies provide strong initial evidence for a link between A3B and mutations in cancer.

In the same year, the association between APOBEC mutagenesis and cancer was extended to A3A. A3A expression is shown to be increased in tumours and its mRNA levels correlated with the total number of C to T mutations in the exomes of both breast cancer and a combination of 2680 cancers of 14 types (Roberts et al., 2013). A3A is normally expressed in the cytoplasm of monocytes as part of innate immunity, therefore it is expected to ask how A3A generates mutations in the genome. A study suggests that A3A can be imported into the nucleoplasm, but this process is inhibited by a mechanism that limits its nuclear import in monocytes that express it (Land et al., 2013). Cells that express A3A without this mechanism would therefore be susceptible to its mutations in the genome. A more recent study also shows that A3A-induced mutations are unique compared to mutations induced by A3B and that the former accounts for a large portion of all mutations caused by APOBEC (Chan et al., 2015).

1.6 The APOBEC Mutational Signature

Being an enzymatic mutator, it is unsurprising that APOBEC proteins generate mutations in preferred sequence contexts, thus having a mutational signature. Over time, the preferred mutated sequences of various APOBEC proteins have been identified using *in vitro* methods. An outline of preferred sequences is available in Table 1 (Salter et al., 2016). For the discussion of mutations in cancer, we will focus on the A3 family. In brief, all A3 family proteins preferentially deaminate at 5'-TCC motifs with the exception of A3G, which prefers 5'-CCC. It is to note that Table 1 displays TCCW as the preferred motif, where W stands for A or T, because later computational studies have confirmed it (Alexandrov et al., 2013; Nik-Zainal et al., 2012), but the *in vitro* studies of the time did not pick up on this 3' base preference.

Among the mutational signatures identified by Nik-Zainal et al. in their initial computational study (Nik-Zainal et al., 2012), one signature is characterised by a strong enrichment of C>G and C>T mutations in a TCA and TCT context (Nik-Zainal et al., 2012). This signature strongly resembles the signatures defined by previous *in vitro* studies on APOBEC (Salter et al., 2016). It is thus thought that this signature can be attributed to the action of APOBEC family cytosine deaminases, specifically A3 family genes (except A3G.) The authors also observed that in breast cancer, mutations of this signature occur in clustered fashion in regions that they named 'kataegis.' Within kataegis, the genomic distance between mutations of this signature can be an order of magnitude below the average distance between mutations. This feature characteristic of the

APOBEC mutational signature sheds light on the mechanism of sequence recognition and deamination by the APOBEC proteins.

Expanding on the work of Nik-Zainal et al., the group applied their methods to 7042 mutational catalogues of genomes and exomes in 30 cancer types (Alexandrov et al., 2013). With the larger availability of samples for analysis, this study resulted in the extraction of 21 distinct mutational signatures. Two of these signatures are attributed to APOBEC mutagenesis: signatures 2 and 13. Both signatures account for TCW to TTW and TGW mutations, but signature 2 accounts slightly more for TCW to TTW mutations while signature 13 accounts more for TCW to TGW mutations.

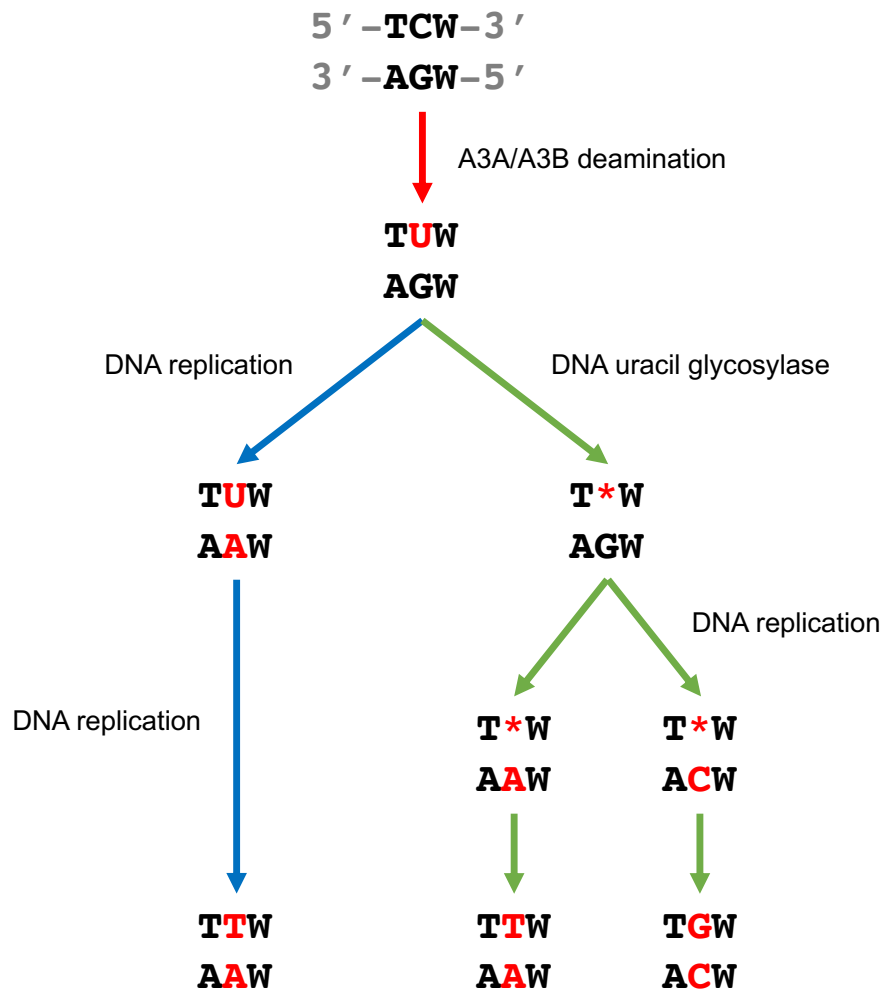
It is thus established that the A3A/A3B mutational signature is defined by a cytosine deamination at TCW motifs resulting in TTW or TGW. This is a multi-step process, however, as APOBEC proteins only deaminate cytosines to uracil (Figure 1.) The resulting TUW motif can be processed in two different ways. If DNA replication occurs before DNA repair mechanisms recognise it, it can lead to a TTW motif being incorporated in one of the daughter strands, as a complementary adenine will pair with the uracil instead of a guanine pairing with cytosine. Alternatively, uracil glycosylase can recognise the uracil in DNA and cleave it from its deoxyribose backbone (Krokan and Bjørås, 2013). DNA replication tends to introduce cytosine or adenine bases as the complementary base to an abasic site (Gibbs and Lawrence, 1995). Thus, if the DNA is replicated with an abasic site in place of the original cytosine, it can lead to a transition into

a thymine or guanine, pairing with the wrongly incorporated adenine or cytosine on the complementary strand.

Since then, a research group has been established to curate data about somatic mutations in cancer, including mutational signatures, and update them with the most current research (Forbes et al., 2017). The Catalogue of Somatic Mutations in Cancer (COSMIC) currently lists 30 consensus mutational signatures defined by trinucleotide contexts, along with information about aetiology and cancer types in which they are found.

For the purpose of simplicity, I will hereon use 'APOBEC mutational signature' to refer to the signature attributed to A3 family proteins (except A3G) in a broad sense. This is to refer to the pattern of the mutations induced by A3 proteins regardless of the method that we use to look at it. It is important to know that different APOBEC proteins in fact act on different sites, thus yielding different mutational signatures, but the vast majority of mutations attributed to APOBEC proteins are of the A3 family pattern.

Figure 1: APOBEC3A/3B Mutagenesis and Downstream Effects



Mechanisms by which APOBEC3A and 3B generate TCW to TTW and TGW mutations. Initially, A3A/A3B deaminates a cytosine in the TCW context, where W stands for A or T. The resulting U can be incorporated directly in daughter DNA or recognised by uracil glycosylase and cleaved. The abasic site then leads to the preferential incorporation of A or C on the complementary strand by DNA replication, and results in the original C mutating into T or G.

1.7 APOBEC and Hypoxia

Although APOBEC proteins, and A3B in particular, have been linked to cancer for a number of years, the mechanism of their induction in cancer is still a clouded matter. Some insight for the regulation of APOBEC gene expression is available in the literature, but much of it is elusive, especially regarding the expression of APOBEC in cancer (Covino et al., 2018). Most of the current evidence regarding the regulation of APOBEC expression concerns their role in innate immunity and their expression in immune cells. However, there is growing evidence that hypoxia may be implicated in the promotion of APOBEC expression, as I will discuss below. We thus hypothesise that hypoxia plays a role in APOBEC-mediated mutagenesis in cancer considering that hypoxia is a prominent feature of tumours.

The first evidence of APOBEC proteins being induced by hypoxia dates back to a study on the effects of hypoxia on gene expression in hepatocytes (Sonna et al., 2003). The authors described an increased expression of A1 and A3C in hepatocytes. In macrophages, RNA editing of succinate dehydrogenase is observed under hypoxia, but the initial study describing this phenomenon could not associate it to APOBEC proteins (Baysal et al., 2013). More recently, in the same study that showed A3A RNA deamination in monocytes and macrophages, the authors link succinate dehydrogenase RNA editing to A3A and observe an increase in overall cytoplasmic C-to-U RNA deamination by A3A in macrophages under hypoxia that is equal to the extent that IFN-1 can increase RNA cytosine deamination (Sharma et al., 2015). This study is performed in cell culture

expressing A3A exposed to 24h of hypoxia. A later study transfected and transiently overexpressed A3A in an immortalised cell line to evaluate the targets of its RNA deamination activity (Sharma et al., 2017). The authors found that the genes targeted were involved in numerous pathologies including breast cancer, Parkinson's disease, ALS, and more. These models provide sound but limited evidence for the induction of APOBEC mutagenesis in hypoxia, guiding us to further investigate this association in cancer.

1.8 Objectives of the Present Study

Here, we use computational methods to investigate the link between APOBEC and hypoxia in cancer by analysing RNA-seq and mutation data from 6723 patient samples of 13 cancer types (Table 1) provided by The Cancer Genome Atlas (TCGA): <http://cancergenome.nih.gov/>. We take interest in comparing the association between APOBEC and hypoxia across the 13 cancer types as well as across subtypes of breast cancer.

First, we assess whether APOBEC genes are expressed in TCGA data, and whether they are differentially expressed in tumour versus normal samples. To evaluate the hypoxia status of samples, we use a hypoxia signature derived from microarray data using gene co-expression networks (Buffa et al., 2010). This hypoxia signature was validated to be accurate and prognostic in multiple cancers.

Then, we set out to find whether hypoxia is associated with increased expression and activity of other APOBEC genes, especially A3B, the mutational activity of which is strongly associated with breast cancer. Evidence for the increase in expression of other A3B under hypoxia would also prompt us to look for mutations caused by A3B in the genome because of its exclusive nuclear localisation.

To investigate the association between APOBEC and hypoxia beyond gene expression, we turn our attention to the prevalence of its mutational signature. We present four methods for analysing the APOBEC mutational signature, two methods based on the stringent TCW to TTW/TGW definition of the signature

and two methods based on the mathematical definition. We compare the methods and discuss their advantages and shortcomings.

Finally, we perform gene co-expression analysis for APOBEC genes to find the genes for which the expression correlates the most strongly with the expression of APOBEC genes. We then look for enriched cellular pathways in these co-expressed genes. The same analysis is performed for the hypoxia signature and we investigate the overlap of genes and pathways obtained by these methods between APOBEC and hypoxia.

Investigating APOBEC has great clinical significance. On one hand, knowing that A3B is an enzymatic mutator in many cancers, understanding it better would enable therapies to potentially be built around it. Pharmacological inhibition of A3B can have the potential to reduce the mutation load in cancers, slowing their progression towards potentially a more aggressive cancer. On the other hand, we can potentially classify clinical cases by APOBEC status. In conjunction with a better understanding of APOBEC, we can classify patients into different treatment groups considering the effects of elevated APOBEC. The question that we address is related to how APOBEC is induced. Understanding this aspect is crucial to designing any kind of preventive measure for APOBEC-related mutagenesis. In the case of hypoxia specifically, which is common in solid tumours, understanding its mechanism of APOBEC induction can shed light on how APOBECs can potentially activate in certain cancers or certain regions of the tumour that are less vascularised or more metabolically demanding. With this knowledge, we would then be able to both design therapies against APOBEC

mutagenesis and stratify patients into treatment groups based on their cancer's hypoxia measure.

2. Methods

2.1 Data Source

Gene expression data is taken as normalized RNA-seq by expectation maximisation (RSEM) mRNA counts. Data generated by the Cancer Genome Atlas (TCGA) Research Network: <http://cancergenome.nih.gov/>. We retrieve 6723 samples across 13 cancer types detailed in Table 2. Samples are clinical tumour samples from patients and corresponding normal tissue samples when available. Tumour samples are heterogeneous within each cancer type; we did not select for stage, grade, treatment, age, sex, or ethnicity. RSEM is an RNA-Seq quantification method allowing for accurate read counts in de-novo assemblies (Li and Dewey, 2011). RSEM presents gene expression as mRNA abundance, which is then normalized to the 75th percentile of all genes in each sample. All TCGA data is downloaded from the Broad Institute of MIT and Harvard via FireBrowse (version of 2016-01-28; doi: 10.7908/C11G0KM9).

A pan-cancer hypoxia gene signature is obtained from a previous study (Buffa et al., 2010). The signature was derived from microarray data using gene co-expression networks built around previously validated hypoxia genes (seeds). Genes highly and consistently co-expressed with seeds are selected and shown to be prognostic across multiple cancer types. The list of genes in the Buffa et al. signature is available in Table S1. We assess the validity of this signature in the mRNA expression data sets using sigQC, a software package that computes

several metrics used to assess validity (Dhawan et al., 2017). sigQC calculates, for each data set, the expression of the signature, its variability, the correlation between its mean expression, its median expression, and the first principal component of the expression, the effects of data standardisation on the signature, and the level of self-correlation in the signature. To be considered valid, a signature should be significantly expressed, contain variability among samples, display correlation between its mean, median, and first principal component, be independent of standardisation techniques, and be highly self-correlated. Here, we use two sigQC metrics to evaluate the usability of the Buffa signature to our 13 data sets: the expression self-correlation and the correlation between the mean, median, and PC1 of the expression. The other metrics are not taken into account here because the software is still in pre-release stage and exhibits minor errors when used.

APOBEC mutation and amplification data is downloaded from TCGA and analysed through cBioPortal: <http://www.cbioportal.org> (Cerami et al., 2012; Gao et al., 2013). Mutation data includes the proportion of samples exhibiting mutations of any kind in every APOBEC family gene. Amplification data is presented as the correlation between copy number alterations and mRNA expression in each gene.

xCell data for in-silico dissection is retrieved from the xCell platform: <http://xcell.ucsf.edu> (Aran et al., 2017), Feb. 28, 2018. xCell scores vary linearly from 0 to 1 and represent the strength of association of a given cell type's signature to a tissue sample. Indirectly, xCell is used to assess the presence of

infiltrating cell types in tumours through the detection of the infiltrating cells' gene signatures.

Somatic mutation data is obtained from a previous study and downloaded from the Wellcome Trust Sanger Institute:

<ftp://ftp.sanger.ac.uk/pub/cancer/AlexandrovEtAl> (Alexandrov et al., 2013). This data is presented as the prevalence of each of the 96 trinucleotide mutational contexts in TCGA and unpublished samples at the time.

Enrichment scores of mutational contexts are generated by and retrieved from a previous study (Roberts et al., 2013). The mutational enrichment score serves to analyse the mutation count in its genomic context. It takes into account the number of available cytosines, the number of trinucleotide contexts preferred by the APOBEC proteins of interest, and the total number of mutated cytosines in the region surrounding each mutated APOBEC target. It normalises the mutation count for variations in genomic context.

Mutational signatures are curated by and downloaded from the Catalogue of Somatic Mutations in Human Cancer (COSMIC) at the Wellcome Trust Sanger Institute: <https://cancer.sanger.ac.uk/cosmic> (Forbes et al., 2017). This data consists of 30 signatures of enrichment of each of the 96 trinucleotide mutational contexts that can be attributed to different mutational processes.

Table 2: Cancer Types Chosen for Study

TCGA abbreviation	Full name	Sample size
BLCA	Bladder urothelial carcinoma	412
BRCA	Breast invasive carcinoma	1097
COAD	Colon adenocarcinoma	461
HNSC	Head and neck squamous cell carcinoma	528
KIRC	Kidney renal clear cell carcinoma	536
KIRP	Kidney renal papillary cell carcinoma	291
LIHC	Liver hepatocellular carcinoma	377
LUAD	Lung adenocarcinoma	521
LUSC	Lung squamous cell carcinoma	504
PRAD	Prostate adenocarcinoma	498
STAD	Stomach adenocarcinoma	443
THCA	Thyroid carcinoma	507
UCEC	Uterine corpus endometrial carcinoma	548

Cancer types studied. Abbreviations are used throughout the text. Adapted from the United-States National Cancer Institute:

<https://tcga-data.nci.nih.gov/docs/publications/tcga/>.

2.2 Data Pre-Processing

TCGA RSEM data is log2 transformed and separated into tumour and normal samples by TCGA sample ID. Breast invasive carcinoma (BRCA) samples are further sub-divided into four molecular consensus subtypes (Perou et al., 2000; Sorlie et al., 2001): basal-like, HER2-enriched, luminal A, and luminal B. Samples of normal-like subtype are not analysed, as these samples show an expression profile similar to adipose tissue, with a low resemblance to epithelial tissue. These molecular subtypes are a result of unsupervised clustering analysis on the expression of 8102 genes in 65 breast cancer samples in 42 patients (Perou et al., 2000). The result of the clustering is five main subtypes of breast cancer. TCGA samples were classified into these subtypes using the PAM50 algorithm, which uses the mRNA expression levels of a 50-gene signature to perform the classification (Koboldt et al., 2012).

Expression of APOBEC, hypoxia, and anoxia signatures is visualised through boxplots depicting mean expression across cancer types and boxplots for individual gene expression in each cancer type. Two-way analyses of variance are performed for cancer types and sample state (normal/tumour) as independent variables.

2.3 Immune Infiltration

Immune infiltration in TCGA tumour samples is inferred using data from xCell (Aran et al., 2017). xCell is a software package developed to digitally dissect tissue samples into its constituent tissue types using microarray or RNA-seq data. It is useful for detecting immune and stromal infiltration in tumour samples. xCell addresses the shortcomings of using gene set enrichment analysis or deconvolution techniques alone, particularly the inaccuracy in distinguishing similar cell types. It uses a combination of deconvolution and enrichment analyses to provide adjusted enrichment scores for 48 cell types through an algorithm comparing the sample to a gene signature for each cell type. The result is a linear score that represents the strength of a cell type-specific gene signature in a sample, elucidating the presence of the cell type in the sample.

Raw xCell data is plotted as a heatmap. xCell scores in individual samples is shown for each cancer type. Mean scores in samples of each cancer is shown across cancers. Correlation is calculated between each xCell signature and APOBEC or hypoxia, however the statistical power is insufficient to draw accurate conclusions because of the low number of samples with significant immune or stromal infiltration and the low extent of the infiltration.

2.4 Correlation Between APOBEC and Hypoxia Signature Expression

Correlation is assessed in all cancer types by Spearman's correlation test. Results are plotted as a heatmap. Rows and columns are clustered using hierarchical clustering. Correlation between individual APOBEC family member and each hypoxia gene is shown for each cancer type. Correlation between individual APOBEC genes and the mean hypoxia signature is shown across cancers. Scatter plots for the relationship between the mean expression of hypoxia genes and each APOBEC family member are shown.

We then extend the analysis to subtypes of BRCA. All analyses are performed again using only BRCA samples, however this time they are stratified into four different subgroups, one for each consensus subtype of breast cancer, excluding the normal-like subtype. Self-correlation is calculated as Spearman's rho within the APOBEC family for all cancer types.

To correct for immune infiltration, the previous analysis is conducted, adjusting for selected xCell cell type signatures. The selected cell type signatures are those for which 20% or more of samples in the TCGA cohort of each cancer type exhibit an enrichment of 0.2 or more. Spearman's correlation is used partial to the selected cell types for each cancer type. We show that correction for immune infiltration is unnecessary for correlation analysis as the result is insignificantly different from the previous, uncorrected result.

Principal components are calculated for APOBEC family mRNA expression in all cancer types. The components are filtered for standard deviation higher than 1. Genes with over 50% of samples having no gene expression data and genes with

mean expression lower than the 25th percentile of all genes across all samples of the same cancer type are excluded. Standard deviation is plotted as bar plots for each cancer type. Spearman's correlation between each APOBEC gene and each principal component is calculated and plotted as scatter plots. Spearman's correlation is also calculated between principal components and the mean hypoxia signature expression. Spearman's rho is filtered for FDR-adjusted p-value < 0.05. To determine the degree of involvement of each principal component in hypoxia, a principal component regression model is developed with the principal components calculated and the mean expression of the hypoxia signature.

2.5 APOBEC and Hypoxia Gene Co-Expression Analysis

Spearman's correlation is calculated between each APOBEC 3 family member and every gene reported by TCGA in BRCA. The other APOBEC genes are not analysed here because of a lack of non-null samples and low overall expression levels. Two binary parameters are considered, creating 4 independent analyses: whether correlation is partial for the hypoxia signature and whether genes with low expression are filtered out. Genes with low expression are defined here as genes with mean expression across samples of a given cancer type lower than the median expression of all genes in all samples of the same cancer type. For each analysis, two genes lists are created: the list of genes with a Spearman's rho filtered for FDR-adjusted p-value < 0.05 greater or equal than 0.3 in at least 5 APOBEC genes and a list for which rho is higher than 0.3 in APOBEC3B. In all, 8 gene lists are generated.

The same analysis is performed for the hypoxia signature genes, except only without partial correlation to hypoxia and filtering for significantly expressed genes. For each gene reported by TCGA, the sum of correlation scores, Spearman's rho, between it and each individual hypoxia genes is calculated. (The correlation is calculated between individual gene pairs. The sum is applied to the result of the calculation, the correlation score between the gene of interest and a hypoxia signature gene.) Genes are ranked by the sum of their correlation scores, and the top 500 genes are selected for downstream analysis.

All gene lists are entered into a gene set enrichment analysis tool, GeneCoDis: <http://genecodis.cnb.csic.es> (Carmona-Saez et al., 2007). GeneCoDis searches

databases to find annotations that are most significantly present in a given input list of genes. For this analysis, we restrict the database to KEGG pathways, a curated list of cellular pathways kept up to date with the latest research: <https://www.kegg.jp> (Ogata et al., 1999). GeneCoDis returns the pathways that are most enriched for the genes given, along with the significance of their enrichment and its associated genes. The top 10 pathways for each list of genes is plotted as a circos plot with pathways on one half and their associated genes ranked by expression on the other, showing the relationship between genes and pathways as well as the overlap of genes among pathways. The overlap between genes and pathways enriched in genes co-expressed with APOBEC3B, any 5 APOBEC genes, and the hypoxia signature is plotted as Venn diagrams.

2.6 Analysis of the APOBEC Mutational Signature

The analysis of the APOBEC mutational signature is done by four different methods: mutation counts, mutation enrichment, enrichment of consensus mutational signatures, and enrichment of curated consensus mutational signatures.

For mutation counts, the number of TCW to TTW or TGW mutations in the exome of individual TCGA samples is reported, without strand bias. Here, the exome is used because genome data is limited by a lower number of available samples, thus providing a weaker statistical power. The use of genomes versus exomes is a trade-off between the stronger decomposition power of a larger mutational catalogue for each sample using genomes and the stronger statistical power of a larger sample size with less decomposition power per sample. Exomes are used predominantly in mutational signature work for this reason, but this is rapidly changing, as more and more genomes are being sequenced. Eventually, the compromise will be extinct, but for now, we choose to use exome data, as the single-sample decomposition power is still sufficient and accurate for this type of analysis. The data we use is generated by a previous study by Alexandrov et al. and retrieved from the Sanger Institute:

<ftp://ftp.sanger.ac.uk/pub/cancer/AlexandrovEtAl> (Alexandrov et al., 2013). Each sample thus has a number of mutations in the TCW to TTW or TGW context reported, and this count is used for downstream analyses. It is to note that contrary to some publications using exome data, the exome data that we use is not normalised to the genomic frequency, as we are not interested in the relative

frequency of mutational contexts in the exome compared to the genome; we are interested in the relative frequency of mutations in these contexts. To adjust for the prevalence of contexts affecting the number of mutations in them, we use another metric, mutational enrichment, discussed below. Normalising this way also defeats our purpose of using exome data for a larger statistical power.

Mutational enrichment seeks to portray the prevalence of APOBEC-mediated mutations considering the genomic landscape around each mutation. It is defined by a previous study and presented as a score (Roberts et al., 2013). This score, E , represents the relative enrichment of TCW (or its complementary WGA) mutations in the context of the mutated region. Specifically, it is the ratio of TCW to TTW/TGW mutations to the number of C/G mutations, divided by the ratio of total TCW contexts to total C/G mutations in the region surrounding each mutated cytidine.

$$E = \frac{M_{TCW} \times C_{CG}}{M_{CG} \times C_{TCW}}$$

M_{TCW} denotes the number of mutations in a TCW (or WGA) context. M_{CG} denotes the total number of mutated cytidines and guanines. C_{TCW} denotes the total number of TCW (or WGA) contexts in the 41-nucleotide region surrounding each mutated cytidine or guanine. C_{CG} denotes the total number of cytidines and guanines in the 41-nucleotide region surrounding each mutated cytidine or guanine.

The calculation of this score is done for BLCA, BRCA, KIRC, LUAD, LUSC, and PRAD by Roberts et al., 2013. The data is retrieved from the same publication.

The enrichment of consensus mutational signatures measures the presence of each of the 30 mutational signatures originally published by Alexandrov et al. and now curated by COSMIC (Alexandrov et al., 2013; Forbes et al., 2017). To assess the presence of each of the 30 mutational signatures, we use deconstructSigs, a software package developed for this role (Rosenthal et al., 2016). deconstructSigs allows the analysis of single TCGA samples for the enrichment of pre-defined mutational signatures. deconstructSigs works by fitting a 'least-squares' multiple regression model with a number of exceptions on each sample using all available consensus mutational signatures (or a custom list of defined signatures) to find the optimal linear combination of signatures that best explains the mutational landscape in the sample. Exceptions include the exclusion of any defined signature with a single trinucleotide context that is not present in the sample making up more than 20% of mutations, and the inability of the algorithm to assign negative coefficients in the linear model. The coefficients resulting from the model are then normalised between 0 and 1 and represented as a score, with scores lower than 0.06 being omitted to exclude false positives. The 0.06 figure is obtained empirically by Rosenthal et al. as the optimal threshold for filtering out false positives.

In the first instance, deconstructSigs is run with the 30 COSMIC signatures in all cancer types. The calculated scores are plotted as jitter. The percentage of non-null scores is reported for each cancer type. Since all values between 0 and 0.06 are filtered down to 0 by deconstructSigs, this percentage represents the samples with significant presence of the mutational signature.

Then, `deconstructSigs` is run with a curated list of signatures that differs for each cancer type. The reason for this is because some signatures are only present biologically in certain cancer types. For example, signature 4 is associated with tobacco smoke and is only found empirically in select cancer types. Since `deconstructSigs` is a multiple regression algorithm producing normalised scores, the inclusion of defined signatures that are not biologically present may skew the result of signatures that are biologically present. The list of signatures relevant to each cancer type is manually curated using information provided by COSMIC. Again, the scores are plotted as jitter and the percentage of non-null scores is reported.

Common to all four methods, the mutational signature analysis result, presented as a count, an enrichment score, or a normalised regression coefficient, is plotted in scatter plots against APOBEC3A, APOBEC3B, and hypoxia signature expression, and their correlation is calculated (Spearman's rho, filtered for FDR-adjusted p-value < 0.05).

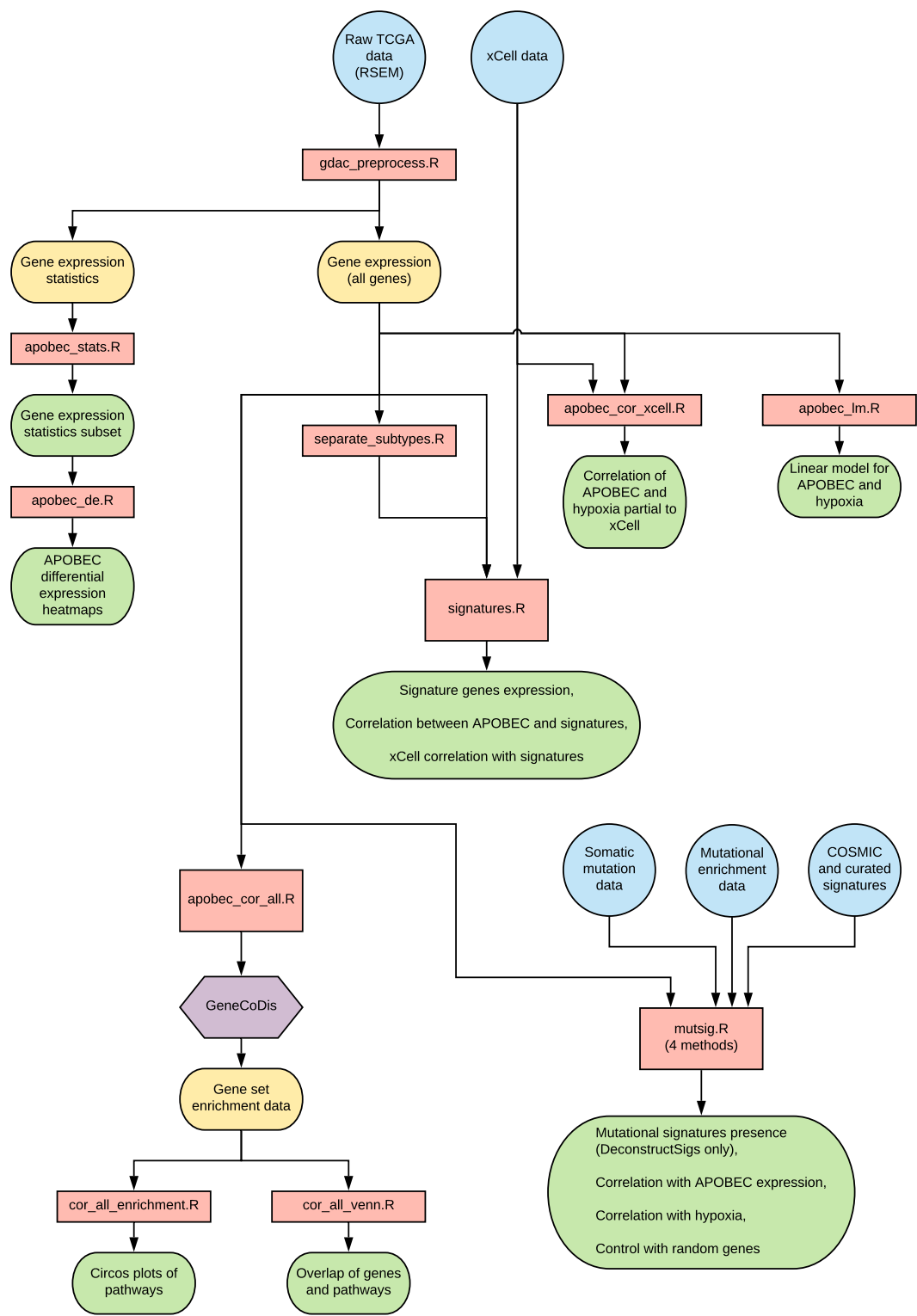
As a control, in each cancer type, 100 genes are selected at random and their individual expression is correlated against the mutational signature analysis result of each method and plotted as boxplots. The correlation scores for APOBEC3A, APOBEC3B, and the mean hypoxia score are overlaid for comparison. A Z-score is calculated for the correlation with APOBEC3A, APOBEC3B, and hypoxia with respect to the distribution of the correlation with random control genes.

2.7 Software Deployment

All analyses with the exception of gene set enrichment analysis using GeneCoDis and analyses performed by previous studies are performed locally in R for Windows (version 3.4.3) running Windows 10 and R for MacOS (version 3.4.3) running MacOS 10.12. Packages used include 'sigQC', 'gplots', 'ggplot2', 'hclust', 'matrixStats', 'parallel', 'doParallel', 'GOplot', and 'deconstructSigs'.

A visual overview of all data and analysis methods used is available in Figure 2.

Figure 2: Pipeline of Analysis Methods



Overview of methods and data used to generate the analyses in this study.

3. Results

3.1 APOBEC is Expressed and Functional Across Cancers

The first step of the analysis is to determine the expression status and activity of APOBEC family members. Gene expression data is obtained as normalised RNA-seq by expectation maximisation (RSEM) values from the Cancer Genome Atlas (TCGA). The samples analysed are clinical tumour and corresponding normal samples retrieved from cancer patients. We analysed 13 cancer types outlined in Table 1. Each cancer type's data set contains a heterogeneous set of tumours that is not selected for sex, age, ethnicity, tumour grade, tumour stage, or treatment. We selected carcinomas for which there were more than 200 samples available in the TCGA data portal. Cervical squamous cell carcinoma and endocervical adenocarcinoma, and ovarian serous cystadenocarcinoma are excluded because of partial data unavailability at the time of retrieval.

3.1.1 APOBEC3 Genes are Significantly and Differentially Expressed Across Cancers

Mean expression of all APOBEC genes (Table 1) is shown across 13 cancer types for both solid tumour samples and normal samples (Figure 3). We observe that mean APOBEC expression varies significantly among cancer types ($p < 0.0001$ two-way ANOVA by cancer type and normal/tumour state), but APOBEC

is not significantly differentially expressed in tumour samples compared to normal ($p = 0.895$ two-way ANOVA by cancer type and normal/tumour state).

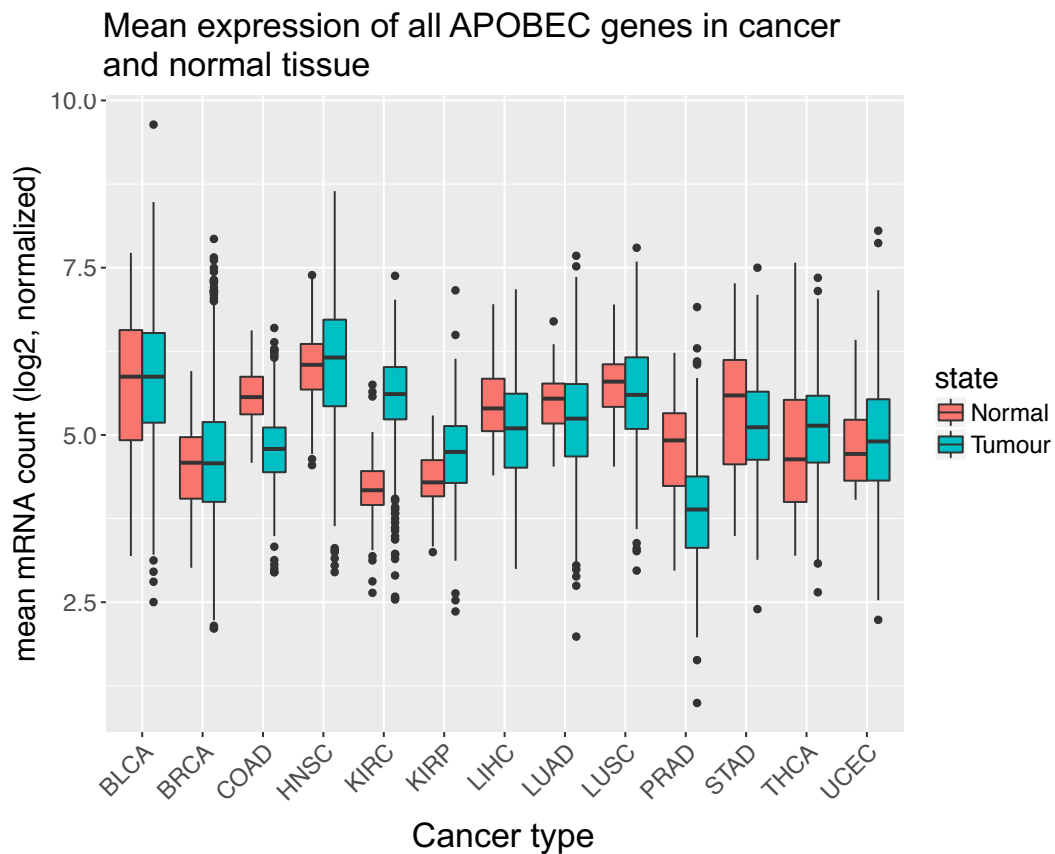
As established previously, APOBEC genes differ largely in function, expression, and regulation. It is unsurprising that the mean expression of all APOBEC genes shows large variations and inconsistencies among cancer types. This visualisation does not, however, give insight into which APOBEC family members are driving the expression up or down. Specific APOBEC family members may be key in acting in cancer, but this effect can be masked by only taking the mean expression of all APOBEC genes. We look at the mean expression of individual APOBEC genes in Figure S1 and observe substantial variations in their expression across all cancer types. We present data for all 13 cancer types, but for this study, we focus on breast cancer for two reasons: APOBEC has already shown to be involved in breast cancer (Burns et al., 2013a, 2013b; Roberts et al., 2013), and the BRCA data set is the largest available on TCGA.

The expression of individual APOBEC genes in breast invasive carcinoma (BRCA; abbreviations in Table 2) is presented in Figure 4. It is clear that multiple APOBEC family members are not significantly expressed in BRCA. APOBEC3 (A3) family genes exhibit the highest expression levels, with APOBEC3A (A3A) and 3B (A3B) being preferentially expressed in tumours. This is in line with previous findings suggesting that A3A and A3B are active in cancer (Roberts et al., 2013). When looking across cancer types, we observe that A3B is preferentially expressed in tumour samples in all cancer types with the exception

of COAD and THCA (Figure 5, $p < 0.0001$ two-way ANOVA). The cause of this discrepancy is unknown.

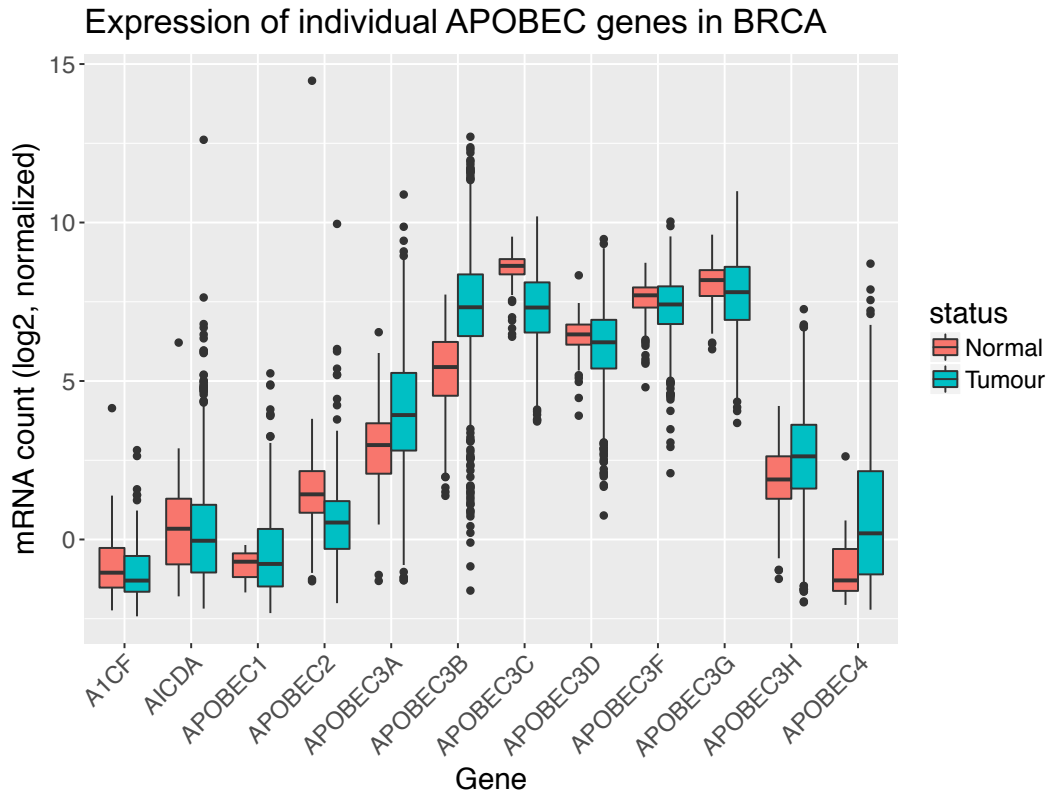
BRCA is chosen for this analysis because APOBEC activity has been reported in breast cancer in the past and because it is the largest TCGA data set, thereby providing the largest statistical power. The trend of APOBEC gene expression is largely conserved among different cancer types, with APOBEC3 family genes being the most highly expressed while other APOBEC genes are lowly expressed in most cancer types, with notable exceptions such as A1CF in KIRC, KIRP, and LIHC. (Figure S1). It is also worth noting that APOBEC3 family genes are consistent in their expression, with relatively low variation among samples. This is not the case for the other APOBEC genes across the 13 cancer types.

Figure 3: Mean Expression of APOBEC Genes across 13 Cancer Types



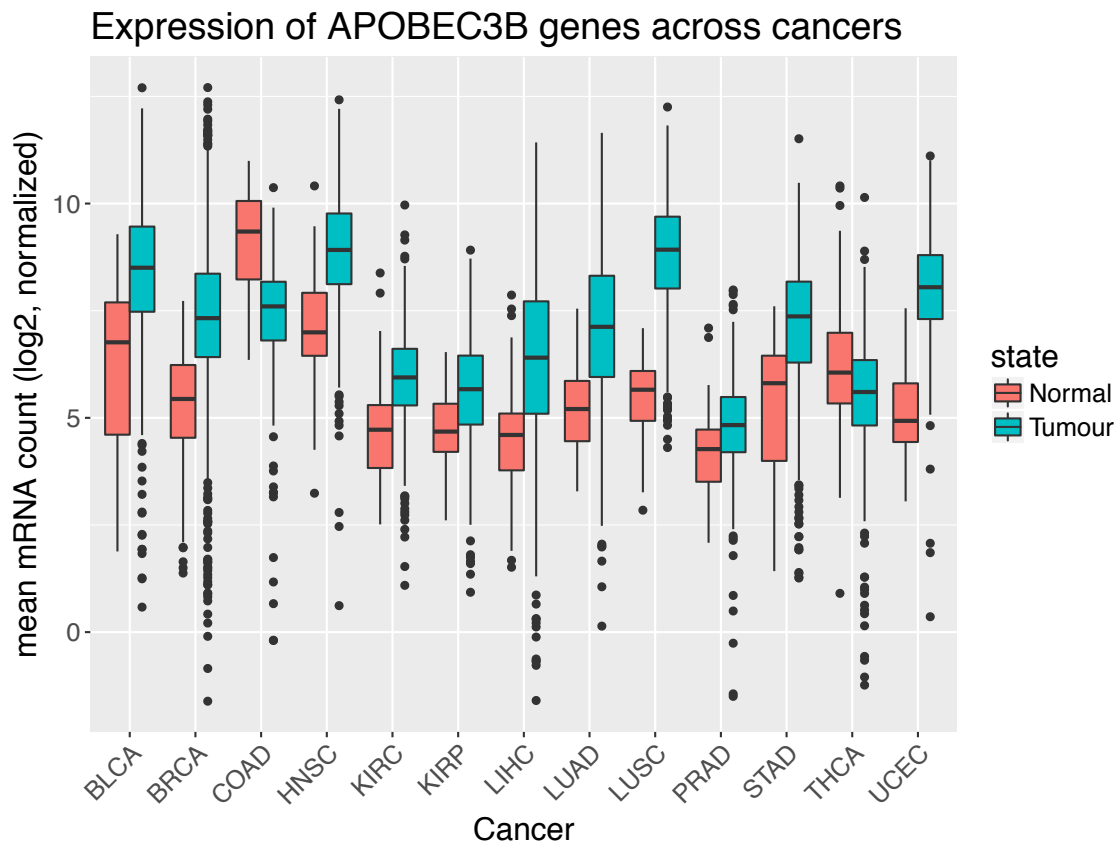
Mean expression of APOBEC genes across 13 cancer types and their associated normal tissues. Data is log 2 transformed, normalised mRNA count from TCGA. The mean of values for all APOBEC genes in each sample is displayed. Boxes represent the range from the first to the third quartiles. Centre line represents the median. Whiskers represent 3rd quartile plus 1.5x the interquartile range or the maximal value, whichever is smaller, and the 1st quartile minus 1.5x the interquartile range or the minimal value, whichever is larger. Dots represent values not contained within whiskers.

Figure 4: Expression of APOBEC Genes in Breast Cancer



Expression of individual APOBEC genes in BRCA, for tumour and normal tissue. Data is log 2 transformed, normalised mRNA count from TCGA. Values for each sample are displayed. Boxes represent the range from the first to the third quartiles. Centre line represents the median. Whiskers represent 3rd quartile plus 1.5x the interquartile range or the maximal value, whichever is smaller, and the 1st quartile minus 1.5x the interquartile range or the minimal value, whichever is larger. Dots represent values not contained within whiskers.

Figure 5: Expression of APOBEC3B across 13 Cancer Types



Expression of APOBEC3B across 13 cancer types and their associated normal tissues. Data is log 2 transformed, normalised mRNA count from TCGA. Boxes represent the range from the first to the third quartiles. Centre line represents the median. Whiskers represent 3rd quartile plus 1.5x the interquartile range or the maximal value, whichever is smaller, and the 1st quartile minus 1.5x the interquartile range or the minimal value, whichever is larger. Dots represent values not contained within whiskers.

3.1.2 Decomposing APOBEC Expression into its Components

To further dissect the involvement of individual APOBEC genes in the family, principal component analysis (PCA) is performed. A principal component is a linear combination of variables from a data set that maximises the variance of the data. In essence, PCA decomposes the data into components of maximum variance, starting by principal component 1 (PC1) containing the maximum variance possible in the data. PC2 is the component orthogonal to PC1 that contains the maximum variance. In other words, PC2 contains the maximum variance that is not already accounted for in PC1. Each subsequent component follows this trend of orthogonality and maximum variance. Each real data point thus has a projection on each principal component and these components have decreasing variance for the data projections.

Principal components for the expression of APOBEC genes are calculated for all cancer types. Components with a standard deviation higher than 1 are retained. Only APOBEC genes with a mean expression higher than the 25th percentile of all genes across all samples of the same cancer type are considered. Genes for which over 50% of samples for with gene expression data not available are discarded. Genes remaining after these filtering steps are the A3 family genes. Coefficients representing each A3 gene for the principal components are displayed in Table 3. These coefficients, taken as absolute values, represent the degree to which each variable, in this case A3 genes, most explains each principal component. We observe that in BRCA, PC1 is mainly driven by A3A and A3H, with A3B, A3C, A3F, and A3G also involved. PC2, on the other hand, is

almost exclusively explained by A3B. We can interpret this result as A3A and A3H having the most varied expression, followed by A3B. In other cancers, the distinction between A3A and A3B, and among A3 family genes in general is much less clear. PC1 is explained by both A3A and A3B, along with other A3 family genes. It is PC2 that sees slightly more distinction between the different genes. This reveals that the APOBEC3 family is highly self-correlated, and we address this question in the next section.

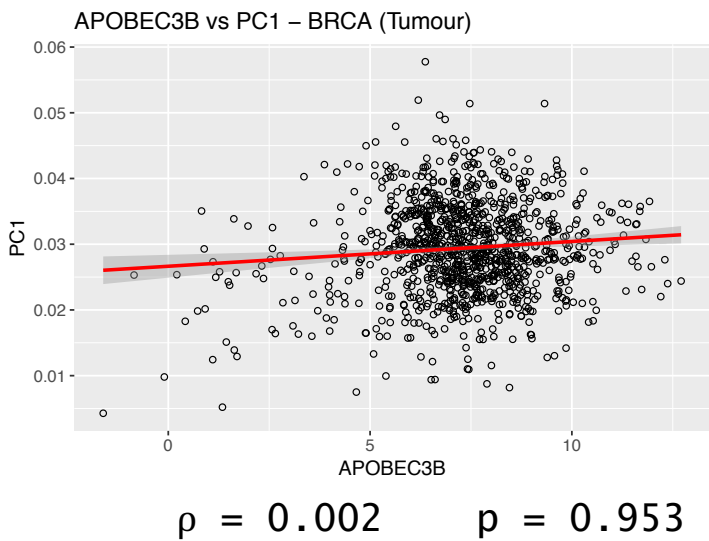
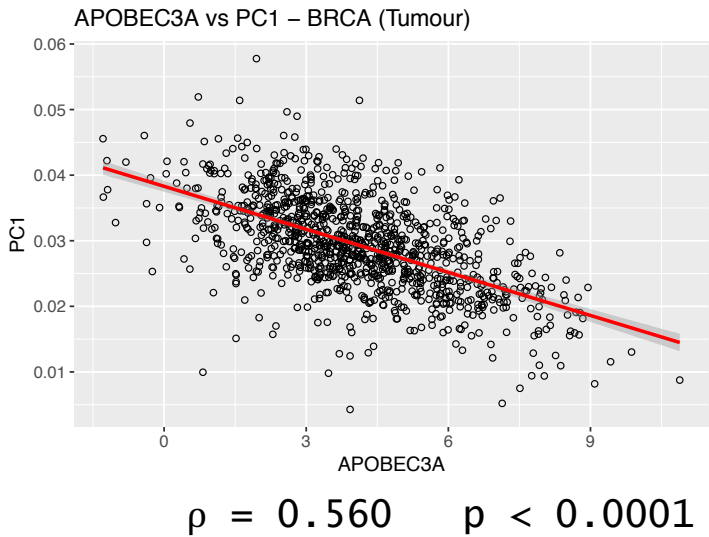
To illustrate the relationship between the principal components and the actual expression of APOBEC genes, we generated scatter plots showing their correlation in BRCA (Figure 6). Both A3A and A3B are correlated with their respective principal components (A3A vs PC1 $\rho = 0.560$; A3B vs PC2 $\rho = 0.802$). This reveals the fact that PC1 is mainly explained by A3A, and PC2 by A3B in BRCA. The clear distinction between A3A and A3B seen in BRCA will be particularly useful in a later section when I will use the results from principal component regression to dissect the involvement of each principal component and APOBEC gene in hypoxia.

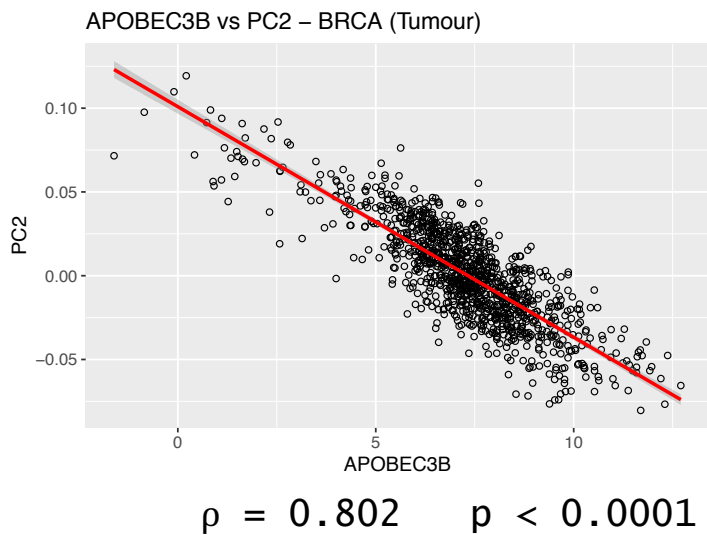
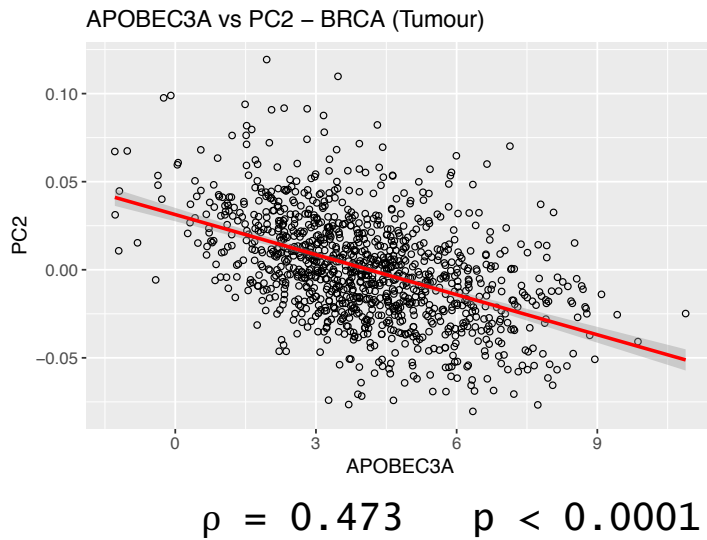
Table 3: Principal Component Coefficients of the Expression of APOBEC3 Genes in Breast Cancer

	PC1	PC2	PC3	PC4	PC5	PC6	PC7
A3A	-6.97E+01	-2.63E+01	-2.47E+01	8.16E-01	-7.02E-01	5.45E-02	-3.40E-15
A3B	4.37E+01	-4.58E+01	1.70E+01	1.76E+00	4.93E-01	1.41E+00	8.74E-14
A3C	4.14E+01	1.65E+01	-1.93E+00	1.31E+01	-1.48E+01	-1.83E+00	6.19E-14
A3D	3.22E+00	1.55E+01	1.05E+00	-1.22E+01	-2.94E+00	1.43E+01	6.11E-14
A3F	4.38E+01	8.06E+00	-2.29E+00	-1.61E+01	-2.61E-02	-1.19E+01	1.03E-13
A3G	5.42E+01	1.65E+01	-5.99E+00	9.70E+00	1.55E+01	1.48E+00	4.94E-14
A3H	-1.17E+02	1.55E+01	1.68E+01	2.93E+00	2.47E+00	-3.55E+00	-3.49E-13

Principal components are calculated for the expression of APOBEC3 family genes in all BRCA samples. Only components with a standard deviation greater than 1 are displayed.

Figure 6: APOBEC3A and 3B Versus Principal Components 1 and 2 of the Expression of All APOBEC Genes



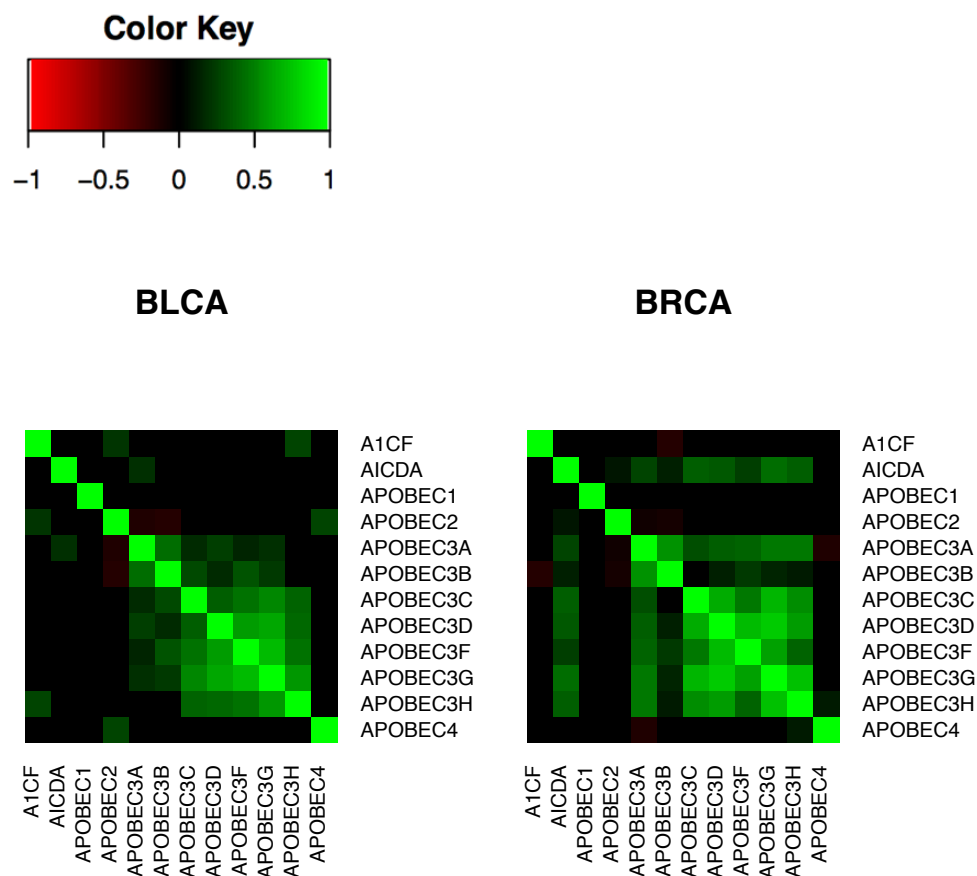


Correlation between APOBEC gene expression and principal components of the expression of all APOBEC genes. Expression data is log 2 transformed, normalised mRNA count from TCGA. Trendline represents least-square linear regression model. Spearman's correlation coefficients are calculated and displayed as absolute values.

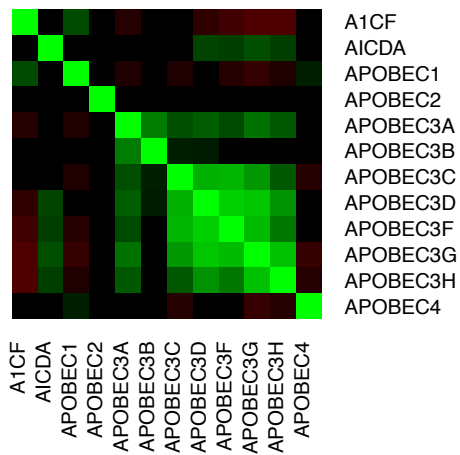
3.1.3 The Expression of APOBEC3 Family Genes is Highly Self-Correlated

Next, I plotted the self-correlation of APOBEC family genes, or the correlation of the expression of APOBEC genes with one another (Figure 7). This technique allows us to identify genes that are co-expressed and provide insight into their regulation. From the plot, we can see that the A3 family is highly self-correlated across cancers. Exceptions include A3B in BRCA, COAD, KIRP, and PRAD, which is only correlated with A3A, and A3A being uncorrelated with any other APOBEC gene in HNSC and KIRC. This trend is expected, as all A3 genes are closely located in the same genomic region. The transcription of all A3 genes may be controlled simultaneously by epigenetic factors and regional promoters that act for all of them. As for A3A and A3B, considering the fact that their function and observed presence, and localisation in the case of A3B, are different from the rest of the A3 family, it is unsurprising that their expression is potentially under different regulation, which leads to its expression as reported by RNA-seq being uncorrelated with others. From here, we can reaffirm that although A3A and A3B are both expressed and active in, their function and regulation are different from the rest of the APOBEC genes, and the two can act in different ways in hypoxia.

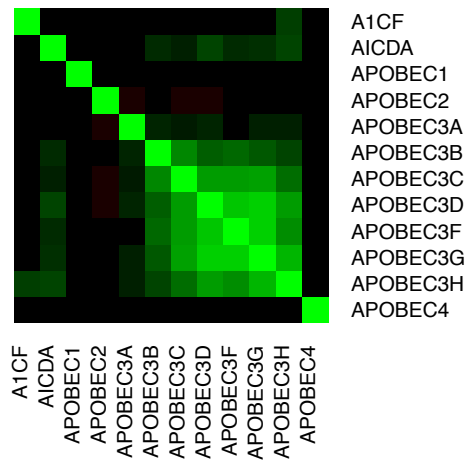
Figure 7: Self-Correlation of the Expression of APOBEC Genes



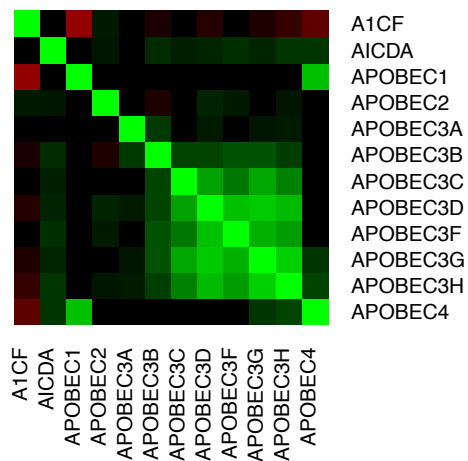
COAD



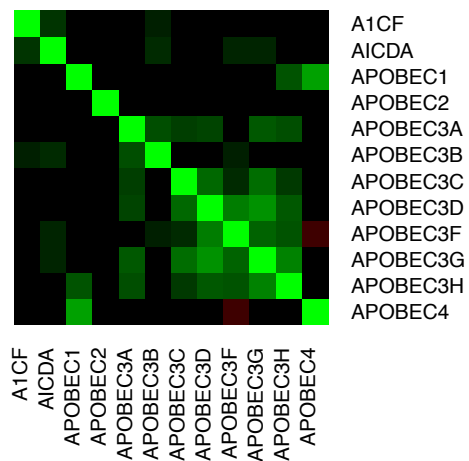
HNSC



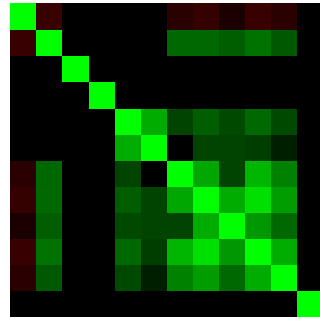
KIRC



KIRP



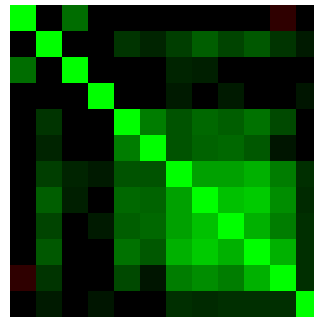
LIHC



A1CF
AICDA
APOBEC1
APOBEC2
APOBEC3A
APOBEC3B
APOBEC3C
APOBEC3D
APOBEC3F
APOBEC3G
APOBEC3H
APOBEC4

A1CF
AICDA
APOBEC1
APOBEC2
APOBEC3A
APOBEC3B
APOBEC3C
APOBEC3D
APOBEC3F
APOBEC3G
APOBEC3H
APOBEC4

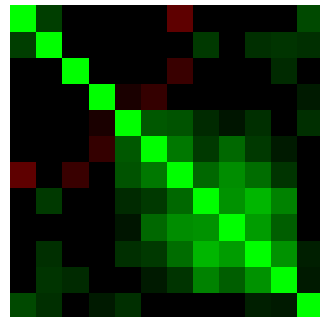
LUAD



A1CF
AICDA
APOBEC1
APOBEC2
APOBEC3A
APOBEC3B
APOBEC3C
APOBEC3D
APOBEC3F
APOBEC3G
APOBEC3H
APOBEC4

A1CF
AICDA
APOBEC1
APOBEC2
APOBEC3A
APOBEC3B
APOBEC3C
APOBEC3D
APOBEC3F
APOBEC3G
APOBEC3H
APOBEC4

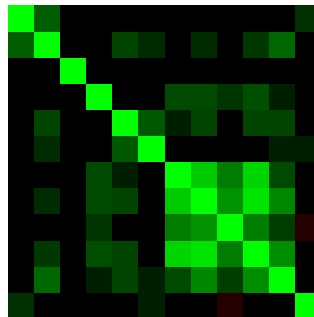
LUSC



A1CF
AICDA
APOBEC1
APOBEC2
APOBEC3A
APOBEC3B
APOBEC3C
APOBEC3D
APOBEC3F
APOBEC3G
APOBEC3H
APOBEC4

A1CF
AICDA
APOBEC1
APOBEC2
APOBEC3A
APOBEC3B
APOBEC3C
APOBEC3D
APOBEC3F
APOBEC3G
APOBEC3H
APOBEC4

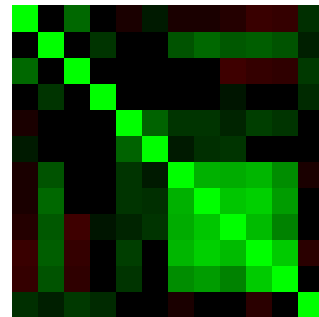
PRAD



A1CF
AICDA
APOBEC1
APOBEC2
APOBEC3A
APOBEC3B
APOBEC3C
APOBEC3D
APOBEC3F
APOBEC3G
APOBEC3H
APOBEC4

A1CF
AICDA
APOBEC1
APOBEC2
APOBEC3A
APOBEC3B
APOBEC3C
APOBEC3D
APOBEC3F
APOBEC3G
APOBEC3H
APOBEC4

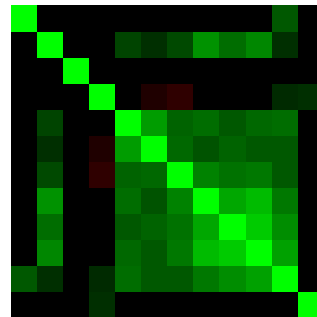
STAD



A1CF
AICDA
APOBEC1
APOBEC2
APOBEC3A
APOBEC3B
APOBEC3C
APOBEC3D
APOBEC3F
APOBEC3G
APOBEC3H
APOBEC4

A1CF
AICDA
APOBEC1
APOBEC2
APOBEC3A
APOBEC3B
APOBEC3C
APOBEC3D
APOBEC3F
APOBEC3G
APOBEC3H
APOBEC4

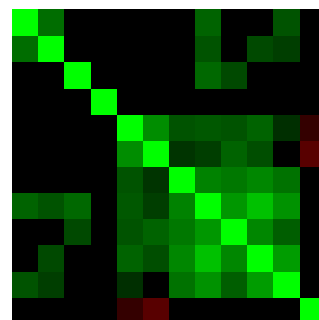
THCA



A1CF
AICDA
APOBEC1
APOBEC2
APOBEC3A
APOBEC3B
APOBEC3C
APOBEC3D
APOBEC3F
APOBEC3G
APOBEC3H
APOBEC4

A1CF
AICDA
APOBEC1
APOBEC2
APOBEC3A
APOBEC3B
APOBEC3C
APOBEC3D
APOBEC3F
APOBEC3G
APOBEC3H
APOBEC4

UCEC



A1CF
AICDA
APOBEC1
APOBEC2
APOBEC3A
APOBEC3B
APOBEC3C
APOBEC3D
APOBEC3F
APOBEC3G
APOBEC3H
APOBEC4

A1CF
AICDA
APOBEC1
APOBEC2
APOBEC3A
APOBEC3B
APOBEC3C
APOBEC3D
APOBEC3F
APOBEC3G
APOBEC3H
APOBEC4

Self-correlation in the expression of APOBEC genes for 13 cancer types. Values displayed represent Spearman's rho, filtered for $p < 0.05$, calculated using all samples from TCGA. Green represents positive correlation, red, negative. Correlations that are not significant are shown in black.

3.1.4 *The APOBEC Mutational Signature is Present in Cancer*

The above analysis shows us that APOBEC genes, namely A3A and A3B, are expressed at significant levels in cancer. We see that A3B, the most active member in breast cancer, is preferentially expressed across nearly all cancers. Decomposing the expression of APOBEC into principal components, we see that the highest variance in expression comes from A3A and A3B in breast cancer. Lastly, we observe that the expression of the APOBEC3 family of genes is highly self-correlated in many cancers.

However, the previous analyses do not provide insight into the activity of a protein product. Measures other than expression are needed to assert whether the end APOBEC protein is functional and active. Since APOBEC proteins can induce mutations in single-stranded DNA and RNA, we can search for traces of their mutational activity. Furthermore, some APOBECs are localised to the nucleus, thus we can examine the genomic DNA of cells for signs of APOBEC-induced mutations. A3A and A3B are the main active cytosine deaminases and they have a selective preference for TCW trinucleotide sequences (with the underlined base mutated; W stands for A or T). By analysing the prevalence of cytosine deamination at TCW motifs, we can paint a picture of the state of APOBEC activity in a cell. Previous studies have confirmed the existence of APOBEC-induced mutations in breast cancer using measures based on this technique (Roberts et al., 2013). We reproduce their results using current data here. Roberts et al. calculated an enrichment score for APOBEC mutations. We plot this score against A3A and A3B expression in Figure 22 (section 3.7.2).

3.2 APOBEC is not Significantly Mutated in Breast Cancer

It is important to assess whether APOBEC is itself mutated in TCGA samples before conclusions are drawn about its functional properties. Even if overexpressed, a mutated and dysfunctional APOBEC may not itself induce mutations in the genome and cause phenotypic effects. We investigate if the APOBEC genes are mutated in TCGA samples using cBioPortal, a platform for simple analysis and visualisation of cancer genomics data (Cerami et al., 2012; Gao et al., 2013).

Retrieving TCGA mutation data and plotting it using cBioPortal, we see that all cancer types show limited mutation in APOBEC genes (Table 4). In fact, only A1CF in LUSC and A1 in UCEC are mutated in more than 3% of samples. The vast majority of cancer/gene pairs have mutations in fewer than 1% of samples. COAD seems to have slightly more mutated APOBEC genes than the others, but the only gene mutated in more than 2% of samples is A1. Importantly, all A3 family genes show low levels of mutation across all cancer types.

Table 4: Mutations in APOBEC Genes

	BLCA	BRCA	COAD	HNSC	KIRC	KIRP	LIHC
AICDA	0.49	0.18	1.82	0.38	0.2	0	0.54
A1CF	0.97	0.65	1.59	1.72	0.59	0	1.08
APOBEC1	0.49	0.28	2.05	0.19	0.2	0.35	0
APOBEC2	0.49	0.18	1.14	0	0	0	0.27
APOBEC3A	0.24	0.18	1.14	0	0	0	0.27
APOBEC3B	1.22	0.46	1.14	0.96	0.2	0.35	0.27
APOBEC3C	0.49	0.09	1.37	0.19	0	0	0
APOBEC3D	0.73	0	0.91	0.38	0	0	0
APOBEC3F	0.24	0.09	1.14	0	0	0	0
APOBEC3G	0.4	0.28	1.82	0.19	0	0	0.27
APOBEC3H	0.97	0	0.68	0.19	0	0	0.27
APOBEC4	1.7	0	1.59	0	0.39	0	0.54

	LUAD	LUSC	PRAD	STAD	THCA	UCEC
AICDA	0.71	1.03	0.2	0.91	0	0
A1CF	2.12	3.08	1.21	1.82	0.2	1.75
APOBEC1	0.35	2.05	0.62	1.36	0	3.51
APOBEC2	0.18	0.41	0.2	0.91	0.2	0
APOBEC3A	0.53	0.41	0	0.23	0	0
APOBEC3B	0.35	1.23	0.81	1.14	0	0
APOBEC3C	0.18	0.21	0	0.68	0	1.75
APOBEC3D	1.06	0.62	0	0.68	0	0
APOBEC3F	0	1.23	0	0.68	0	1.75
APOBEC3G	0.88	0.62	0.2	0.68	0.2	0
APOBEC3H	0.71	0.41	0	0.23	0	0
APOBEC4	0.88	0.41	0	0.45	0	0

Percentage of TCGA samples with a mutation in each of the APOBEC genes.

Data provided by TCGA and retrieved using cBioPortal.

3.3 APOBEC Amplification Does not Drive the Differences in Expression Between Breast Cancer and Normal Breast Samples

An increase in mature mRNA detected by RNA-seq can be due to a multitude of factors, notably the upregulation of transcription or, for example, the amplification of the gene. Here, we argue that although amplification can be seen in APOBEC, it is not likely to be the cause of the change in expression.

We measure amplification by the copy number of APOBEC genes. This data is provided by TCGA and retrieved and processed using cBioPortal. The Spearman's correlation between copy number and mRNA count in BRCA is displayed in Table 5. Correlation between mRNA count and copy number is an indication that the increase in copy number translates to an increase in mRNA. APOBEC 1, 2, and 4 show a low, but significant level of correlation between their mRNA count and amplification. However, we recall from the previous section that these genes are lowly expressed, and the effect seen here may not be significant. The APOBEC3 family appears to show significant correlation between copy number and expression. This may explain the relatively high expression of APOBEC3 genes as measured by mRNA count compared to other APOBEC genes. However, this cluster of genes is located in a very near region thus amplification would tend to increase the expression of all members of the family. The heterogeneity that we observe indicates that while the genomic APOBEC3 gene cluster may be amplified, amplification cannot account for the differences in expression among its family members. With amplification being a result of selection, it suggests that the APOBEC3 family is selected for in tumours. This

mirrors the evolutionary history of this family of genes, which relatively recently has diverged from AID. Nevertheless, cellular regulation is involved in determining differential transcription of the family members since we observe differing mRNA levels for members of the APOBEC3 family. As an example, A3B has been shown to be inducible by the NF- κ B signalling pathway in hepatocytes (Matsumoto et al., 2006) multiple human cancers (Leonard et al., 2015; Maruyama et al., 2016). Conversely, A3A and A3G can be induced by IFN- α (Covino et al., 2018).

Table 5: Correlation Between Somatic Copy Number and Normalised mRNA Count of APOBEC Genes in Breast Cancer

	Spearman's rho Normalized mRNA count vs somatic copy number
AICDA	0.01
A1CF	0.1
APOBEC1	0.11
APOBEC2	0.1
APOBEC3A	0.13
APOBEC3B	0.24
APOBEC3C	0.17
APOBEC3D	0.16
APOBEC3F	0.29
APOBEC3G	0.11
APOBEC3H	0.04
APOBEC4	0.15

Spearman's correlation coefficient between the normalised mRNA count and somatic copy number of individual APOBEC genes in breast cancer. mRNA expression and copy number data provided by TCGA. Correlation calculated by cBioPortal.

3.4 Hypoxia is Present in Tumours and Correlates with APOBEC3A and APOBEC3B Expression

The mechanism of induction of APOBEC expression in cancer is currently unclear (Covino et al., 2018). Evidence is limited in both the number of studies conducted and the scope of the studies. However, some evidence currently suggests that hypoxia can induce the expression of APOBEC genes (Baysal et al., 2013; Sharma et al., 2015, 2017). Specifically, one study shows that A3A can be induced by hypoxia to the same extent as immune cytokines can in macrophages (Sharma et al., 2015). This leads us to believe that hypoxia may also be responsible for the induction of APOBEC gene expression in cancer.

3.4.1 A Hypoxia Gene Signature is Highly and Preferentially Expressed in Tumours

In order to investigate the role of hypoxia in the possible regulation of APOBEC genes, we must first assess the level of hypoxia present in TCGA samples. To do so, we use a previously validated gene signature for hypoxia (Buffa et al., 2010). This signature consists of a list of 52 genes that are preferentially expressed in hypoxic conditions across cancers. It was generated using microarray expression data in a gene co-expression network analysis. Namely, starting with 10 hypoxia ‘seed’ genes obtained from manual curation, other genes are recruited based on their co-expression with these seeds. The co-expression network is allowed to reach a steady state, where the most interconnected genes (‘hubs’) are selected to become the signature. This signature was validated in

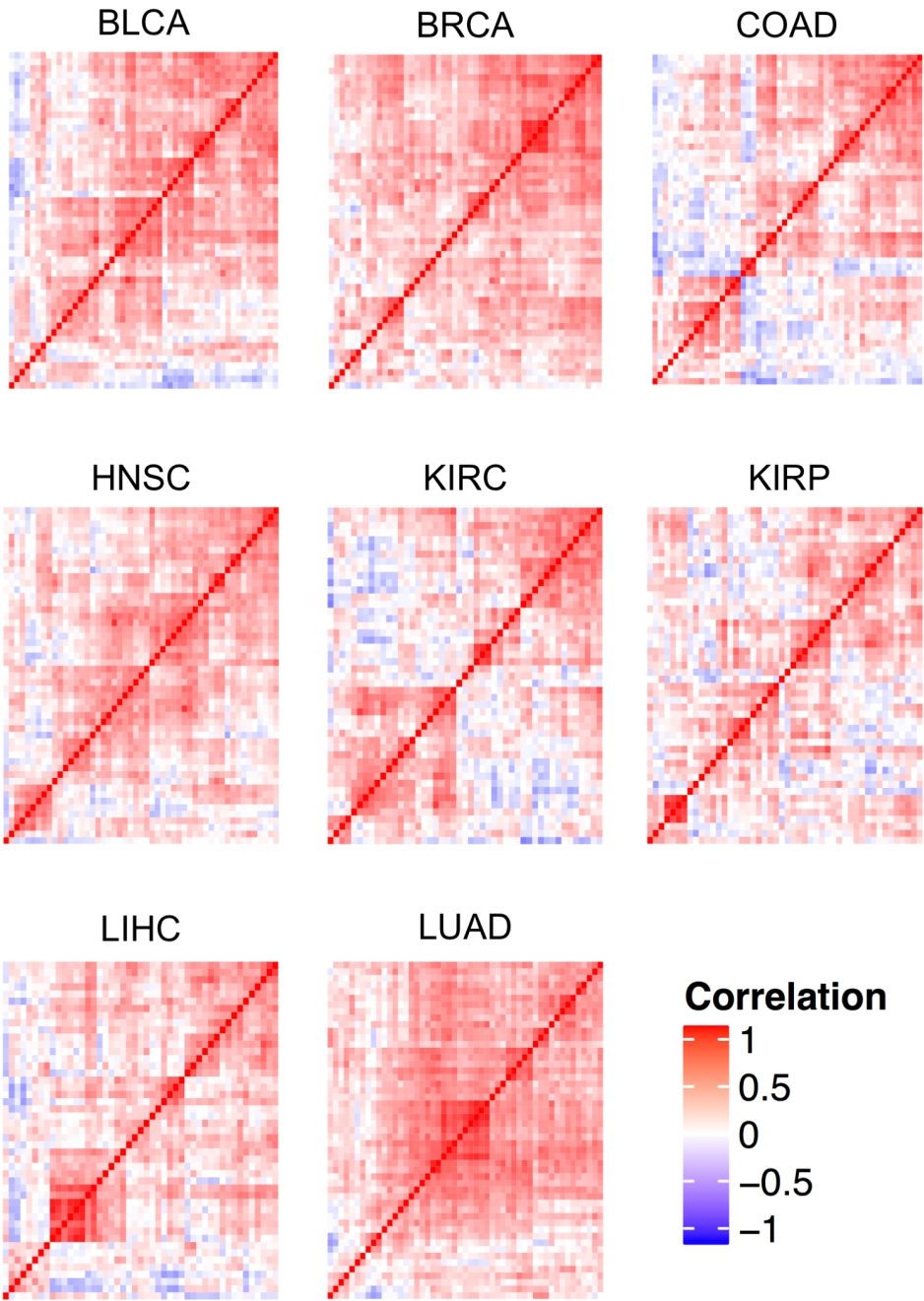
breast cancer only and may give rise to artefacts when applied to other cancer types. Although this signature has not been validated as a hypoxia signature in all cancers, it has been found to be robust and prognostic across multiple cancer types.

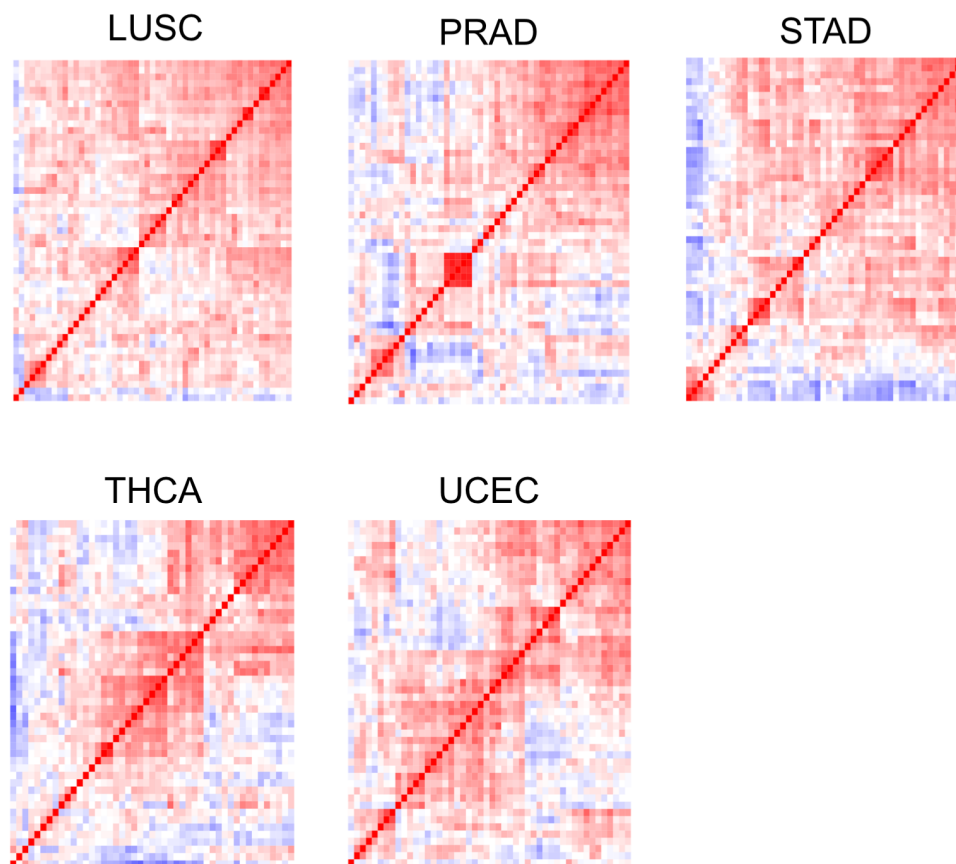
To assess the validity of this signature in our data sets, we use sigQC to perform a series of computations that describe the applicability of a signature in a data set (Dhawan et al., 2017). sigQC calculates many metrics to assess the validity of a signature. Two important metrics are the degree of self-correlation in the expression of the signature, referred to as 'compactness', and the degree of correlation between the mean, median, and PC1 of the expression of the signature. These correlation metrics should be high for a signature to be valid in a given data set. We apply sigQC on the Buffa signature in the 13 cancer data sets used here. A summary of the self-correlation in the signature is presented in Figure 8, with the full figures in Figure S2, while the correlation between mean, median, and PC1 of the expression is presented in Figure 9. We see from the sigQC analysis that the signature is highly self-correlated and shows strong correlation between mean, median, and PC1 of its expression in most data sets. BRCA, HNSC, and LUAD show especially strong correlation. For COAD, KIRC, PRAD, and THCA, there is some low and inverse correlation in the expression of several genes and a lower correlation between mean, median, and PC1 of the expression compared to other cancer types. We must keep this in mind when interpreting results from these data sets.

Here, we also show that this gene signature is over-expressed in the TCGA samples. The expression of the mean hypoxia signature is displayed in Figure 10. For every TCGA sample, the mean of the expression of each individual gene in the Buffa signature is taken and plotted. We observe that tumour samples show a higher expression of the mean hypoxia signature overall and across all cancer types ($p < 0.0001$ two-way ANOVA). This result is expected, as hypoxia is known to be strongly manifested in tumours.

Although this signature has been shown to be prognostic in multiple cancer types, it has its limitations. Notably, this is a signature to the transcriptional response to hypoxia, which can vary across cancer types. In addition, post-transcriptional factors are not considered here. A study is underway to investigate hypoxia signatures per cancer type and adjusting for cancer subtypes. I took part in the project, but further work is required to use these signatures for analysis.

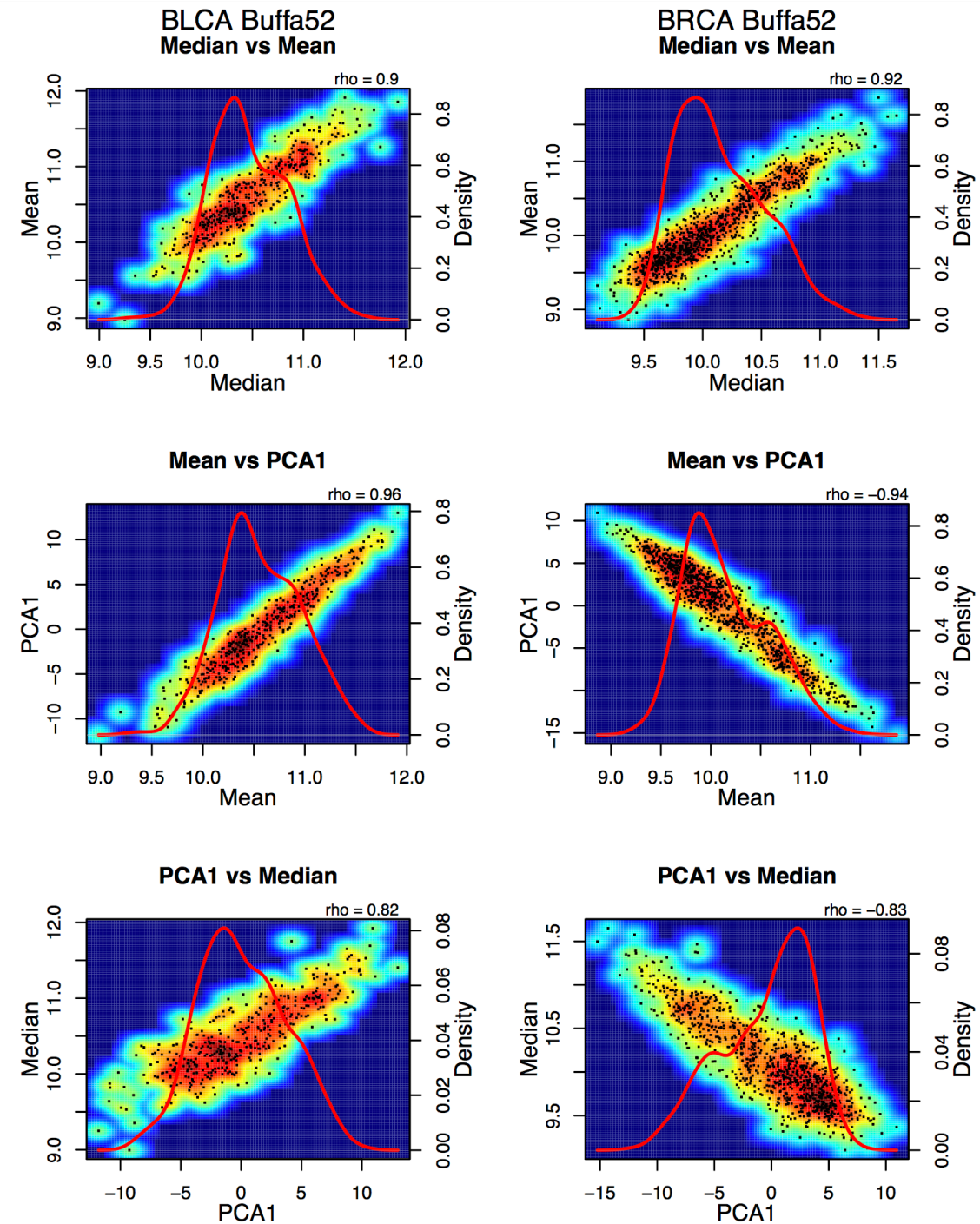
Figure 8: Overview of Self-Correlation of Expression in the Buffa Hypoxia Signature Across 13 Cancer Types



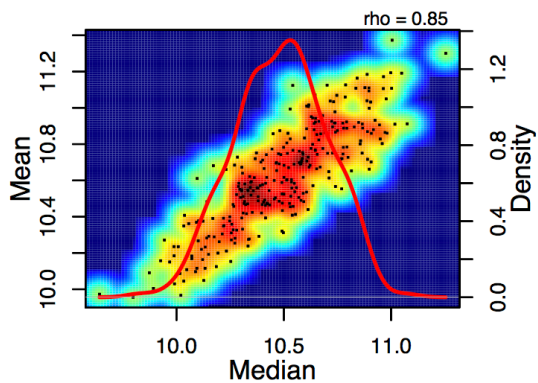


Self-correlation of the Buffa hypoxia signature in terms of mRNA expression across 13 cancer types. Correlation calculated as Spearman's rho. A highly self-correlated signature is a prerequisite for usability in a given data set. Full-size figures complete with dendrograms and gene names available in Figure S2. Figure generated by sigQC.

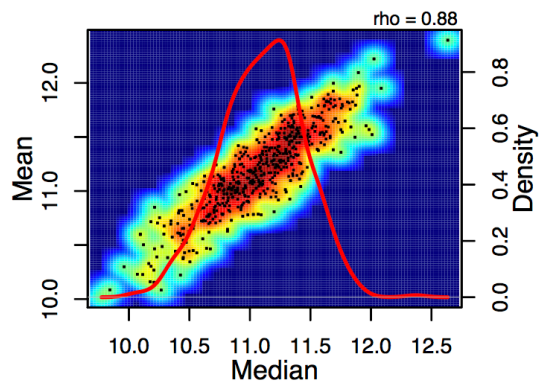
Figure 9: Correlation Between Mean, Median, and PC1 of the Expression of the Buffa Signature across 13 Cancer Types



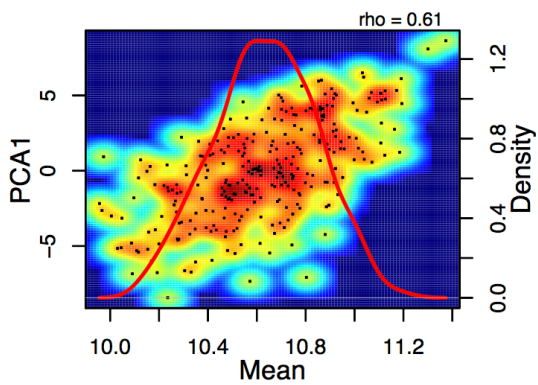
**COAD Buffa52
Median vs Mean**



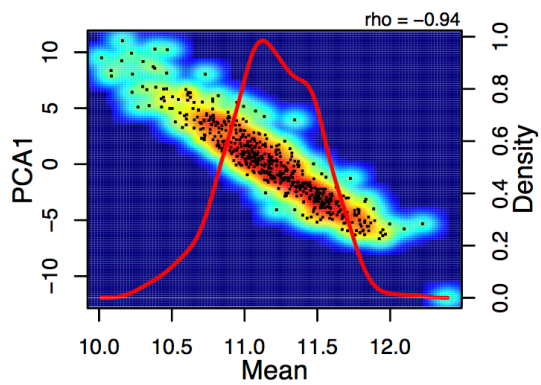
**HNSC Buffa52
Median vs Mean**



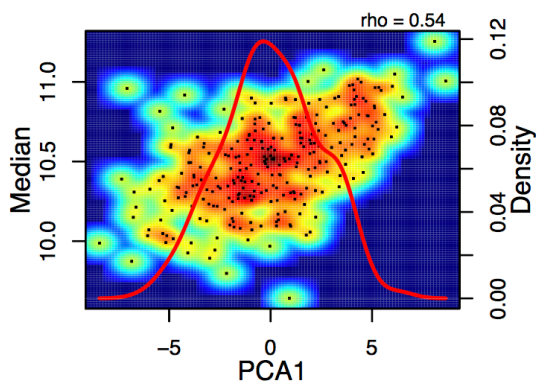
Mean vs PCA1



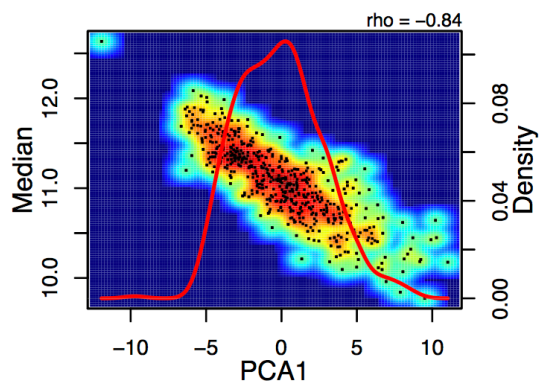
Mean vs PCA1



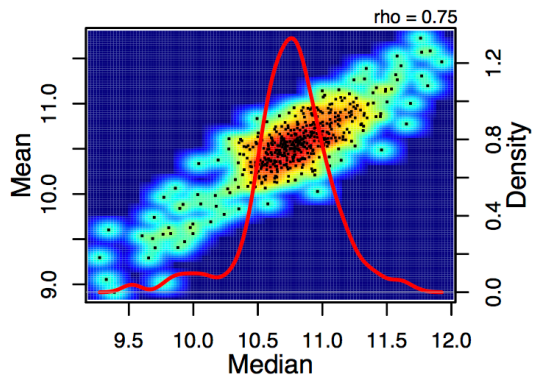
PCA1 vs Median



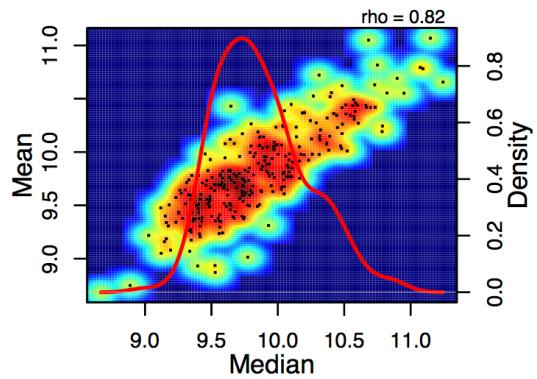
PCA1 vs Median



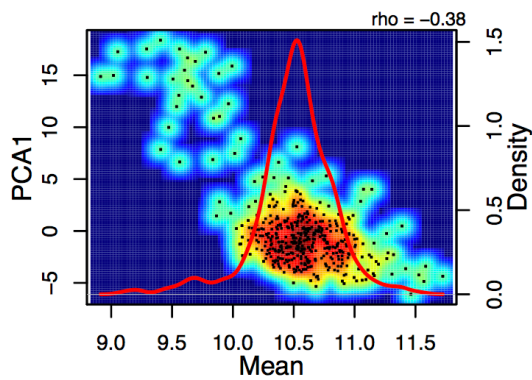
KIRC Buffa52
Median vs Mean



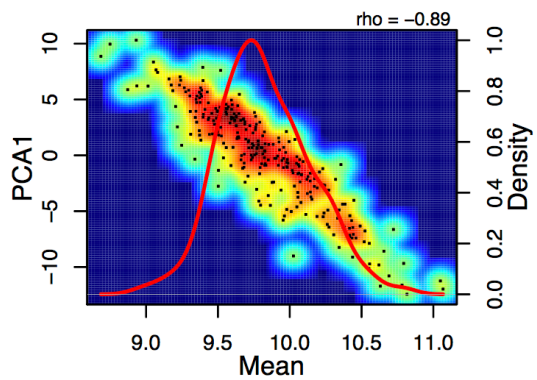
KIRP Buffa52
Median vs Mean



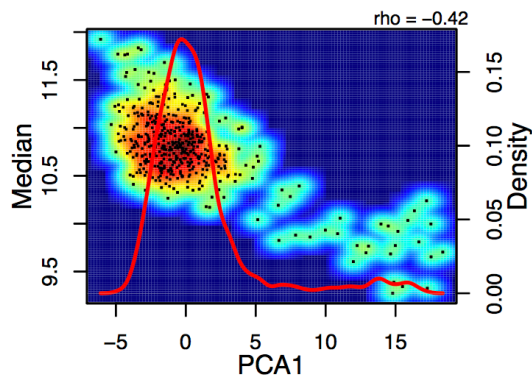
Mean vs PCA1



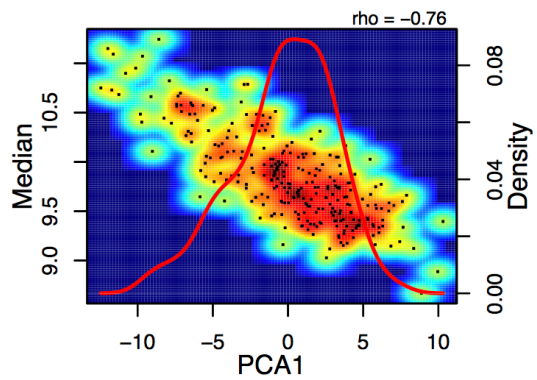
Mean vs PCA1



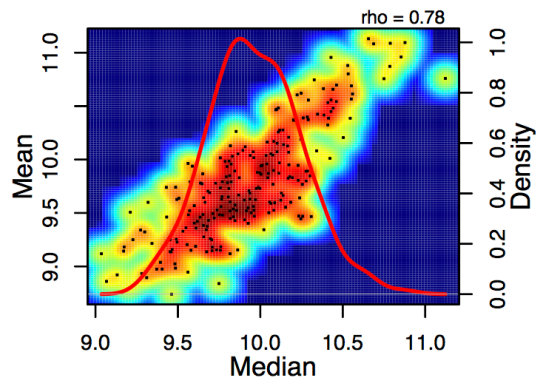
PCA1 vs Median



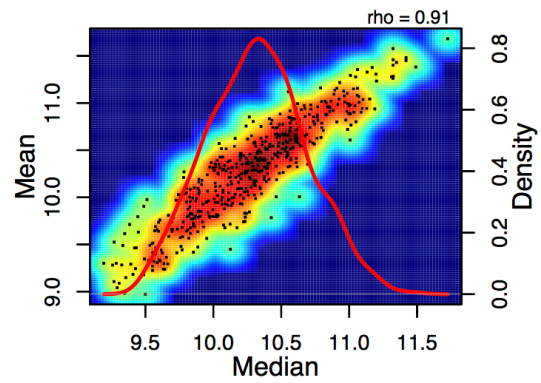
PCA1 vs Median



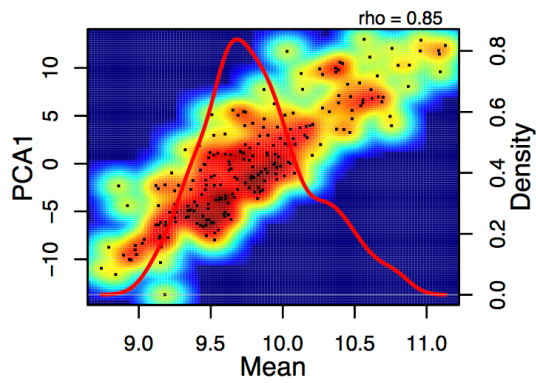
LIHC Buffa52
Median vs Mean



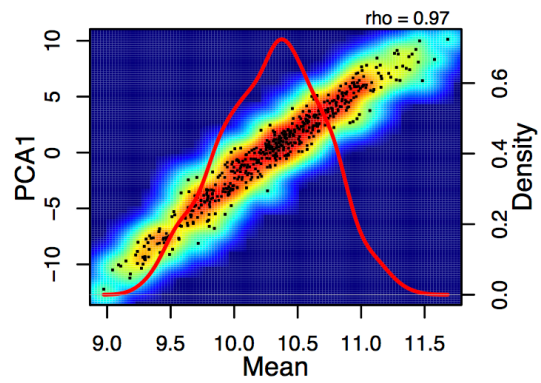
LUAD Buffa52
Median vs Mean



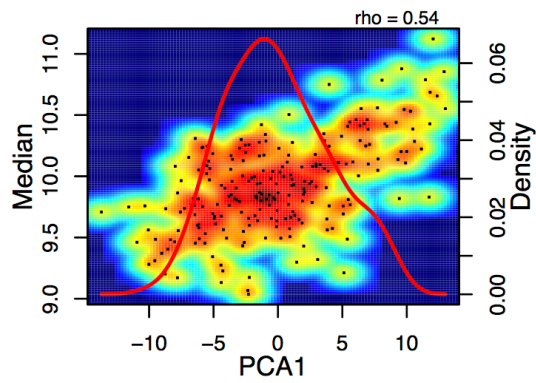
Mean vs PCA1



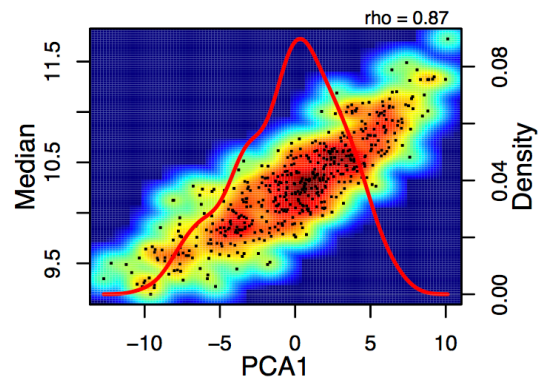
Mean vs PCA1



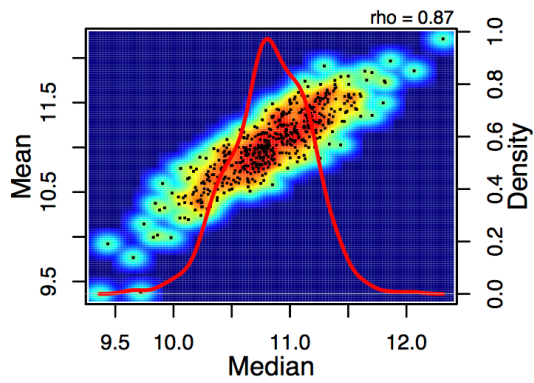
PCA1 vs Median



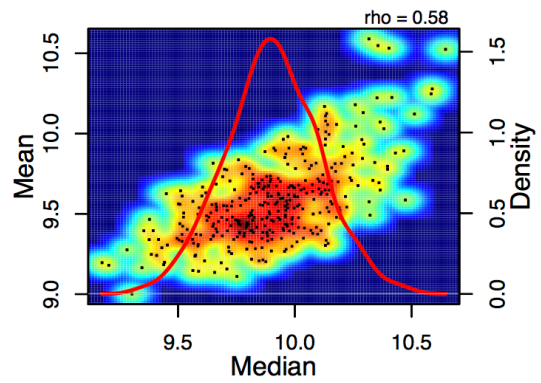
PCA1 vs Median



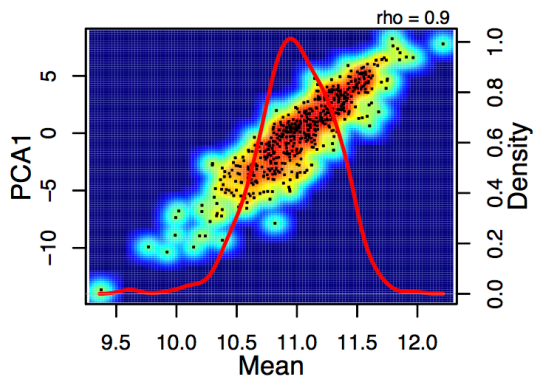
LUSC Buffa52
Median vs Mean



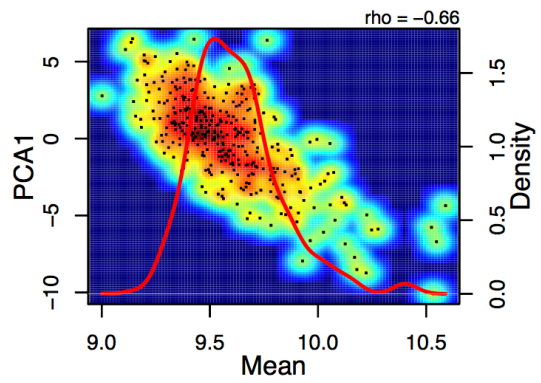
PRAD Buffa52
Median vs Mean



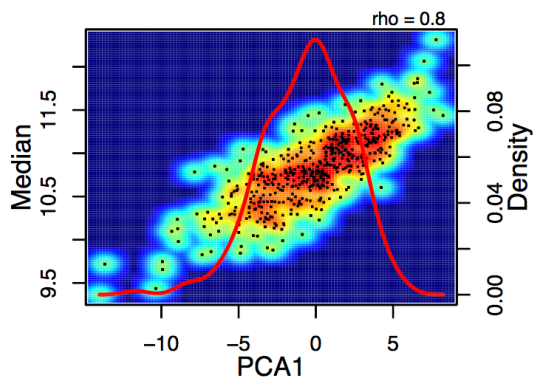
Mean vs PCA1



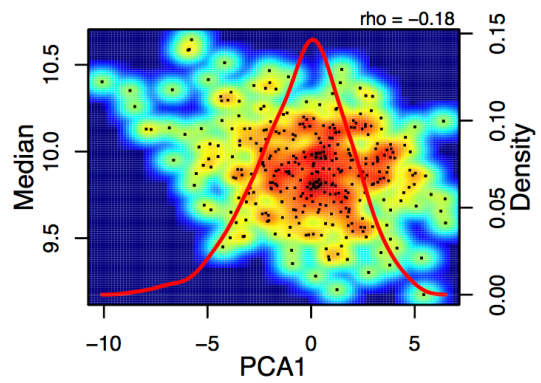
Mean vs PCA1



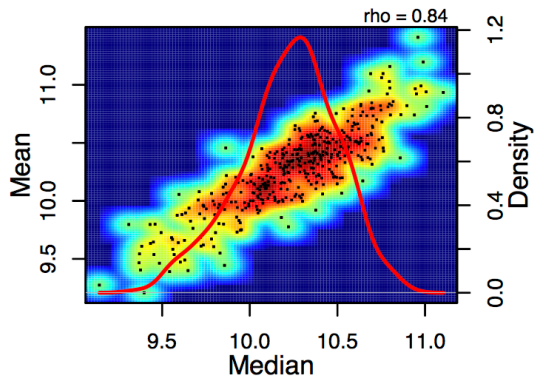
PCA1 vs Median



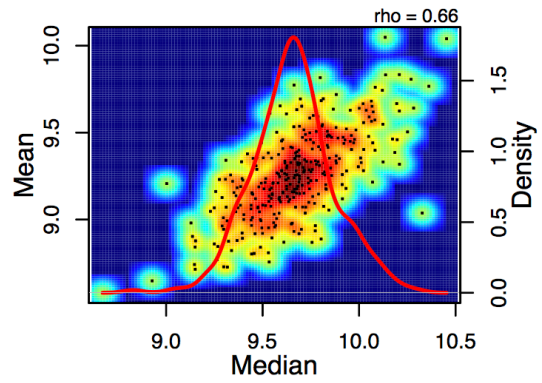
PCA1 vs Median



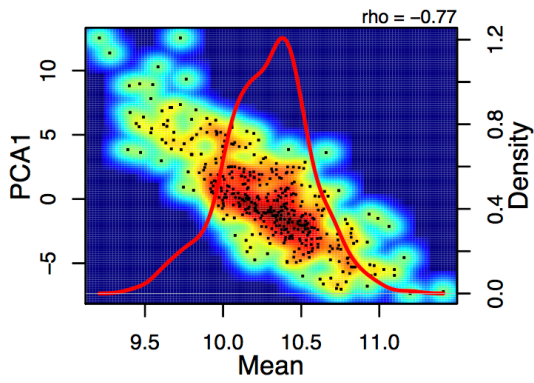
**STAD Buffa52
Median vs Mean**



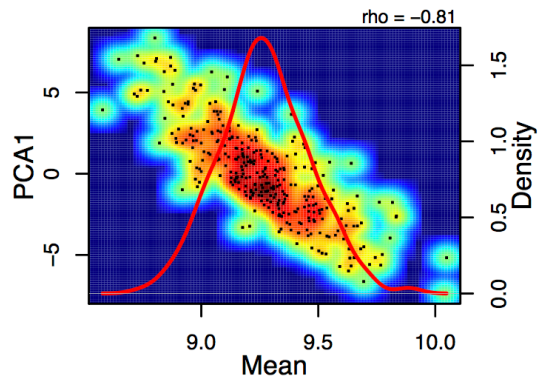
**THCA Buffa52
Median vs Mean**



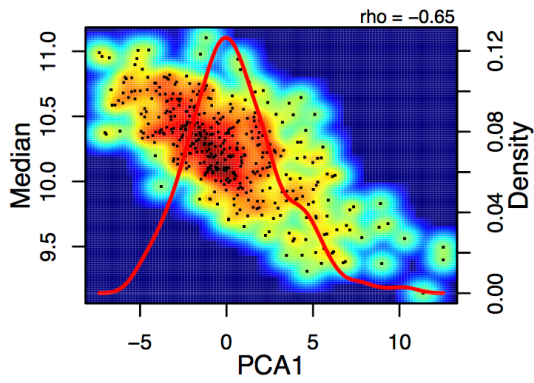
Mean vs PCA1



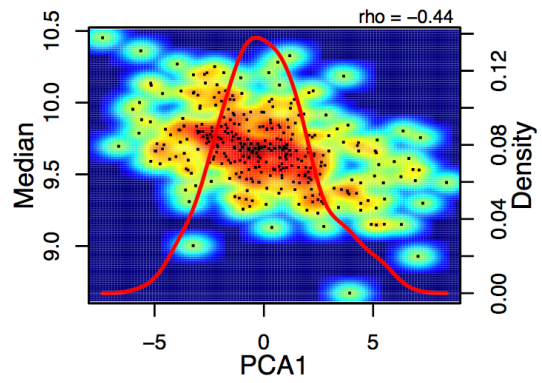
Mean vs PCA1

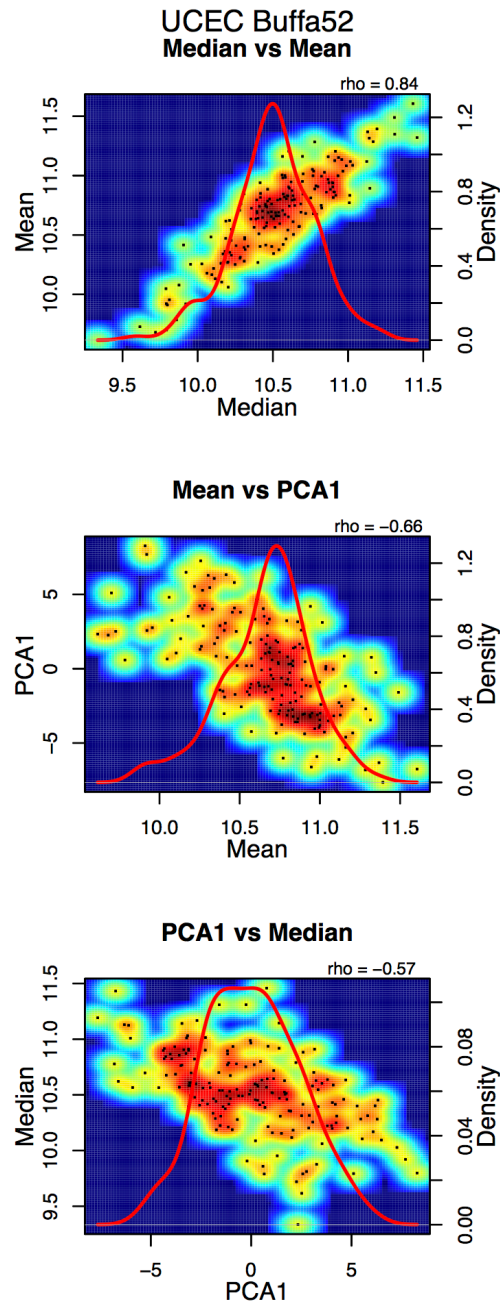


PCA1 vs Median



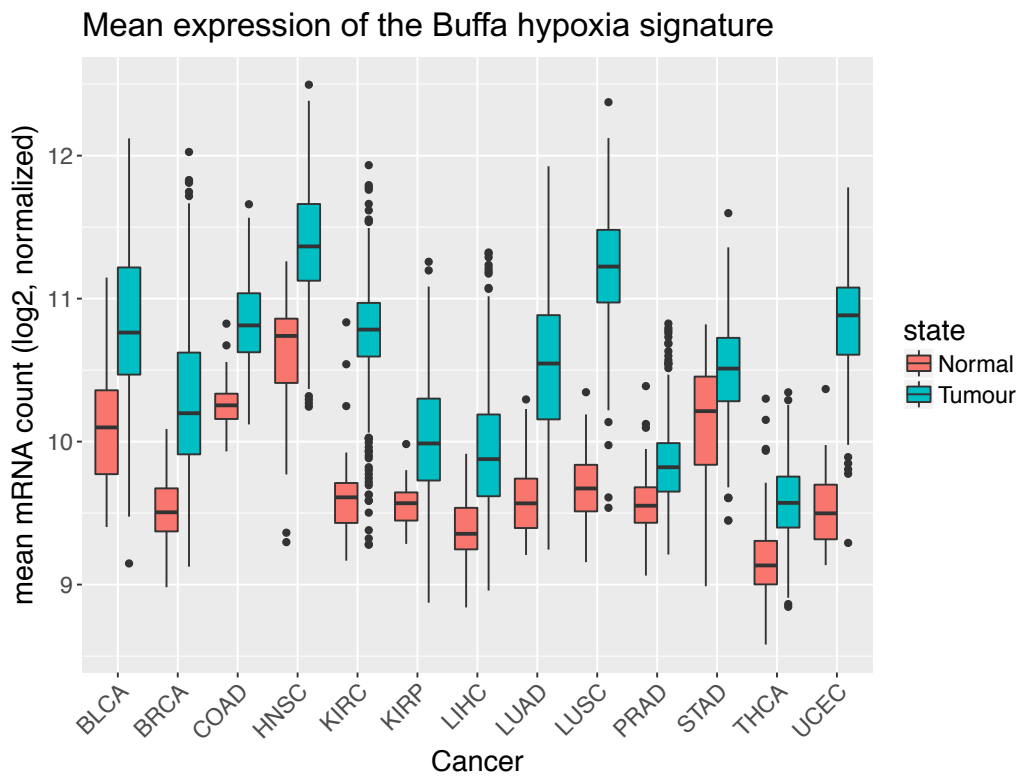
PCA1 vs Median





Relationship and correlation between mean and median, mean and PC1, and median and PC1 of the expression of the Buffa signature across 13 cancer types. Correlation calculated as Spearman's rho. A high degree of correlation between the mean, median, and PC1 of the expression of the signature is a pre-requisite for usability in a given data set. The red line represents the density of samples along the x axis. Figure generated by sigQC.

Figure 10: Mean expression of the Buffa et al. hypoxia signature across 13 cancer types



Mean expression of the Buffa (Buffa et al. 2010) hypoxia signature across 13 cancer types. Data is log 2 transformed, normalised mRNA count from TCGA. The mean of values for all Buffa signature genes in each sample is displayed. Boxes represent the range from the first to the third quartiles. Centre line represents the median. Whiskers represent 3rd quartile plus 1.5x the interquartile range or the maximal value, whichever is smaller, and the 1st quartile minus 1.5x the interquartile range or the minimal value, whichever is larger. Dots represent values not contained within whiskers.

3.4.2 *APOBEC3A and APOBEC3B Expression Correlates with the Hypoxia Signature*

Knowing that TCGA samples express the hypoxia signature genes, we then turn our attention to the association between hypoxia and APOBEC on the gene expression level. We calculate the Spearman's correlation between the mean hypoxia signature expression and the expression of every APOBEC gene, in all 13 cancer types. Spearman's rho is plotted as a heatmap (Figure 11). Results are filtered for $p < 0.05$.

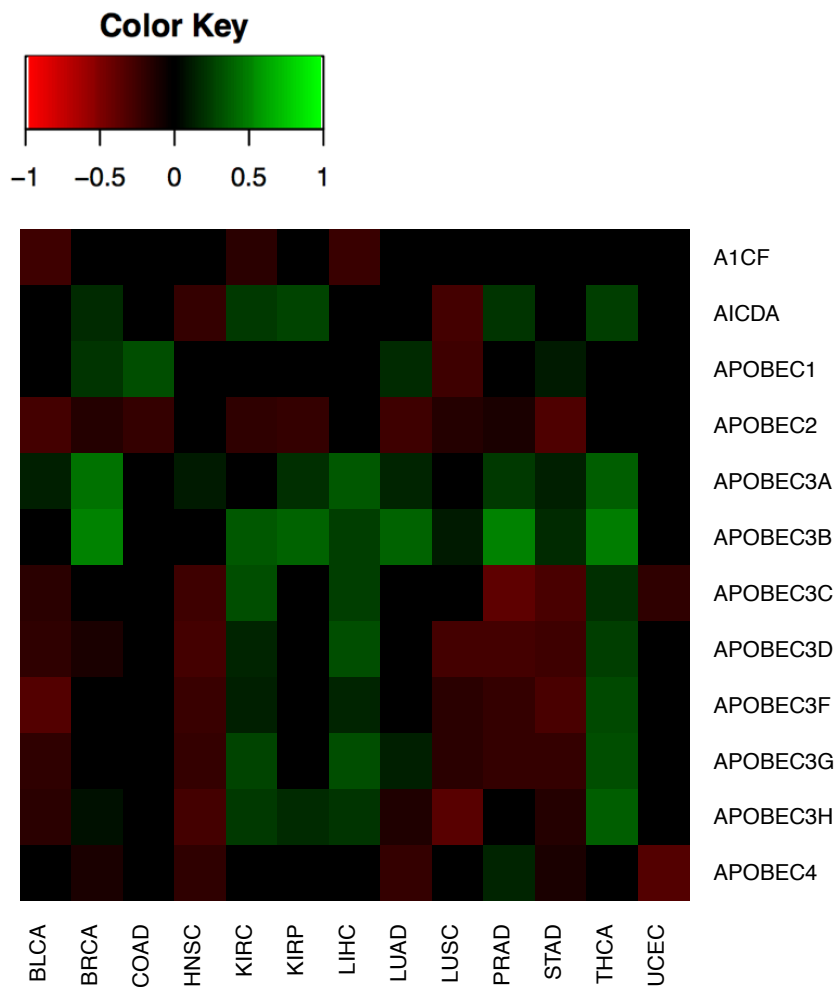
We observe that A3A and A3B are significantly correlated with the hypoxia signature in many cancers, including BRCA, KIRC, KIRP, LIHC, LUAD, PRAD, and THCA. The rest of the A3 family shows significant correlation with KIRC, LIHC, and THCA. As expected, APOBEC genes outside of the A3 family show limited correlation with hypoxia, if any, since they are not significantly expressed to begin with in the majority of cancer types. It is also expected that A3A and A3B are the most strongly correlated across cancers since they are the ones that are thought to be most involved in cancer. In the case of BRCA, where A3A and A3B activity has been widely reported, we see the strongest correlation with hypoxia. We illustrate the relationship between A3A and A3B expression and the expression of the mean hypoxia signature on a per-sample basis in BRCA using scatter plots and compare tumour versus normal samples (Figure 12). This visual representation shows the spread and cluster of samples that shape the relationship between the two, as we can visually see the correlation, especially in A3B. Here, we also see the differential expression of hypoxia, A3A, and A3B

in tumour samples, with the spread of the scatter shifted up and to the right for tumour. Since the Buffa signature has been validate in breast cancer, we can say that it reflects hypoxia in general. A3A and A3B expression thus correlates with hypoxia status in breast cancer.

We then perform the same analysis in BRCA on individual hypoxia genes (Figure 13). Again, we observe strong correlation between A3A and A3B and most genes in the hypoxia signature. This is performed to verify that the correlation with individual hypoxia genes does not follow a wide, spread-out distribution with large inconsistency and to identify any outlier gene in the signature.

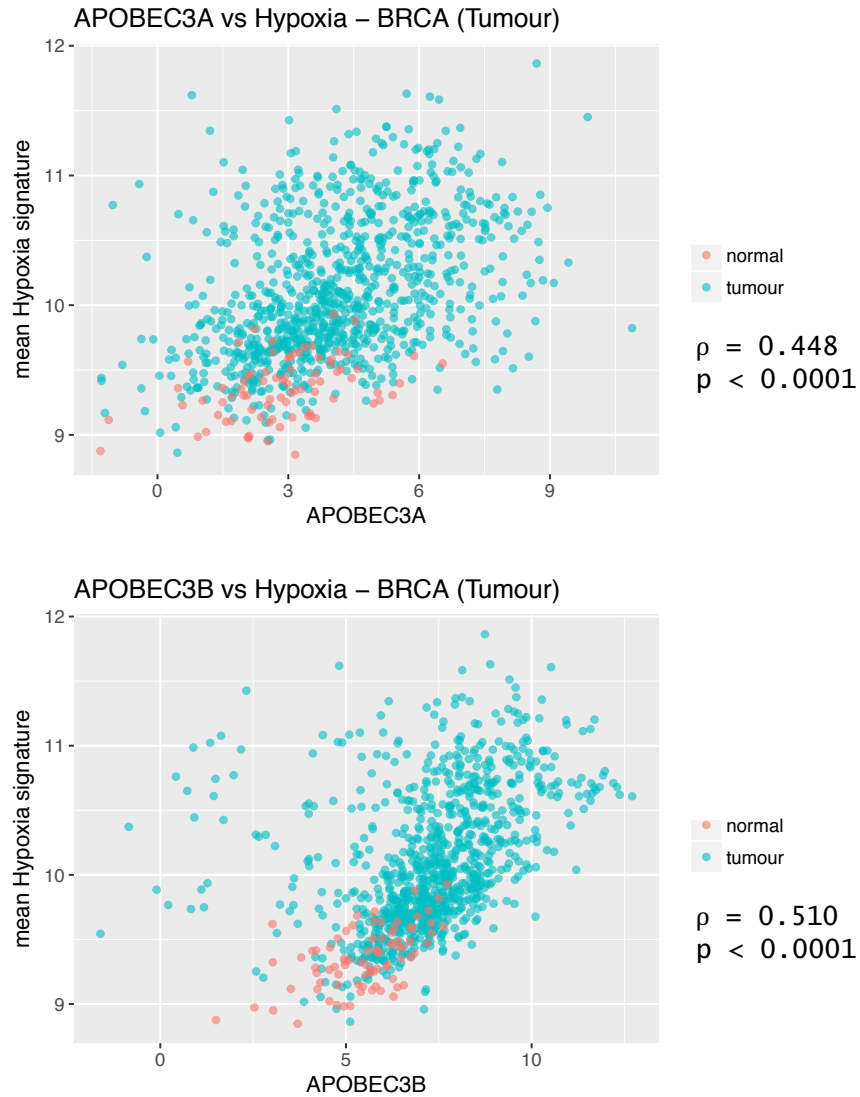
This provides preliminary in silico evidence for the role of hypoxia in APOBEC. A3A and A3B are the most active APOBEC genes in cancer. They are preferentially expressed in tumours and their expression correlates with hypoxia levels.

Figure 11: Correlation between the Expression of APOBEC Genes and the Mean Buffa Hypoxia Signature



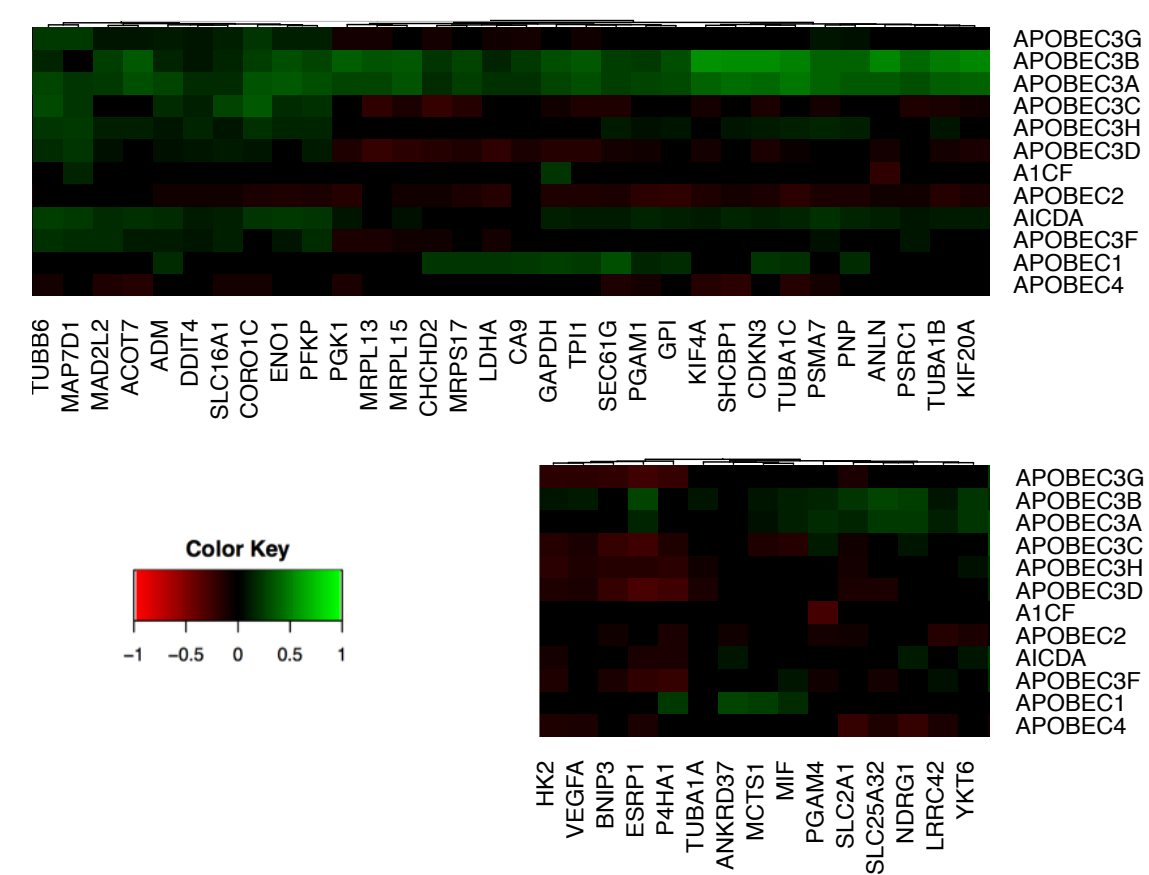
Correlation of APOBEC genes and the mean Buffa hypoxia signature. Spearman's correlation is calculated between the mean expression of the Buffa hypoxia signature and the expression of each APOBEC gene, in 13 cancer types. Values displayed are Spearman's rho, filtered for $p < 0.05$. Data from TCGA.

Figure 12: Relationship between the Expression of the Hypoxia Signature and APOBEC3A/3B in Breast Cancer



Correlation between the mean expression of the Buffa hypoxia signature and A3A/A3B in normal and tumour BRCA samples. Values represent log 2 transformed, normalised mRNA count from TCGA. Spearman's rho is calculated and displayed alongside p values

Figure 13: Correlation between Individual APOBEC Genes and Individual Eenes in the Buffa Hypoxia Signature in Breast Cancer



Correlation of APOBEC genes and the Buffa hypoxia signature genes in BRCA. Spearman's correlation is calculated between the expression of genes from the Buffa hypoxia signature and the expression of each APOBEC gene, in BRCA. Values displayed are Spearman's rho, filtered for $p < 0.05$. Data from TCGA.

3.4.3 A Principal Component Regression Model to Explain the Relationship between APOBEC and Hypoxia

We take a second approach to study the relationship between hypoxia and APOBEC: principal component regression. Previously, we used principal component analysis to obtain the components of highest variance in the expression of APOBEC genes. In contrast, now we use the previously generated principal components for a regression analysis. Here, a least-square linear regression model is fitted with the principal components of APOBEC expression as opposed to the APOBEC expression directly, such is the case in simple linear regression. The principal components have already been calculated in the first section. The main advantage of using principal component regression instead of a simple linear regression is the reduction of the multicollinearity effect. In essence, multicollinearity occurs when two or more of the regressed variables have a reasonable degree of correlation among themselves, which we see in the APOBEC 3 family, as we recall from the previous sections. This leads to the model being less accurate at predicting the involvement of individual variables in the regression. Since principal components are orthogonal to one another, multicollinearity is absent. In addition, principal component regression reduces the number of dimensions in the data by eliminating components with insignificant variation, and in turn reducing the number of parameters in the regression. As a whole, the exclusion of parameters with low variance leads to a more accurate regression model.

The result of the model is available in Table 6. We first need to assess the validity of the model. Looking at the p values of our model in BRCA, we see that all 6 principal components included are significant. Looking at the residuals distribution (Figure S4), we see that it is fairly normal with a slight tail in the positive end. Then, we look at the residuals Q-Q plot (Figure S4) and see that the data is fairly linear and suitable for a linear regression model. With this, we can conclude that the linear model applied is suitable for the experiment in BRCA. The same qualitative assessment of residuals is done for the other cancer types, and we reach the same conclusion that the linear model is suitable (Figure S4.)

We observe that PC1 dominates as the most important parameter in many cancers. This is unsurprising, as PC1 is explained by many APOBEC genes because of self-correlation in the APOBEC 3 family. In BRCA, PC2 also stands out above the rest since there is a clearer separation between PC1 and PC2, with PC1 being explained by A3A and A3H, while PC2 is explained by A3B. We turn our attention to BRCA both because it is where APOBEC has been the most reported and because the linear model returned significantly large coefficients for multiple principle components. In BRCA, the model shows that PC1 is the most important variable, which means that a combination of A3A, A3H, and other A3 genes explaining the largest variance in APOBEC expression is most accurate at predicting hypoxia, followed closely by a component explained solely by A3B. The distinction of which, of A3A, A3H, or A3B, is the best at predicting hypoxia is not clear, since PC1 is explained by all three, but we can confidently say that a combination of these three genes is an accurate predictor of hypoxia in BRCA. Referring back to the Spearman's correlation calculated previously, we see that

it concords with our current result and suggests that A3H may not be the gene driving PC1 to be an accurate predictor, but rather A3A and A3B, since A3H is not significantly correlated with hypoxia on its own. Overall, we conclude from the principal component regression model that again, A3A and A3B are the best at predicting hypoxia in BRCA, but as for the other cancer types, the result is unclear, as PC1 encompasses too many genes and is the only good predictor according to the model.

Overall, there is sufficient evidence to conclude that the expression of active APOBEC genes is likely to be correlated with hypoxia in multiple cancers, and is especially strong in BRCA, leading us to believe that hypoxia may be driving APOBEC expression, either by selection or active regulation. From here, we look beyond correlational evidence for signs of hypoxia involvement in APOBEC activity.

**Table 6: Principal Component Regression of the Expression of APOBEC3
Genes in Relation to the Buffa Hypoxia Signature**

BLCA

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	10.7877	0.1035	104.267	< 2e-16	***
PC1	0.5640	2.0748	0.272	0.78588	
PC2	-3.0581	0.4714	-6.488	2.57e-10	***
PC3	-0.8618	0.4916	-1.753	0.08037	.
PC4	1.5631	0.4979	3.139	0.00182	**
PC5	-0.2267	0.4776	-0.475	0.63531	
PC6	-0.6829	0.4717	-1.448	0.14847	

BRCA

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	10.6146	0.0704	150.77	< 2e-16	***
PC1	-11.2614	2.3012	-4.89	1.1e-06	***
PC2	-7.9237	0.4247	-18.66	< 2e-16	***
PC3	-1.4237	0.4621	-3.08	0.00212	**
PC4	1.7805	0.4655	3.83	0.00014	***
PC5	-1.0354	0.4739	-2.18	0.02911	*
PC6	-1.4939	0.4261	-3.51	0.00047	***

COAD

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	10.8652	0.0961	113.01	<2e-16	***
PC1	-0.8144	1.6098	-0.51	0.613	
PC2	-0.3729	0.3038	-1.23	0.221	
PC3	-0.2167	0.2995	-0.72	0.470	
PC4	0.6192	0.3448	1.80	0.074	.
PC5	0.1462	0.3046	0.48	0.632	
PC6	-0.0792	0.2974	-0.27	0.790	

HNSC

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	11.0101	0.0834	132.06	< 2e-16	***
PC1	-8.4345	1.8752	-4.50	8.5e-06	***
PC2	0.8697	0.4137	2.10	0.0360	*
PC3	0.1706	0.3888	0.44	0.6610	
PC4	1.1920	0.3913	3.05	0.0024	**
PC5	-0.5743	0.3731	-1.54	0.1244	
PC6	0.3993	0.3837	1.04	0.2985	

KIRC

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	10.7204	0.0987	108.64	< 2e-16	***
PC1	1.1917	2.2608	0.53	0.5983	
PC2	-1.2233	0.3850	-3.18	0.0016	**
PC3	-0.7314	0.3426	-2.13	0.0332	*
PC4	-2.7820	0.3395	-8.20	1.9e-15	***
PC5	1.5562	0.3844	4.05	5.9e-05	***
PC6	-0.4953	0.3354	-1.48	0.1403	

KIRP

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	10.430	0.122	85.29	< 2e-16	***
PC1	-7.050	2.073	-3.40	0.00077	***
PC2	1.490	0.382	3.90	0.00012	***
PC3	-1.093	0.381	-2.87	0.00446	**
PC4	-0.928	0.387	-2.40	0.01695	*
PC5	1.058	0.420	2.52	0.01231	*
PC6	-1.230	0.381	-3.23	0.00139	**

LIHC

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	10.1375	0.1266	80.09	< 2e-16 ***
PC1	-3.9558	2.4131	-1.64	0.10201
PC2	0.0426	0.4537	0.09	0.92530
PC3	-1.2818	0.4533	-2.83	0.00494 **
PC4	0.5103	0.4493	1.14	0.25685
PC5	1.7777	0.4814	3.69	0.00026 ***
PC6	-0.0116	0.5352	-0.02	0.98272

LUAD

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	10.038	0.102	98.12	< 2e-16 ***
PC1	10.919	2.300	4.75	2.7e-06 ***
PC2	-3.835	0.439	-8.74	< 2e-16 ***
PC3	0.362	0.456	0.79	0.43
PC4	3.334	0.495	6.74	4.4e-11 ***
PC5	-3.562	0.443	-8.05	6.0e-15 ***
PC6	-0.310	0.449	-0.69	0.49

LUSC

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	10.6783	0.0814	131.11	< 2e-16 ***
PC1	12.3210	1.8102	6.81	2.9e-11 ***
PC2	-1.9364	0.3557	-5.44	8.3e-08 ***
PC3	0.4469	0.3735	1.20	0.2321
PC4	0.6802	0.3573	1.90	0.0575 .
PC5	-0.9705	0.3658	-2.65	0.0082 **
PC6	0.5734	0.3669	1.56	0.1187

PRAD

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	10.1229	0.0601	168.45	< 2e-16	***
PC1	-6.7058	1.3233	-5.07	5.7e-07	***
PC2	-2.2948	0.2527	-9.08	< 2e-16	***
PC3	1.1580	0.2235	5.18	3.2e-07	***
PC4	-0.4532	0.2231	-2.03	0.04270	*
PC5	0.0111	0.2398	0.05	0.96304	
PC6	0.8878	0.2667	3.33	0.00094	***

STAD

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	10.6792	0.0949	112.55	< 2e-16	***
PC1	-3.9562	1.9136	-2.07	0.039	*
PC2	-2.1995	0.3215	-6.84	2.9e-11	***
PC3	0.1793	0.3494	0.51	0.608	
PC4	0.9819	0.3891	2.52	0.012	*
PC5	-0.4730	0.3308	-1.43	0.154	
PC6	-0.0867	0.3244	-0.27	0.789	

THCA

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	10.0757	0.0643	156.70	< 2e-16	***
PC1	-11.5657	1.4278	-8.10	4.3e-15	***
PC2	-0.0243	0.2435	-0.10	0.92052	
PC3	-0.8006	0.2343	-3.42	0.00068	***
PC4	-0.7631	0.2543	-3.00	0.00282	**
PC5	-0.1854	0.2502	-0.74	0.45900	
PC6	-0.8356	0.2500	-3.34	0.00089	***

UCEC

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	10.764	0.118	91.10	<2e-16	***
PC1	1.042	1.556	0.67	0.5038	
PC2	-1.253	0.391	-3.20	0.0016	**
PC3	-0.251	0.366	-0.68	0.4945	
PC4	0.890	0.364	2.45	0.0154	*
PC5	0.890	0.367	2.42	0.0165	*
PC6	0.257	0.378	0.68	0.4962	

Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Principal components are taken from Table 4 and used as coefficients in a linear regression model with the mean expression of the Buffa hypoxia signature. Coefficient estimates are displayed, along with standard error, t-values, p-values, and significance.

3.5 Immune Infiltration Has a Negligible Effect on APOBEC Presence

A concern one might have when discussing gene expression results arising from solid tumour samples is the purity of the latter. Tumour samples are never formed exclusively of tumour cells and will always contain some degree of cellular impurity, as small as it may be. The infiltration of non-tumour cells must thus be minimised to avoid infiltrations confounding the results. Considering that APOBEC genes are of immune origin, it is especially crucial to ensure that immune cell infiltration, a common occurrence in tumours, does not contaminate the expression and thus confound our analysis. Here, we address this issue by examining the effect that the immune infiltration present in the TCGA samples has on our analysis, and we show that this infiltration is present but does not change significantly our results.

To obtain data about immune infiltration, we use xCell, a recently developed method for in silico dissection of samples (Aran et al., 2017). xCell uses a combination of deconvolution and enrichment analysis techniques to infer the presence of infiltration in the cancer samples starting from the gene signatures of 48 distinct immune cell types. For each sample analysed, xCell assigns a score for the strength of association between the expression profile of the clinical sample, and the expression profile of each of the 48 signatures. In all, xCell scores samples in a cohort for the similarity between its gene expression profile and 48 signatures corresponding to immune infiltrates. It thus infers the extent of infiltration for 48 cell types independently.

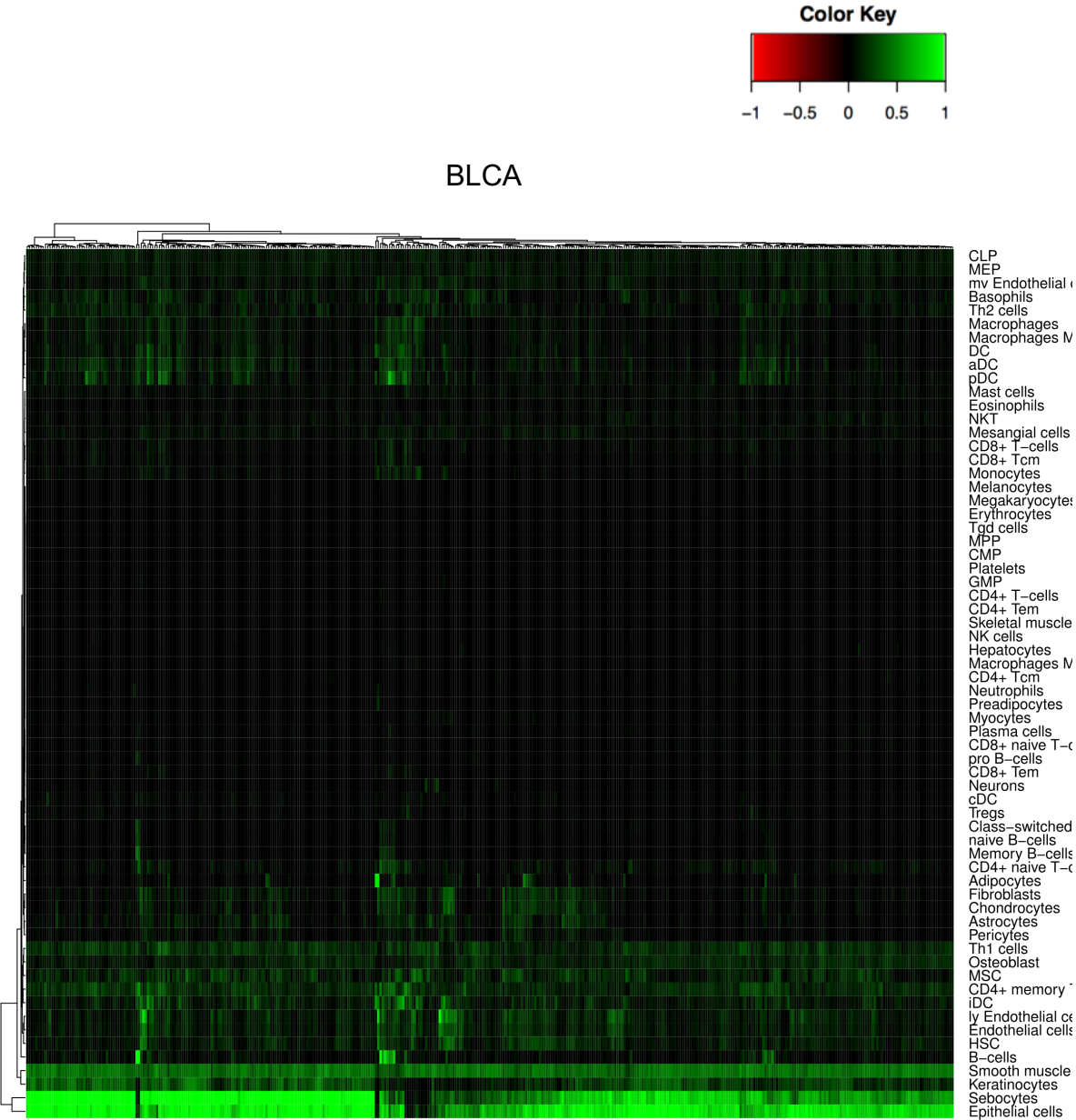
xCell scores for TCGA data have been pre-computed by the xCell team. We acquired the scores directly from the xCell platform and plot them in heatmaps (Figure 14). Upon visual inspection, we notice that few signatures are present in TCGA samples, and few samples have a high presence of multiple signatures. This is expected, as TCGA selects its samples for immune infiltration and contamination, and the published samples are all curated to be reliable for in silico analysis. Nevertheless, some extent of infiltration is present according to xCell, and we correct for this.

Not all signatures are present in a high enough level for us to consider it a true signal instead of noise. To select the relevant signatures from the 48, we set two thresholds: the signature must be present and score more than 0.2 in at least 20% of samples. The thresholds were chosen to filter out signatures for which too few samples are available, rendering the statistical power too low. 0.2 and 20% were chosen after empirical testing. Cell types that are known to not be present in a certain tissue score below 0.2 in general, for example hepatocytes in breast cancer (Figure 14). Conversely, cell types known to be present in a tissue score well above this threshold, for example sebocytes in head and neck cancer (Figure 14). The signatures are selected as such for each cancer type.

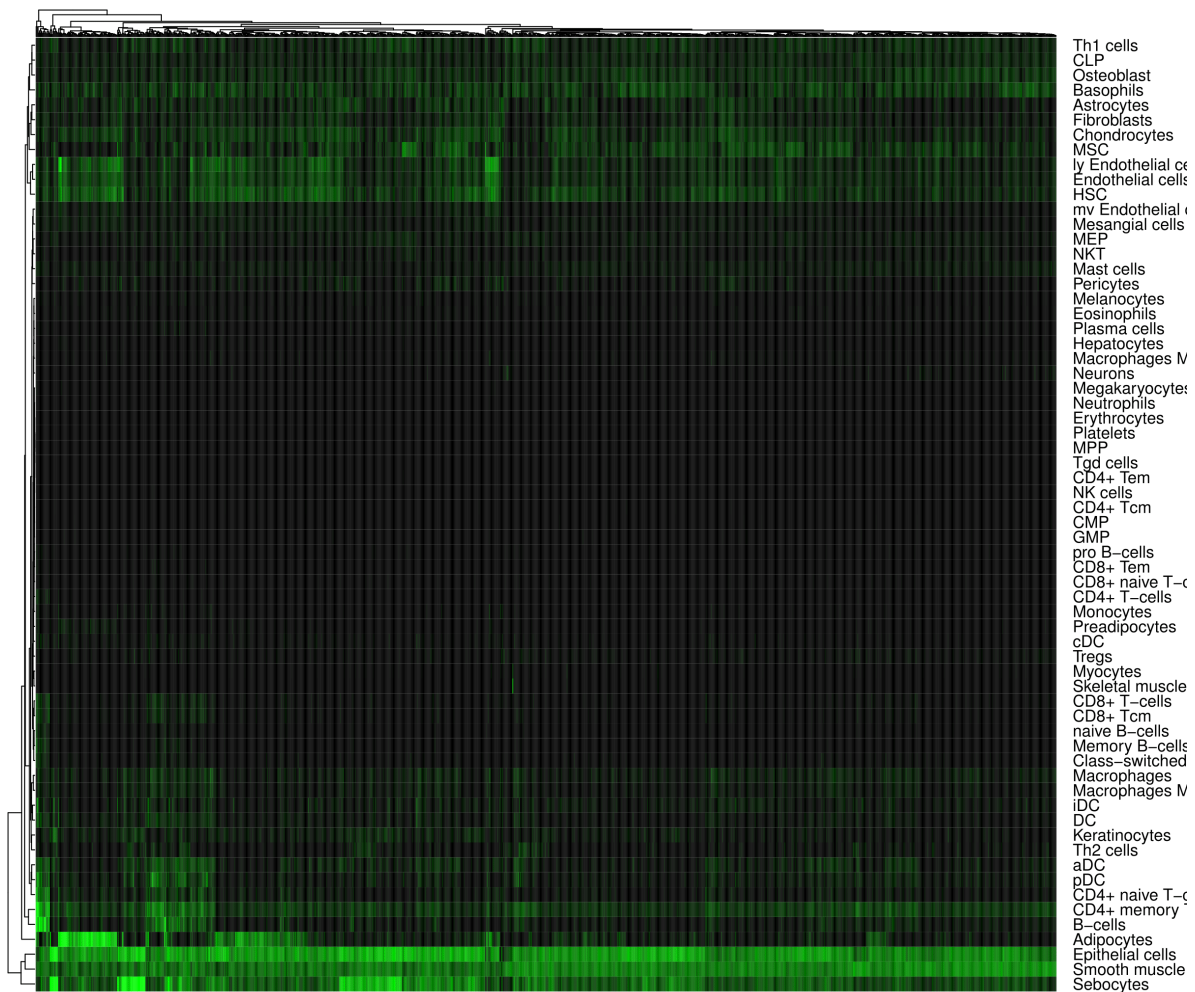
The correlation between the mean expression of the Buffa hypoxia signature and each APOBEC gene is calculated again, using Spearman's correlation partial to all the relevant signatures for each cancer type. Partial correlation calculates the correlation between two random variables while controlling for a third. This effectively eliminates the bias created when the third variable may be correlated

with both tested variables. Partial correlation disregards the component of the correlation between two variables that is due to a third variable. The results are filtered for p values < 0.05 and displayed as a heatmap in Figure 15. The difference between the partial correlation result and the original result from Figure 15 is shown in Figure S4. As we observe, the difference in Spearman's correlation score is minimal, with only isolated cases showing a small difference. Overall, the result from the partial correlation tells the same story as our original impartial calculation. We can thus conclude that the low level of immune infiltration in TCGA data has some, but overall negligible effects on our analysis and the conclusions drawn from them.

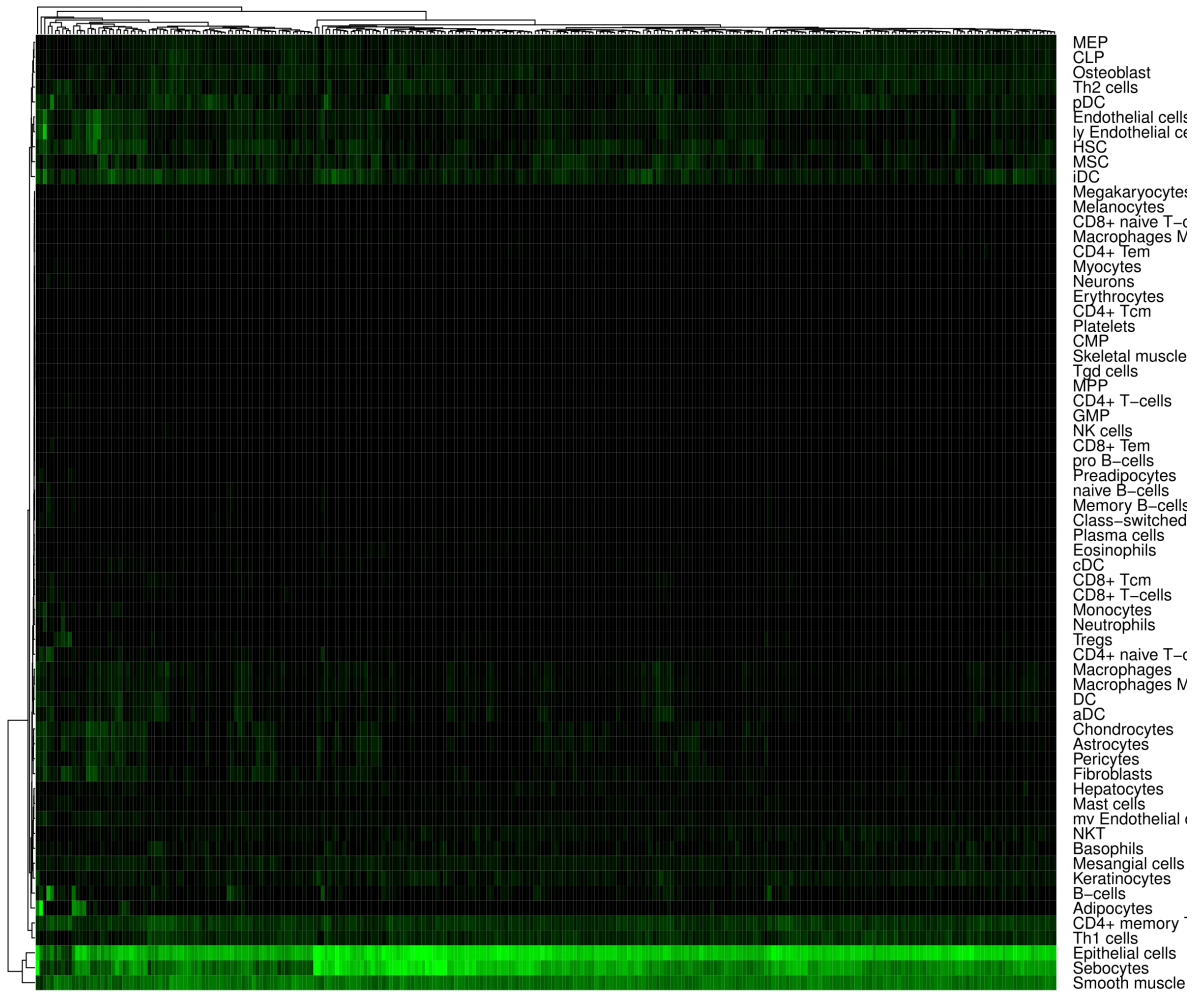
Figure 14: xCell Scores for TCGA Samples of 13 Cancer Types



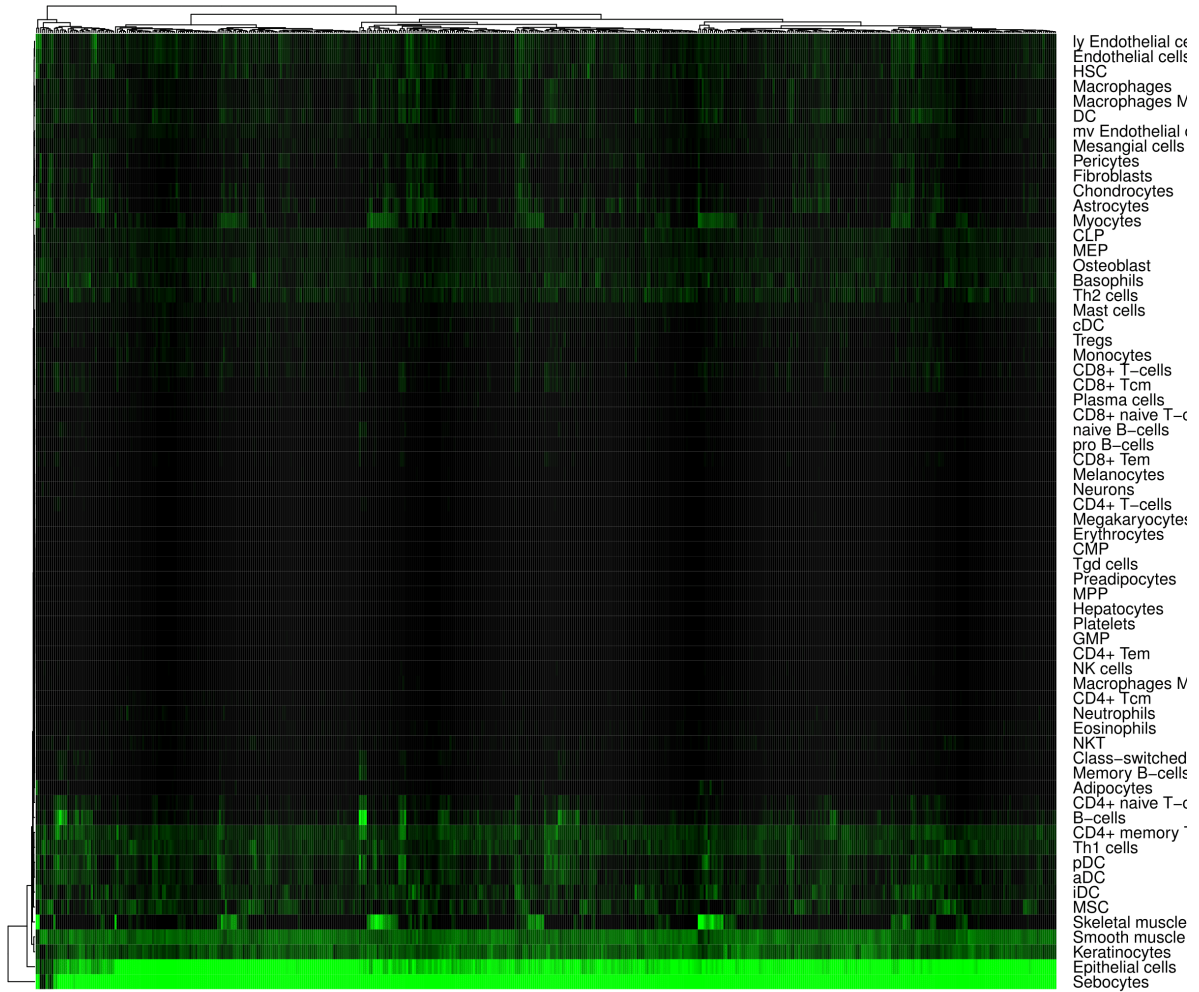
BRCA



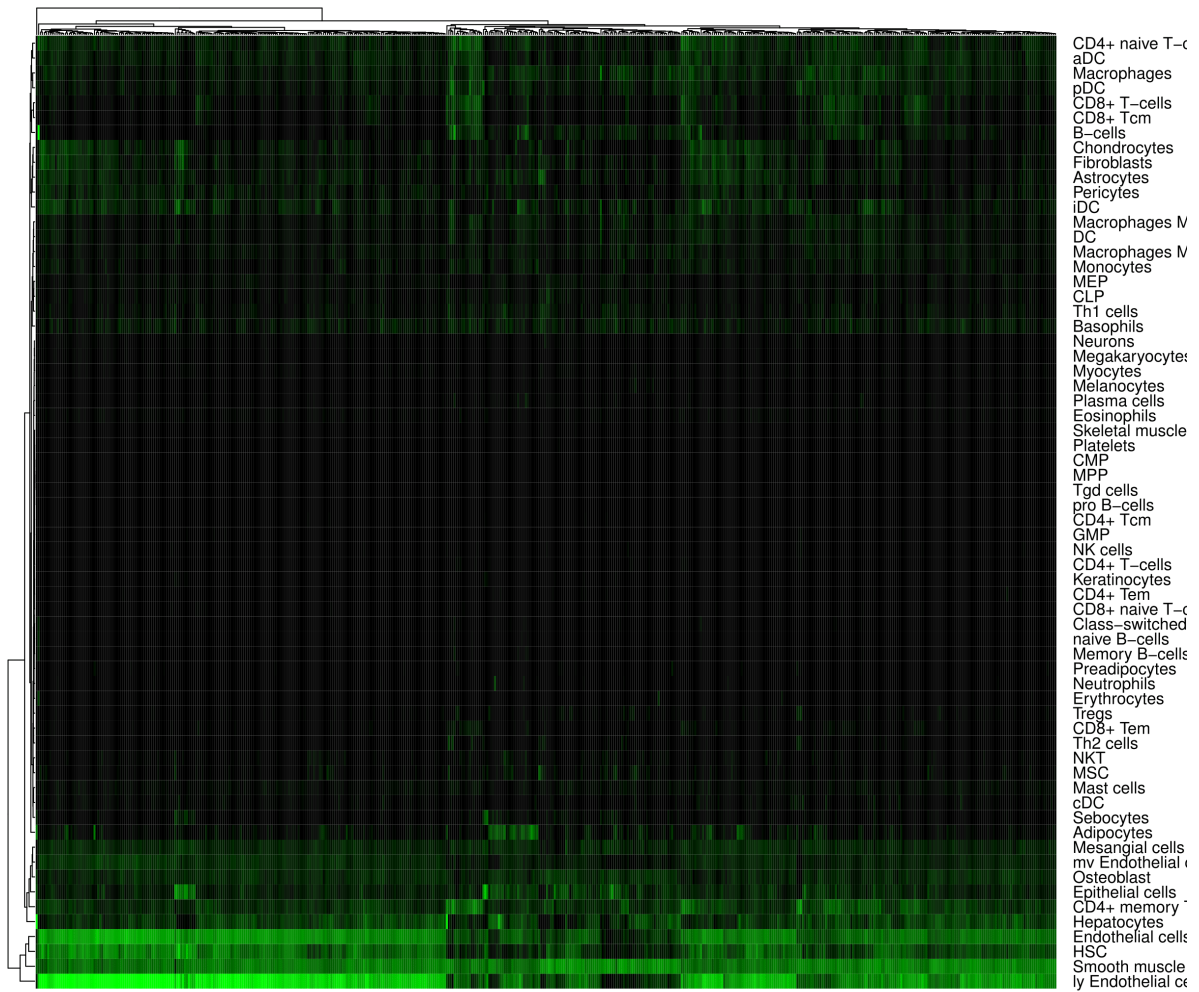
COAD



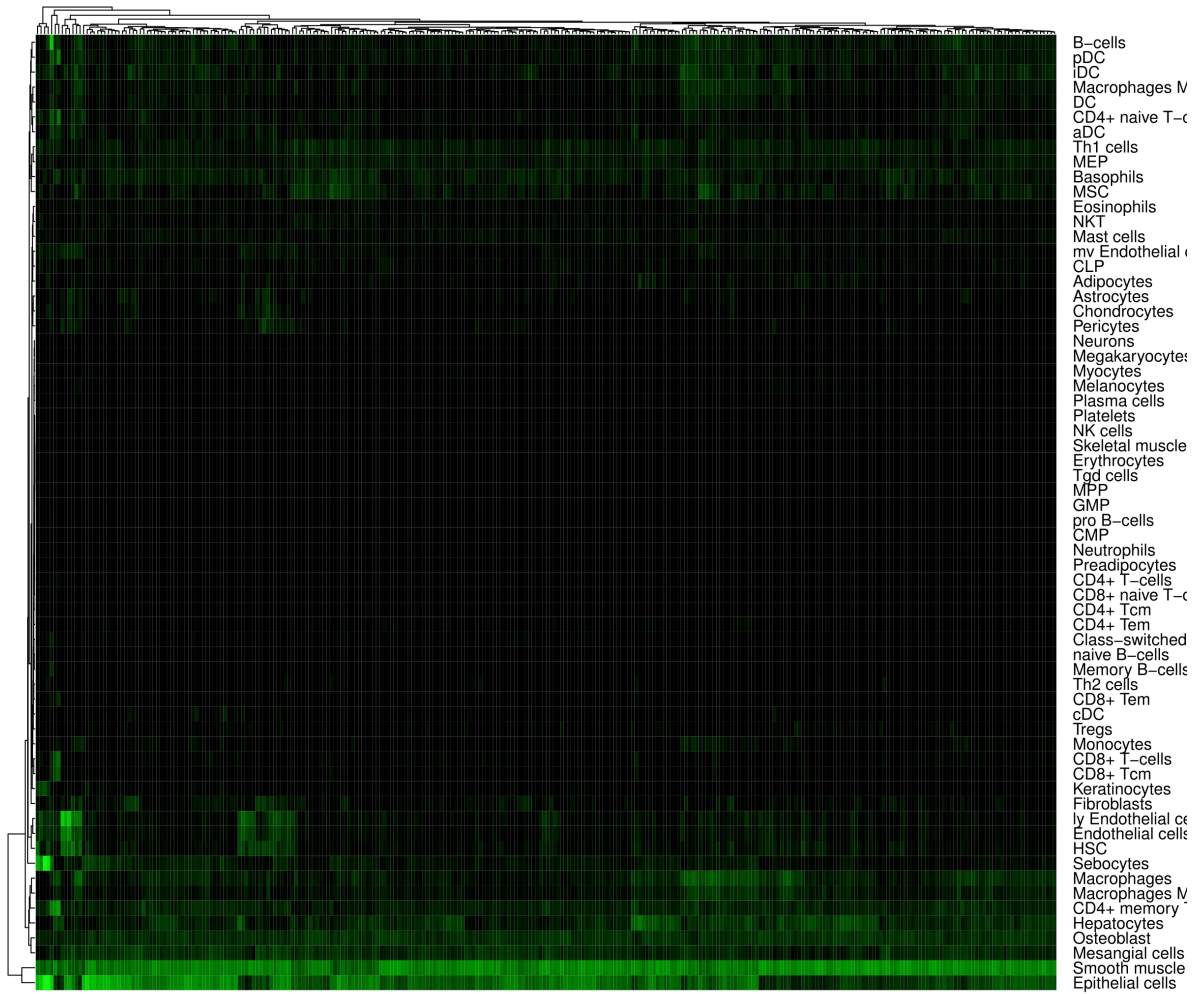
HNSC



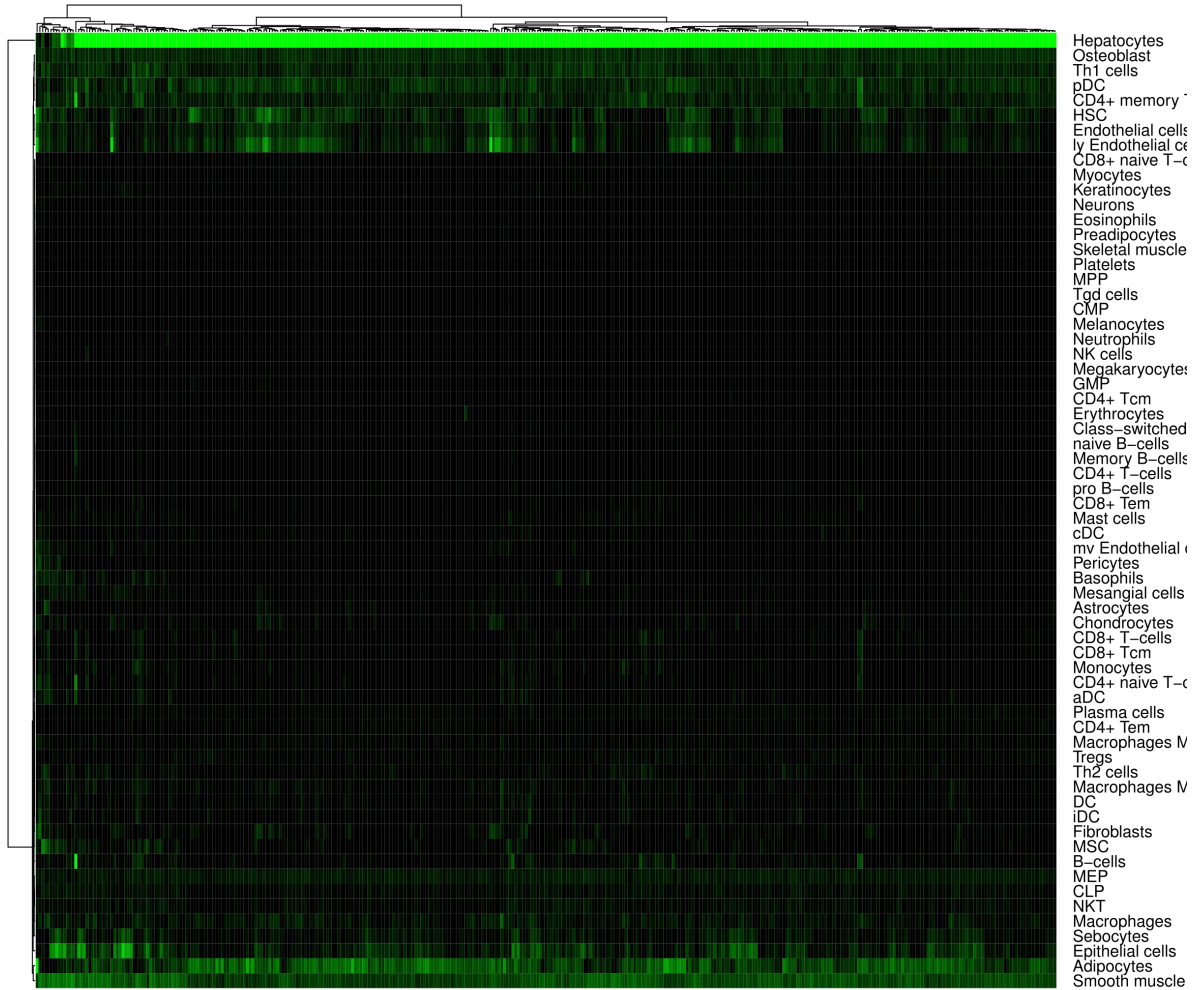
KIRC



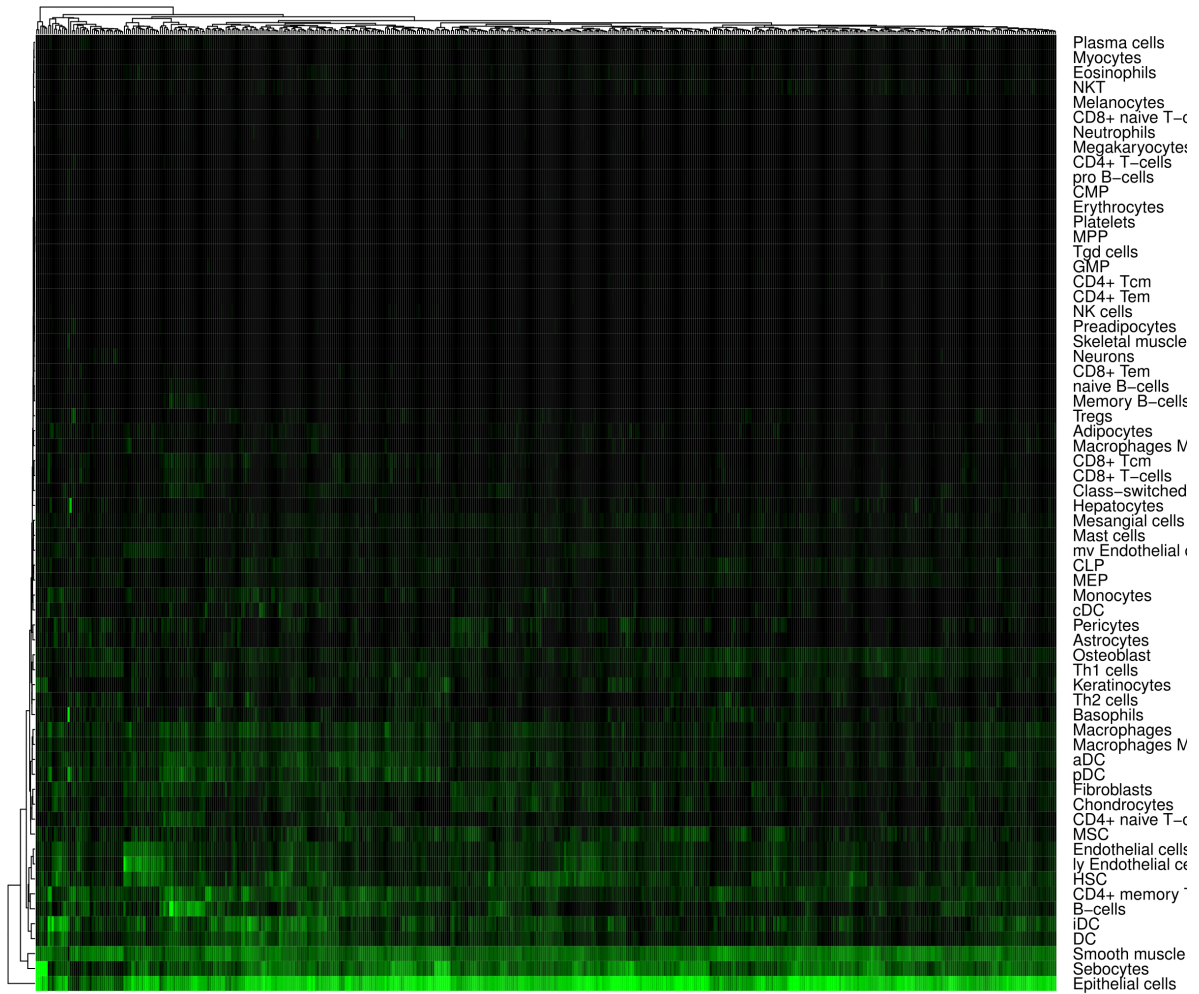
KIRP



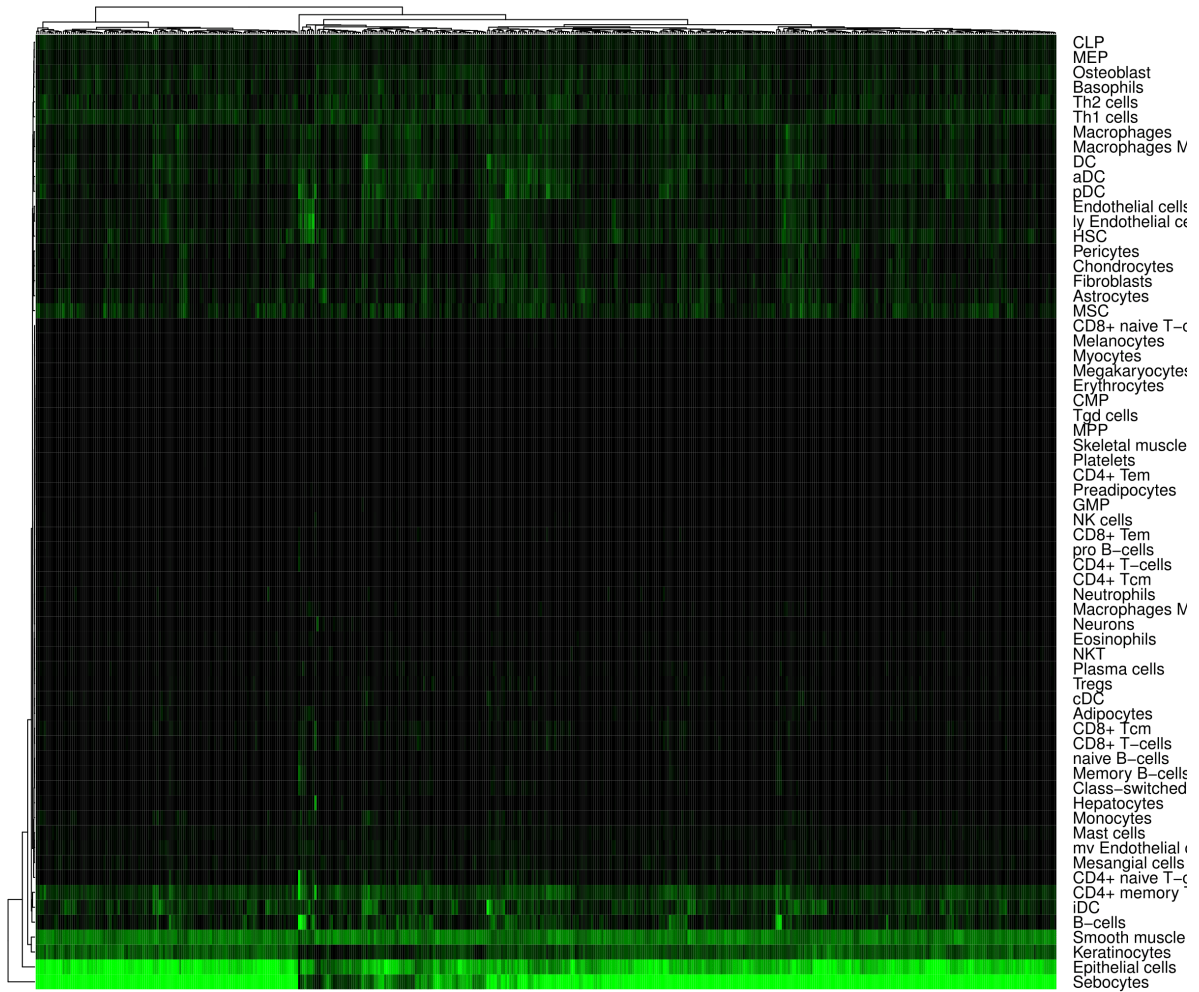
LIHC



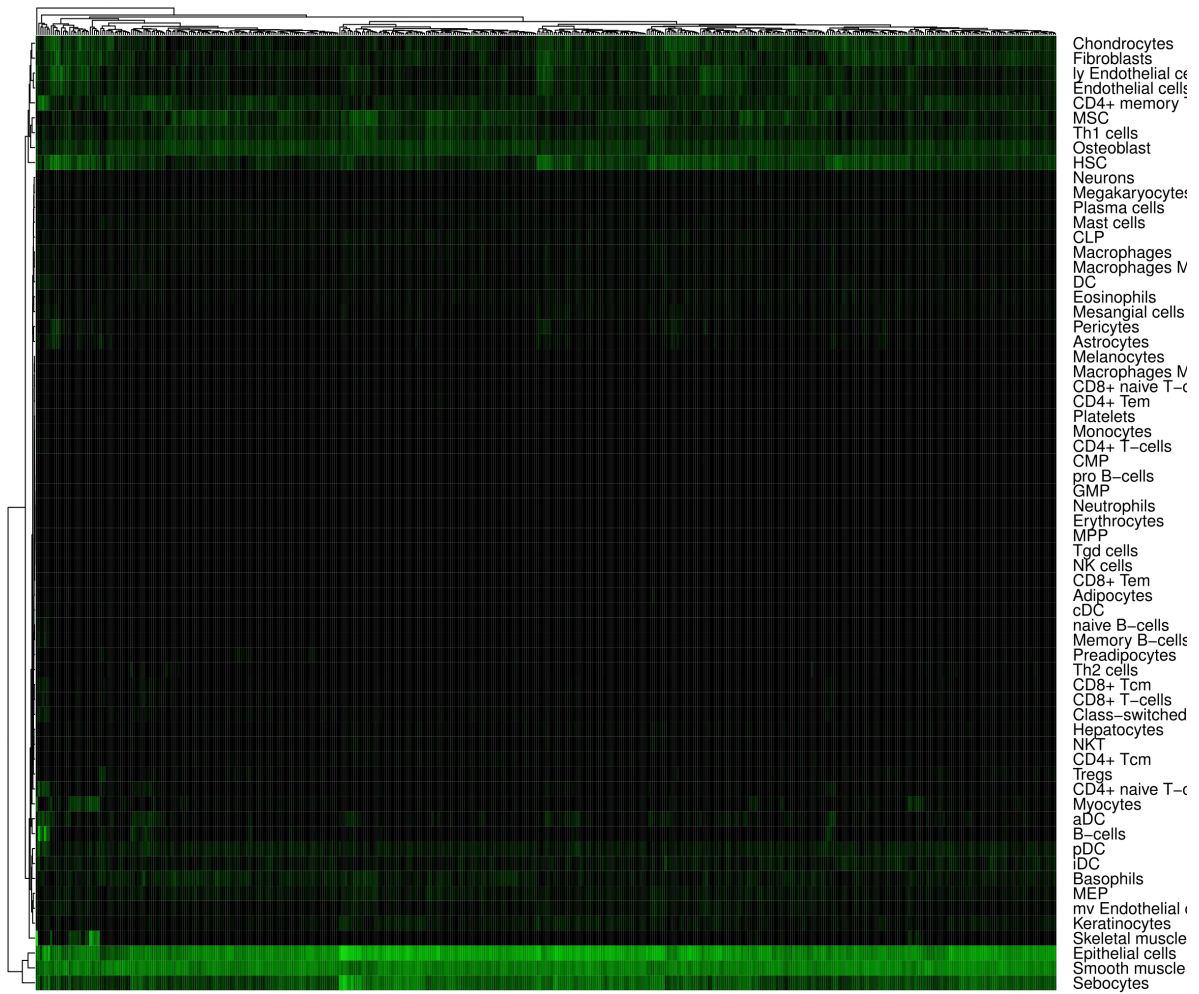
LUAD



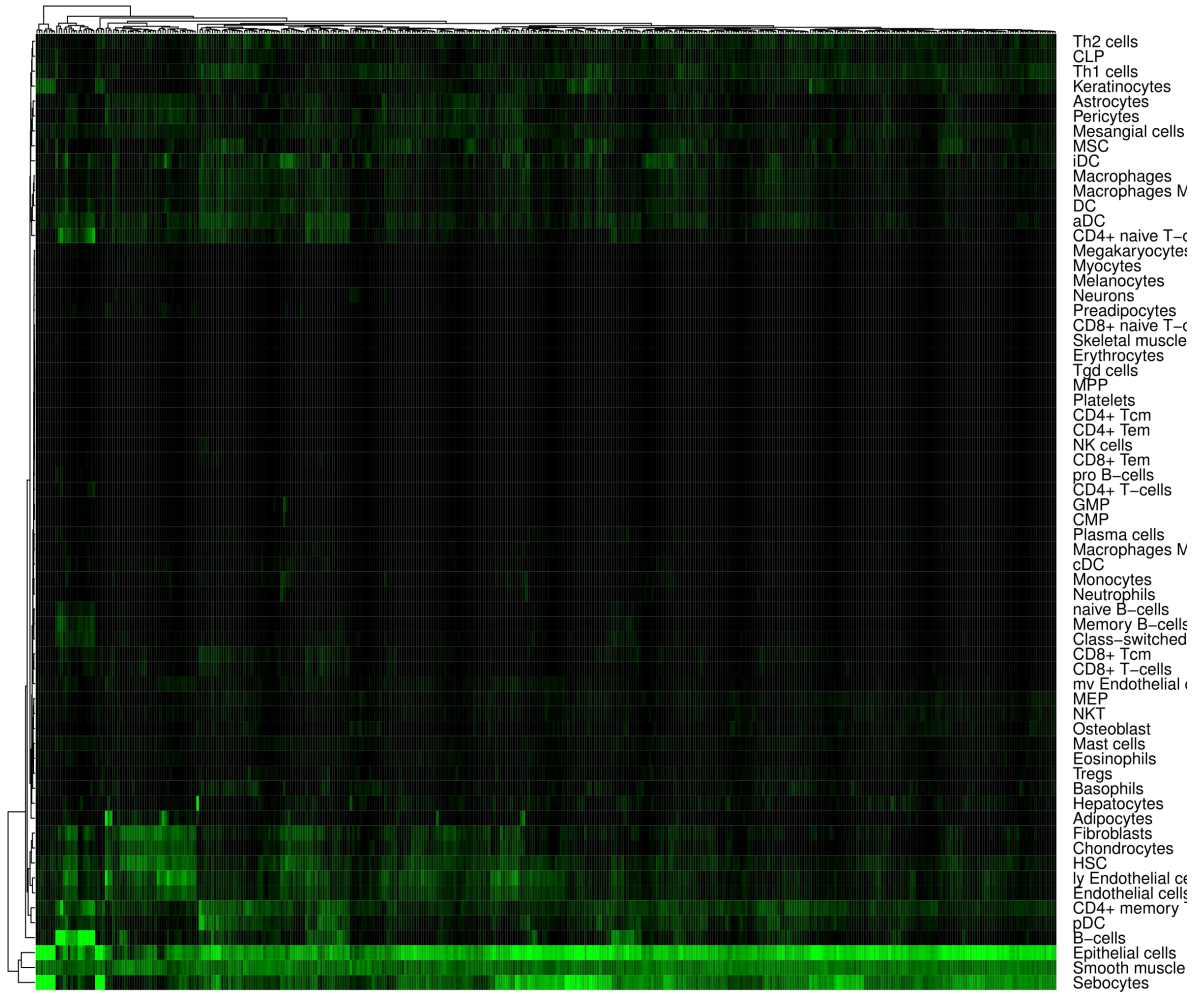
LUSC



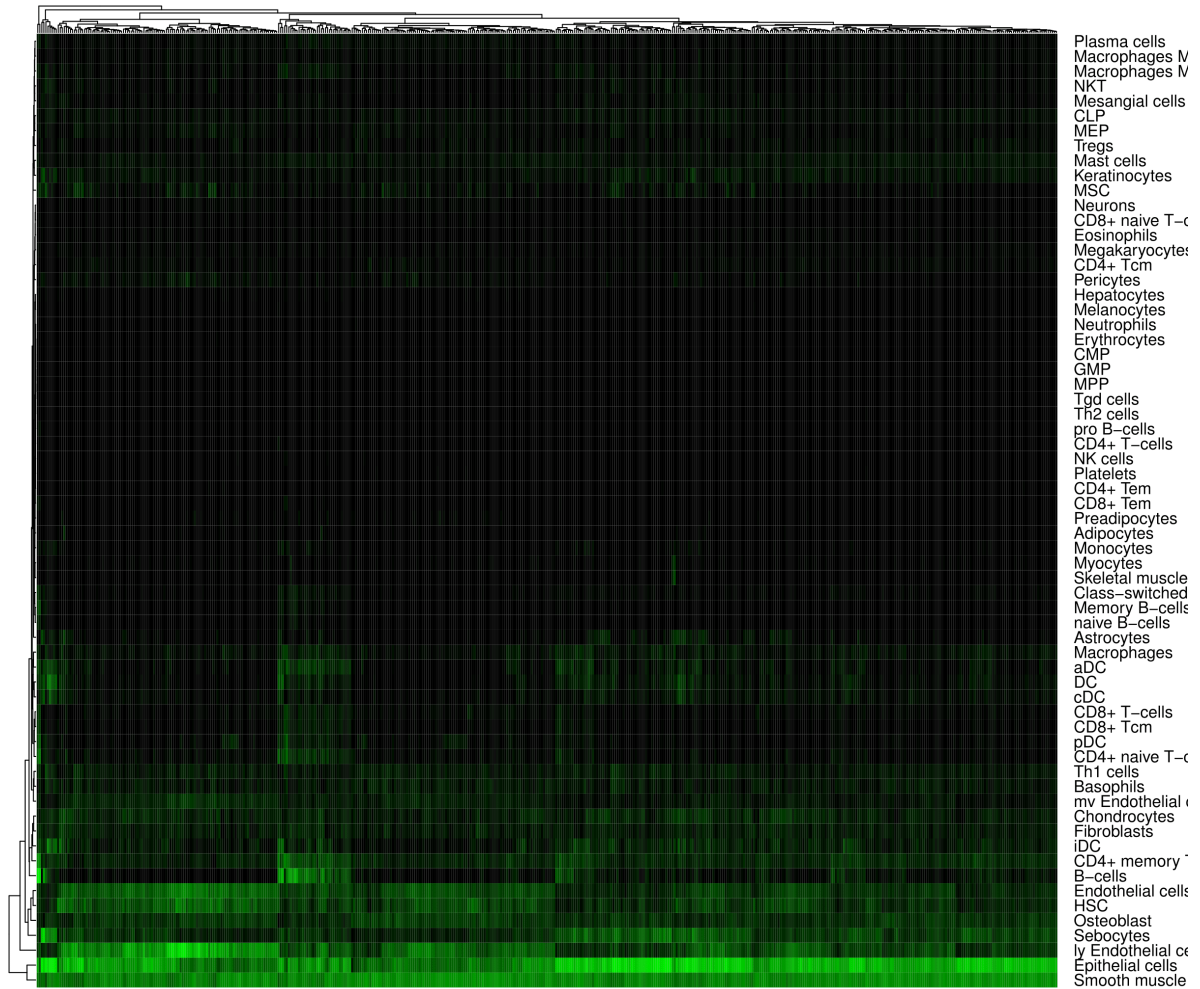
PRAD



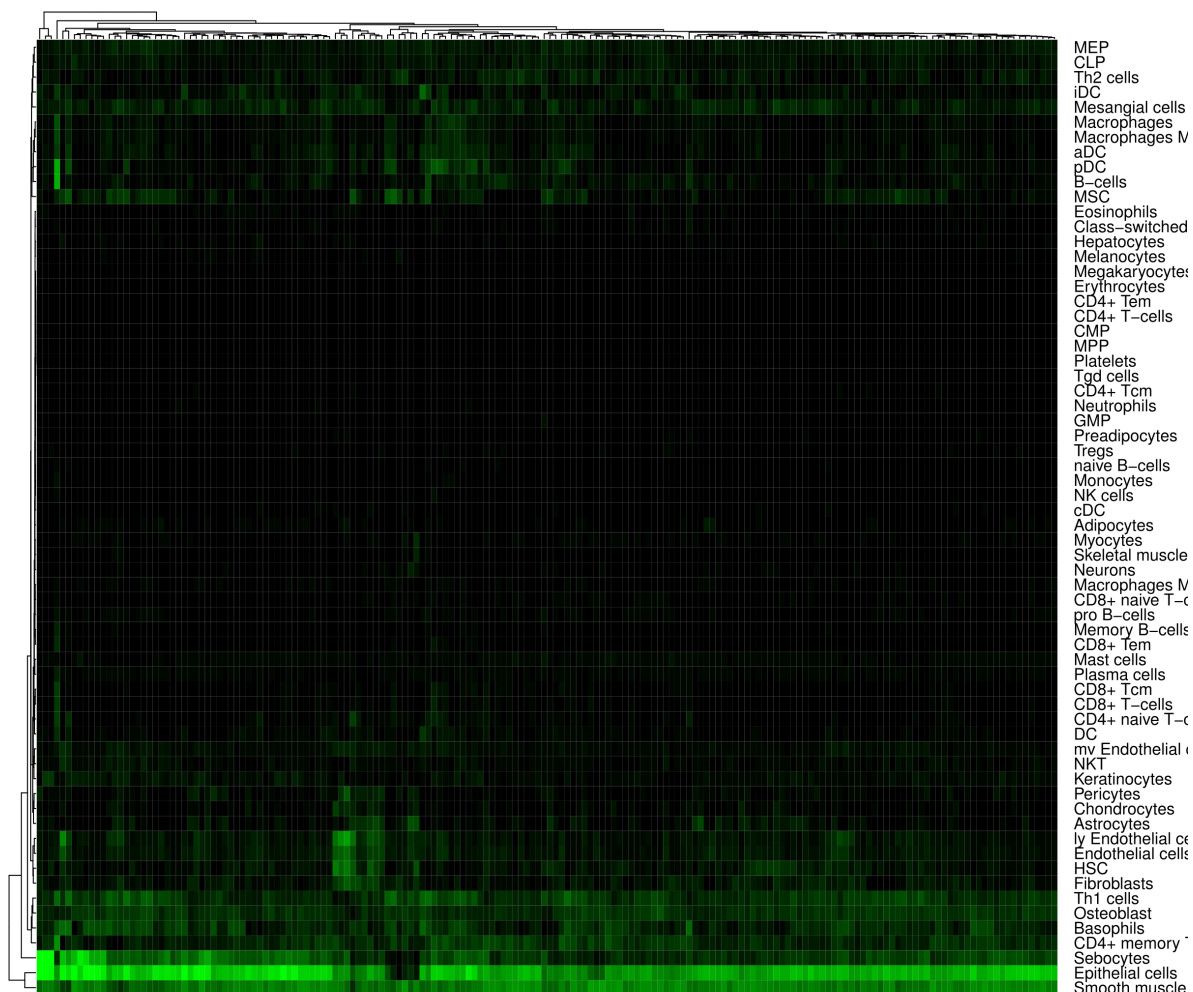
STAD



THCA

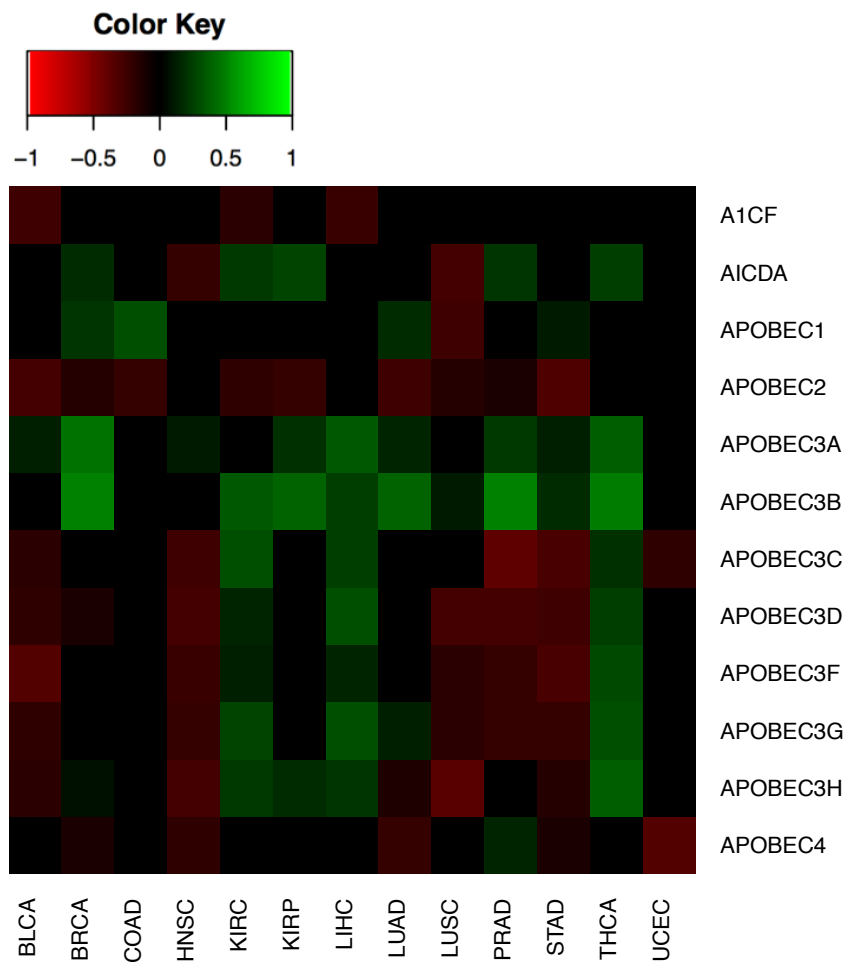


UCEC



xCell scores of 48 cell types for TCGA samples. Samples and cell types are clustered with k-means clustering. Data retrieved from the xCell platform.

Figure 15: Correlation Between APOBEC Genes and the Mean Buffa Hypoxia Signature Partial to Selected xCell Signatures



Correlation of APOBEC genes and the mean Buffa hypoxia signature partial to selected xCell signatures. xCell signatures are selected if 20% of more of samples exhibit a score greater or equal to 0.2. Spearman’s correlation is calculated between the mean expression of the Buffa hypoxia signature and the expression of each APOBEC gene, partial to the selected xCell signatures. Values displayed are Spearman’s rho, filtered for $p < 0.05$. Gene expression data from TCGA.

3.6 Subtypes of Breast Cancer Show Different Relationships between APOBEC and Hypoxia

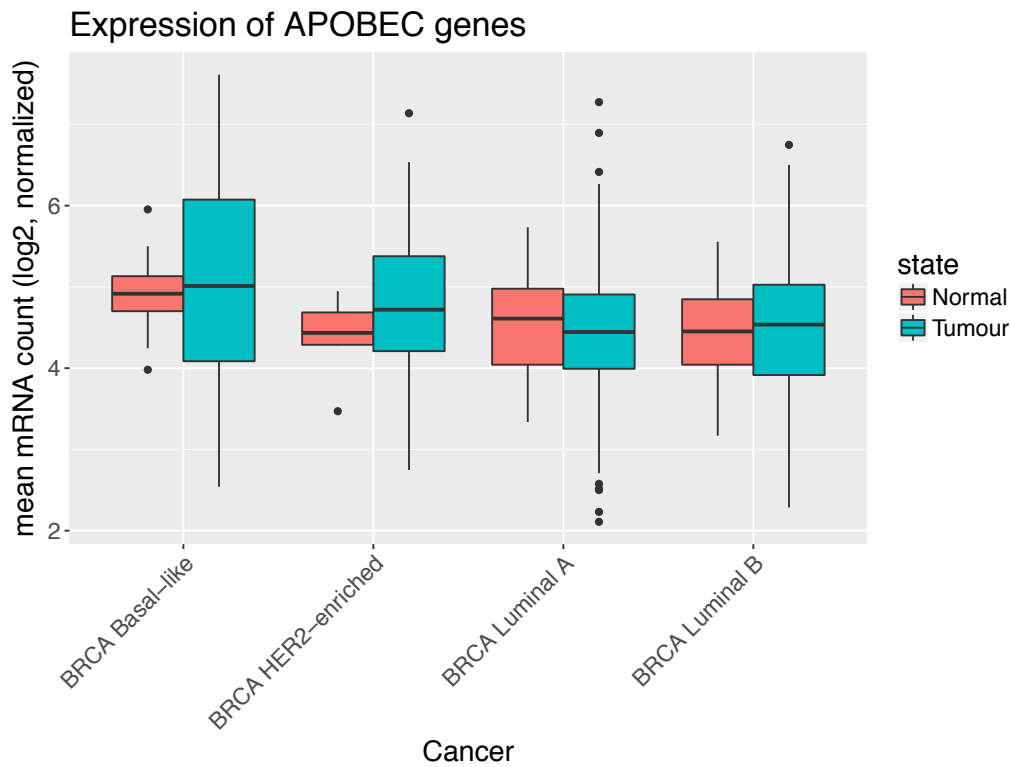
We further divide BRCA samples according to their subtype. We stratify samples into four subtypes by their molecular subtype classification: basal-like, HER2-enriched, luminal A, and luminal B (Perou et al., 2000; Sorlie et al., 2001). We perform the same analyses as previously done to observe that subtypes of breast cancer differ in terms of the association between APOBEC and hypoxia.

First, we plot out the mean expression of APOBEC genes in breast cancer subtypes (Figure 16). We see a lack of significant difference among cancer subtypes, just as we did among cancer types. Because of the lack of significant expression in several APOBEC genes as described earlier, we plot out the expression of A3A and A3B individually (Figure 17). Here, we see that although both A3A and A3B are preferentially expressed in tumours across all four subtypes, the degree to which they differ. Notably, the highly hypoxic basal-like subtype (Rakha et al., 2009; Tan et al., 2009) shows a large differential expression of both A3A and A3B in tumours. In addition, the HER2-enriched subtype, which overexpresses APOBEC 3 genes as previously reported (Kanu et al., 2016), also shows a large differential expression of A3A and A3B. In terms of hypoxia, again, we observe a large differential mean expression of the Buffa signature between normal and tumour samples in all four subtypes (Figure 18). The largest differential is found in the basal-like subtype, as expected, as it is known to be more hypoxic than other subtypes and to highly express HIF targets, hence it is expected to express this hypoxia signature (Koboldt et al., 2012).

Then, we calculate the Spearman's correlation between the mean expression of the Buffa hypoxia signature and individual APOBEC genes in the four subtypes (Figure 19). We notice that the correlation seen in A3A and A3B when considering BRCA as a whole is now only seen in the luminal subtypes. We generate scatter plots to visualise the relationship between A3A, A3B, and the hypoxia signature (Figure 20). We see that there is no correlation in the basal-like subtype. This could be a consequence of the fact that all samples are highly hypoxic. For Spearman's correlation to be significant, the data must have a certain spread. Here, all of our samples have a high mean hypoxia signature expression. The same effect can be seen in the HER2-enriched subtype, but for A3B. HER2, along with PTEN, have been shown to directly induce A3B activity in vitro, and the HER2 subtype has been reported to highly express APOBEC 3 family genes (Kanu et al., 2016). This is seen here in the A3B vs hypoxia plot, as all samples are clustered to the right, indicating high A3B expression. As for A3A, there is simply no correlation in the HER2-enriched subtype. With the luminal subtypes having a large range in the expression of both the mean hypoxia signature, A3A, and A3B, we see significant correlation.

Overall, we see that the relationship between hypoxia and APOBEC is heterogeneous among breast cancer subtypes. Luminal subtypes show a clear sign of correlation. Basal-like cancers all express hypoxia genes at a high level thus the correlational evidence is less clear, but APOBEC genes are expressed, while HER2-enriched cancers all highly express A3B.

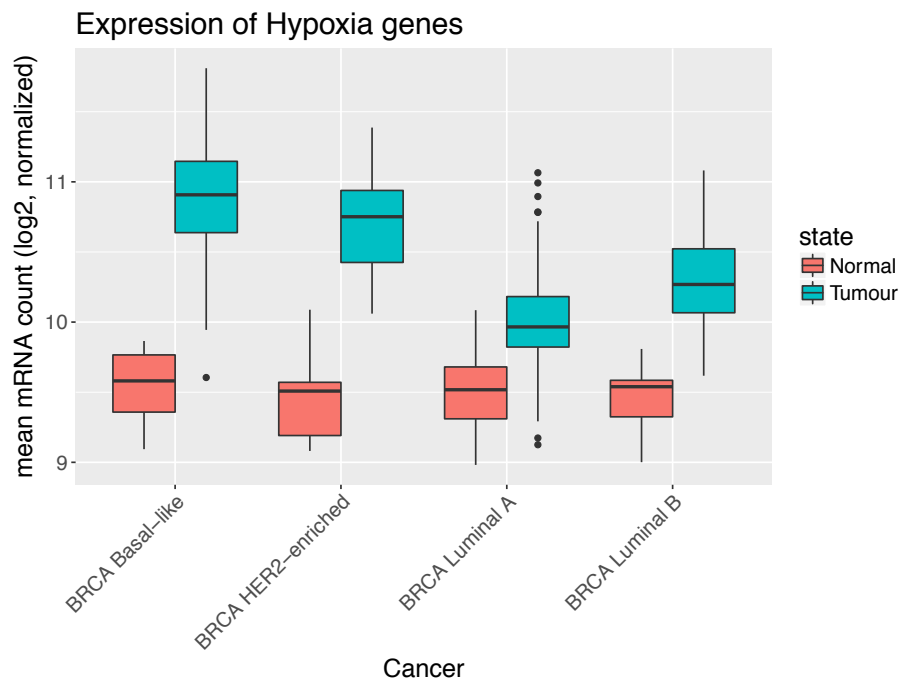
Figure 16: Mean Expression of APOBEC Genes in Breast Cancer Subtypes



Mean expression of APOBEC genes in four subtypes of BRCA. Subtypes represent the PAM50 classification of breast cancers. Data is log 2 transformed, normalised mRNA count from TCGA. The mean of values for all APOBEC genes in each sample is displayed. Boxes represent the range from the first to the third quartiles. Centre line represents the median. Whiskers represent 3rd quartile plus 1.5x the interquartile range or the maximal value, whichever is smaller, and the 1st quartile minus 1.5x the interquartile range or the minimal value, whichever is larger. Dots represent values not contained within whiskers.

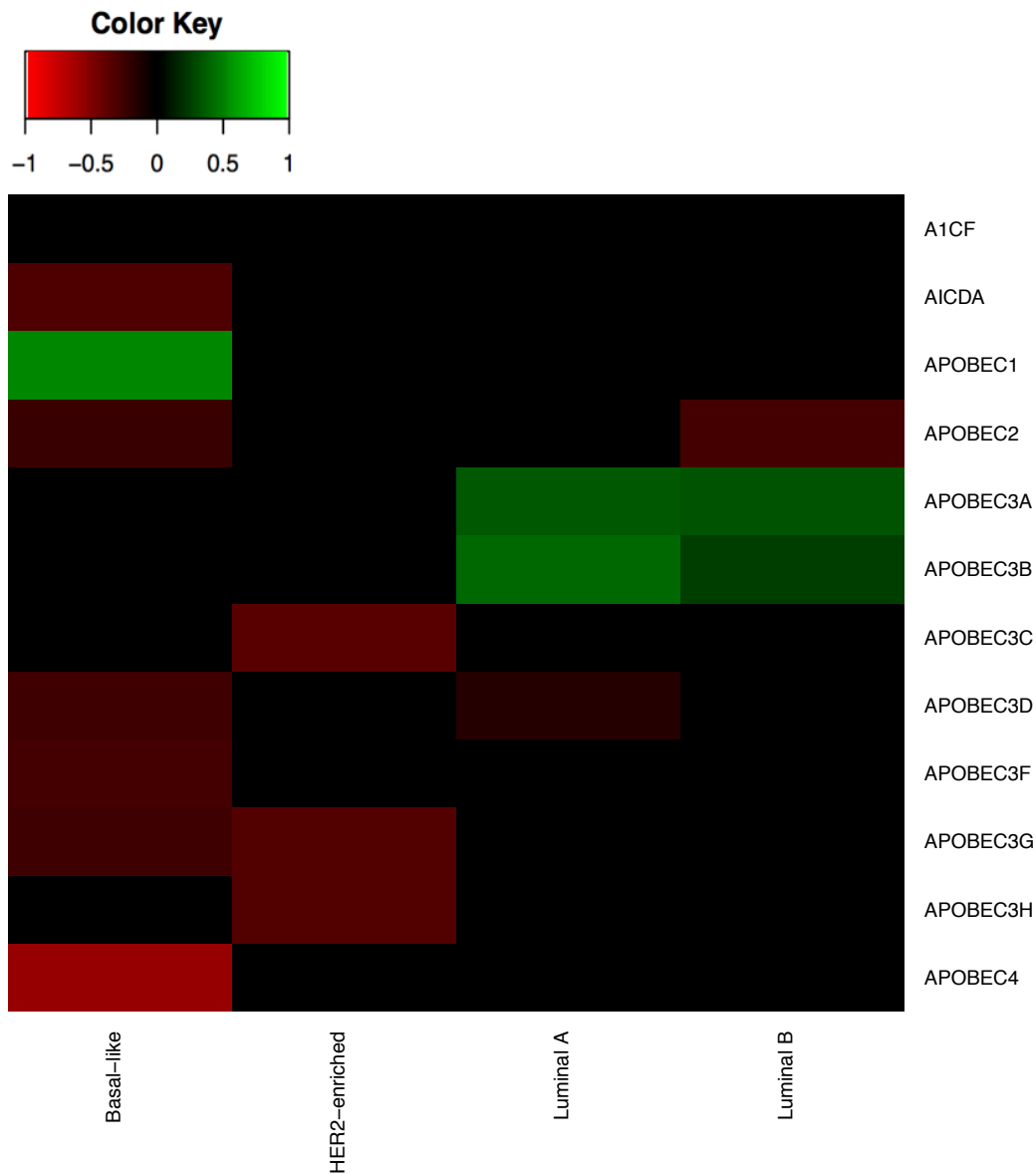
Expression of APOBEC3A and APOBEC3B in four PAM50 subtypes of BRCA. Data is log 2 transformed, normalised mRNA count from TCGA. Values in individual samples are displayed. Boxes represent the range from the first to the third quartiles. Centre line represents the median. Whiskers represent 3rd quartile plus 1.5x the interquartile range or the maximal value, whichever is smaller, and the 1st quartile minus 1.5x the interquartile range or the minimal value, whichever is larger. Dots represent values not contained within whiskers.

Figure 18: Mean Expression of the Buffa Hypoxia Signature across Breast Cancer Subtypes



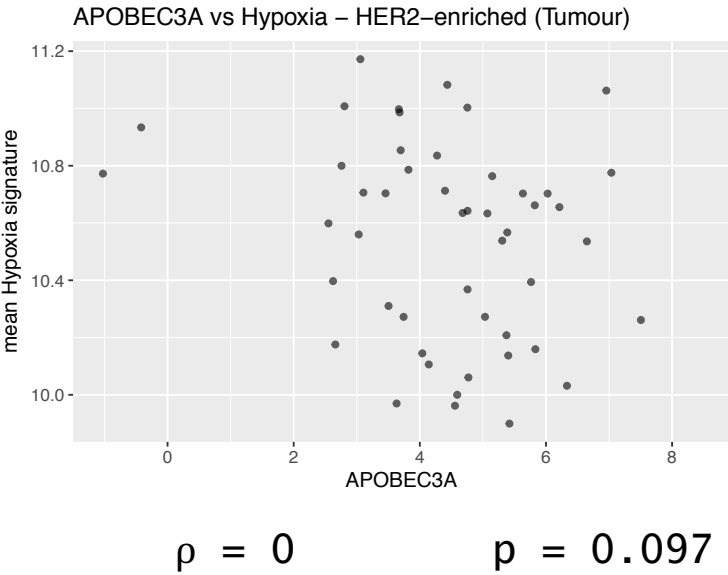
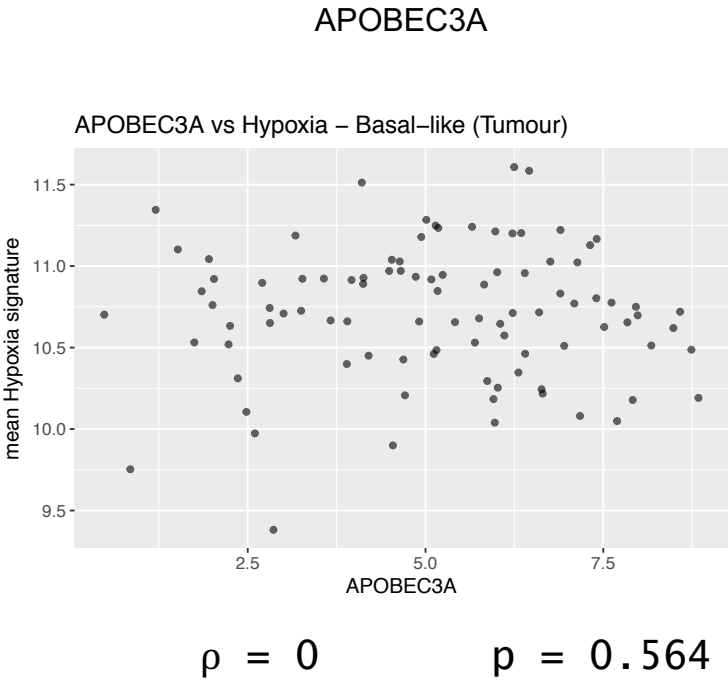
Mean expression of the Buffa et al. hypoxia signature in four PAM50 subtypes of BRCA. Data is log 2 transformed, normalised mRNA count from TCGA. The mean of values for all Buffa hypoxia signature genes in each sample is displayed. Boxes represent the range from the first to the third quartiles. Centre line represents the median. Whiskers represent 3rd quartile plus 1.5x the interquartile range or the maximal value, whichever is smaller, and the 1st quartile minus 1.5x the interquartile range or the minimal value, whichever is larger. Dots represent values not contained within whiskers.

Figure 19: Correlation between APOBEC Genes and the Mean Buffa Hypoxia Signature in Subtypes of Breast Cancer

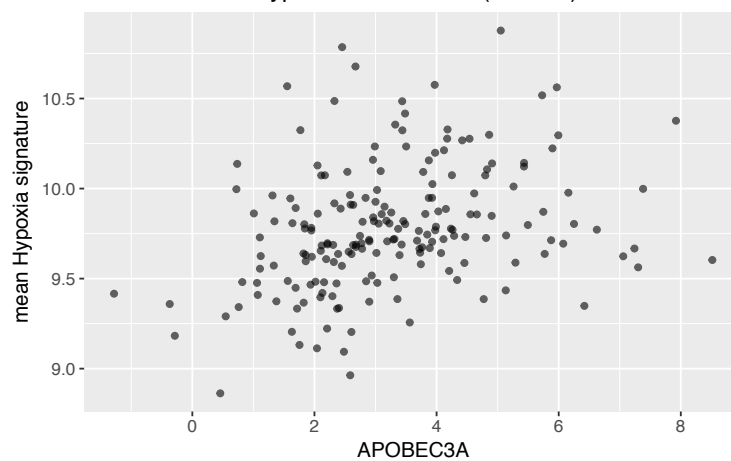


Correlation of APOBEC genes and the mean Buffa hypoxia signature in four subtypes of BRCA. Spearman’s correlation is calculated between the mean expression of the Buffa hypoxia signature and the expression of each APOBEC gene. Values displayed are Spearman’s rho, filtered for $p < 0.05$. Gene expression data from TCGA.

Figure 20: Relationship between the Expression of the Hypoxia Signature and APOBEC3A/3B in Subtypes of Breast Cancer

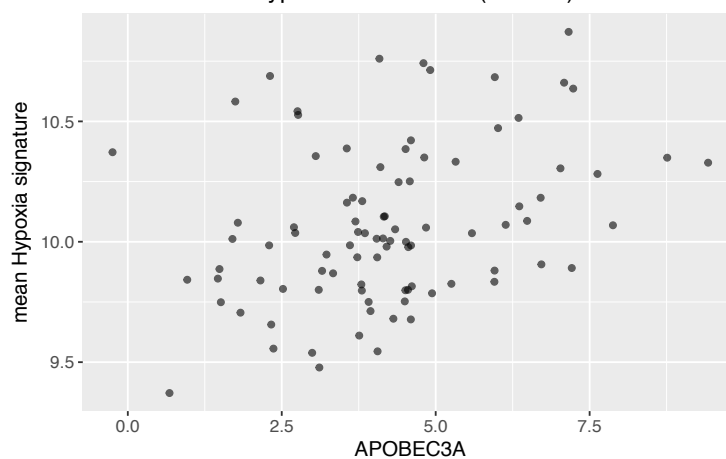


APOBEC3A vs Hypoxia – Luminal A (Tumour)



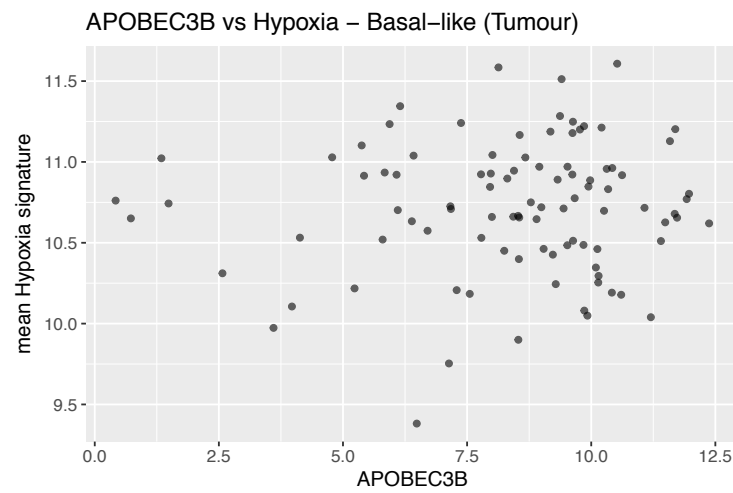
$$\rho = 0.352 \quad p < 0.0001$$

APOBEC3A vs Hypoxia – Luminal B (Tumour)



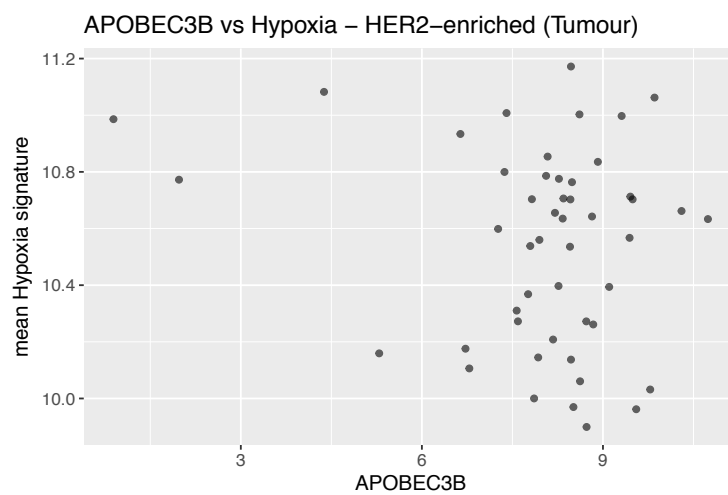
$$\rho = 0.324 \quad p < 0.005$$

APOBEC3B



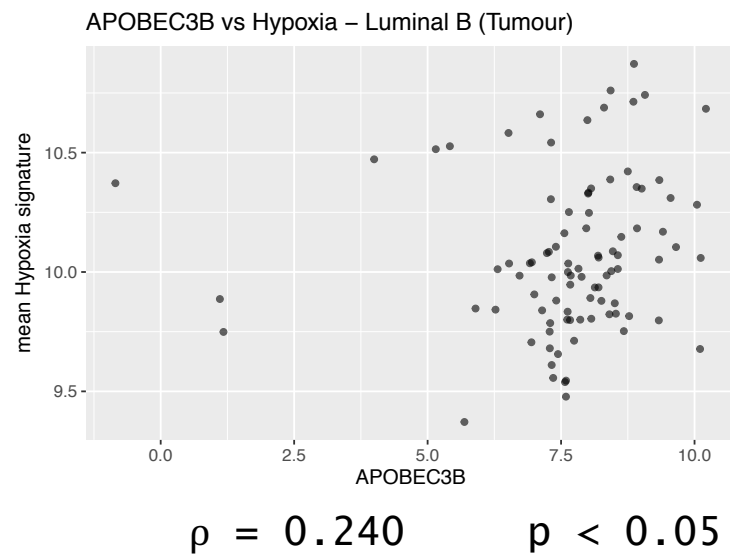
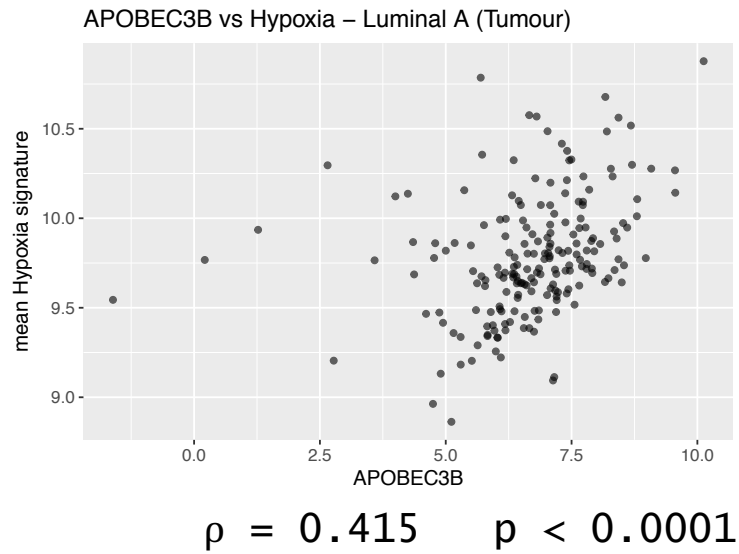
$$\rho = 0$$

$$p = 0.818$$



$$\rho = 0$$

$$p = 0.558$$



Correlation between the mean expression of the Buffa hypoxia signature and A3A/A3B in four subtypes of BRCA. Values represent log 2 transformed, normalised mRNA count from TCGA. Spearman's rho is calculated and displayed alongside p values.

3.7 The APOBEC Mutational Signature is Present in Multiple Cancers

While the expression status of a gene provides strong evidence for its presence, it does not provide evidence for its activity and proper function. RNA-seq data from TCGA is taken at the mature mRNA level, with multiple levels of regulation left before a functional protein product is made. Because of this, functional assays are preferable when available. A direct measure of APOBEC activity is its mutational signature – that is evidence of APOBEC mutation in the genomic DNA. APOBEC 1, 3A, 3B, 3C, 3D, and 3F cytosine deaminases have a preferred binding motif: TCW (W represents an adenine or a thymine), where the underlined cytosine is mutated. They perform a C-to-U deamination on TCW cytosines. Since uracil is not normally present in DNA, DNA repair proteins recognise the resulting TUW motif and yields TTW and TGW motifs left in the genomic DNA (Figure 1). This mutational pattern has been validated and confirmed to be present in multiple cancers (Burns et al., 2013b).

From this, we have a ‘mutational signature’ – a signature that can trace a mutational pattern to its root cause. Mutational signatures have been developed using various techniques for various mutation sources. In this case, the consensus TCW to TTW/TGW mutational signature is attributed to APOBEC cytosine deamination. Thus, we can utilise the prevalence of this signature in a sample’s genome or exome to infer the degree of APOBEC activity. In addition, a positive correlation between the mutational signature and APOBEC expression is evidence that the increased expression is manifested in mutations, suggesting a functional implication of increased APOBEC expression. There is currently no

method for the analysis of the APOBEC mutational signature that is fully validated and noise free, therefore we present four methods for measuring APOBEC activity this way and provide a comparison.

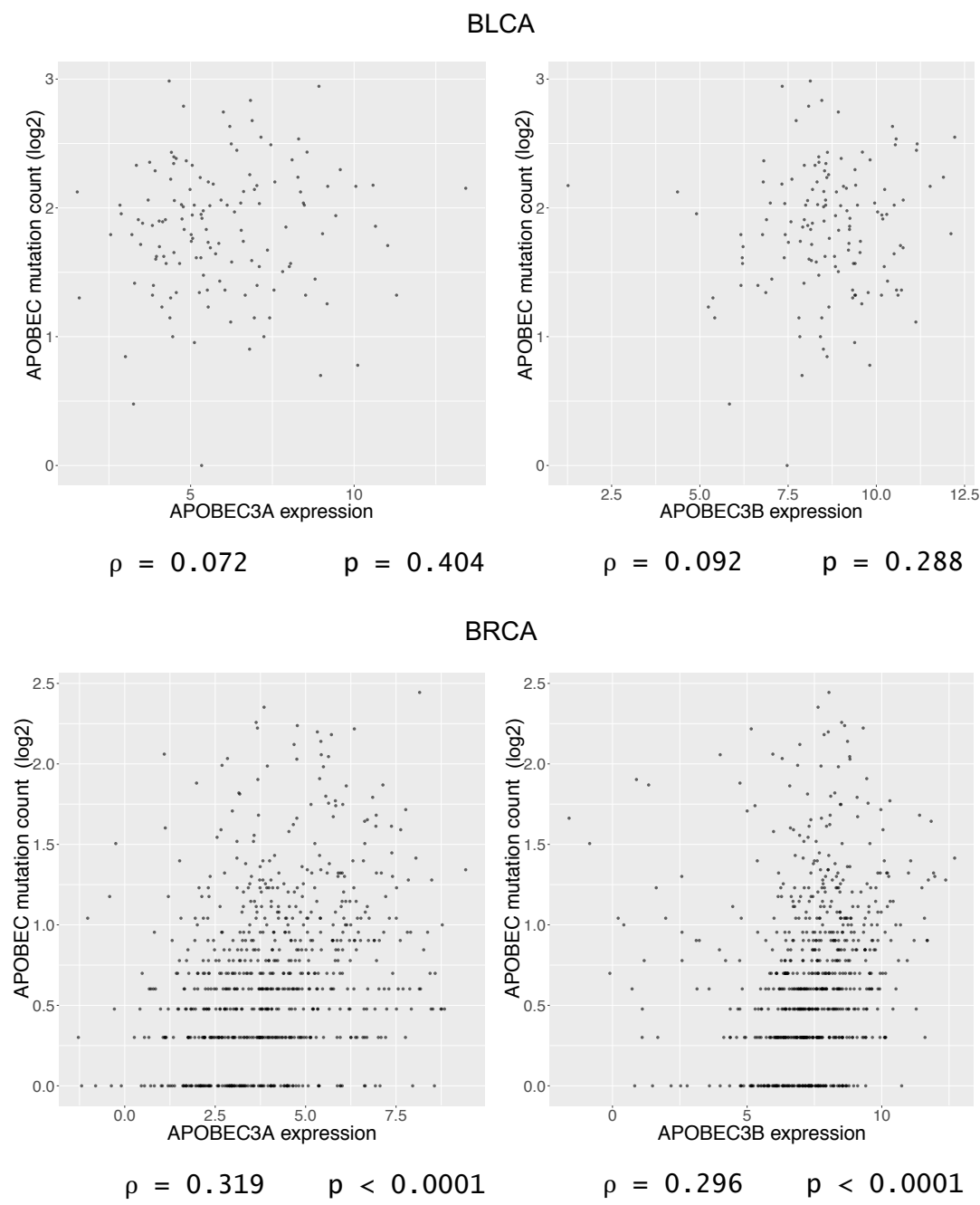
Because of limitations in data availability, we restrict the analysis of mutational signatures to subsets of the 13 cancer types analysed previously. LIHC and STAD are excluded for the lack of mutational catalogue data. COAD and UCEC are excluded because there are too few samples for statistical significance.

3.7.1 Count of Mutations at a Consensus Motif

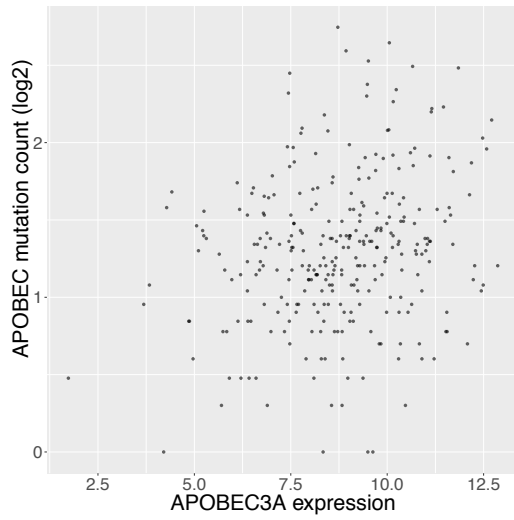
As a first measure of APOBEC-mediated mutagenesis, we count the number of TCW to TTW or TGW mutations in exomes. To visualise the occurrence of these mutations, we plotted the number of APOBEC mutations against the expression of A3A and A3B, the main active APOBECs (Figure 21), and calculated the Spearman's correlation score between them and associated p values. This graphic shows that TCW mutations indeed exist in various cell lines, and the correlation scores indicate that both A3A and A3B expression correlate with TCW mutations in BRCA, HNSC, KIRP, LUAD, and THCA. It has already been reported that both A3A and A3B expression correlate with the number of TCW mutations in BRCA and in all cancer types combined (Roberts et al., 2013). This new data expands on the previous work by providing evidence for which cancer types specifically have A3A and A3B expression correlated with mutation.

This method uses the raw count for the number of TCW to TTW or TGW mutations in the samples' exomes. Exomes are used instead of genomes because of wider availability of data giving us better statistical power. Using genomes increases the resolution of the analysis on a per-sample basis, but this benefit is overshadowed by the roughly 10 times fewer samples available across all cancers. Since exons are relatively conserved regions of the genome, the prevalence of available TCW contexts to be mutated should not change dramatically, thus a normalisation step for TCW availability is not absolutely essential in theory, although it is performed in the next section for comparison.

Figure 21: Relationship between the APOBEC3A and 3B and the Number of APOBEC-Induced Mutations

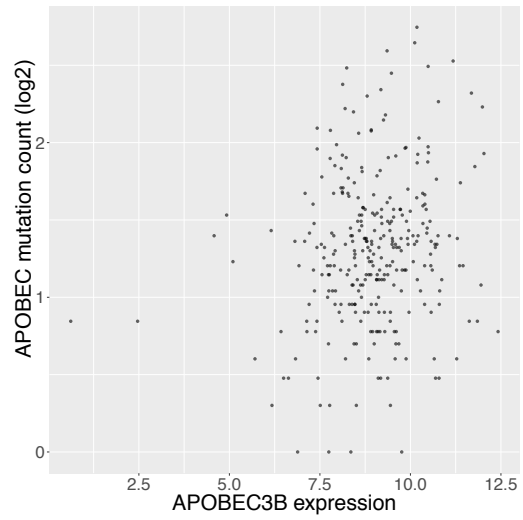


HNSC



$\rho = 0.198$

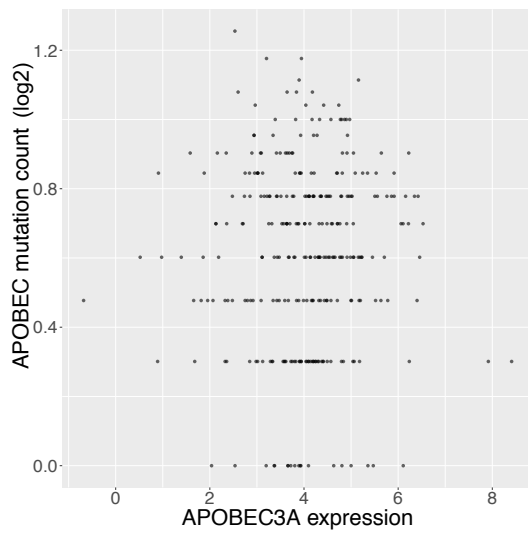
$p < 0.001$



$\rho = 0.149$

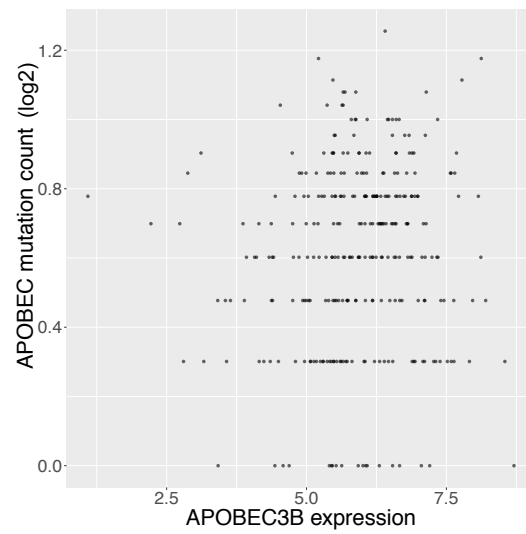
$p < 0.05$

KIRC



$\rho = 0.016$

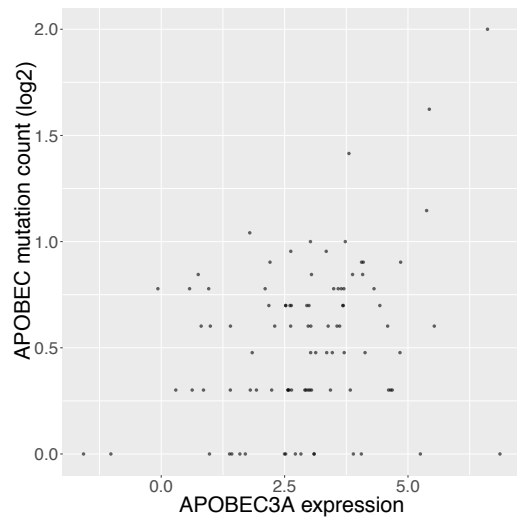
$p = 0.785$



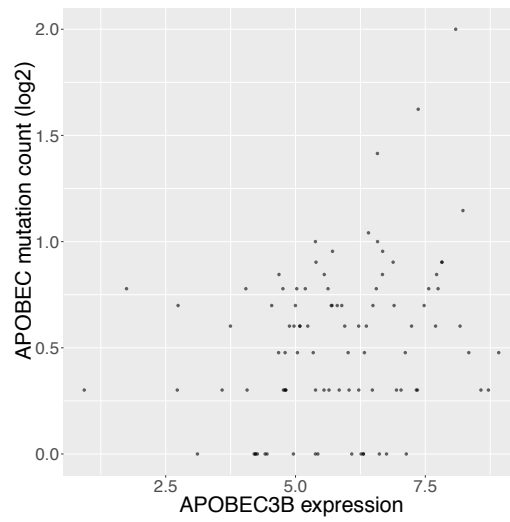
$\rho = 0.106$

$p = 0.067$

KIRP

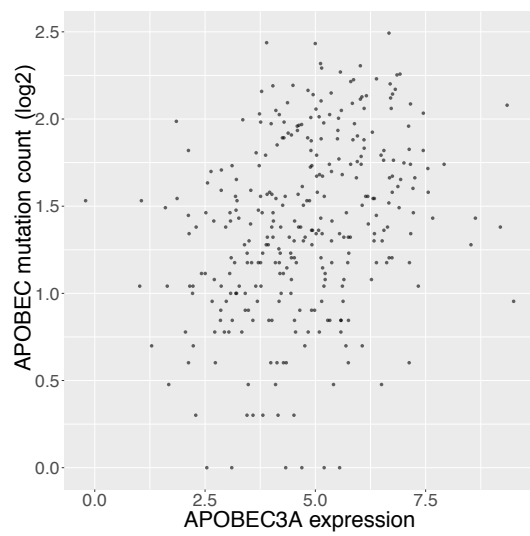


$\rho = 0.247$ $p < 0.05$

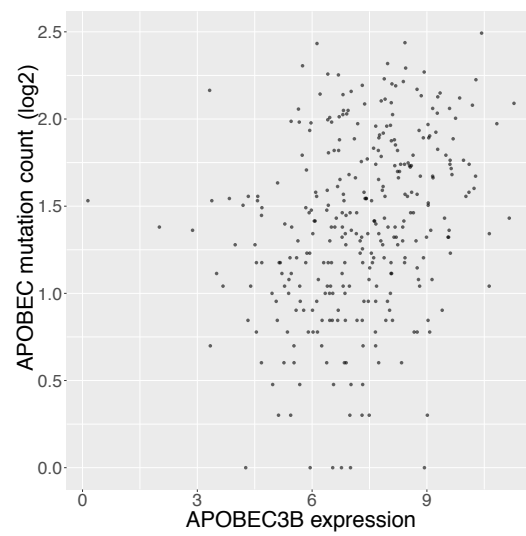


$\rho = 0.260$ $p < 0.05$

LUAD

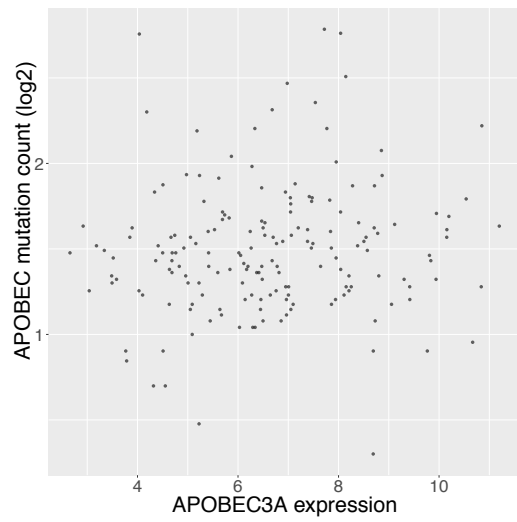


$\rho = 0.337$ $p < 0.0001$



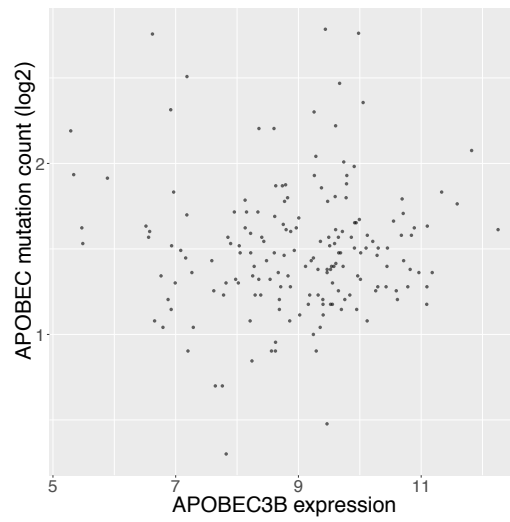
$\rho = 0.326$ $p < 0.0001$

LUSC



$$\rho = 0.126$$

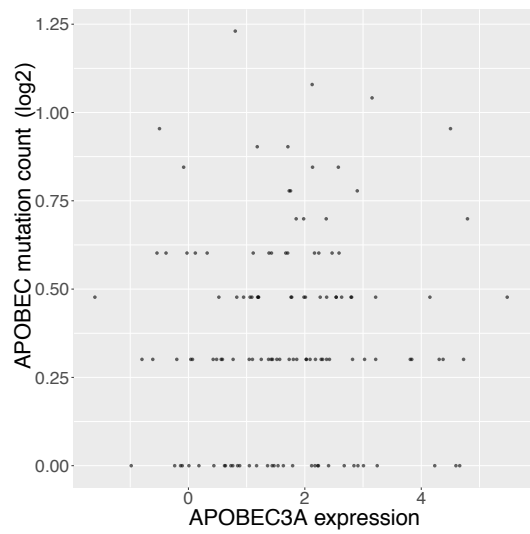
$$p = 0.096$$



$$\rho = 0.066$$

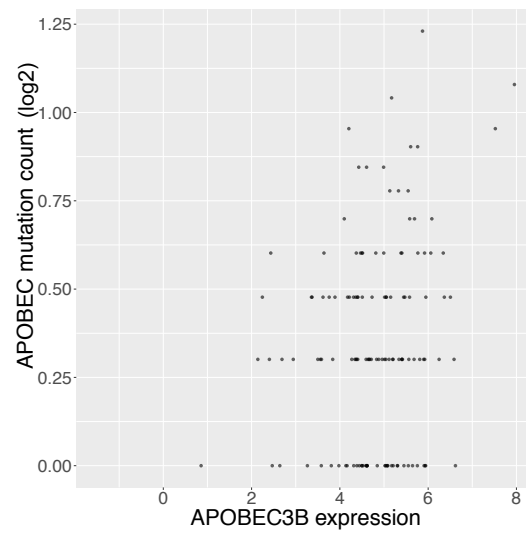
$$p = 0.386$$

PRAD



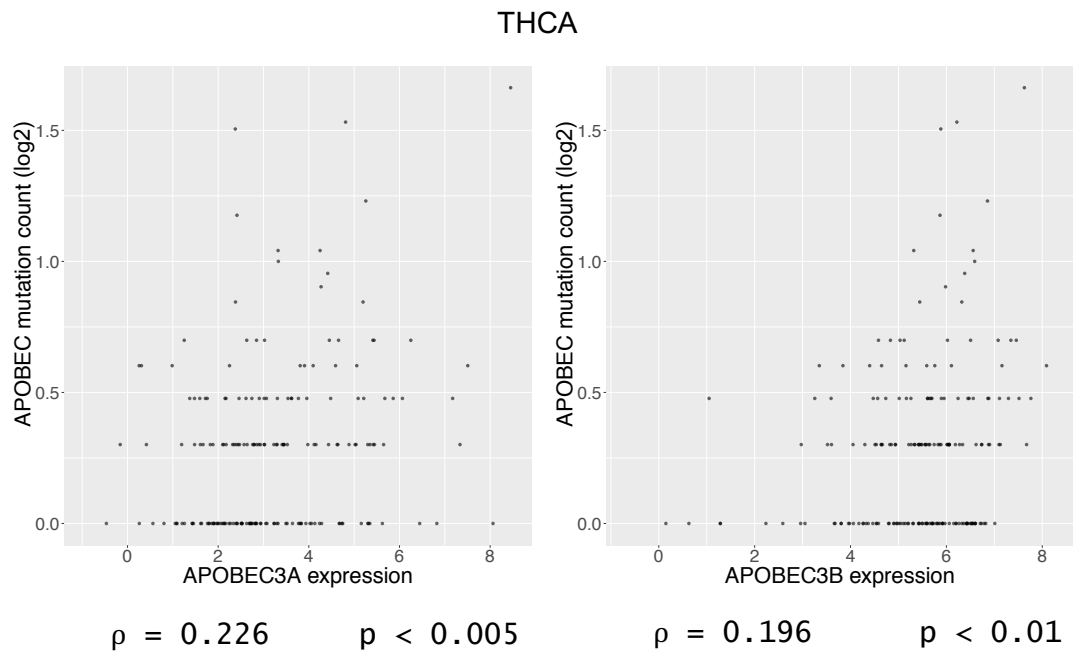
$$\rho = 0.064$$

$$p = 0.478$$



$$\rho = 0.156$$

$$p = 0.077$$



Correlation between the number of APOBEC mutations and A3A/A3B expression in 9 cancer types. APOBEC mutations are defined as TCW to TTW or TGW mutations. Mutation values are log 2 transformed counts of APOBEC mutations in the samples' exomes. Expression values are log 2 transformed, normalised mRNA counts from TCGA. Spearman's rho is calculated and displayed alongside p values.

3.7.2 *Enrichment of Mutations at a Consensus Motif*

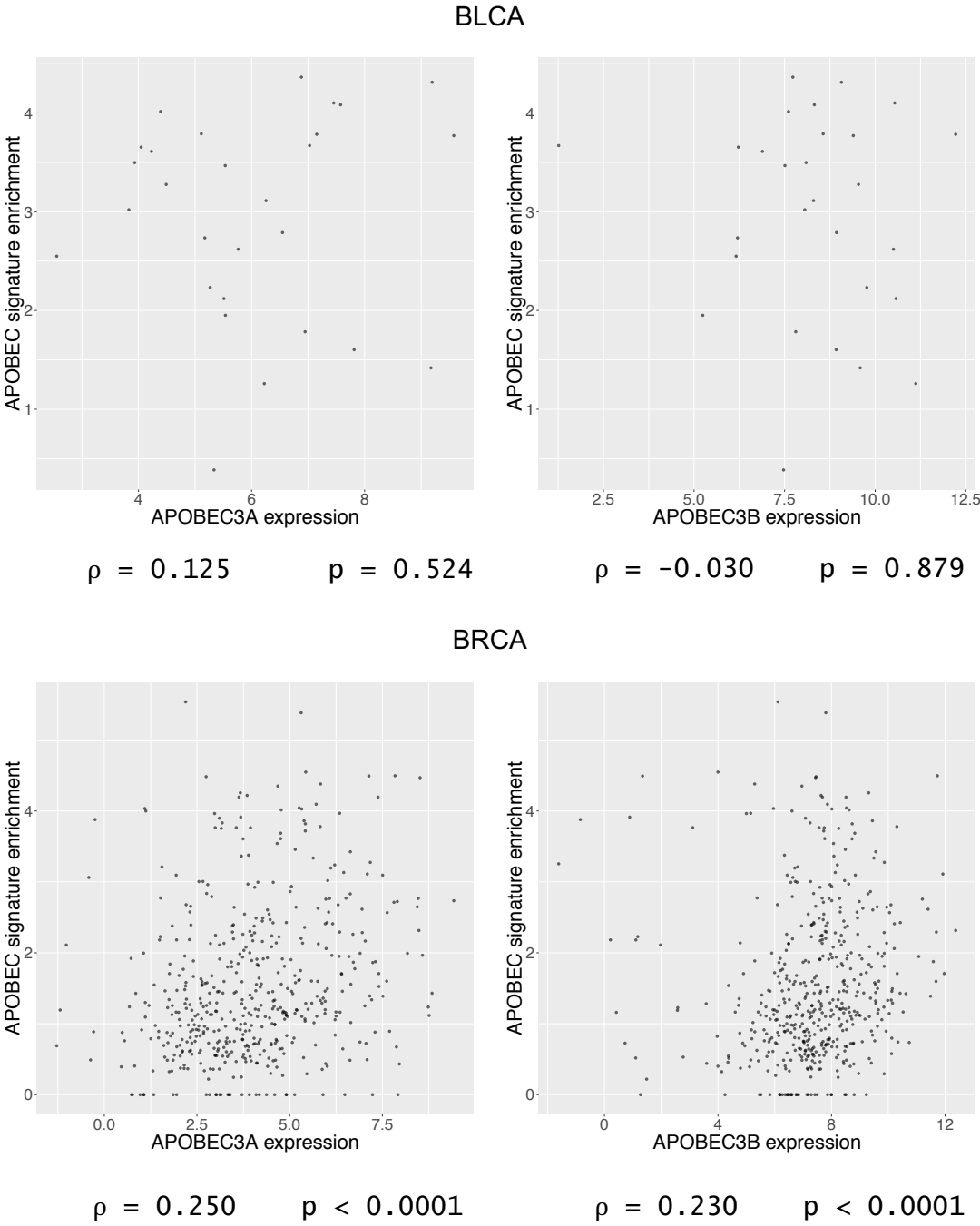
Although the number of deamination mutations at TCW motifs is an indicator of APOBEC activity, it does not take into account the genetic landscape of each cell. The availability of TCW motifs, the activity of other cytosine deaminases, and the amount of spontaneous cytosine deamination can all affect the number of TCW mutations we observe. To take these factors into account, we use a metric derived by a previous study, known as the enrichment of the TCW to TTW/TGW mutation (Roberts et al., 2013). This enrichment metric represents the ratio of TCW to TTW/TGW mutations to the number of C/G mutations, divided by the ratio of total TCW contexts to total C/G mutations in the region surrounding each mutated cytosine. Effectively, this score normalises the mutation counts for the availability of TCW contexts and C/G mutations caused by other factors. We use data provided by Roberts et al. in their publication for our analysis. They performed their analyses on TCGA data of 6 cancer types that are relevant to us: BLCA, BRCA, KIRC, LUAD, LUSC, and PRAD.

We plot out Roberts' enrichment scores against the expression of A3A and A3B as scatter plots and calculate the Spearman's correlation score for the 6 cancers and associated p values (Figure 22). We see recurring patterns with significant correlation with A3A in BRCA, LUAD, and LUSC, and with A3B in BRCA, LUAD, LUSC, and PRAD.

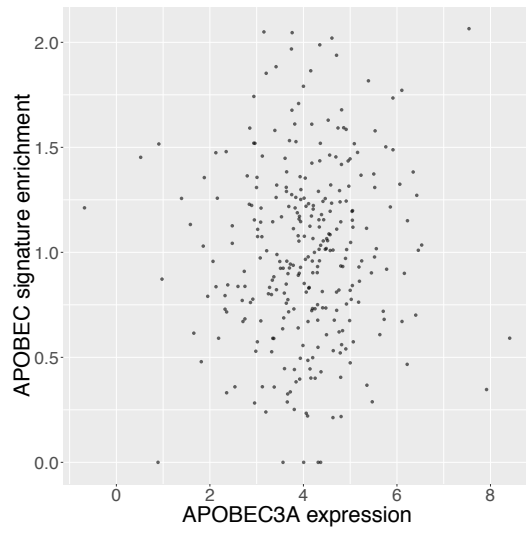
Although this method takes into account more factors than the sheer number of mutations, it does not necessarily imply that it is absolutely superior. Two notable drawbacks are to be mentioned. First, the inclusion of more variables intrinsically

includes more noise and error since each variable has its error margin and associated noise. This noise may offset the benefit of controlling for another variable. Second, the limited number of samples and cancer types compared to data drawn directly from TCGA drives the statistical power down. It would be possible to re-compute the enrichment metric for every TCGA sample, as Roberts et al. did in their original publication, but this work is time- and computational power-consuming, thus omitted in the present study.

Figure 22: Relationship between the APOBEC3A and 3B and the Enrichment of APOBEC-Induced Mutations

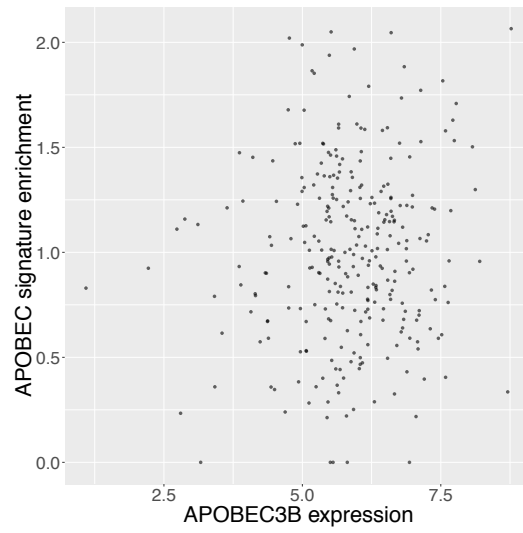


KIRC



$\rho = 0.062$

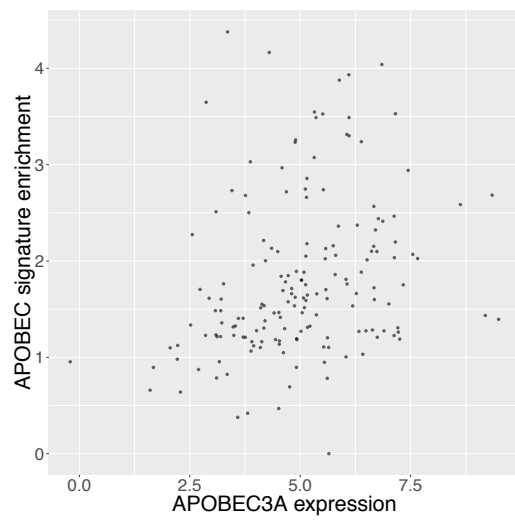
$p = 0.288$



$\rho = 0.040$

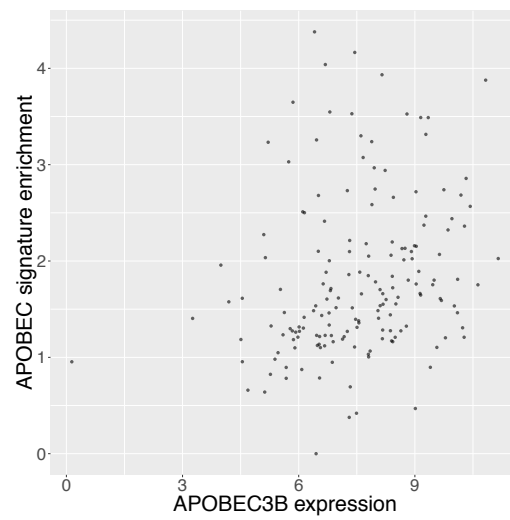
$p = 0.497$

LUAD



$\rho = 0.342$

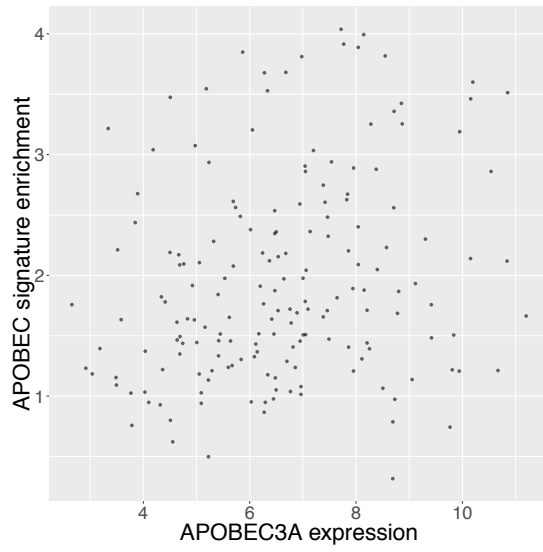
$p < 0.0001$



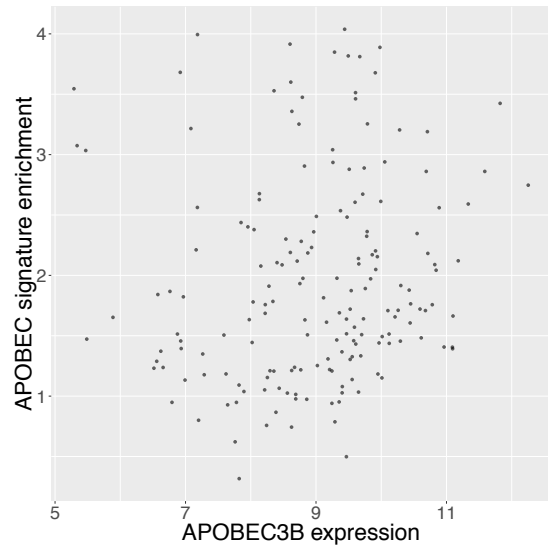
$\rho = 0.319$

$p < 0.0001$

LUSC

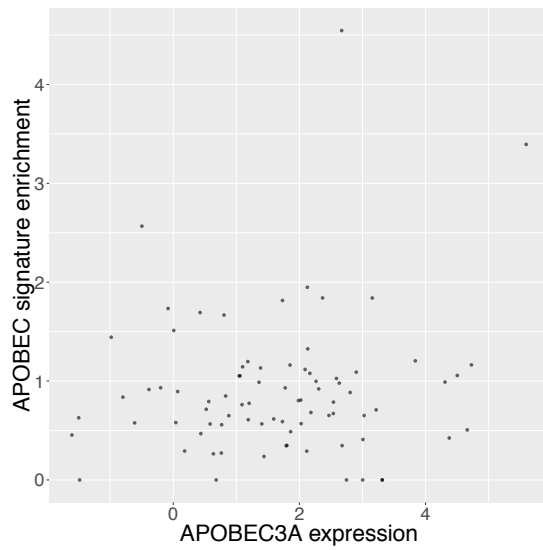


$\rho = 0.237$ $p < 0.005$

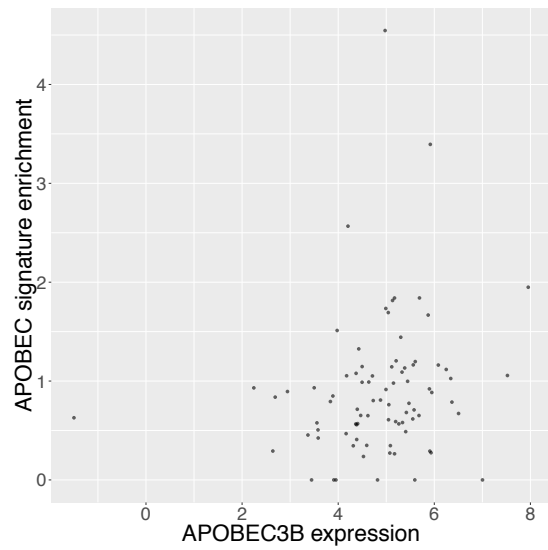


$\rho = 0.199$ $p < 0.01$

PRAD



$\rho = 0.044$ $p = 0.696$



$\rho = 0.239$ $p < 0.05$

Correlation between the enrichment of APOBEC mutations and A3A/A3B expression in 9 cancer types. Enrichment values represent the enrichment score defined in Roberts et al., 2013 and explained in the main text. Expression values are log 2 transformed, normalised mRNA count from TCGA. Spearman's rho is calculated and displayed alongside p values.

3.7.3 Enrichment of Consensus Mutational Patterns in Samples

Since the TCW motif was derived from an empirical hypothesis, the two first methods for analysing the presence of an APOBEC mutational signature are thus hypothesis driven. Both the counts and the enrichment methods' main caveat is their reliance on the simple definition of the APOBEC mutational signature as a TCW to TGW or TTW mutation. A de novo method for extracting mutational signatures from large data sets using mathematical techniques was developed in recent years and has become a staple tool in the analysis of mutational signatures (Nik-Zainal et al., 2012). Using this method, the authors analysed TCGA and unpublished samples at the time to derive 22 consensus mutational signatures (Alexandrov et al., 2013). Recent work has expanded this list to 30 signatures and these signatures are being constantly curated and maintained by the Wellcome Trust Sanger Institute under the Catalogue Of Somatic Mutations In Cancer (COSMIC) project. (Forbes et al., 2017). These new signatures differ from those previously generated empirically by providing not only information about the trinucleotide context surrounding the mutated base, but also the relative frequency of occurrence of mutations in each of the 96 possible trinucleotide contexts. These signatures can be assigned to a variety of causes, two of which are APOBEC-mediated cytosine deamination: signatures 2 and 13.

With the consensus signatures, we set out to find whether they are present in TCGA samples, and whether the APOBEC signatures are enriched and correlated with APOBEC expression. We use deconstructSigs, a tool comparing a given sample's mutational profile to consensus mutational signatures

(Rosenthal et al., 2016), to analyse the extent to which these 30 consensus signatures are present in TCGA samples. `deconstructSigs` uses a multiple regression model to fit each consensus signature to the mutational profile of a sample and assigns an enrichment score based on the coefficient of the signature in the regression model, normalised between 0 and 1. `deconstructSigs` was used on TCGA data with all 30 current signatures. This way, we are not relying on a simple definition of APOBEC mutagenesis as one trinucleotide substitution, but rather one that is the pattern that most accurately reflects what occurs in thousands of samples statistically. The analysis is performed on all tumour samples of 9 selected cancer types using the 30 mutational signatures validated by COSMIC (Forbes et al., 2017).

The raw scores obtained from the algorithm are plotted as jitter in Figure 23. The percentage of samples for which the score is non-null for each signature is shown in Figure 24. A point to note is that a null score here represents a strictly positive regression coefficient between 0 and 0.06; `deconstructSigs` implements a filtering step to reduce any final result lower than 0.06 to 0 because it is considered noise and unreliable. This threshold is chosen empirically by the authors (Rosenthal et al., 2016).

From the jitter plots, we see that signature 1 is present across cancers. Signature 1 is the signature for the spontaneous deamination of methylated cytosines from C to T (Forbes et al., 2017). This phenomenon is associated with aging and is normally present throughout all types of cancer. Signature 1 serves as a

qualitative positive control, as it should always be present. If signature 1 is absent, the parameters of the algorithm need adjustment.

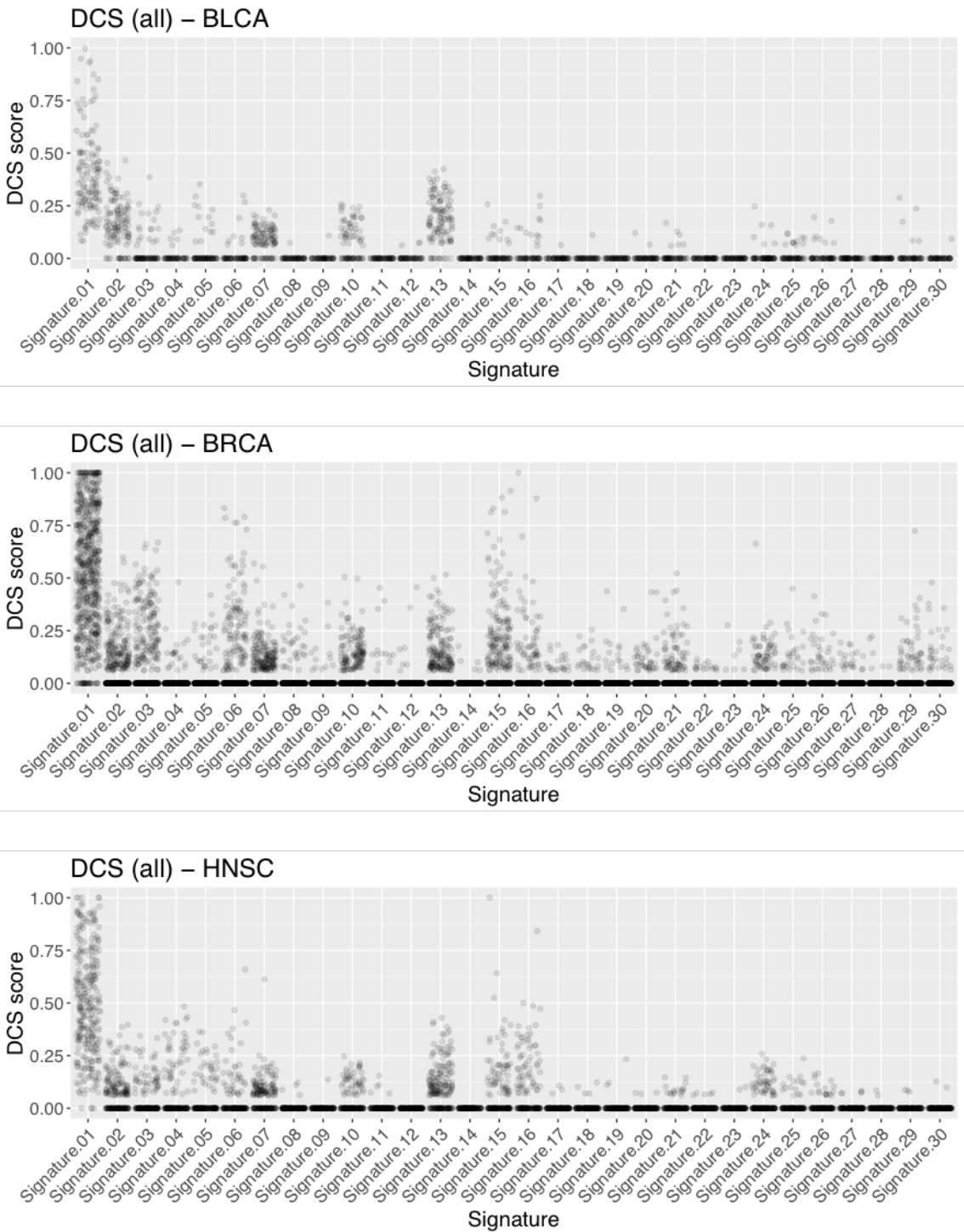
The APOBEC signatures, 2 and 13, are present in significant quantity in BLCA, BRCA, HNSC, LUAD, and LUSC. Looking at the percentage of samples with a non-null presence of each signature, we see that nearly all signatures except signature 1 are present in no more than half of the samples. Although this may seem alarming, it is expected for this method, as even when generating the mutational signatures, Alexandrov et al. reported low proportions of samples carrying the signature for every signature except signature 1 (Alexandrov et al., 2013). Despite the low numbers, signatures 2 and 13 stand out as dominant over the others in terms of presence in BLCA, BRCA, and HNSC by being among the signatures with the highest proportion of samples carrying them. Together, this data suggests that APOBEC mutagenesis is present in the mentioned cancers and that further analysis can be conducted with deconstructSigs data alongside expression data.

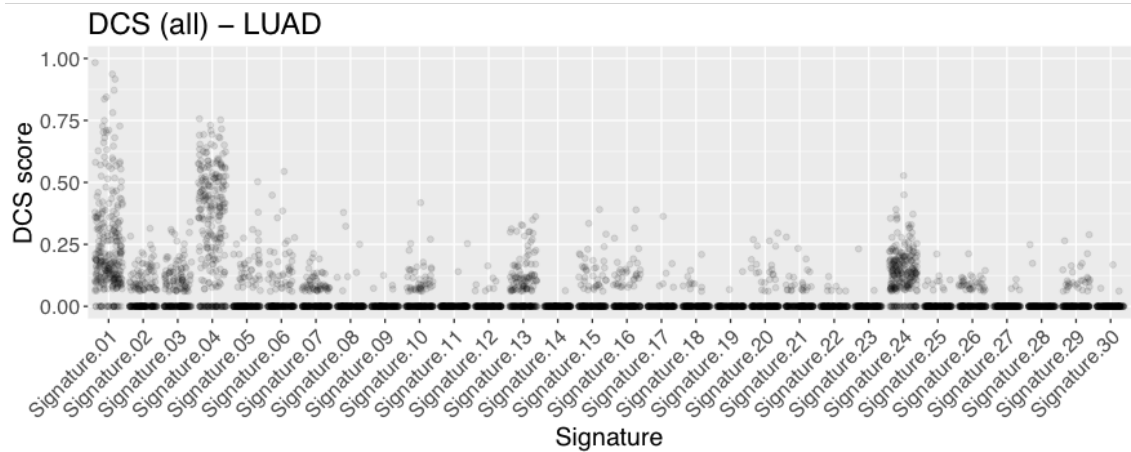
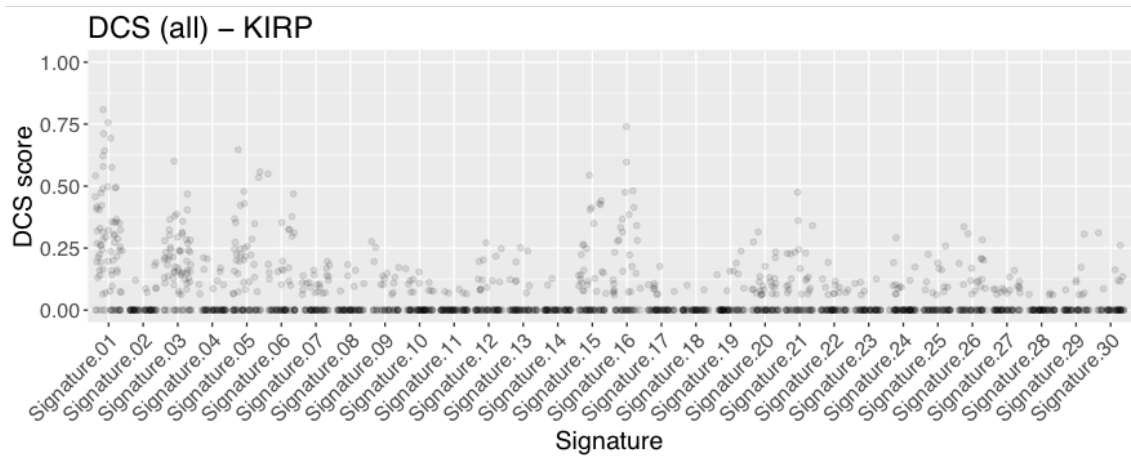
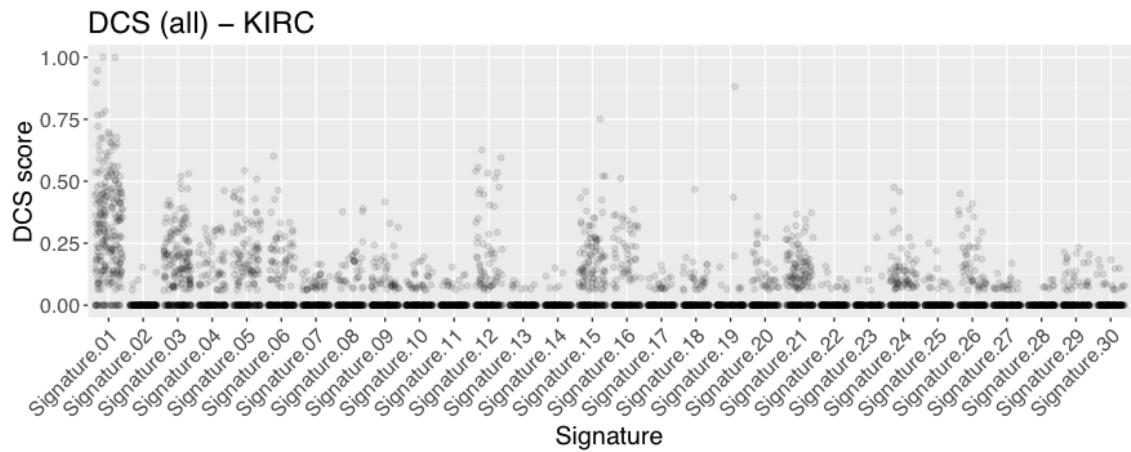
Note that COSMIC reports the presence of signatures 2 and 13 in several more cancer types. The first reason behind this is that COSMIC reports even a small presence in a few samples. This is not wrong for their purposes, but for our analysis, a select few samples showing a signature is not relevant. The second reason is because deconstructSigs adds another layer of noise by using data generated by COSMIC to search in TCGA. COSMIC reports that they successfully derived the two signatures from samples of a certain cancer type. We try to fit the raw data from the samples to the derived signatures using

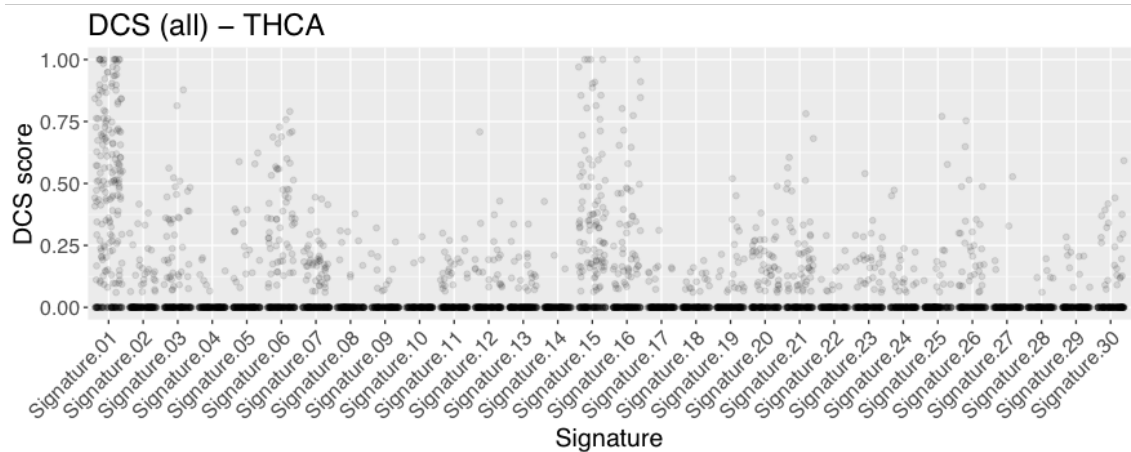
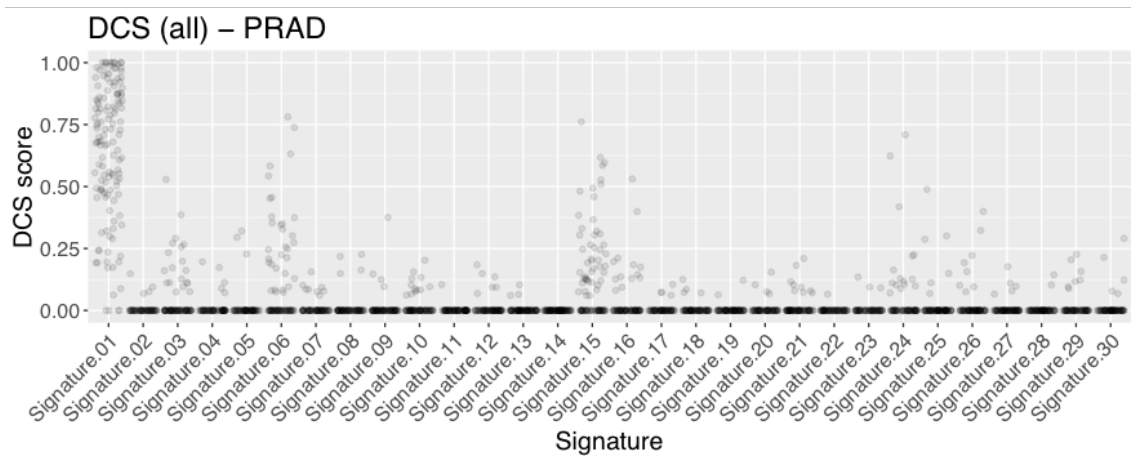
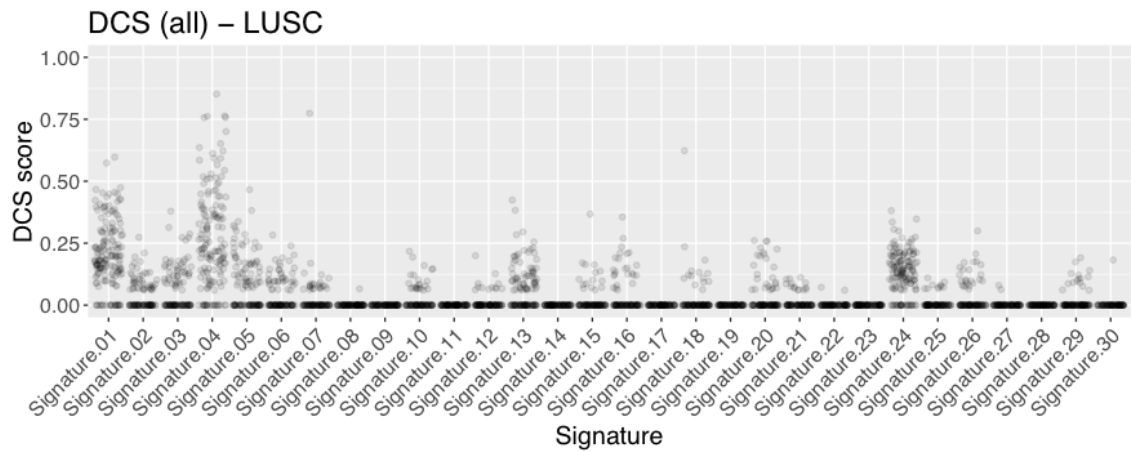
regression. In other words, COSMIC uses a forward discovery process while we use a reverse validation process with what COSMIC generated. This may lead to differences in reported cancer types when the presence is small or limited to a small number of samples.

As with counts and enrichment, we plot the deconstructSigs scores against A3A and A3B expression as scatter plots and calculate the Spearman's correlation coefficient and associated p values (Figure 25). For the deconstructSigs scores, we use the maximum of signatures 2 and 13. The reason for this is because signature 2 mainly accounts for most TCW to TTW mutations while signature 13 accounts for TCW to TGW. To group these together, we take the maximum of the two. deconstructSigs scores correlate with A3A expression in BRCA, HNSC, KIRP, LUAD, and LUSC, and with A3B expression in BRCA, KIRP, and LUAD. We can see that these are recurrent cancer types compared to the results of previous analyses using the counts or the enrichment of TCW to TTW or TGW mutations. We would not expect a large difference however, since the COSMIC signatures for APOBEC largely reflect the consensus mutations used in the previous sections. It is to note that signatures 2 and 13 are not present in relatively high proportions in KIRP, with less than 3% of samples showing non-null presence. Thus, the correlation observed between the deconstructSigs scores and APOBEC expression in KIRP may be inaccurate due to the high proportion of samples being null.

Figure 23: deconstructSigs Scores for the 30 COSMIC Signatures

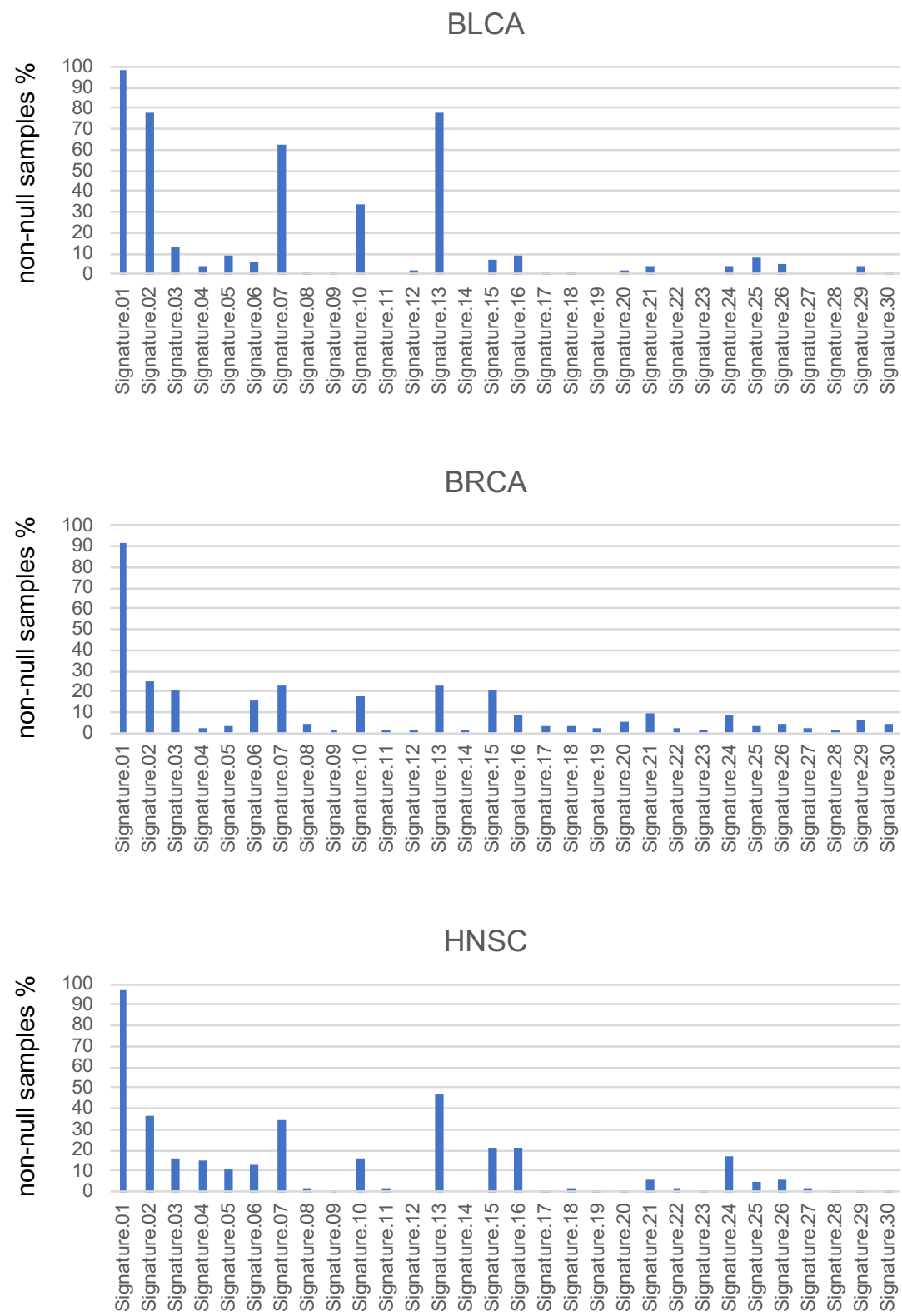




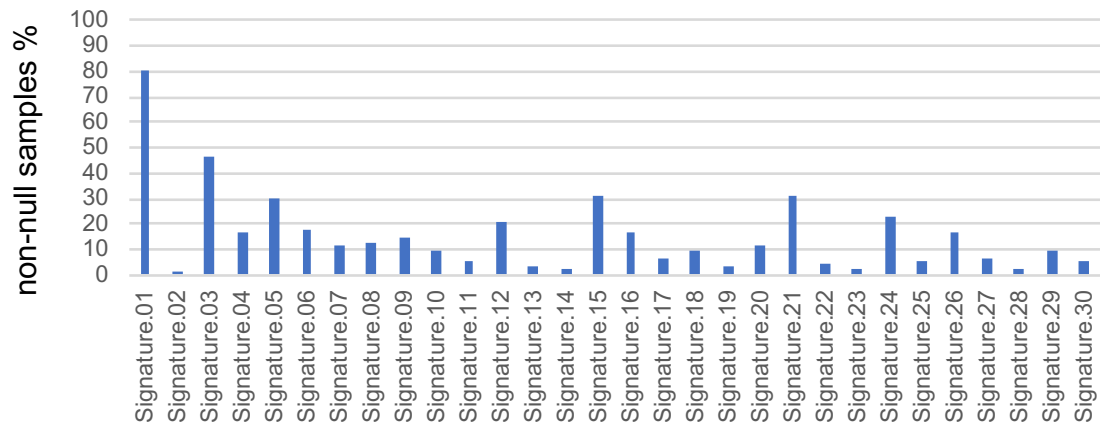


Scores from the deconstructSigs (DCS) algorithm using all 30 COSMIC mutational signatures. Values represent the level of enrichment of a mutational signature in single samples.

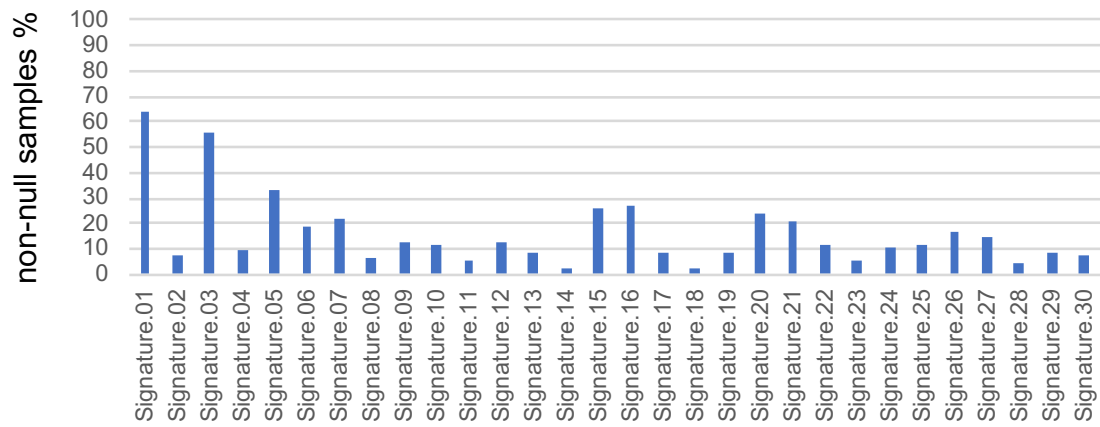
Figure 24: Percentage of Non-Null deconstructSigs Scores for the 30 COSMIC Signatures



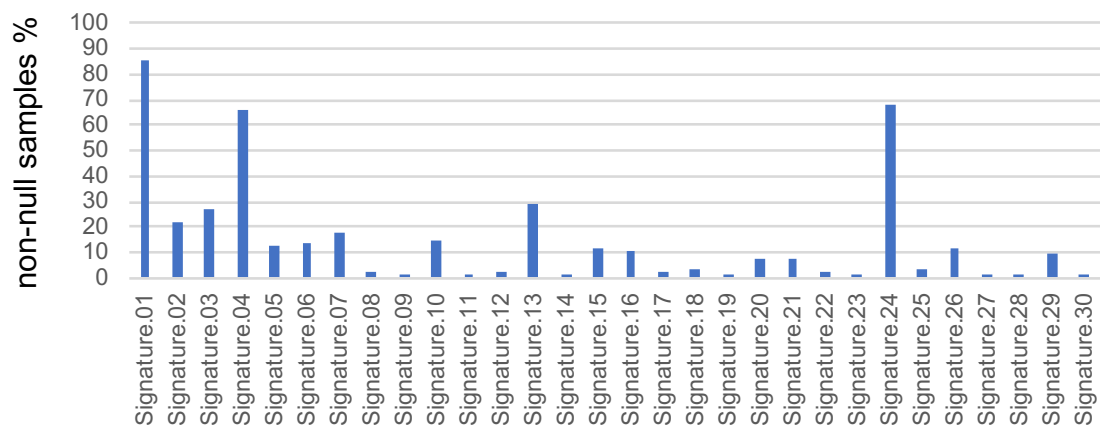
KIRC



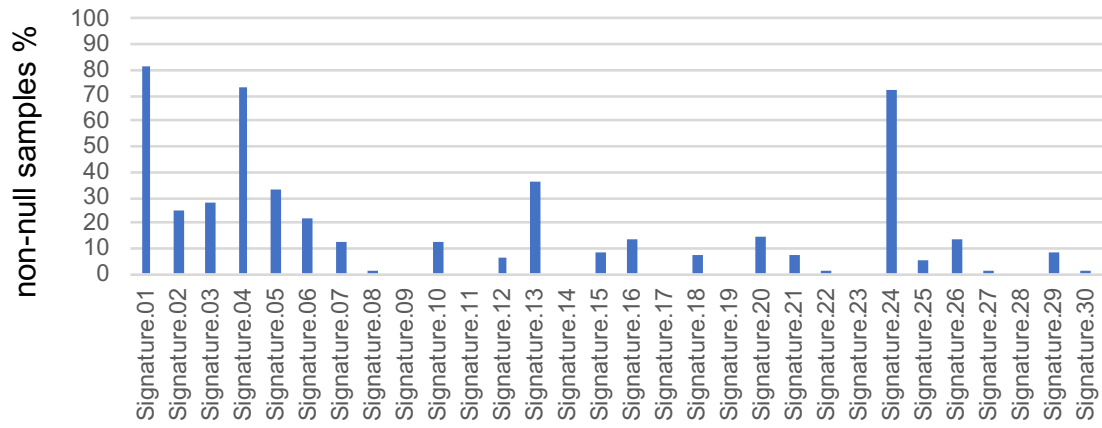
KIRP



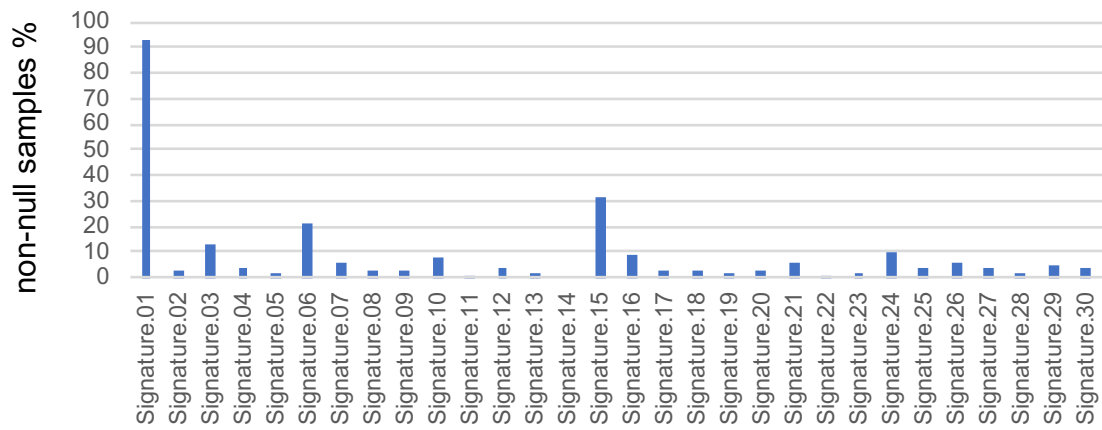
LUAD



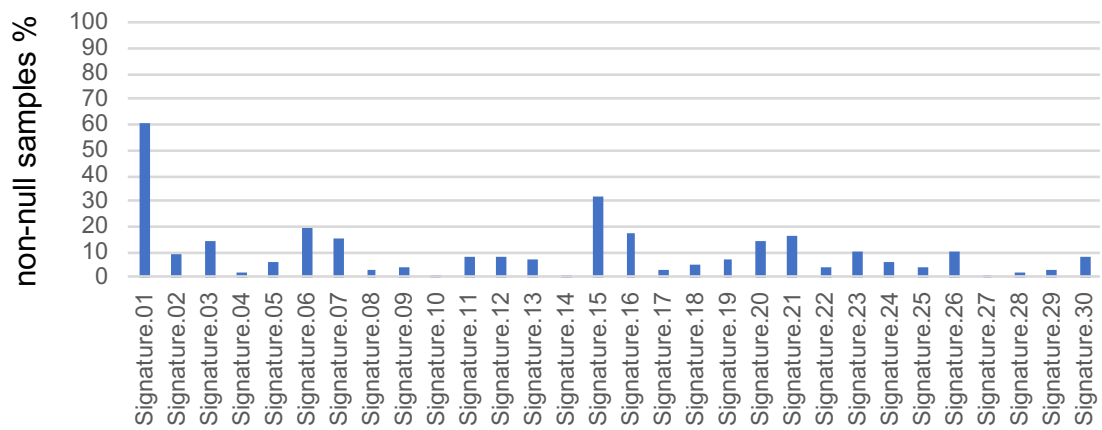
LUSC



PRAD

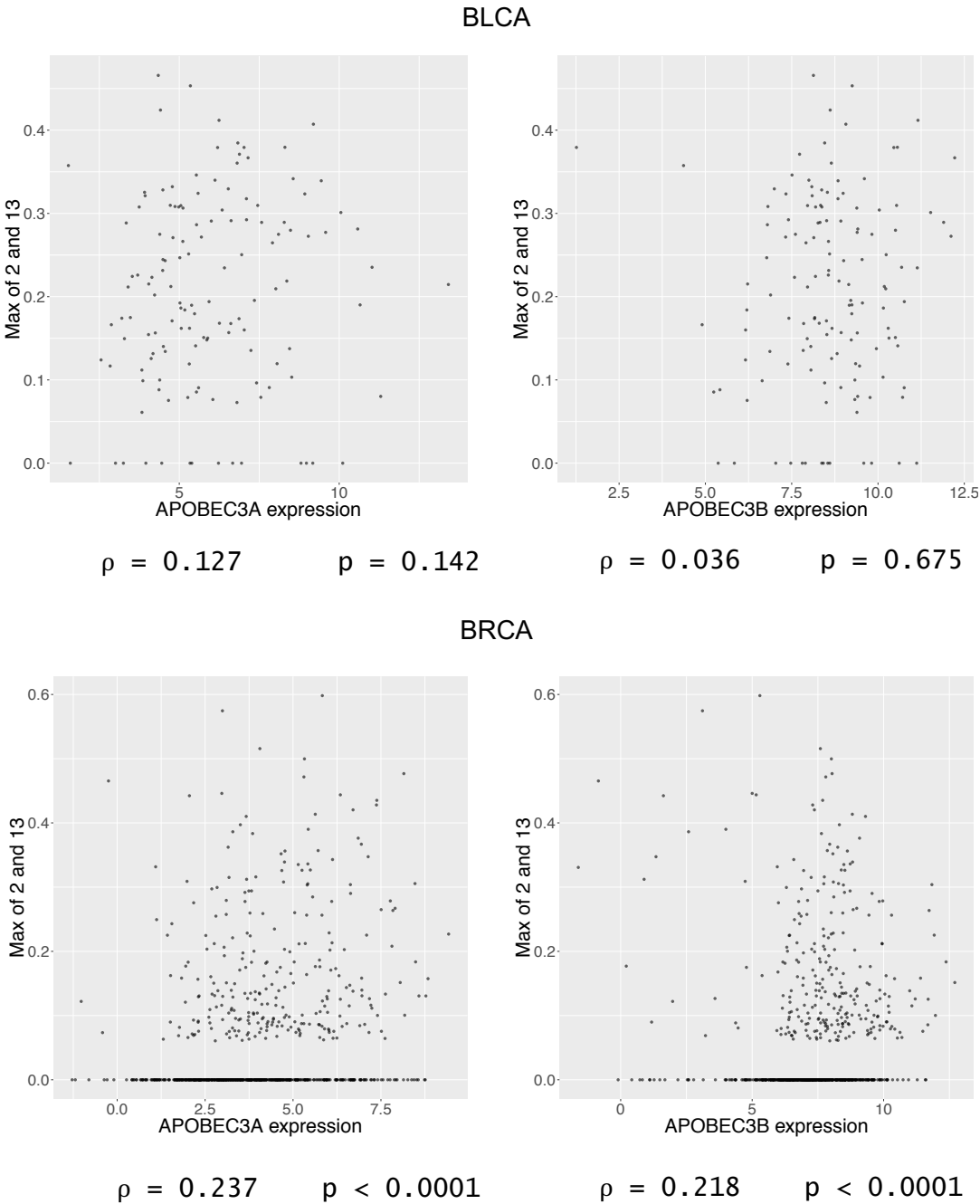


THCA

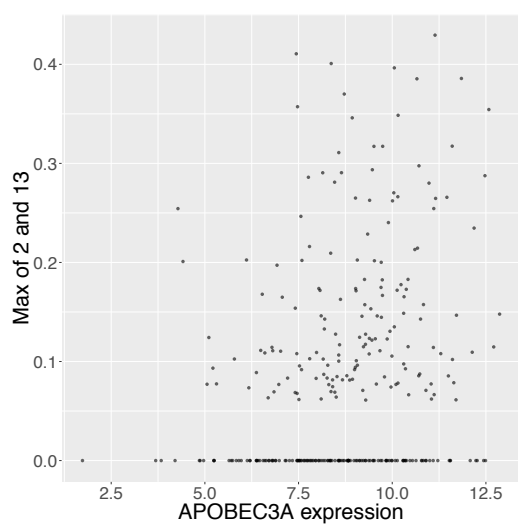


Percentage of samples with a non-null deconstructSigs score for each COSMIC signature. Scores are the result of the deconstructSigs algorithm using all 30 COSMIC mutational signatures.

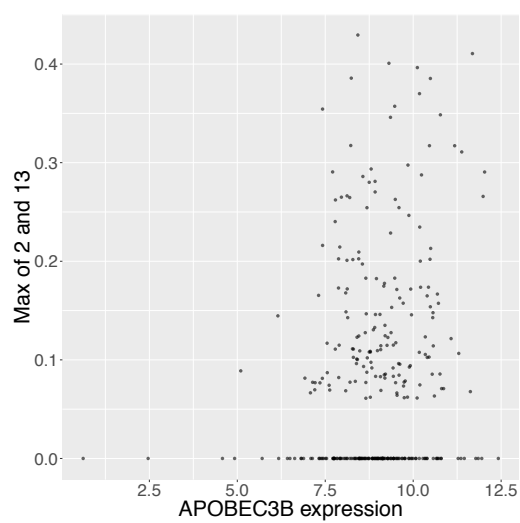
Figure 25: Relationship between the APOBEC3A and 3B and deconstructSigs Scores



HNSC

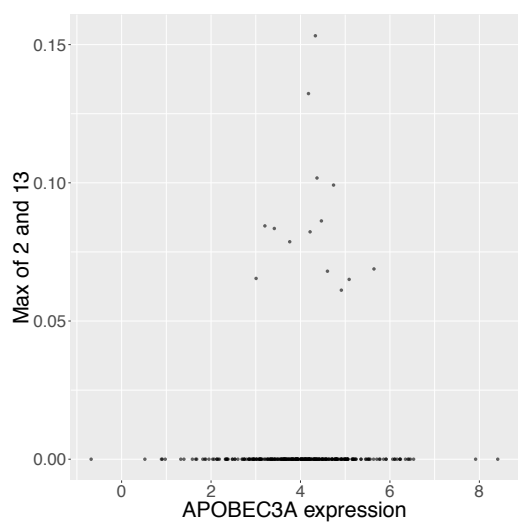


$\rho = 0.233$ $p < 0.0001$

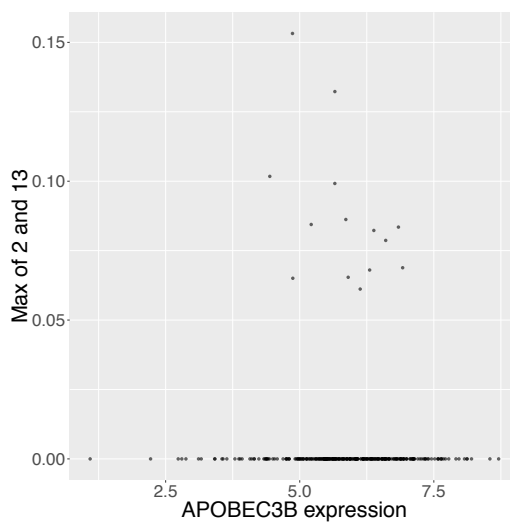


$\rho = 0.125$ $p < 0.05$

KIRC

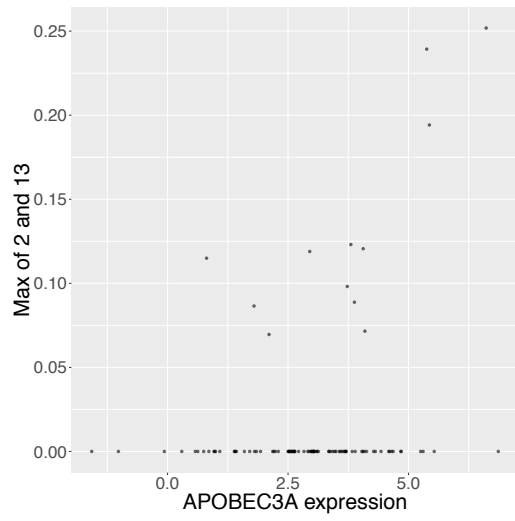


$\rho = 0.065$ $p = 0.259$

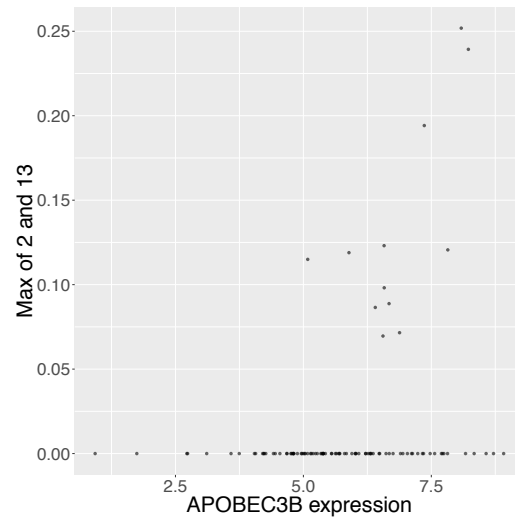


$\rho = -0.021$ $p = 0.719$

KIRP

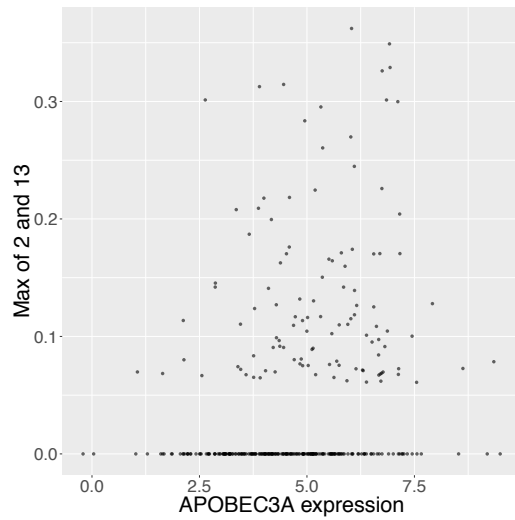


$\rho = 0.202$ $p < 0.05$

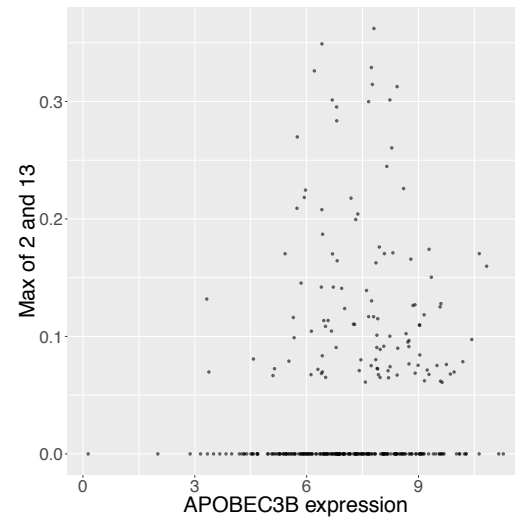


$\rho = 0.304$ $p < 0.005$

LUAD

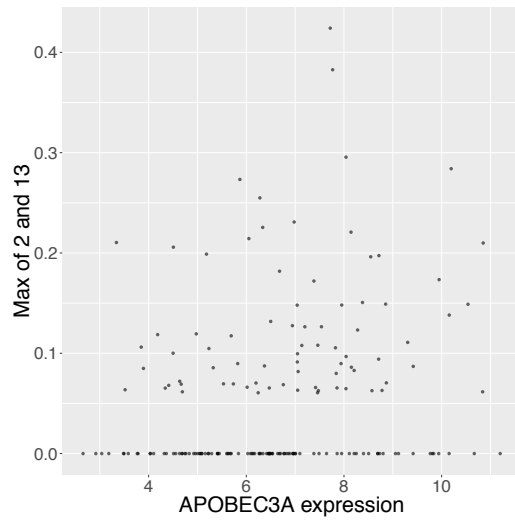


$\rho = 0.276$ $p < 0.0001$

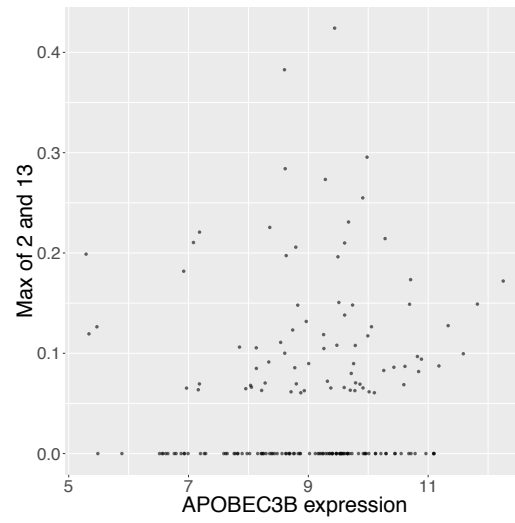


$\rho = 0.171$ $p < 0.005$

LUSC

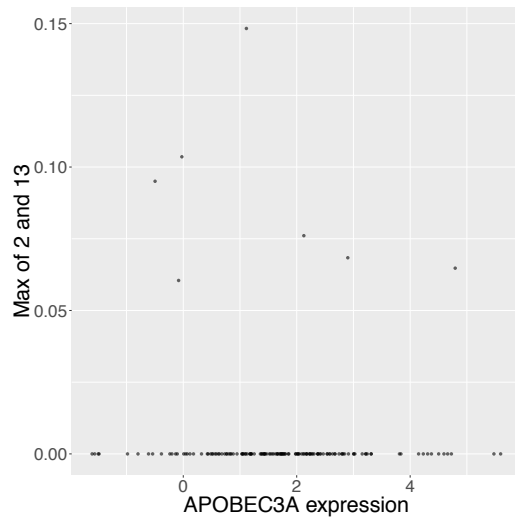


$\rho = 0.229$ $p < 0.005$

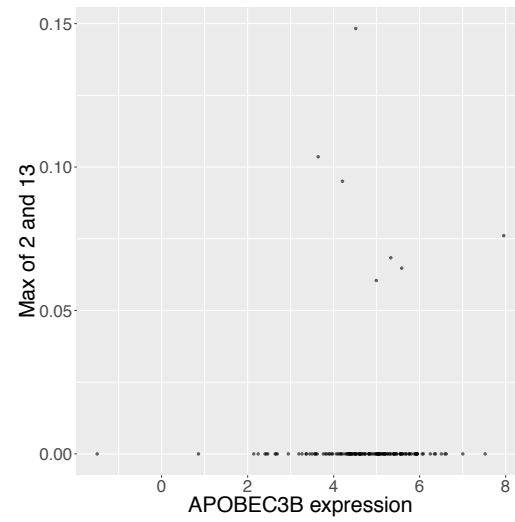


$\rho = 0.135$ $p = 0.075$

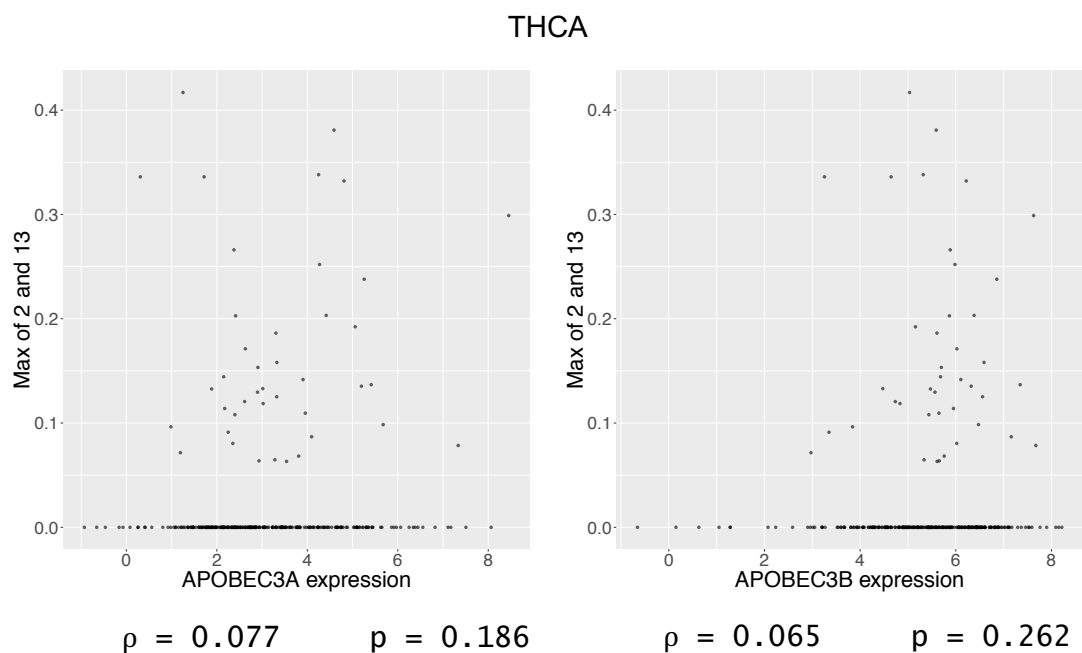
PRAD



$\rho = -0.054$ $p = 0.522$



$\rho = 0.028$ $p = 0.731$



Correlation between the deconstructSigs scores and A3A/A3B expression in 9 cancer types. deconstructSigs scores are the maximum of signatures 2 and 13 in an analysis generating scores for all 30 COSMIC signatures. Mutation data from Alexandrov et al., 2013. Expression values are log 2 transformed, normalised mRNA count from TCGA. Spearman's rho is calculated and displayed alongside p values.

3.7.4 *Enrichment of Curated Mutational Patterns in Samples*

Since deconstructSigs is based on a multiple regression algorithm, it will initially consider all 30 COSMIC signatures when attempting to fit a model. Although this makes mathematical sense, we know a priori that some signatures are not present in some cancers and thus should not be considered in the regression. Most signatures are, in fact, not present in the majority of cancers. The inclusion of these signatures for every cancer potentially adds noise to our end result, and here we address this issue. Using details about the presence of mutational signatures in different cancers available on the COSMIC platform, we manually curate lists of mutational signatures to use for the 9 cancer types analysed (Table 7). deconstructSigs is then used with these curated lists of signatures to analyse the same 9 cancers' mutational profiles.

Raw scores from deconstructSigs are plotted as jitter in Figure 26. The non-null sample percentage for each signature is shown in Figure 27. Here again, signature 1 is highly present across all cancer types. Signatures 2 and 13 are present in BLCA, BRCA, HNSC, LUAD, and LUSC. With the removal of many signatures from the input to deconstructSigs, and by association, parameters in the regression model, more samples show a non-null score across all signatures. This is expected, as with the use of strictly positive normalised coefficients in a regression model, the fewer the number of coefficients, the larger the coefficients overall. Here, coefficients are normalised between 0 and 1 with a cut-off at 0.06. Fewer coefficients would mean that each coefficient would have a larger probability of being above the cut off, resulting in more samples with non-null final

scores. Comparing the percentage of BRCA samples with non-null deconstructSigs values between using all 30 COSMIC signatures and using only curated ones, every signature showed a higher proportion of non-null samples using the curated signatures with the sole exception of signature 13, which showed only 0.14 percentiles less using the curated signatures. It is thus safe to say that using only curated signatures effectively decreases the number of samples showing a null deconstructSigs score.

In line with the previous methods, scatter plots for the deconstructSigs scores in relation to A3A and A3B expression are generated (Figure 28), along with the Spearman's correlation coefficients and associated p values. Using deconstructSigs with curated lists yields significant correlation with A3A in BRCA, HNSC, KIRP, LUAD, and LUSC, and with A3B in BRCA, KIRP, and LUAD. Comparing with using all signatures, we see the same cancer types showing significant correlation between deconstructSigs scores and both A3A and A3B. As with the analysis using all 30 signatures, KIRP has a relatively low percentage of samples with non-null presence of signatures 2 and 13. However, the percentage here is higher than before, with 10% of samples having a non-null value. Nevertheless, the result for KIRP must still be taken with this in consideration.

With fewer null samples skewing the overall distribution of the results, we observe a slight overall decrease in the correlation between deconstructSigs scores and A3A and A3B expression in different cancers when compared to using all signatures. The only exceptions to this are KIRP with A3A and LUAD with A3B.

This raises the question as to whether using curated signatures truly helped us obtain a better signal.

While using curated signatures decreases the percentage of null samples, it does not necessarily imply that it leads to a better analysis. One argument against the use of curated signatures is in fact the attenuation of filtering by using the 0.06 threshold. The threshold is set empirically by the authors of `deconstructSigs`. They used 10 signatures for analysis and found that 0.06 best filters out false positive results while retaining true positives. By using fewer signatures, or coefficients, we are effectively evading the filtering of false positives and potentially introducing them into our data. This is because with fewer coefficients, each coefficient would be more represented overall. While this practice may increase the statistical power of any test done by not having null values skewing the distribution, the values in place of null may not be accurate. Although, since the threshold was originally developed by empirical testing with 10 signatures, one can argue that by using 30 signatures, 0.06 may be too stringent as a threshold since the higher the number of signatures, the lower each signature's score will be on average.

Nevertheless, the paradigm behind using curated signature still holds – we should not include signatures that are not biologically present into the calculation. We may lose accuracy in the lower values that are close to the threshold, but that is not to say that the entire analysis is erroneous. Besides, filtering is always a trade-off between type 1 and type 2 errors. The authors of `deconstructSigs` found a reasonable threshold for specific conditions leading to specific percentages of

false positives and false negatives. It does not indicate that this threshold is universally ideal. Ideally, the user should define the threshold for his or her usage specifically using information about the samples, the number of signatures, and the tolerance for type 1 and type 2 errors. However, this process would require much additional empirical testing or a theoretical approach to set a threshold. This is beyond the scope of the current study.

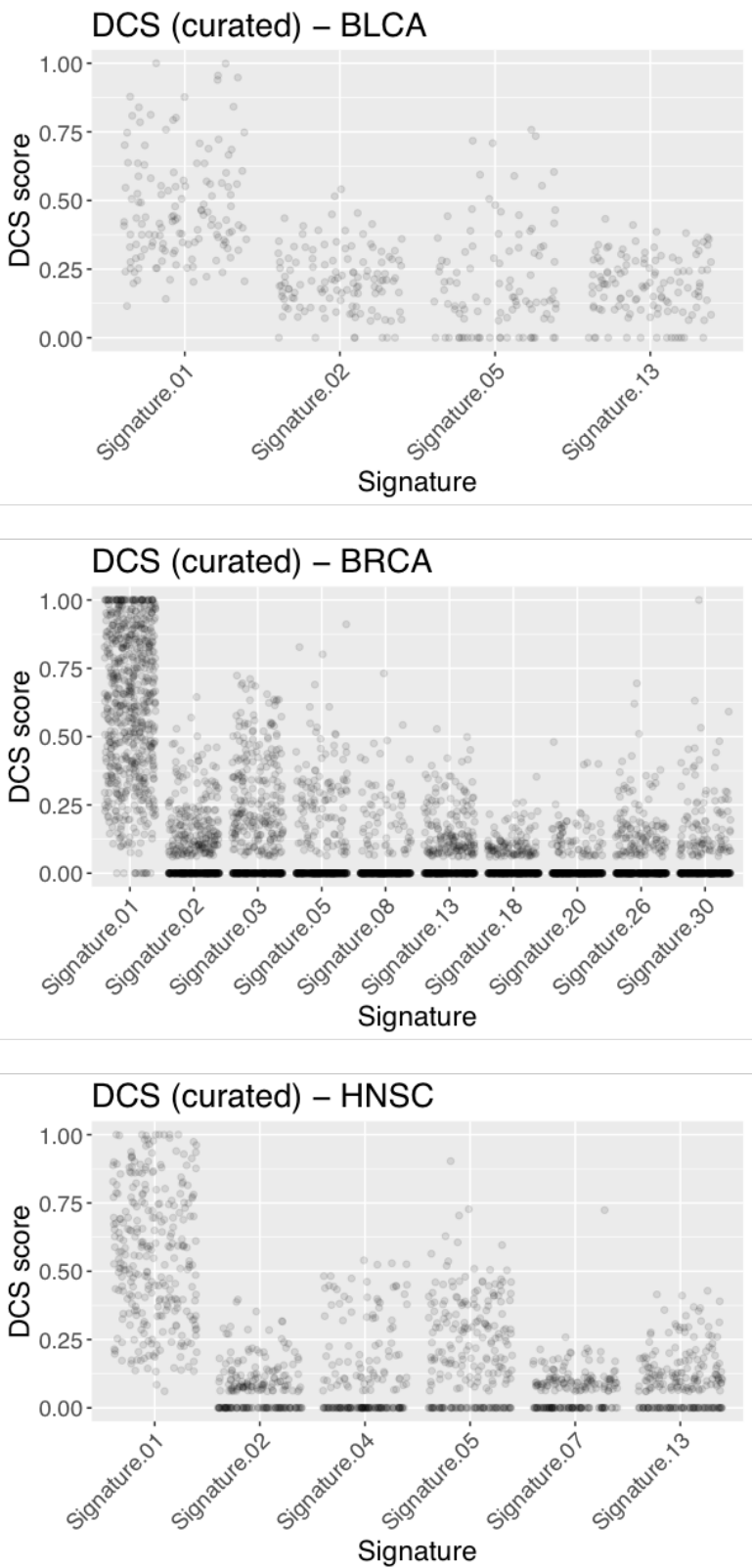
What we conclude here is that the use of curated signatures with deconstructSigs as it is relatively attenuates the effect of filtering when compared to using all signatures. This leads to the inclusion of more false positive results in the deconstructSigs scores, which in turn leads to a weaker correlation between the scores and APOBEC expression. There is the potential for this method to be more accurate than using all 30 signatures, but the filtering must be adjusted for this.

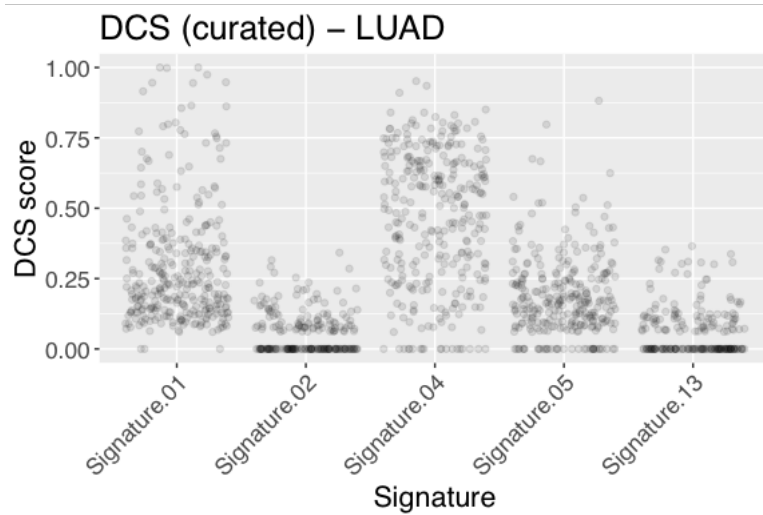
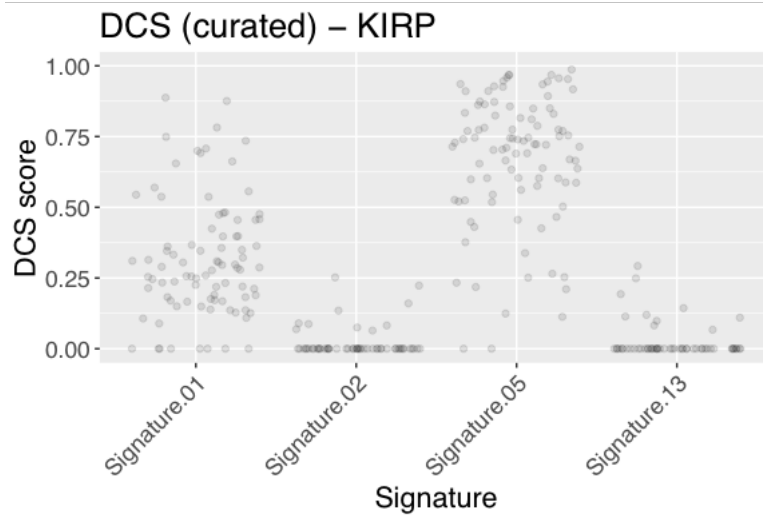
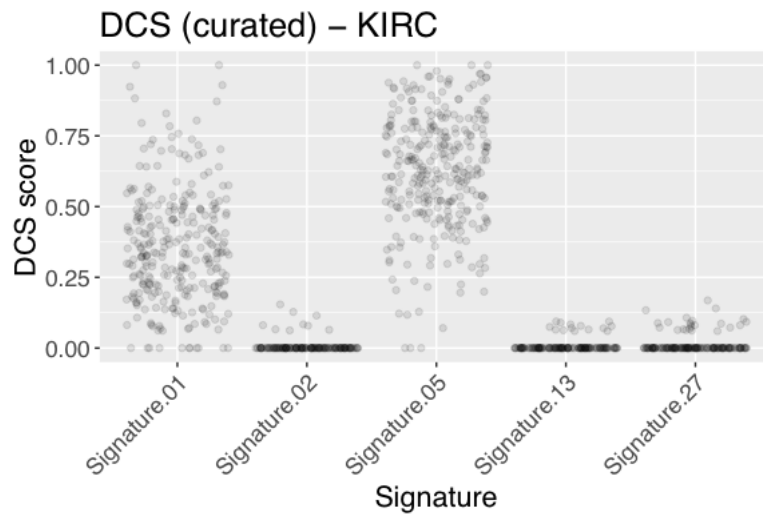
Table 7: COSMIC Mutational Signatures Selected for Analysis

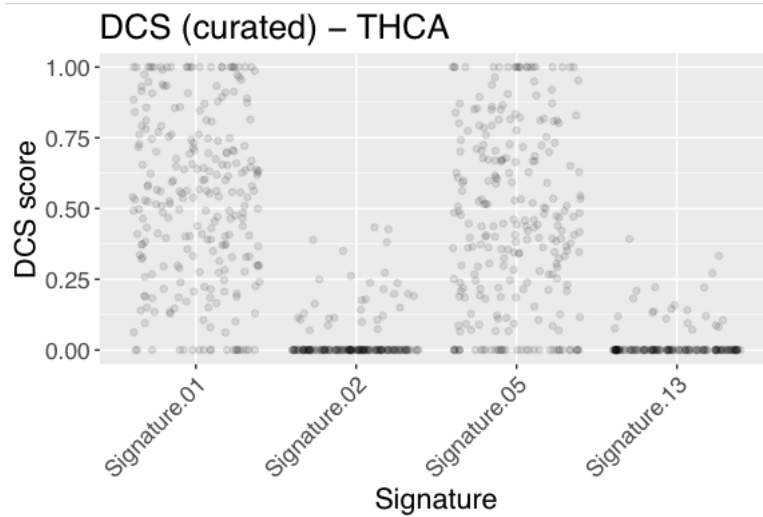
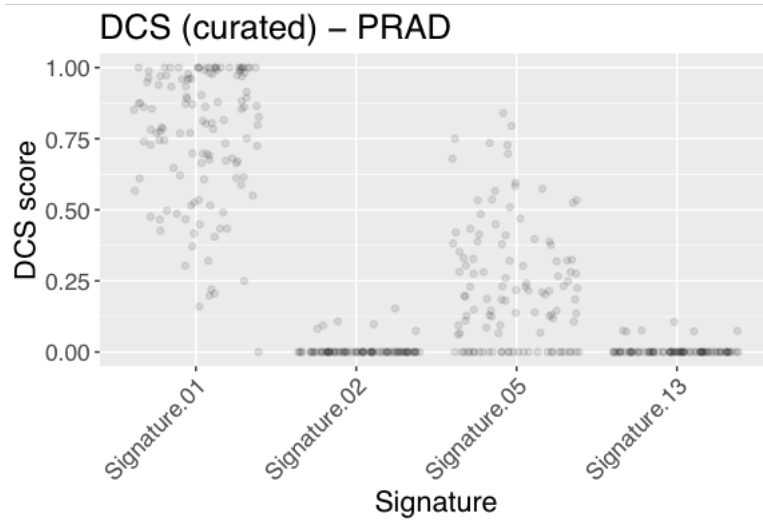
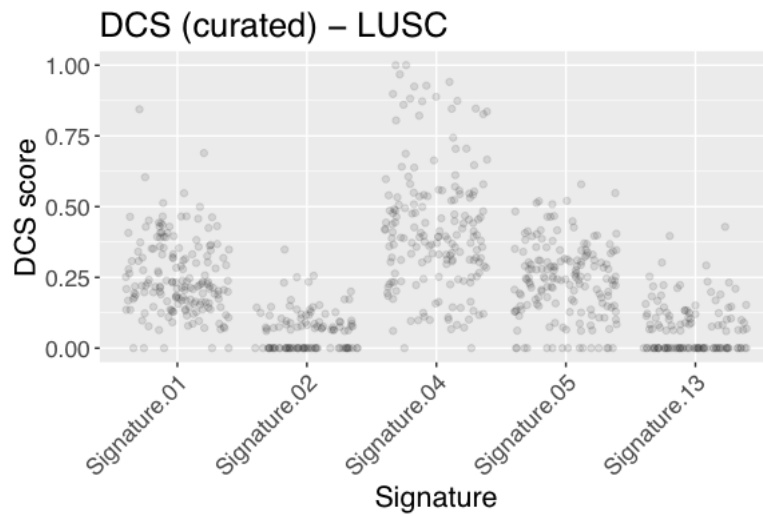
	Selected signatures
BLCA	1, 2, 5, 13
BRCA	1, 2, 3, 5, 8, 13, 18, 20, 26, 30
HNSC	1, 2, 4, 5, 7, 13
KIRC	1, 2, 5, 13, 27
KIRP	1, 2, 5, 13
LUAD	1, 2, 4, 5, 13
LUSC	1, 2, 4, 5, 13
PRAD	1, 2, 5, 13
THCA	1, 2, 5, 13

COSMIC signatures selected manually based on their reported presence in each cancer type. These curated lists of signatures are used in the analysis of mutational signatures using deconstructSigs with only relevant signatures in each cancer type. We curated these lists using information consolidated and provided by the COSMIC platform. The exception this is signatures 2 and 13, which are analysed in all cancers, even if no presence is reported by COSMIC.

Figure 26: deconstructSigs Scores for COSMIC Signatures Curated for each CancerType

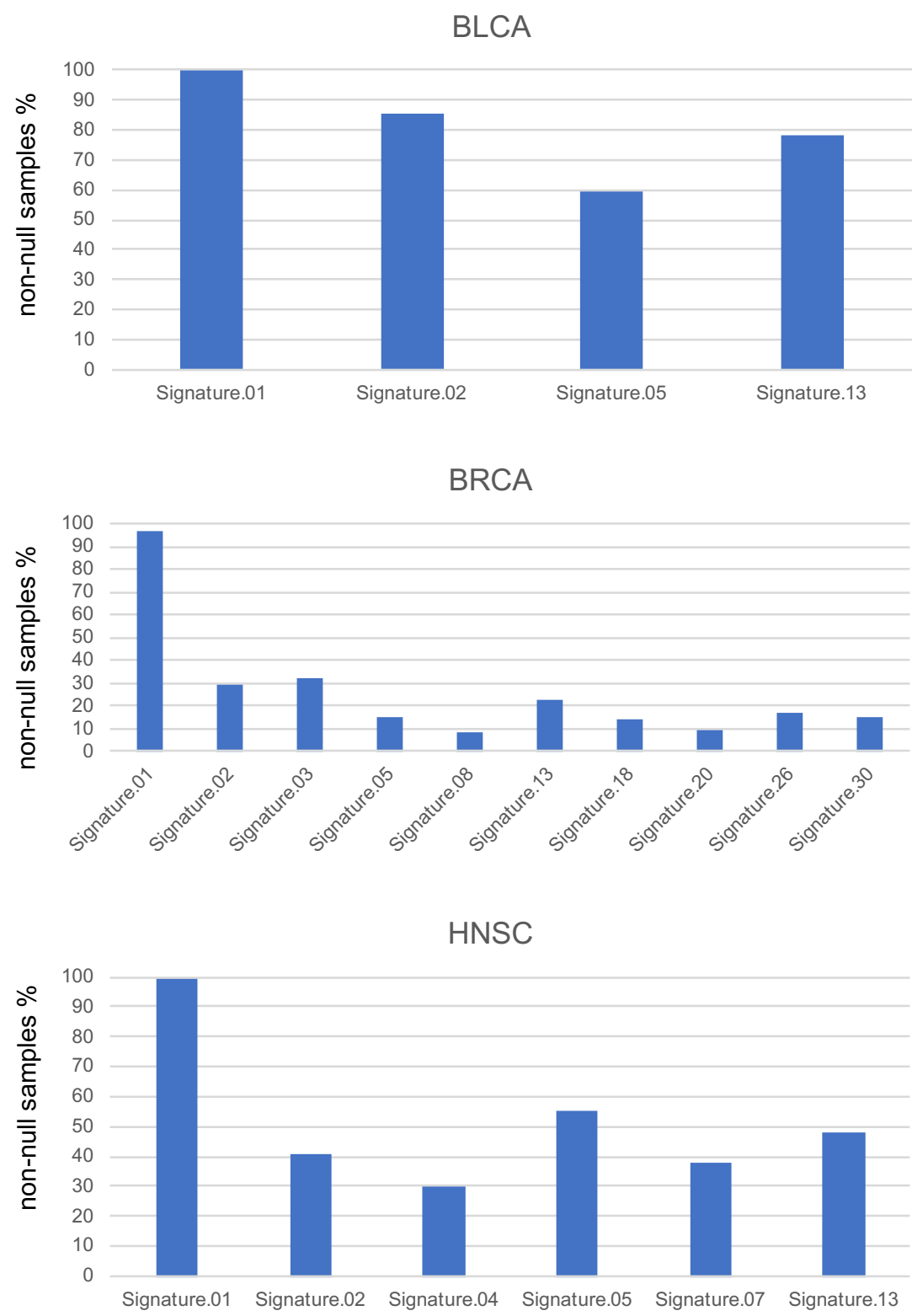


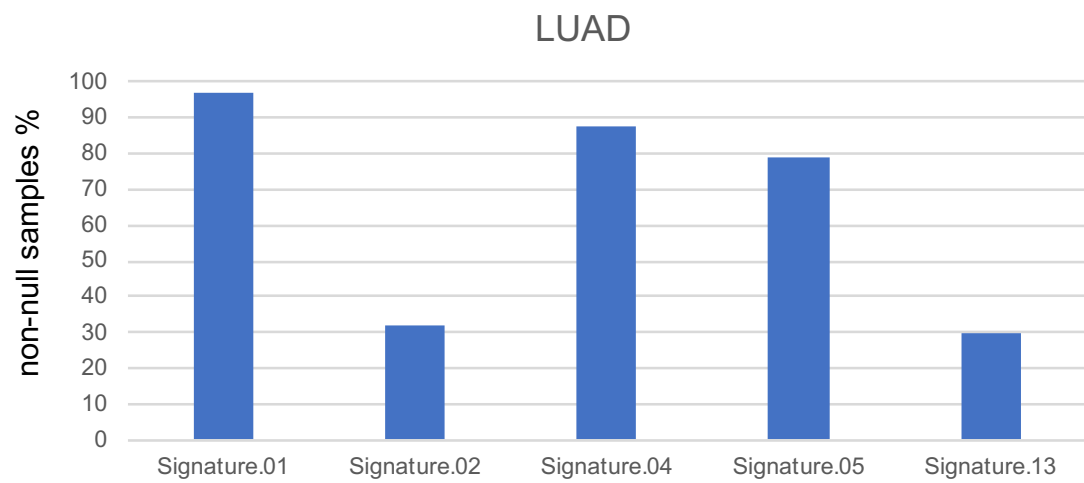
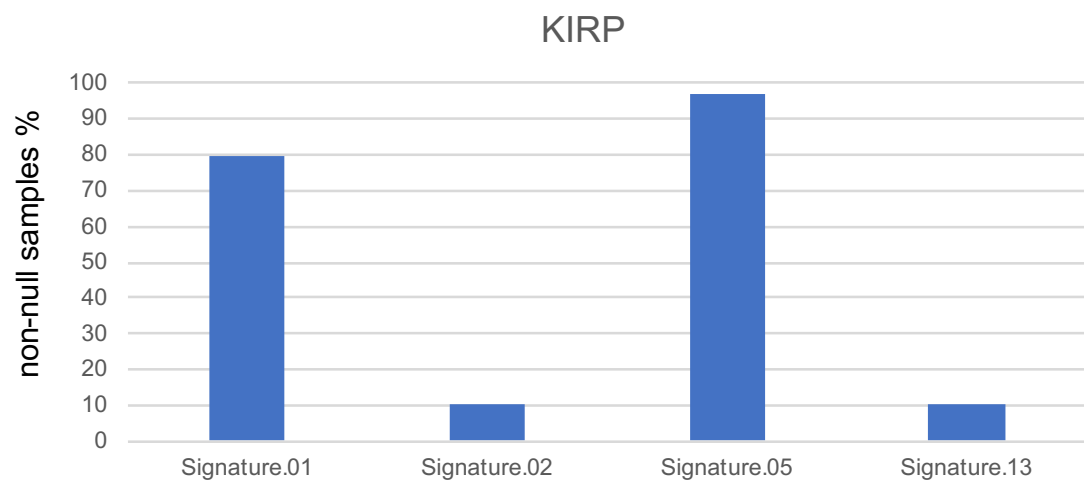
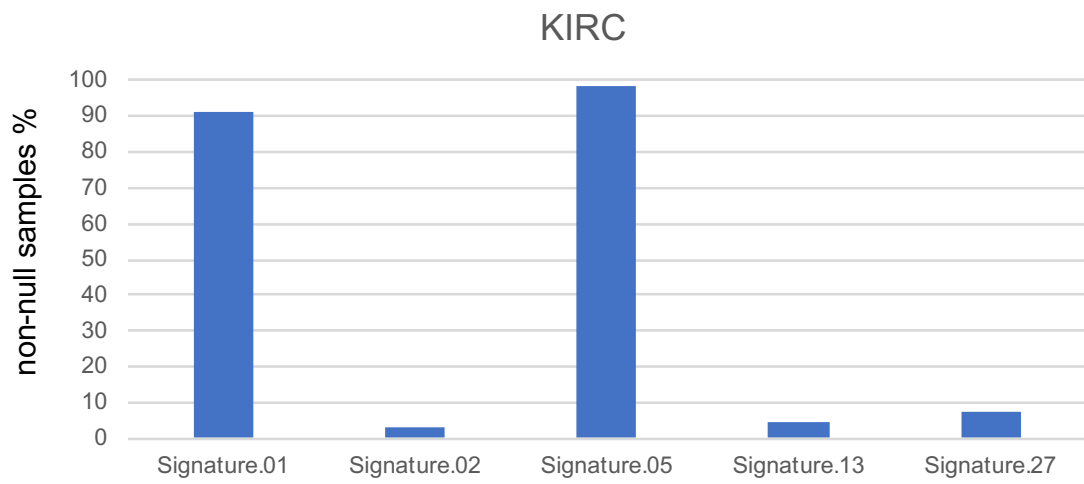


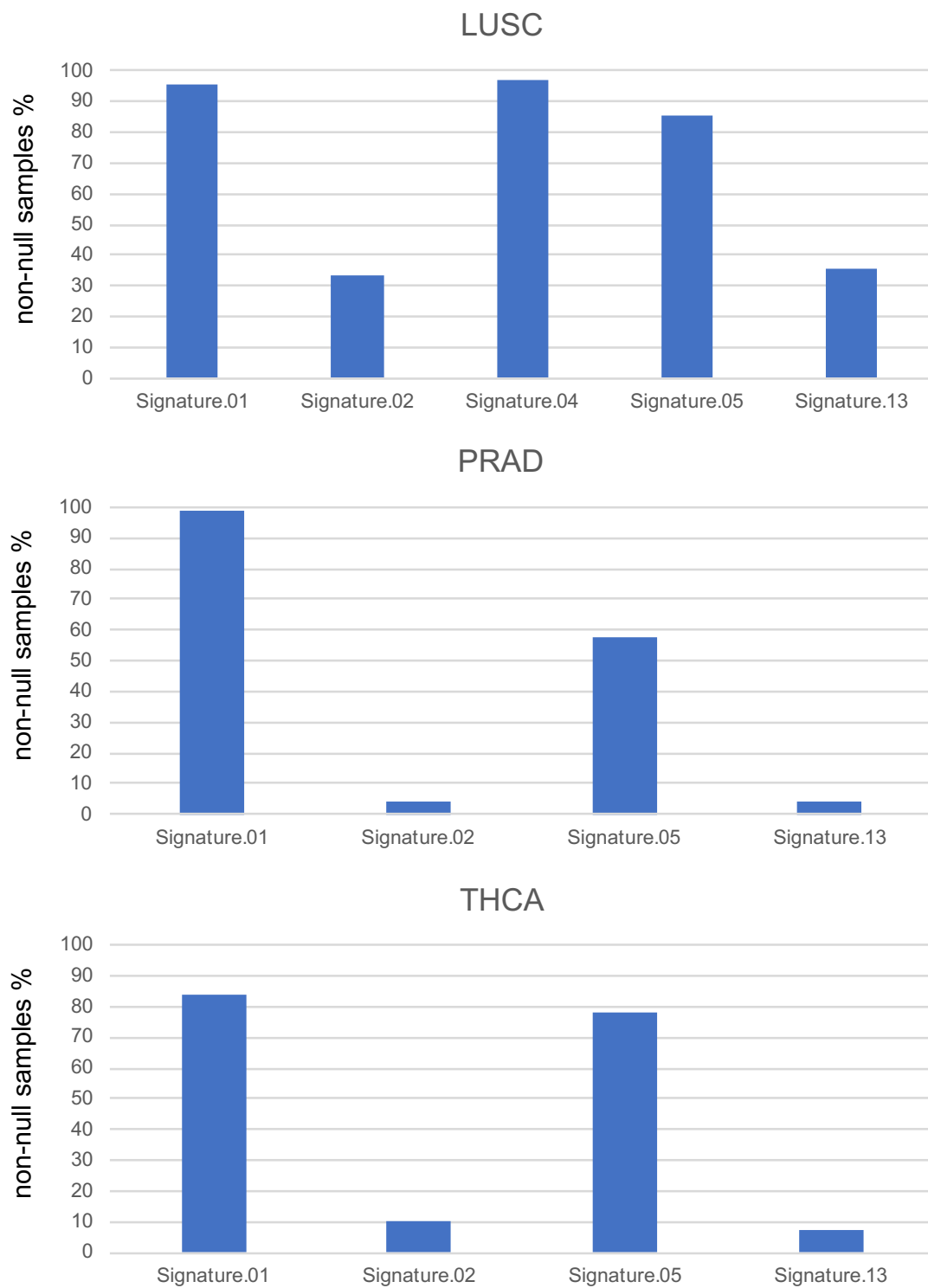


Scores from the deconstructSigs (DCS) algorithm using only manually curated COSMIC mutational signatures for each cancer type. Values represent the level of enrichment of a mutational signature in single samples.

Figure 27: Percentage of Non-Null deconstructSigs Scores for the 30 COSMIC Signatures

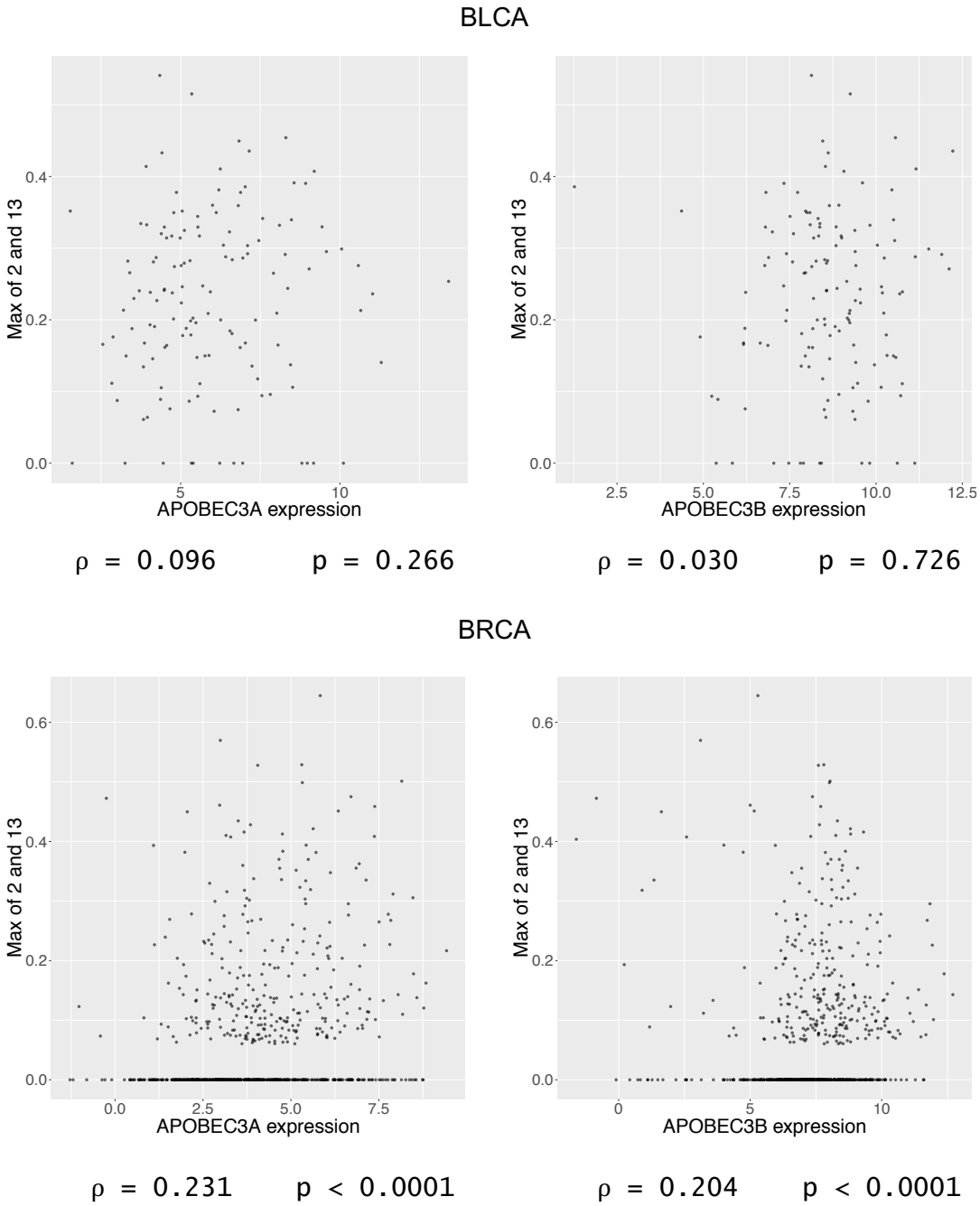




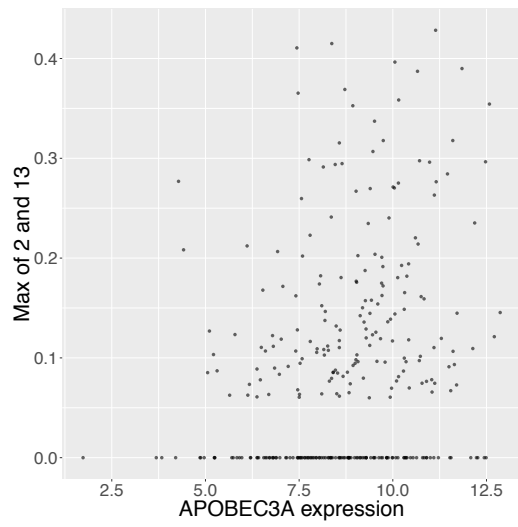


Percentage of samples with a non-null deconstructSigs score for curated COSMIC signature in each cancer type. Scores are the result of the deconstructSigs algorithm using only curated COSMIC mutational signatures.

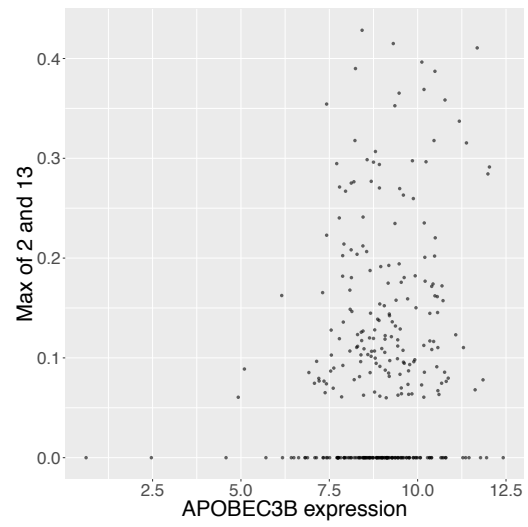
Figure 28: Relationship between the APOBEC3A and 3B and deconstructSigs Scores Using only Curated Signatures



HNSC

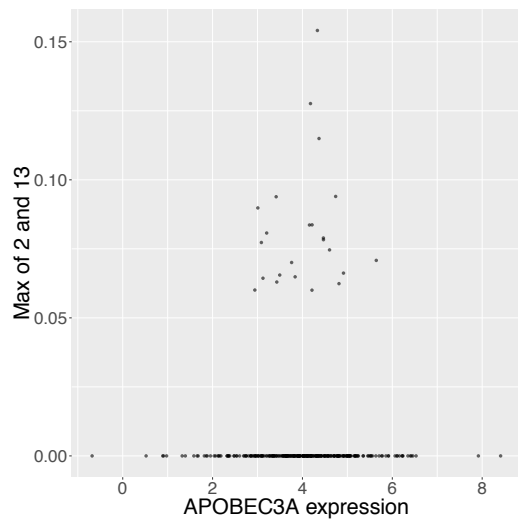


$\rho = 0.227$ $p < 0.0001$

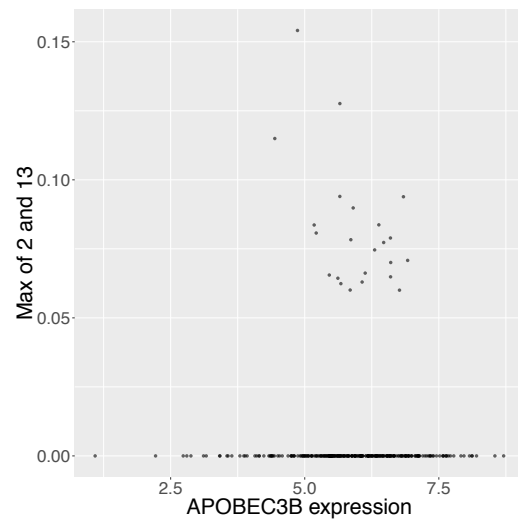


$\rho = 0.127$ $p < 0.05$

KIRC

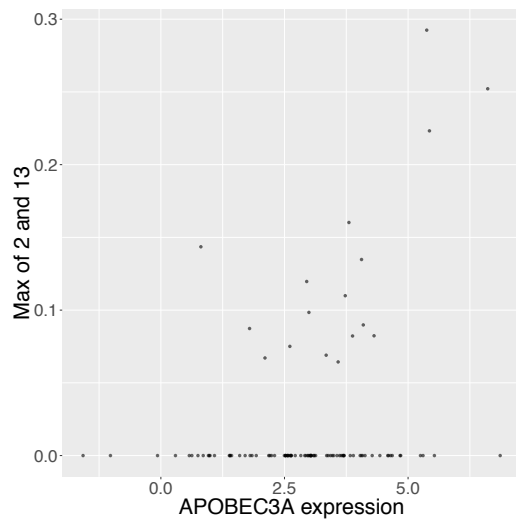


$\rho = -0.002$ $p = 0.973$

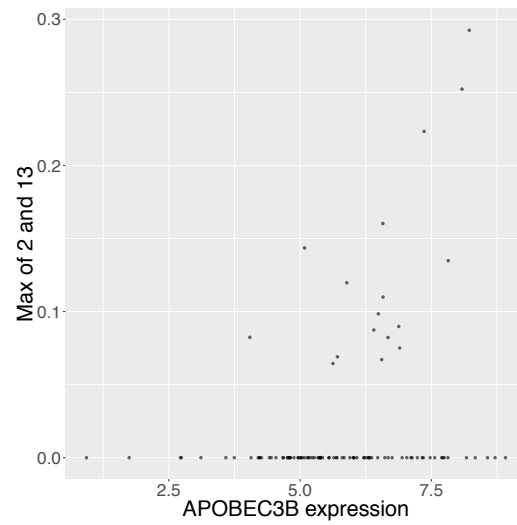


$\rho = 0.011$ $p = 0.851$

KIRP

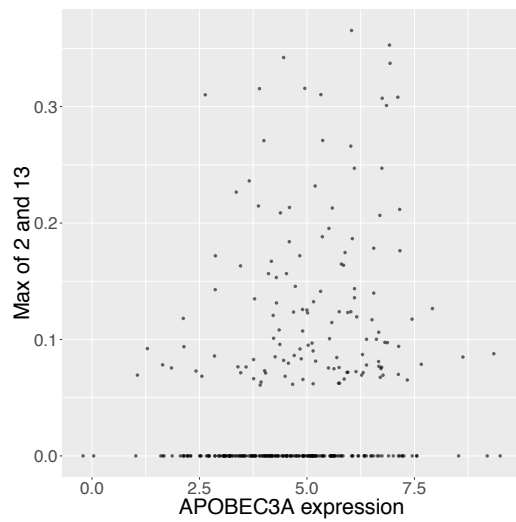


$\rho = 0.220$ $p < 0.05$

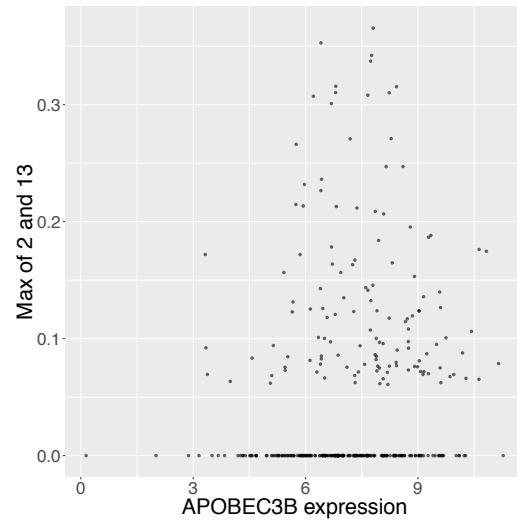


$\rho = 0.272$ $p < 0.01$

LUAD

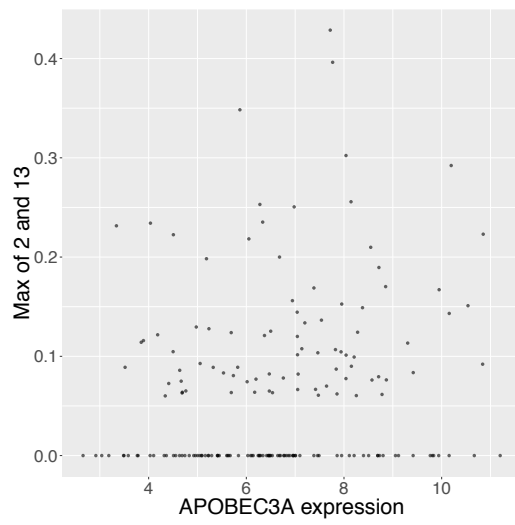


$\rho = 0.267$ $p < 0.0001$

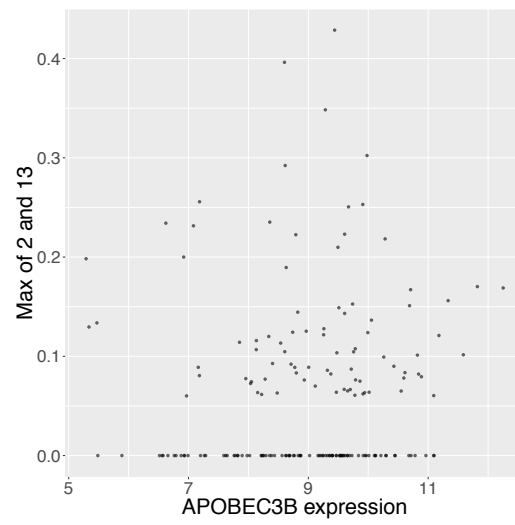


$\rho = 0.179$ $p < 0.005$

LUSC

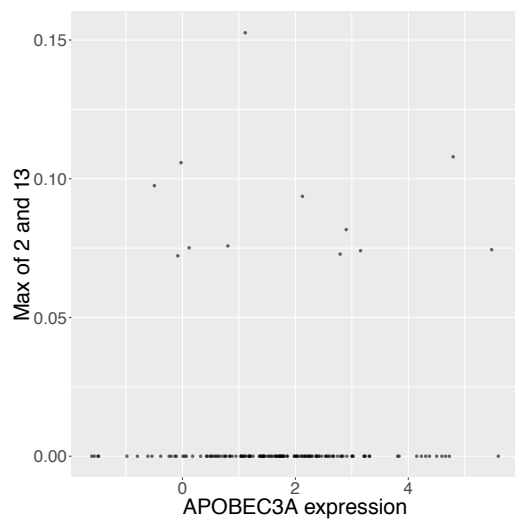


$\rho = 0.183$ $p < 0.05$

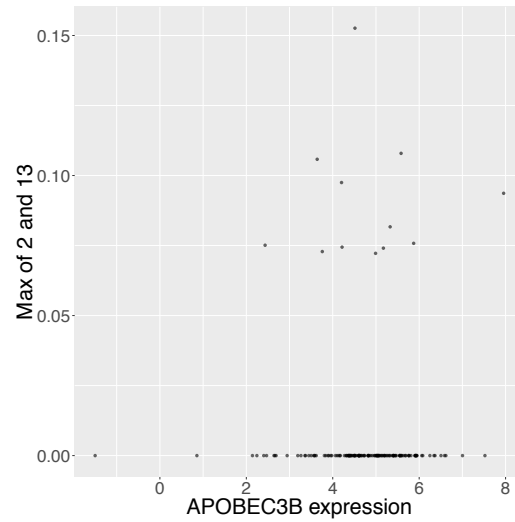


$\rho = 0.111$ $p = 0.142$

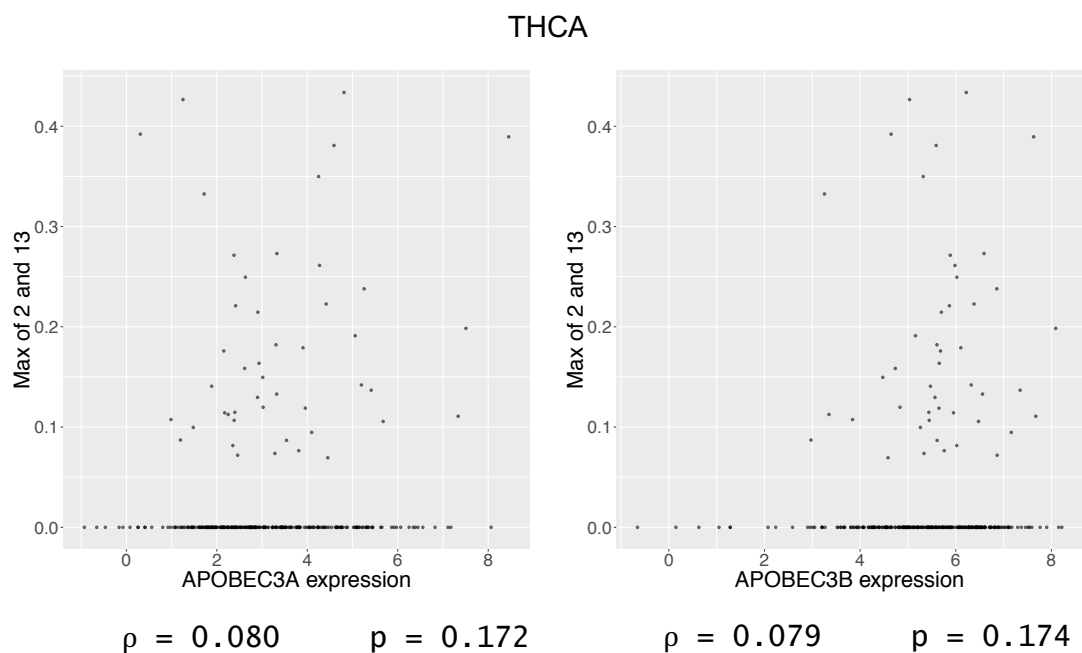
PRAD



$\rho = 0.006$ $p = 0.942$



$\rho = -0.022$ $p = 0.793$



Correlation between the deconstructSigs scores and A3A/A3B expression in 9 cancer types. deconstructSigs scores are the maximum of signatures 2 and 13 in an analysis generating scores only for selected COSMIC signatures. Mutation data from Alexandrov et al., 2013. Expression values are log 2 transformed, normalised mRNA count from TCGA. Spearman's rho is calculated and displayed alongside p values.

3.7.5 Comparing the Four Methods

Overall, despite the differences between the four methods, they all suggest that the APOBEC mutational signature is present and correlated with A3A and A3B mRNA count in multiple cancers. Particularly, BRCA and LUAD show significant correlation with both A3A and A3B using all four methods.

As an additional control, we randomly select 100 genes in TCGA data and calculate their correlation with the mutational signature result from each method. This verifies if the distribution of correlation scores using random genes is compact and centred around 0. A wide or heavily skewed distribution is indicative of the method generating noise or bias, as randomly selected genes should not have a high chance of correlating with the APOBEC mutational signature. The distributions are plotted in Figure 29 with A3A and A3B expression overlaid (A3A in green, A3B in blue). Z statistics are calculated for A3A and A3B in all cancers and for all methods (Figure 30). The vast majority of significant correlation results from the previous sections are validated here by exhibiting a high positive deviation from the mean of the random distribution. Only three results failed to show a deviation of more than 1.5 times the standard deviation of the random distribution: PRAD vs A3B using the enrichment of TCW mutations and KIRP vs A3A in both deconstructSigs methods. This effectively controls for the possibility that a significant correlation seen in previous sections is due to random chance alone. We know that this is not the case in the majority of our results.

A summary of the results from the mutational signatures analyses is available in Table 8. This table briefly shows whether A3A or A3B expression is significantly

correlated with the mutational signature according to the four methods used. It considers the controls that may exclude a statistically significant correlation from being interpreted as biologically significant, namely whether a mutational signature is present at all and whether the correlation seen is different enough from that seen in random genes.

From here, after excluding the three results in the last control, we see that the BRCA and LUAD mutational signatures still show significant correlation with both A3A and A3B expression across all four methods. HNSC's signature shows significant correlation with the expression of A3A in all methods except the enrichment of TCW mutations. The same methods lead to significant correlation between the expression of A3B and the mutational signature in KIRP. The mutational signature in LUSC contains significant correlation with the expression of A3A in all methods except the count of TCW mutations. Since we cannot judge which method is objectively superior, we treat a result that appears in multiple methods as being more significant. Thus, we use these methods to cross-validate one another. Simplifying the interpretation, we can say that BRCA and LUAD show the most evidence of correlation between APOBEC mRNA expression and its mutational signature, followed by HNSC, KIRP, and LUSC. In turn, this is direct evidence of APOBEC mutagenesis in these cancers and that the increase in APOBEC expression in these cancers drives APOBEC mutations.

Combined with expression data from earlier analyses, we obtain a clear picture of which cancers have the greatest APOBEC activity. Among cancers for which the mutational signature correlates with A3A expression, BRCA and HNSC have

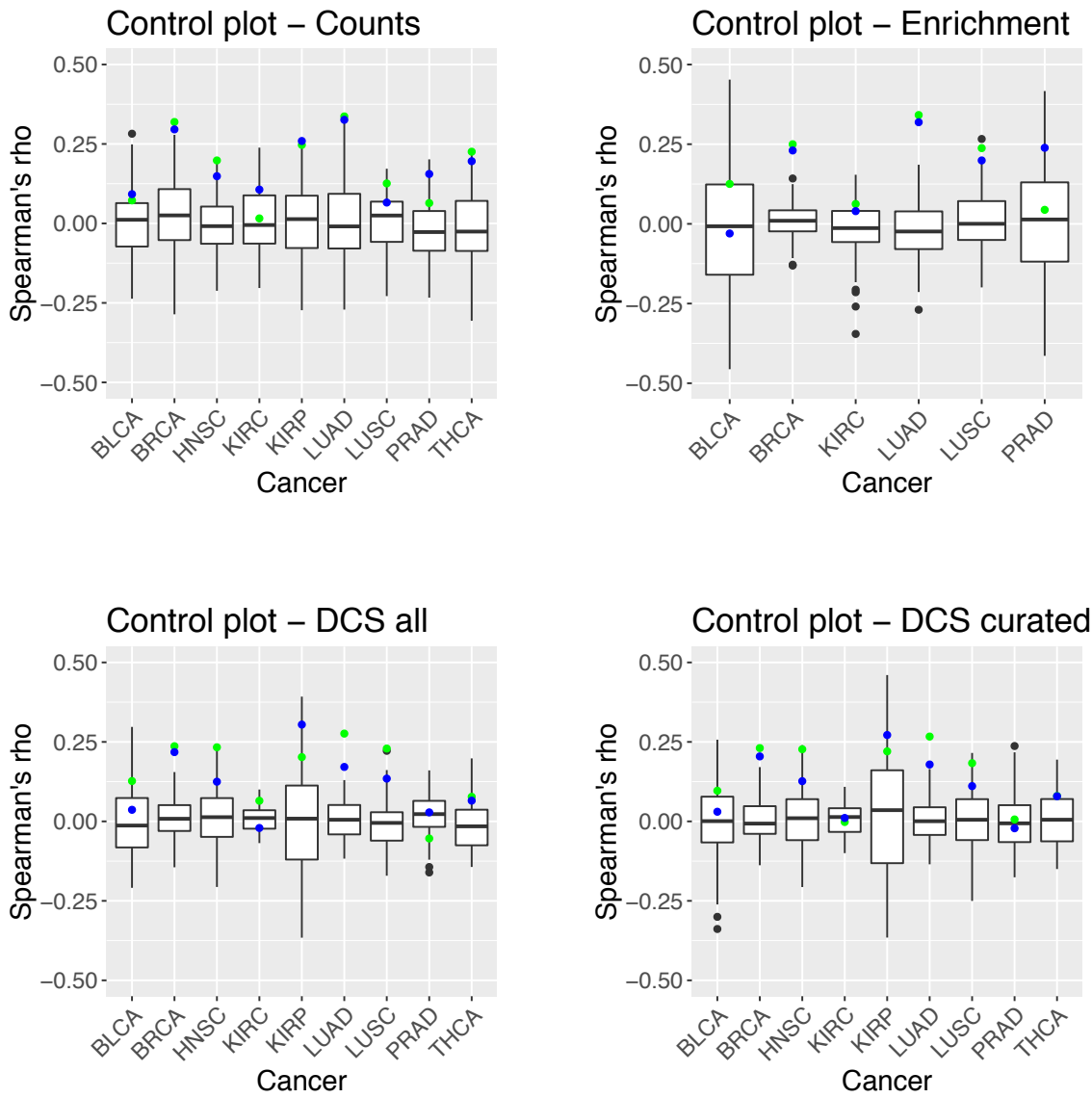
positive differential expression between tumour and normal samples. While LUAD and LUSC show correlation, they have a negative differential expression between tumour and normal samples, meaning that A3A is less expressed in tumour samples. This does not invalidate the fact that a higher expression of A3A drives APOBEC mutations however. It only indicates that A3A is underexpressed in tumours and may have a lower overall activity. As for A3B, it is overexpressed in all cancers exhibiting a significant correlation between the mutational signature and its mRNA expression. This leads us to believe that BRCA, LUAD, and KIRP are the most relevant cancer types to analyse in terms of APOBEC involvement. They exhibit high and preferential expression of active APOBECs in tumour samples and this expression correlates with APOBEC mutagenesis.

An earlier study has identified the cancer types in which TCW to TTW or TGW mutations are more prevalent (Burns et al., 2013b). The study lists BLCA, CESC, LUSC, LUAD, HNSC, and BRCA. Of these six cancer types, four are shown here to have high correlation between the APOBEC mutational signature and the expression of A3A and/or A3Bm and this was confirmed by at least three methods for estimating the mutational signature. BLCA did not show positive correlation and CESC was not included in the analysis because of data availability. This suggests that, using the empirical definition of the APOBEC mutational signature as TCW to TTW or TGW mutations, the cancer types with the highest prevalence of the signature also shows correlation between the signature and APOBEC expression. It further validates the fact that the signatures are real and caused by APOBEC mutagenesis. The reason why the correlation is not seen in cancers

with a lower presence of the mutational signature may be because of the lack of statistical power causing the data to be lost in background noise.

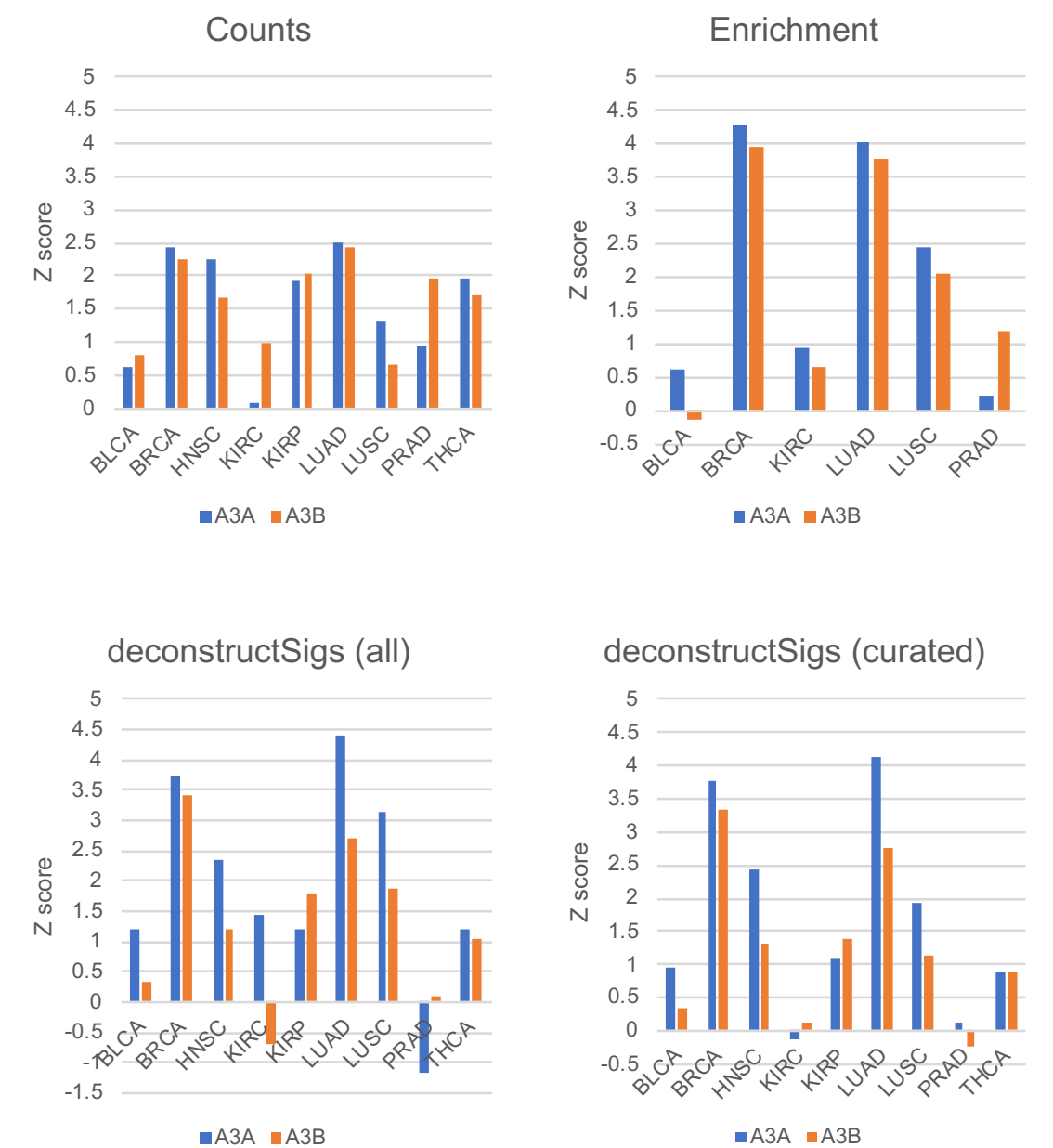
The effect is preserved when using the COSMIC signatures as well. We observe that among the five cancers with a significant presence of signatures 2 and 13, only BLCA does not show high correlation between the deconstructSigs scores and APOBEC expression. This effectively provides more evidence the association between signatures 2 and 13 and APOBEC since their presence correlates with the expression of active APOBEC genes in cancer.

Figure 29: Correlation between the Expression of Random genes and the APOBEC Mutational Signature



Correlation between the expression of randomly selected genes and the APOBEC mutational signature as reported by four methods (DCS is abbreviation for deconstructSigs). 100 genes are selected at random. For each cancer type and each method of measuring the APOBEC mutational signature, the correlation between the expression of the randomly selected genes and the mutational signature is calculated. Values represent Spearman's rho. The green marker shows the correlation between A3A and the mutational signature. The blue marker shows the same for A3B. Boxes represent the range from the first to the third quartiles. Centre line represents the median. Whiskers represent 3rd quartile plus 1.5x the interquartile range or the maximal value, whichever is smaller, and the 1st quartile minus 1.5x the interquartile range or the minimal value, whichever is larger. Dots represent values not contained within whiskers.

Figure 30: Z Scores for A3A and A3B Relative to the Distribution of Random Genes in Terms of Correlation with the APOBEC Mutational Signature



Z scores calculated for A3A's and A3B's correlation with the APOBEC mutational signature relative to the correlation of 100 random genes with the signature, using four methods. Bars represent Z scores.

Table 8: Summary of Correlation between the Expression of A3A and A3B and the Presence of the APOBEC Mutational Signature

	Counts	Enrichment	DCS all	DCS curated
BLCA				
BRCA	A/B	A/B	A/B	A/B
HNSC	A/B		A/B	A/B
KIRC				
KIRP	A/B		A/B	A/B
LUAD	A/B	A/B	A	A/B
LUSC		A/B	A	A
PRAD		B		
THCA	A/B			

Summary table showing which cancer types exhibit significant correlation between the expression of A3A/A3B and the presence of the APOBEC mutational signature using four analysis methods. ‘A’ indicates significant correlation between A3A and the presence of the mutational signature, ‘B’ for A3B, and ‘A/B’ for both A3A and A3B. Empty boxes denote an absence of correlation. Black boxes indicate that data is unavailable due to the analyses not being performed. Yellow boxes indicate that the mutational signature is not significantly present. Red boxes indicate that the correlation seen is not significantly different from the correlation between random genes and the APOBEC mutational signature. Orange boxes indicate a mix of both previous points.

3.8 The APOBEC Mutational Signature Correlates with Hypoxia in Breast Cancer

I have described the correlation between hypoxia and expression in previous sections. Taking the hypoxia analysis one step further, we calculate the correlation between the APOBEC mutational signature and the mean hypoxia signature. We consider the results from the four methods used in the previous section to infer the mutational signature and perform the same correlation calculations as previously done for APOBEC expression, but now with the expression of the mean hypoxia signature. Scatter plots for each cancer type and method combination show the relationship between the mutational signature and hypoxia in the analysed cancer types (Figure 31). Spearman's correlation coefficients and associated p values are displayed alongside.

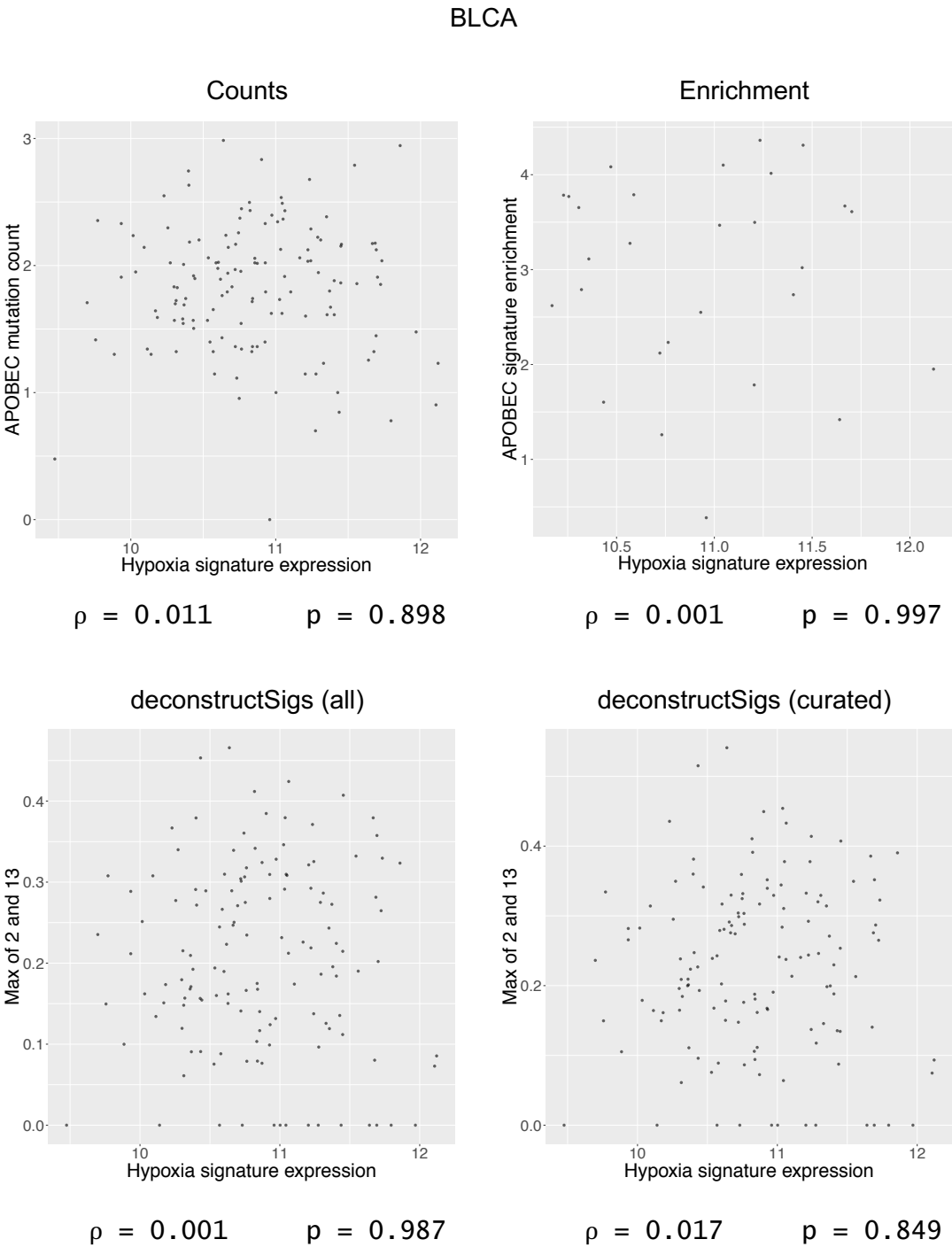
We observe that using counts of TCW mutations yields the highest number of cancer types showing significant correlation with hypoxia, with 6 cancer types: BRCA, KIRC, KIRP, LUAD, PRAD, and THCA. Enrichment of TCW mutations yields two cancer types: BRCA and LUAD. Both deconstructSigs methods yield BRCA, KIRP, and THCA. However, KIRP and THCA contain a low percentage of samples with non-null presence of signatures 2 and 13 using both methods, often under 10%. These results must be taken with this in consideration since a high proportion of null samples can heavily skew the results. We can thus only confidently affirm that BRCA has an APOBEC mutational signature that is significantly correlated with hypoxia using deconstructSigs.

As with A3A and A3B, we overlay the correlation coefficients for hypoxia over the distribution of coefficients for randomly selected genes (Figure 32) and calculate the Z-statistics (Figure 33). We observe that all of our results deviate from the mean of the random distribution by more than 1.5 times the standard deviation except the results for KIRP. Across the three methods for which KIRP showed a positive correlation with hypoxia, its correlation coefficient lies closer to the mean than 1.5 times the standard deviation. When looking at the distribution of correlation for random genes in KIRP compared to other cancer types, it is clear that the variation is large in this cancer type. The reason behind this can only be hypothesised at this time and are beyond the scope of this study.

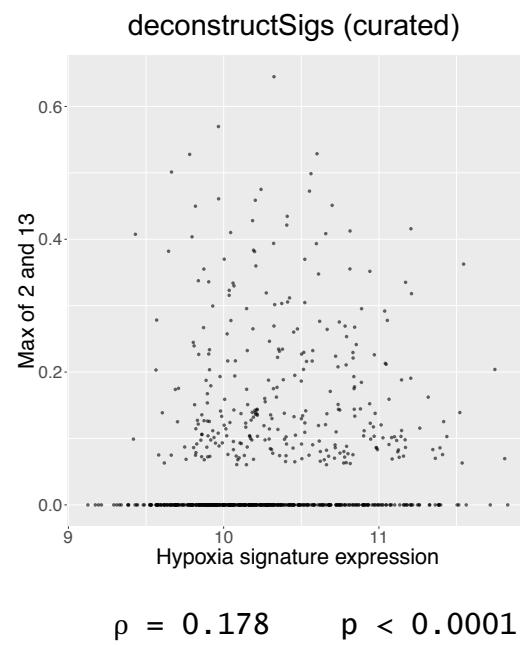
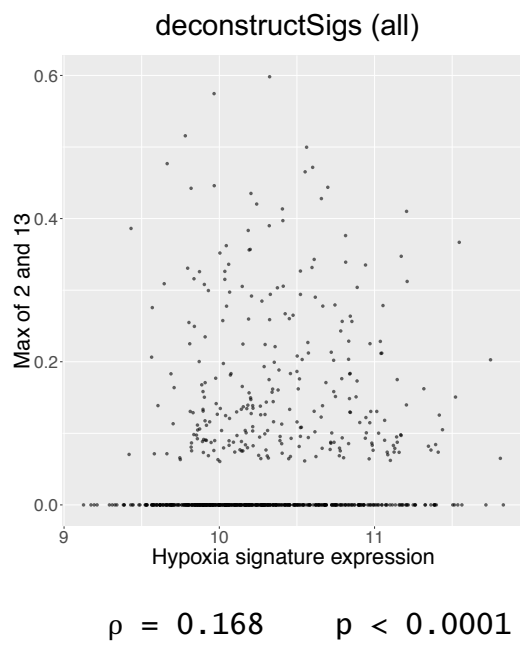
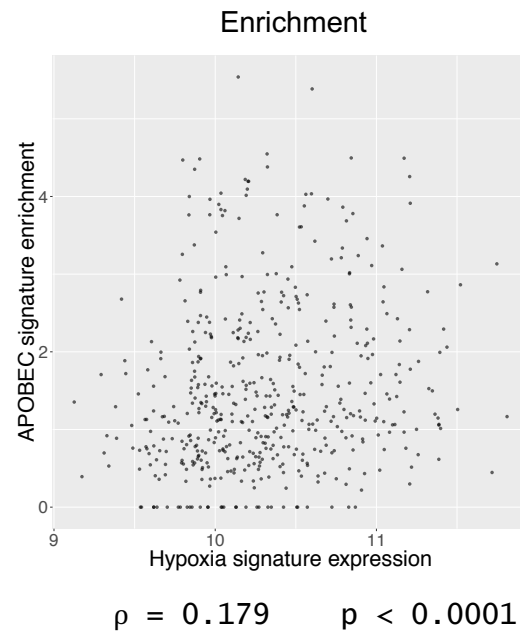
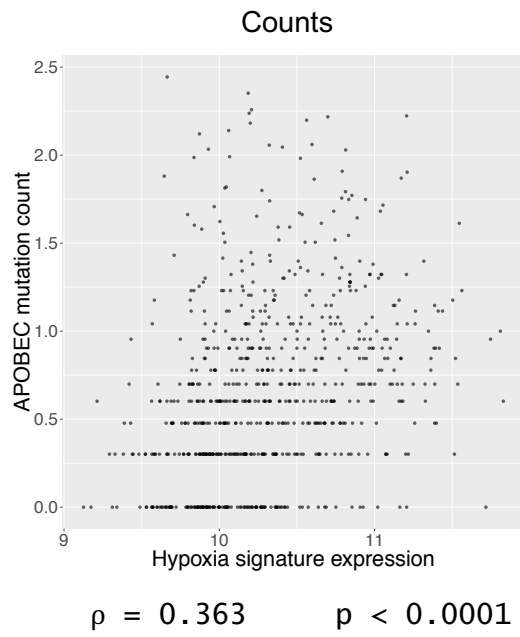
From this, we can confidently conclude that there is correlation between the APOBEC mutational signature and hypoxia in BRCA. There is some evidence for this correlation in other cancer types, however the evidence comes mainly from one method of calculation for the mutational signatures: by using the counts of TCW mutations. With the other methods failing to cross-validate this result, it is difficult to draw a firm conclusion.

Thus, we provide mechanistic evidence for hypoxia-related APOBEC mutagenesis in BRCA. Hypoxic BRCA samples see an increase in APOBEC mutagenesis when measured by all four methods in addition to exhibiting an increased expression of active APOBEC genes.

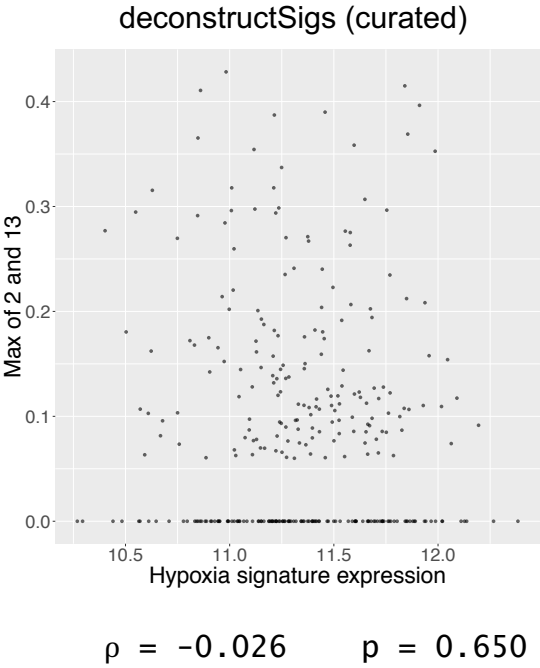
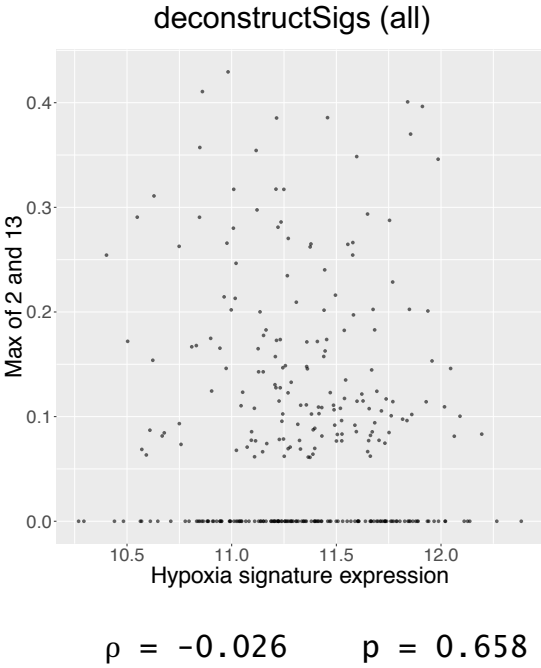
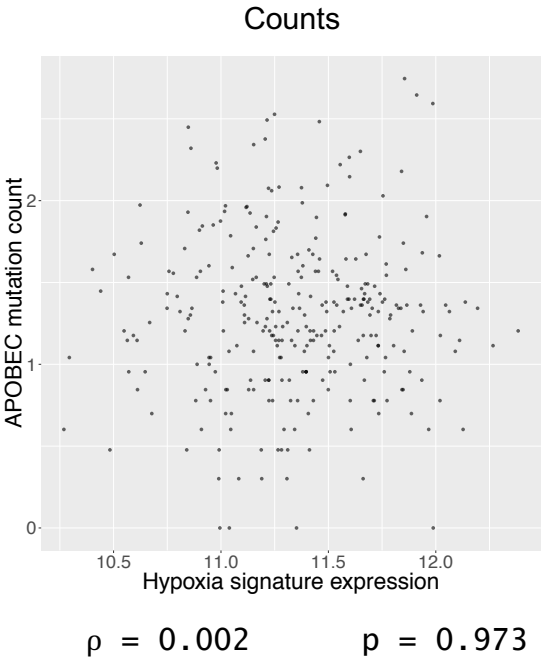
Figure 31: Relationship between Hypoxia and the APOBEC Mutational Signature



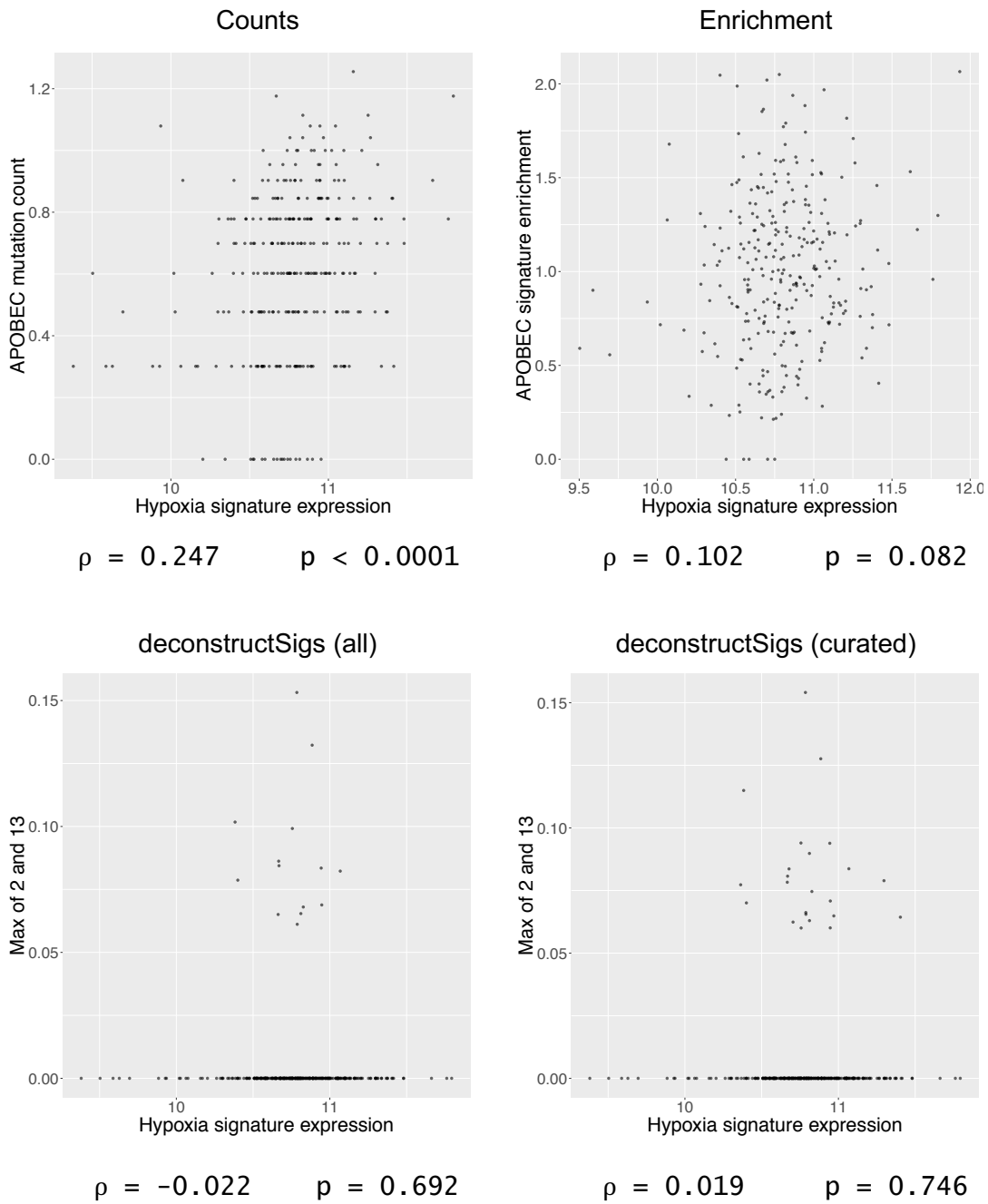
BRCA



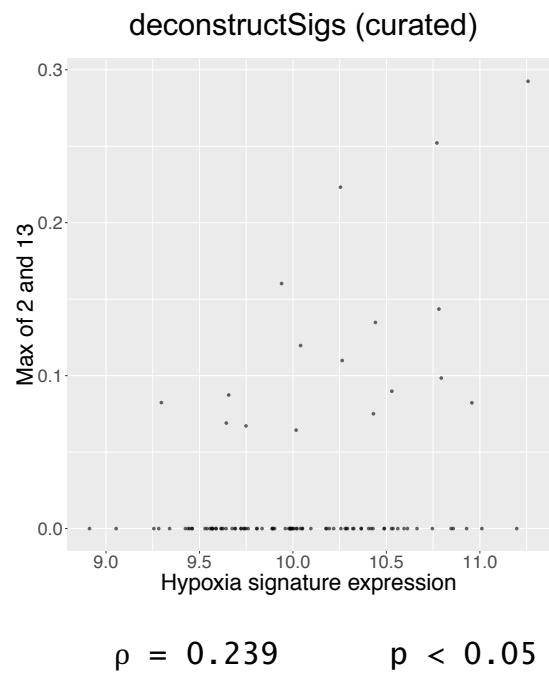
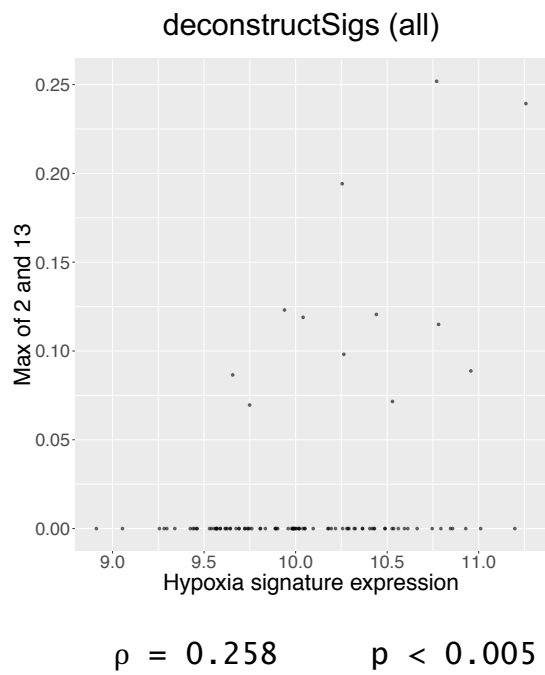
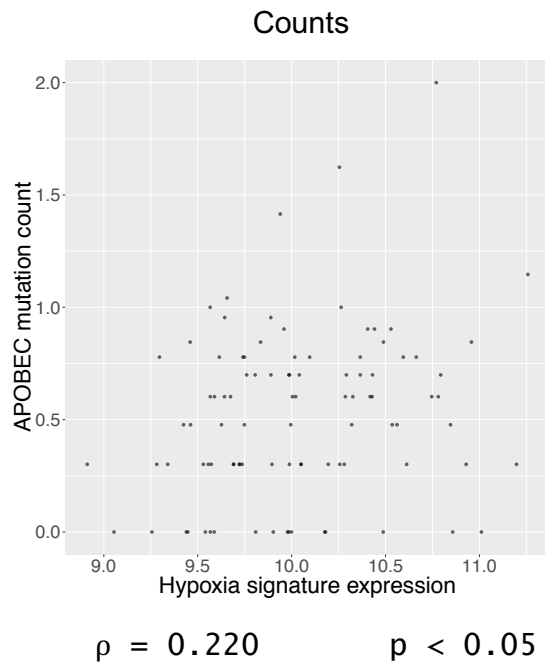
HNSC



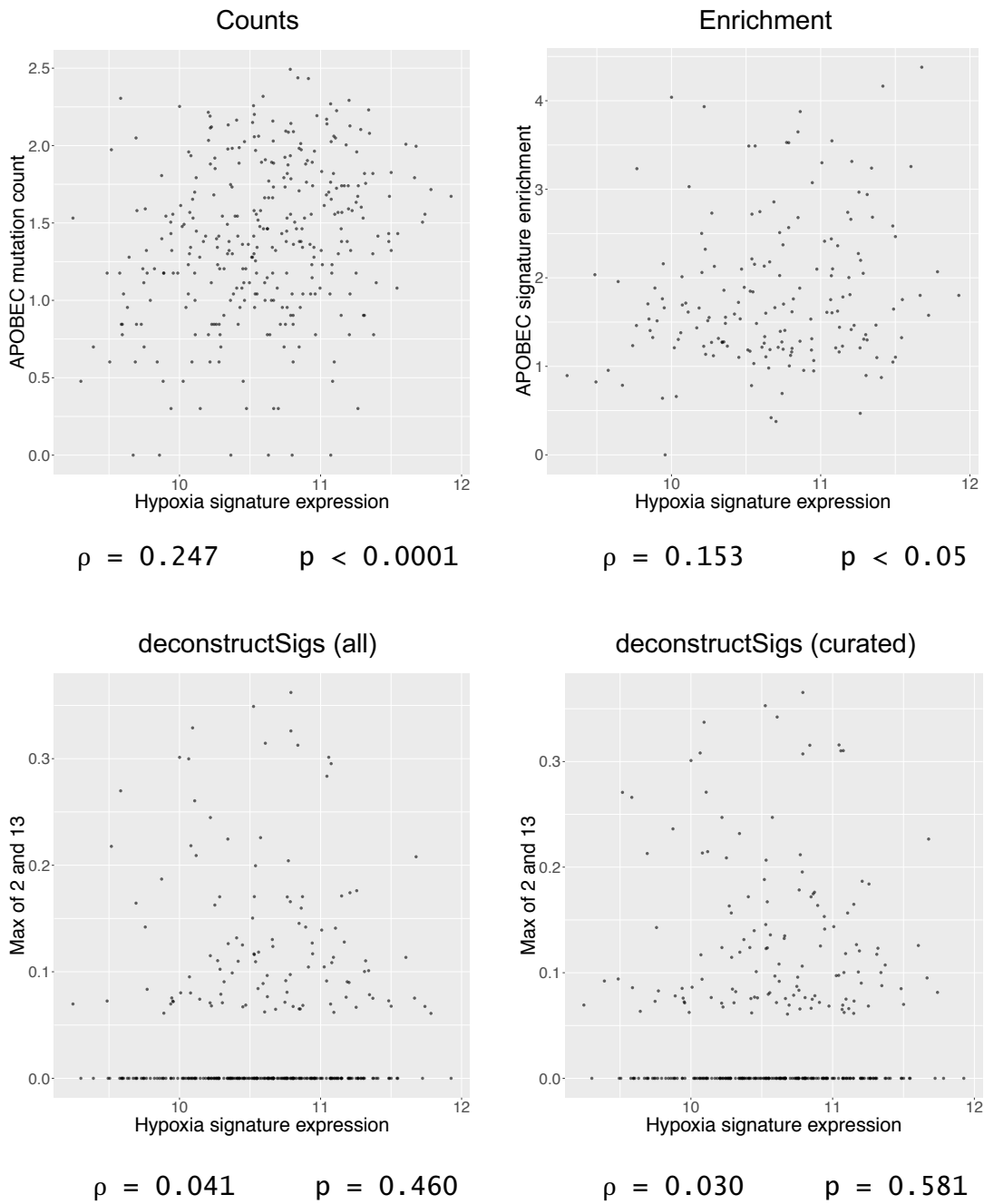
KIRC



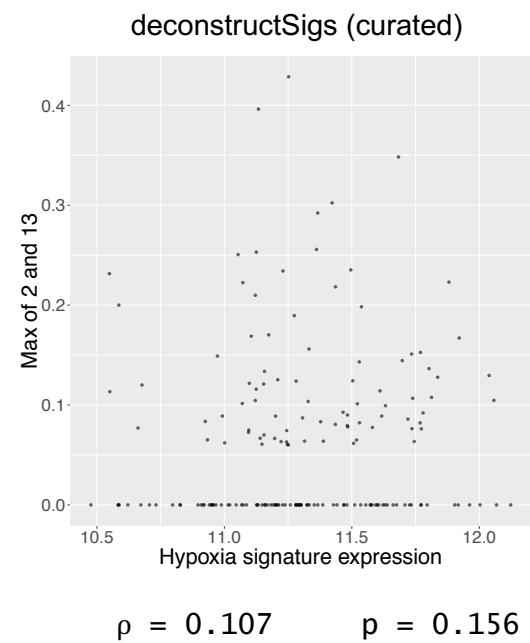
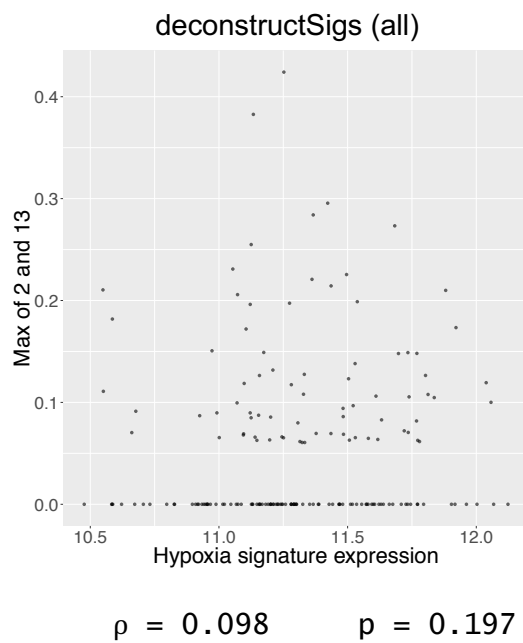
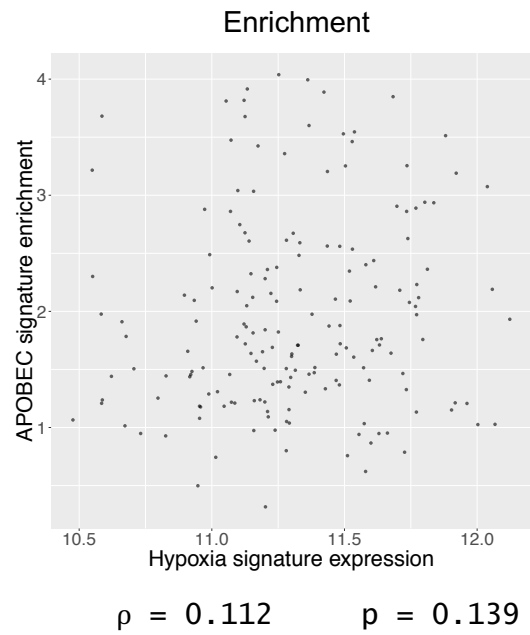
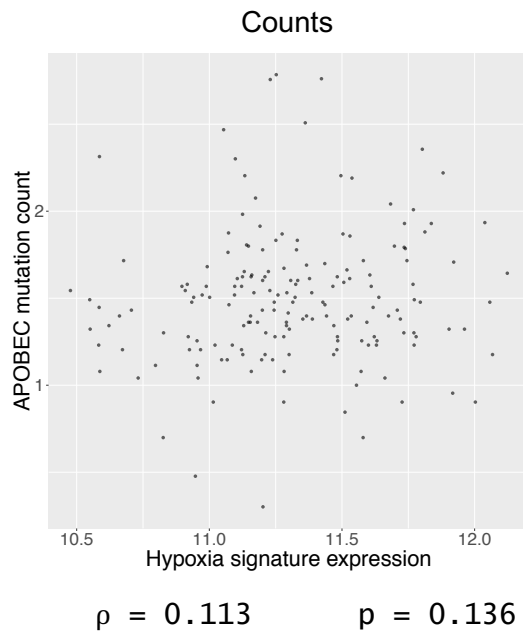
KIRP



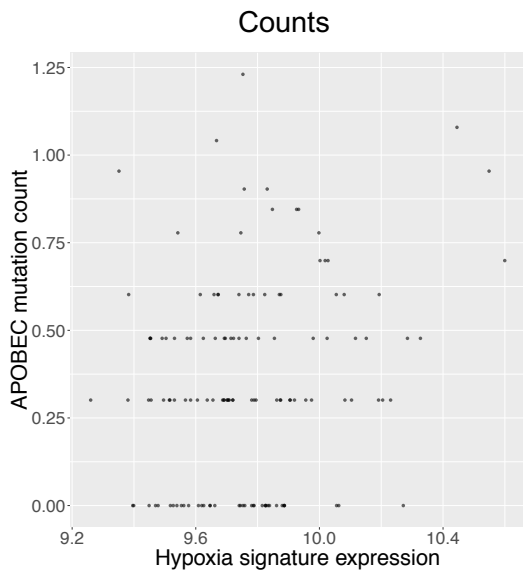
LUAD



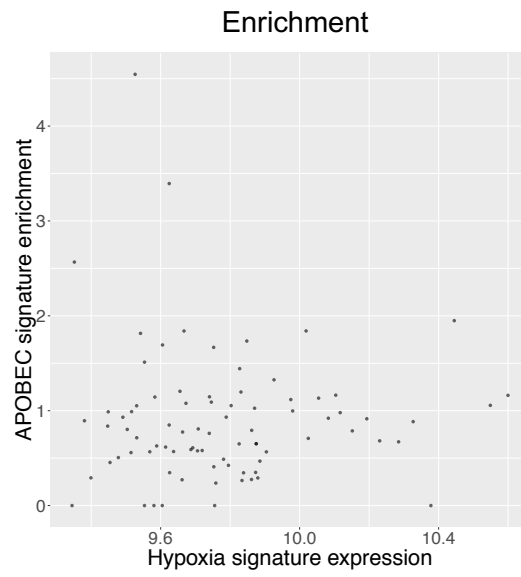
LUSC



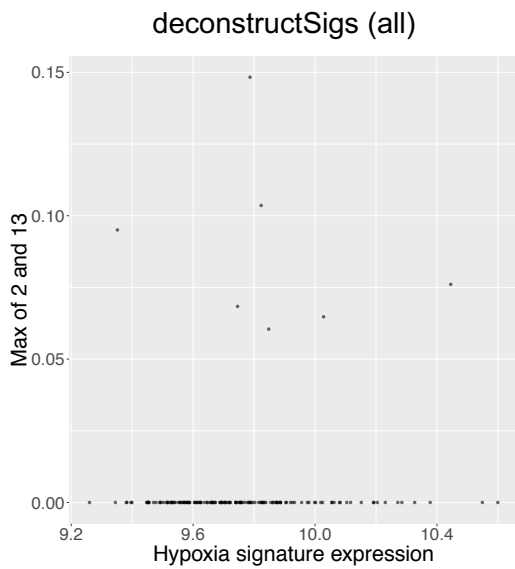
PRAD



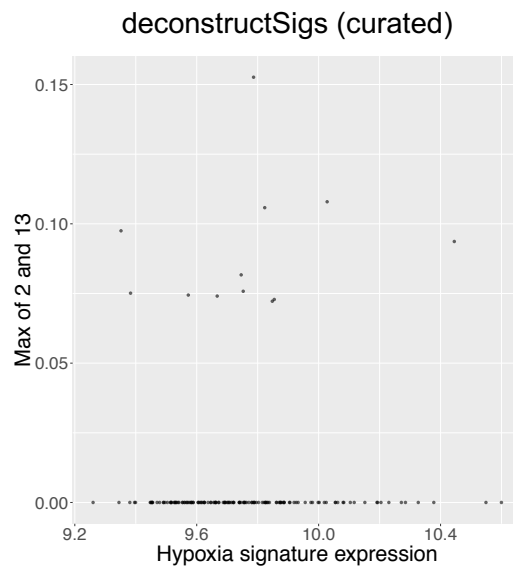
$\rho = 0.183$ $p < 0.05$



$\rho = 0.086$ $p = 0.437$

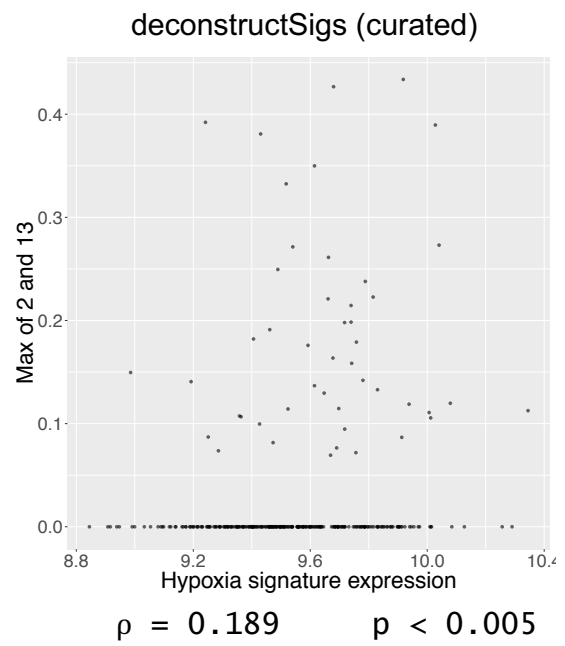
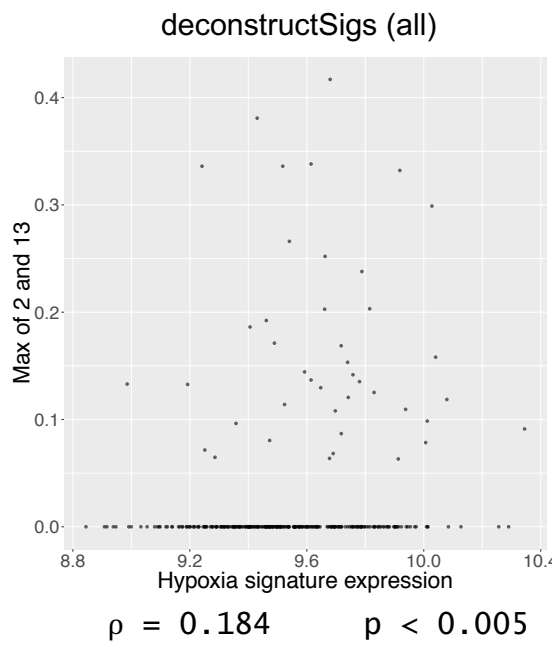
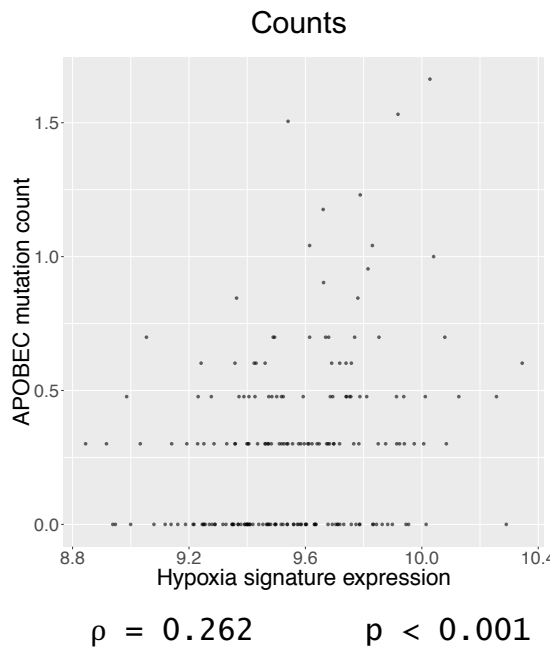


$\rho = 0.097$ $p = 0.237$



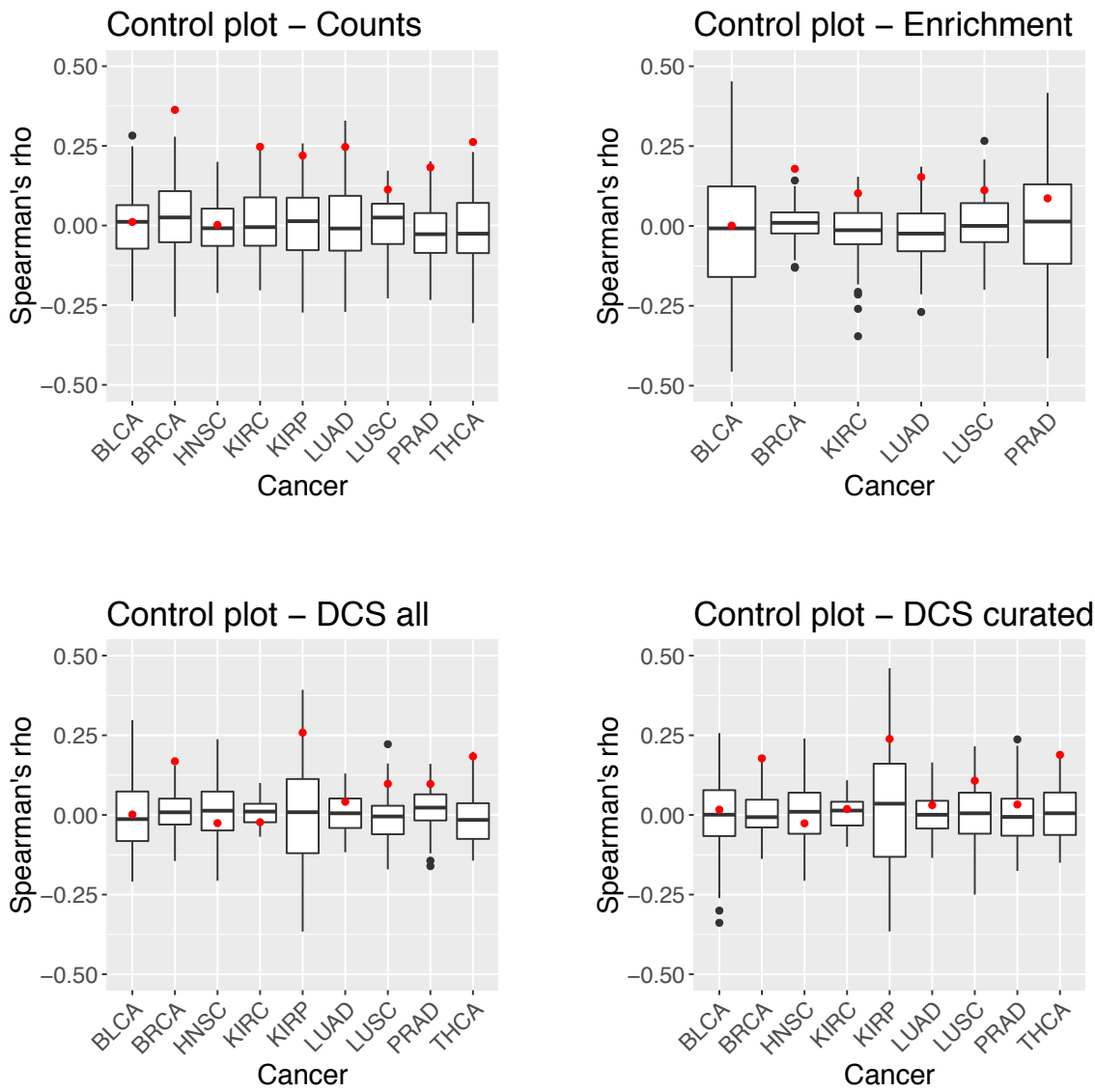
$\rho = 0.033$ $p = 0.691$

THCA



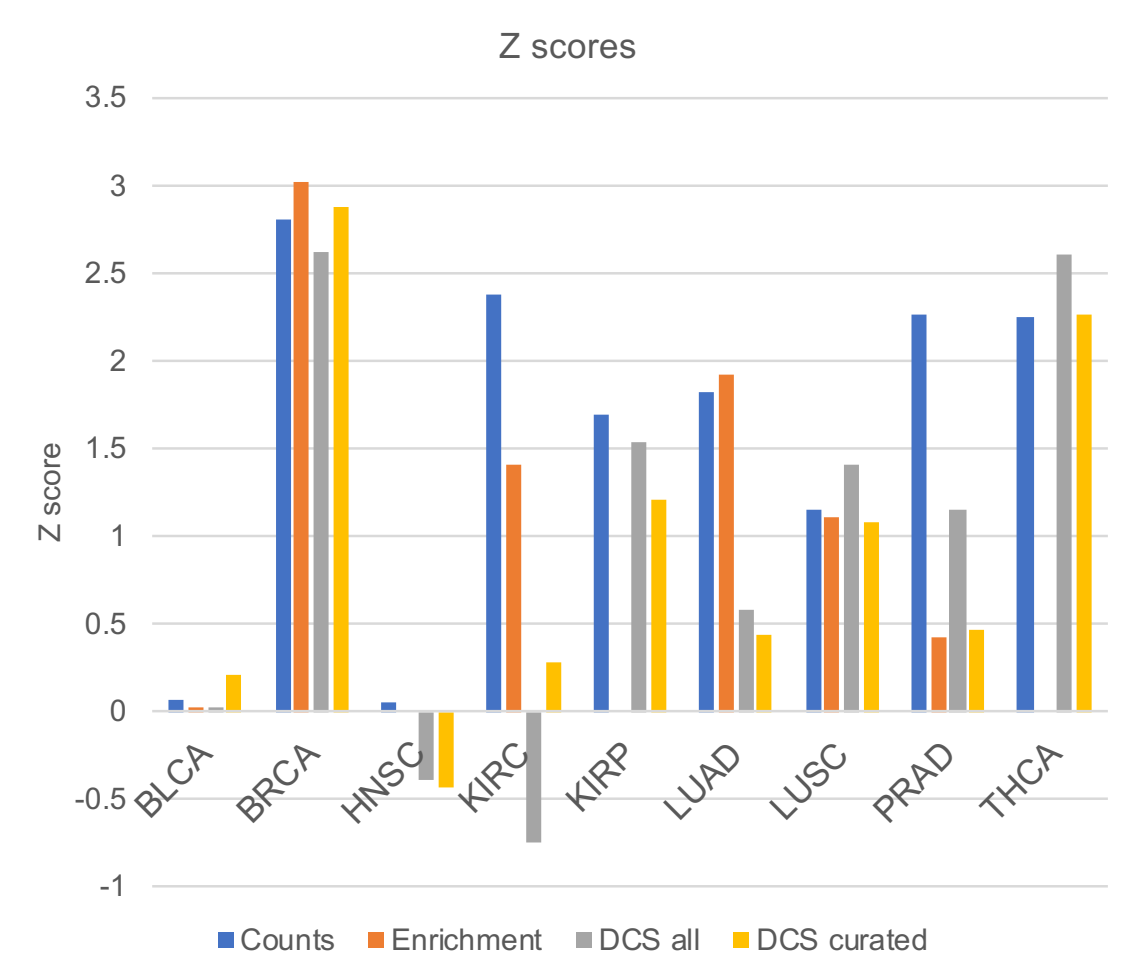
Correlation between the mean expression of the Buffa hypoxia signature and the APOBEC mutational signature, as measured by 4 methods. Count values are log 2 transformed counts of the TCW to TTW or TGW mutation in sample exomes. Enrichment values are the enrichment score defined by Roberts et al, 2013 and discussed in the main text. deconstructSigs scores are the maximum of signatures 2 and 13. Mutation data from Alexandrov et al., 2013. Expression values are log 2 transformed, normalised mRNA count from TCGA. Spearman's rho is calculated and displayed alongside p values.

Figure 32: Correlation between the APOBEC Mutational Signature and the Expression of Random Genes in Relation to the Correlation with the Mean Expression of the Buffa Hypoxia Signature



Correlation between the expression of randomly selected genes and the APOBEC mutational signature as reported by four methods (DCS is abbreviation for deconstructSigs). 100 genes are selected at random. For each cancer type and each method of measuring the APOBEC mutational signature, the correlation between the expression of the randomly selected genes and the mutational signature is calculated. Values represent Spearman's rho. The red marker shows the correlation between the mean expression of the Buffa signature and the mutational signature. Boxes represent the range from the first to the third quartiles. Centre line represents the median. Whiskers represent 3rd quartile plus 1.5x the interquartile range or the maximal value, whichever is smaller, and the 1st quartile minus 1.5x the interquartile range or the minimal value, whichever is larger. Dots represent values not contained within whiskers.

Figure 33: Z Scores for the Mean Buffa Signature Relative to the Distribution of Random Genes in Terms of Correlation with the APOBEC Mutational Signature



Z scores calculated for the Buffa signature mean expression's correlation with the APOBEC mutational signature relative to the correlation of 100 random genes with the signature, using four methods. Bars represent Z scores.

3.9 APOBEC3B and Hypoxia are Associated with Common Pathways

To assess the functional association between APOBEC and hypoxia, we turn to the analysis of enriched cellular pathways. First, we determine the genes that are co-expressed with APOBEC and hypoxia. Then, we use gene set enrichment analysis to find the pathways that are enriched for each list of co-expressed genes. We examine the association between the pathways, the genes in our lists that are present in the pathways, and the expression levels of these genes. Finally, we examine the overlap between the genes and pathways associated with APOBEC and hypoxia.

We calculate the Spearman's correlation coefficient between each APOBEC3 family gene and every gene reported in TCGA data sets for breast cancer. Results are filtered for FDR-adjusted p values < 0.05 . We retain lists of genes with Spearman's correlation coefficient greater or equal to 0.3. Genes with lower expression than the median expression of all genes across all samples in BRCA are discarded. Thus, we have a list of co-expressed genes for every APOBEC 3 gene. The same analysis is performed again, this time using Spearman's correlation partial to the Buffa hypoxia signature. This is done to obtain only the genes co-expressed with APOBEC genes independently of hypoxia. From these lists, we compile a list of genes that are considered co-expressed with at least five APOBEC3 genes and retain the list of genes that are co-expressed with A3B. From here, we see that in the analysis adjusting for hypoxia, no gene is co-expressed with A3B. This indicates that the association between A3B and hypoxia is strong and that there is no gene co-expressed with A3B independently

of hypoxia. Thus, for downstream analysis, we do not adjust for the hypoxia signature. Our final lists are displayed in Table S2.

For the hypoxia genes, we calculate the Spearman's correlation coefficient between each gene in the Buffa signature and every gene reported by TCGA in BRCA. Results are filtered for FDR-adjusted p values < 0.05 . For every gene in TCGA, we sum the Spearman's correlation coefficients calculated for each of the 52 genes in the Buffa signature. From the list of summed coefficients, we take the top 500 genes (Table S3).

The three resulting gene lists (co-expressed with A3B, 5 or more APOBECs, and hypoxia) are used as input for gene set enrichment analysis. Gene set enrichment analysis consists of finding pre-determined sets of associated genes in an input list. It returns the most highly enriched gene sets found along with significance levels. We can apply this to our study by using pathways as the gene sets, thus finding the most enriched cellular pathways in a given list of genes. We perform this analysis using the GeneCoDis platform: <http://genecodis.cnb.csic.es> (Carmona-Saez et al., 2007). The pathways in the analysis are the KEGG pathways: <https://www.kegg.jp> (Ogata et al., 1999). The KEGG database constantly maintained and kept up to date with the latest findings.

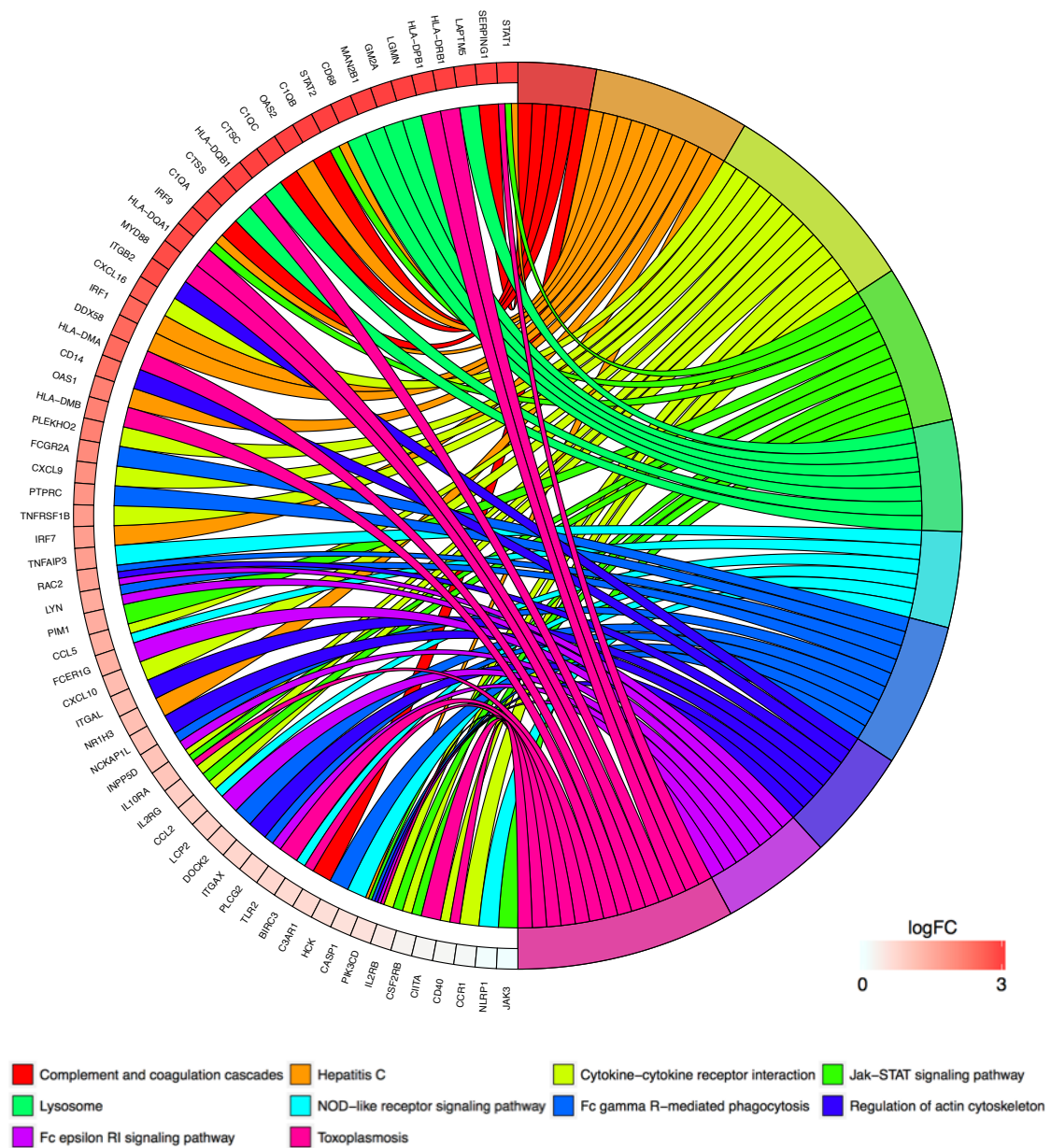
The 10 most enriched pathways for each gene list are displayed alongside the genes in the lists that are found in them, ranked by their expression level (Figures 34-36). Among the pathways enriched in the genes co-expressed with the hypoxia signature are major metabolic pathways such as oxidative phosphorylation, glycolysis/gluconeogenesis, and the metabolism of various

macromolecules. For genes co-expressed with A3B, we see various pathways with a link of significance to APOBEC, DNA damage, and cancer: DNA replication, mismatch repair, spliceosome, and the p53 signalling pathway. We also see metabolic pathways being enriched, some found in the pathways enriched in hypoxia, notably purine and pyrimidine metabolism. As for genes co-expressed with at least 5 APOBEC 3 genes, the enriched pathways are mostly in the immune response, the fundamental role of APOBECs.

We can further observe the overlap between genes and pathways associated with APOBEC and hypoxia by generating Venn diagrams for genes and pathways overlapping among the three groups (Figures 37-38). We observe significant overlap both in terms of genes and of pathways between A3B and hypoxia, while the group of other APOBECs is completely disjoint from the other two. Looking at the pathways in each group, it is clear that other APOBEC genes are functionally distinct. The most enriched pathways are immune response pathways, suggesting that their major role in breast cancer is still the immune response, whereas A3B is associated with metabolism and DNA hypermutation. This is further evidence that A3B functions differently compared to other APOBEC 3 family members, and associates with hypoxia.

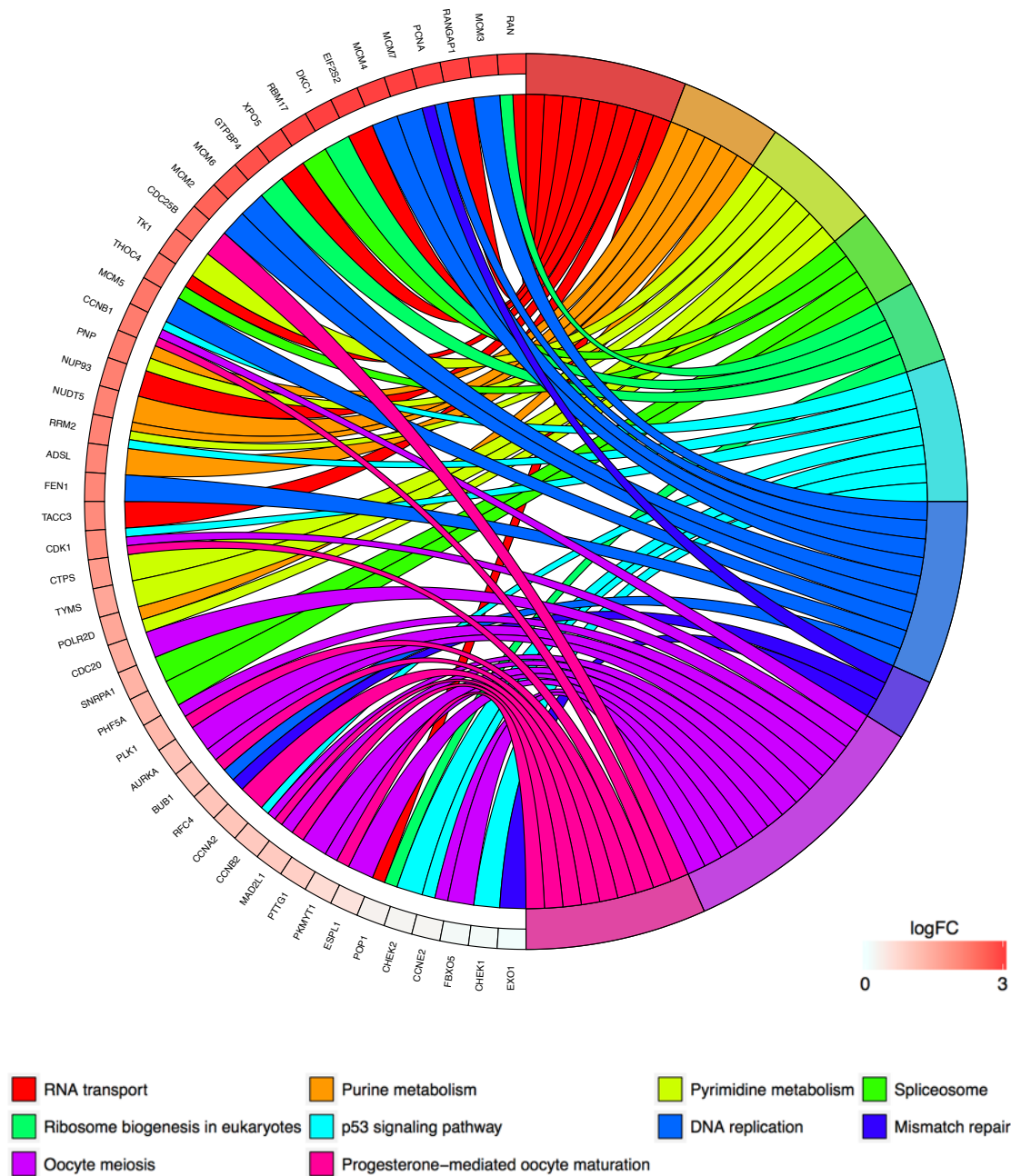
Overall, we conclude that there is a functional relationship between A3B and hypoxia. The genes co-expressed with them as well as the pathways associated with them overlap significantly. On the flip side, other APOBEC genes tend to activate immune pathways and are completely dissociated functionally from both A3B and hypoxia.

Figure 34: Top Enriched Pathways in Genes Co-Expressed with 5 or more APOBEC3 Family Genes in Breast Cancer



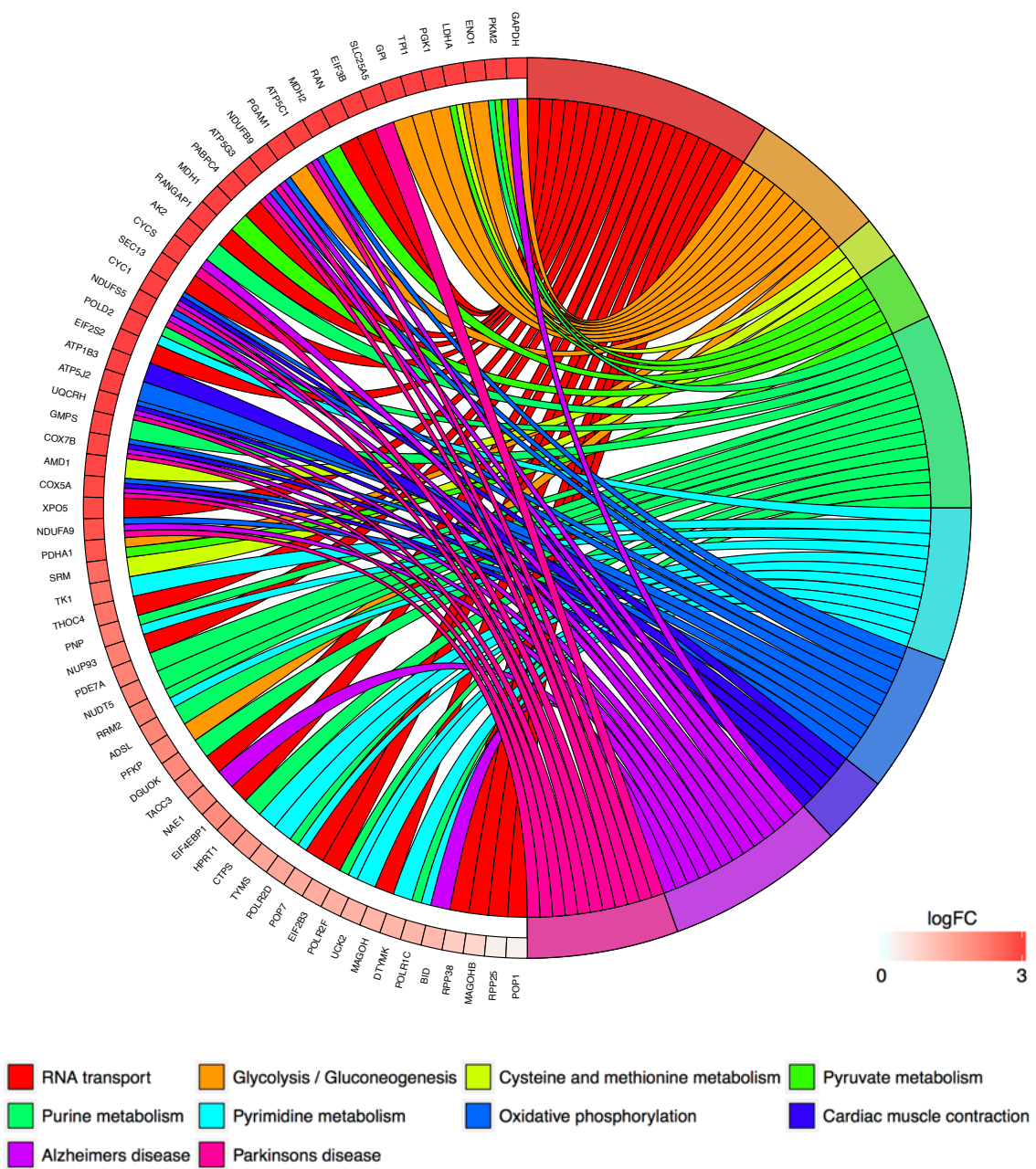
The top 10 enriched pathways in genes co-expressed with 5 or more APOBEC3 family genes in BRCA, along with the genes that are part of the pathways, ranked by expression. Spearman's rho is calculated between every gene in TCGA and all APOBEC3 genes. All genes with $\rho \geq 0.3$ and $p < 0.05$ in 5 or more APOBEC3 genes are used for enrichment. Enrichment is calculated using the GeneGoDis platform with the KEGG pathways. Expression data (logFC) is presented as the log 2 fold change relative to the median expression of all genes in all samples in BRCA. Expression data is normalised mRNA count from TCGA.

Figure 35: Top Enriched Pathways in Genes Co-Expressed with APOBEC3B in Breast Cancer



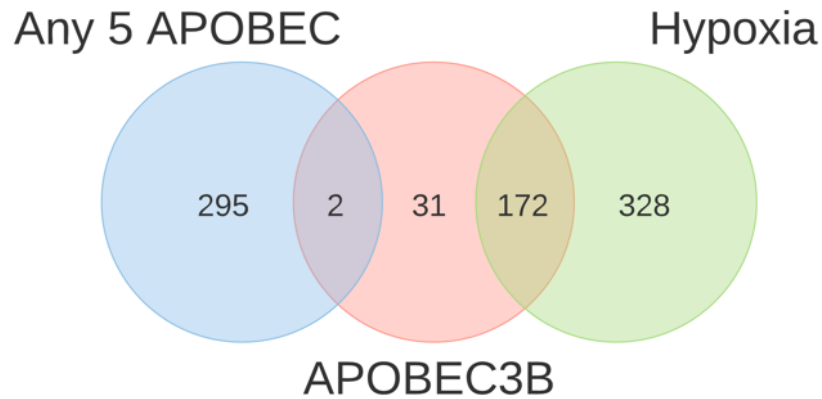
The top 10 enriched pathways in genes co-expressed with APOBEC3B in BRCA, along with the genes that are part of the pathways, ranked by expression. Spearman's rho is calculated between every gene in TCGA and APOBEC3B. All genes with $\rho \geq 0.3$ and $p < 0.05$ are used for enrichment. Enrichment is calculated using the GeneGoDis platform with the KEGG pathways. Expression data (logFC) is presented as the log 2 fold change relative to the median expression of all genes in all samples in BRCA. Expression data is normalised mRNA count from TCGA.

Figure 36: Top Enriched Pathways in Genes Co-Expressed with Hypoxia Genes in Breast Cancer



The top 10 enriched pathways in genes co-expressed with hypoxia genes in BRCA, along with the genes that are part of the pathways, ranked by expression. The Spearman's rho is calculated between every gene in TCGA and every gene in the Buffa signature. TCGA genes are ranked by the sum of rho for all genes in the Buffa signature. The top 500 genes are used for enrichment. Enrichment is calculated using the GeneGoDis platform with the KEGG pathways. Expression data (logFC) is presented as the log 2 fold change relative to the median expression of all genes in all samples in BRCA. Expression data is normalised mRNA count from TCGA.

Figure 37: Overlap between Genes Correlated with any 5 APOBEC3 Genes, APOBEC3B, and Hypoxia in Breast Cancer in Terms of Genes



Genes overlapping between any five A3 genes and hypoxia:

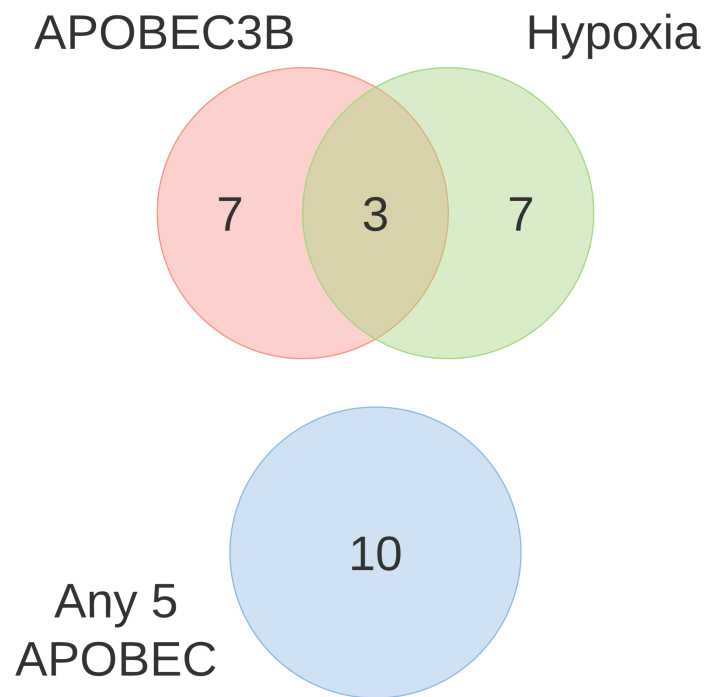
SH3BP1 WARS

Genes overlapping between APOBEC3B and hypoxia:

ANLN	CDCA5	E2F1	HMGA1	MKI67	POP1	STIL	AARS	NFKBIL2
ARHGAP11A	CDCA8	ECE2	HMMR	MRPL15	PPIL5	STMN1	ADSL	PBK
ASF1B	CDK1	EIF2C2	HSPA14	MTHFD2	PPPDE2	SUV39H2	C15orf23	PDXP
ASPM	CDT1	EIF2S2	IMPA2	MYBL2	PRC1	TACC3	C16orf75	PRIM2
ATP5C1	CENPE	EPR1	IQGAP3	NCAPG	PSMB2	TCF19	C9orf140	RACGAP1
AURKA	CENPN	ESPL1	IRAK1	NCAPH	PSMD14	THOC4	CCDC99	RANGAP1
BIRC5	CEP55	EXO1	KIF11	NDC80	PTTG1	TK1	CCT5	RAN
BUB1B	CHAF1B	EZH2	KIF20A	NEK2	RANBP1	TPX2	CDC6	SEPHS1
BUB1	CHEK1	FAM136A	KIF23	NIP7	RBM17	TRIP13	CENPF	TOMM22
C16orf61	CHEK2	FAM83D	KIF2C	NUDT5	RCC1	TROAP	DKC1	UHRF1
C20orf24	CKS1B	FANCA	KIF4A	NUF2	RECQL4	TTK	DRG1	XPO5
CBX2	CKS2	FEN1	KIFC1	NUP93	RFC4	TUBA1B	FANCI	ZWINT
CCNA2	CTPS	FOXM1	KPNA2	NUSAP1	RIPK2	TUBA1C	FBXO5	
CCNB1	DBF4	GINS1	LMNB1	PCNA	RPN1	TUBB	GPSM2	
CCNB2	DDX39	GMNN	LMNB2	PHF5A	RRM2	TYMS	ILF2	
CDC123	DEPDC1	GPR172A	MAD2L1	PKMYT1	SLC25A19	UBE2C	KIAA0101	
CDC20	DLGAP5	GTPBP4	MCM2	PLK1	SLC25A5	UBE2S	MCM3	
CDC25B	DNAJB11	H2AFZ	MCM5	PNP	SLC7A5	UBE2T	MCM4	
CDC7	DNAJC9	HDGF	MCM6	POC1A	SNRPA1	VRK1	MCM7	
CDCA3	DONSON	HJURP	MELK	POLR2D	SPAG5	YARS	NCAPD2	

Overlap of genes correlated with any 5 APOBEC3 genes, APOBEC3B, and the Buffa hypoxia signature. For APOBEC genes, Spearman's rho is calculated between every gene in TCGA and all APOBEC3 genes. All genes with $\rho \geq 0.3$ and $p < 0.05$ in 5 or more APOBEC3 genes are taken. The same is done for only APOBEC3B. For the hypoxia genes, the Spearman's rho is calculated between every gene in TCGA and every gene in the Buffa signature. TCGA genes are ranked by the sum of rho for all genes in the Buffa signature. The top 500 genes are taken.

Figure 38: Overlap between Genes Correlated with any 5 APOBEC3 Genes, APOBEC3B, and Hypoxia in Breast Cancer in Terms of Pathways



Pathways overlapping between APOBEC3B and hypoxia:

- RNA transport
- Purine metabolism
- Pyrimidine metabolism

Overlap of pathways enriched for in genes correlated with any 5 APOBEC3 genes, APOBEC3B, and the Buffa hypoxia signature. Enrichment performed with GeneGoDis and KEGG pathways.

4. Discussion

4.1 Expression of APOBEC Family Genes in Patient Samples across Cancer Types

First, we asked if APOBEC genes were expressed in cancer, and in which cancer types. The first analysis performed shows a great heterogeneity in the expression of APOBEC genes in cancer samples. In fact, the APOBEC3 family of genes is consistently expressed at high levels across all cancer types analysed while A3B is preferentially expressed in tumour versus normal samples of multiple cancers, including bladder, breast, head and neck, kidney, liver, lung, prostate, stomach, and uterine cancers. One of the original studies linking A3B to mutations in cancer also reported a differential expression of A3B between tumour and normal samples in the same cancer types using a slightly different approach (Burns et al., 2013b). The authors normalised gene expression reported by RNA-seq to the expression of TBP, a housekeeping gene commonly used as control in qPCR. The increased expression of A3B is expected in cancer, as its mutational signature has also been found in multiple cancer types (Burns et al., 2013a, 2013b; Roberts et al., 2013). It would be logical that an increased expression of A3B ultimately leads to more mutations fitting its signature. Also, considering that A3B is constitutively located in the nucleus after translation, more functional protein should lead to more mutations in the genome, perhaps up to a saturation point that is not reached biologically. If the rate of mRNA processing and

translation remain unchanged, it would be expected that an increased expression of A3B mRNA would lead to an increase in A3B-induced mutations.

We then investigated the presence of structure in the expression of gene in the APOBEC family. Using principal component analysis, we decompose the APOBEC expression pattern across cancers and samples into its components of highest variance. This shows us that most of the variance in the expression of APOBEC genes can be explained by A3A, A3B, and A3H in breast cancer. In addition, we observe that A3A and A3B vary independently, as they are associated with different principal components. However, this effect is lost in many other cancer types, as the expression of most APOBEC genes are explained mainly by PC1. This is an evidence of a highly self-correlated expression, and we do indeed observe a higher degree of self-correlation in cancers other than breast. Particularly, A3B expression does not correlate with that of any other APOBEC gene in breast cancer except for A3A. Looking back at the principal component analysis, we see that PC1 accounts for variations in A3A, A3B, and A3H, while PC2 accounts for variations in A3B. Taking the principal component analysis and self-correlation together, we can say that although there is correlation between A3A and A3B expression, there is a component of A3B expression that is independent from the expression of A3A and all other APOBEC genes.

It is expected that the expression of A3 family genes show some level of self-correlation since they are located in the same genomic cluster. From our analysis of expression in relation to gene amplification, we see that A3 genes show signs

of correlation between the two measures. This is a sign that the increase in A3 expression may be driven at least to some extent by amplification of the genomic cluster and that the degree of amplification balancing with the regulation of transcription together control the overall mRNA levels of A3 genes detected by RNA-seq.

On the other hand, we can already see when decomposing the signal by PCA that in fact, A3A expression and A3B expression are associated with different components of variance. This can be explained by different regulation mechanism acting on their mRNA levels. The component of A3B expression that varies alongside A3A expression might be driven by gene amplification while the component that varies independently could be driven by transcriptional regulation.

In fact, studies on the regulation of APOBEC gene expression have already provided evidence that the expression of different members of the A3 family is controlled by different regulatory mechanisms despite the fact that the genes are in the same genomic cluster (Covino et al., 2018). In monocytes and macrophages, the expression of A3A and A3G is mainly controlled by interferon-mediated pathways while the A3B proximal promoter region shows binding sites for NF- κ B. This suggests that in normal circumstances, A3A and A3B have their transcription activated by different mechanisms. However, the activation of APOBEC genes in cancers may be under multiple mechanisms of regulation. For example, it is known that double-stranded breaks in DNA can lead to an increase in APOBEC mutagenesis in vivo (Kanu et al., 2016). This can lead to APOBEC genes being more active in cancer in general, possibly resulting from an

upregulation of A3A and A3B. However, one should consider that an increase in double-stranded breaks also leads to an increase in APOBEC binding sites, as ssDNA is exposed and available for binding. Therefore, the increase in APOBEC mutagenesis may reflect the increased or modified action of the same amount of protein. Later research also revealed that the expression of A3 genes is controlled in a number of ways by TP53, a well-known tumour suppressor gene (Menendez et al., 2017). Specifically, the study showed that TP53 promotes the expression of A3A but not A3B. Another study examined the regulation of APOBEC expression in cancer cell lines with induced double-stranded breaks and found that cells with mutant TP53 showed higher expression of A3B while A3A is preferentially activated by viral infection (Middlebrooks et al., 2016), possibly through its ability to be activated by interferons.

To investigate further the link between APOBEC and TP53 in clinical samples, we stratified BRCA samples by TP53 status into mutated and wild-type categories and analysed the expression of A3A and A3B, finding they were both expressed at high level expressed at a higher level in the mutant samples. In the case of A3B, this is in line with previous findings, however the results are more controversial for A3A and do not align with the previous report stating that A3A expression is promoted by TP53. On the other hand, in these cancers A3A might be activated under different mechanisms that are indirectly caused by the loss of functional TP53.

Taking the studies referenced above together with our results from PCA, amplification and self-correlation analyses, we see that the findings converge.

Since A3A and A3B are located in the same gene cluster, effects of amplification can affect their expression simultaneously, leading to signs of self-correlation. However, this does not account for all of the variance in A3B expression, as reported by PCA. Therefore, regulatory mechanisms might be involved, with the effects of TP53 being a possible candidate. Together, amplification, TP53 activity, and IFN or NF- κ B signalling can all contribute in different ways to the expression of A3 genes, and a different combination of the three factors will yield different levels of correlation between the expression of various A3 genes. With the diversity among sequenced tumours, it is not surprising to see variance on these fronts leading to a complex relationship between A3A and A3B that is heterogeneous among samples.

One thing to note about our current analyses is the fact that these are performed on bulk samples containing populations of cells that are heterogeneous. This heterogeneity leads to all cells not necessarily having the same mutations or gene expression patterns. It would be interesting to use microdissected samples for further analysis to obtain information as to in which specific cells APOBEC proteins are expressed. This is outside the scope of this project but would be a potential project for the future.

4.2 The Relationship between APOBEC and Hypoxia on the Gene Expression Level

Our initial hypothesis that hypoxia may be driving APOBEC expression and mutagenesis in cancer is based on evidence in cell models suggesting that A3 family genes can be activated by hypoxia. In particular, three studies stand out. First, the initial study that linked APOBEC to hypoxia aimed to investigate the genes that have their expression altered by hypoxia in hepatocytes (Sonna et al., 2003). Among the genes analysed, A1 and A3C were shown to be upregulated by hypoxia. Although this was not a finding that the authors put much emphasis on since APOBEC mutagenesis was not the scope of their study, this result drove later research into APOBEC and hypoxia. Namely, one study conducted on macrophages in cell culture to assess APOBEC activity showed that A3A is active as an RNA cytosine deaminase and is induced by IFN-1 and hypoxia to the same extent (Sharma et al., 2015). Knowing that A3A activity can normally be induced by IFN-1 in macrophages for the purposes of innate immunity, it is an interesting finding that hypoxia also drives A3A deamination to the same extent and indicates that this increase in activity is biologically significant. The authors then published findings in a later study suggesting that A3A overexpression is linked to cancer (Sharma et al., 2017). In addition, chronic hypoxia drives TP53 mutations through selection (Graeber et al., 1994). Knowing that TP53 suppresses the expression and activity of A3B (Middlebrooks et al., 2016), mutations in TP53 selected for in hypoxia can lift the inhibition on A3B and render it more active. Although limited both in quantity and in scope, taken together, the

evidence shown in these previous studies leads to us asking whether APOBEC is induced by hypoxia in cancer.

Thus, we started our analysis of hypoxia in the TCGA samples by showing that the expression of a hypoxia gene signature previously derived by our group (Buffa et al., 2010), and validated in other studies, is significantly higher in tumour samples of 13 cancer types (for a total number of samples of N=6723) compared to the corresponding normal tissue. Considering that hypoxia is a prominent feature of solid tumours (Vaupel and Mayer, 2007), it is not a surprise that the 13 carcinomas and adenocarcinomas analysed show a higher expression of the hypoxia signature in tumour tissue.

Considering that TP53 mutations are selected for in chronic hypoxia, we investigated the effects of TP53 mutations on the expression of the hypoxia signature. We observe that the hypoxia signature is expressed at high levels in TP53 mutant cases, which is in line with previous studies.

Although the results are as expected, we must be careful in interpreting the biology behind any signature when we apply it to cancer types other than those in which the signatures were derived and validated. Furthermore, the Buffa signature was derived from data collected in microarray experiments, not RNA-seq. Although both methods measure gene expression, they do so in different ways, thus some variability in the reported gene expression is expected between the two. Thus, we use a previously developed signature quality control tool, sigQC (Dhawan et al., 2017), to address these issues and ensure that the Buffa signature maintains the appropriate expression and variability characteristics to

be applied to TCGA data. This analysis confirmed that that genes in the signature were behaving as a coherent set of genes, with a generally high self-correlation. This does not prove that the signature is measuring hypoxia in the same way across cancer types, but does reassure us that the gene in these signature have the expected behaviour in the cancers, and carry a high and coherent signal, as could be anticipated for genes co-regulated under the same phenotype.

We then assessed the correlation between the expression of APOBEC genes and the mean expression of the Buffa hypoxia signature in the clinical samples. This showed that A3A and A3B expression correlates with this signature expression in BRCA, KIRC, KIRP, LIHC, LUAD, PRAD, and THCA, while the entire A3 family shows correlation in KIRC, LIHC, and THCA. Non-A3 genes show little to no correlation in expression with the Buffa signature. In addition, we use a principal component regression to build a model for the relationship between the expression of A3 genes and that of the hypoxia signature. We see that A3A and A3B expression can be used to predict the expression of the hypoxia signature in breast cancer, the cancer type where we have more samples and hence more power to resolve this. Taken together, these results show that A3A and A3B are the main A3 genes that are co-expressed and associated with the hypoxia signature.

Here, A3A expression correlating with the Buffa signature expression concords with previous findings that A3A is inducible by hypoxia. However, this may be coincidental, as the previous studies were conducted in immune cells, studying APOBEC in its canonical role. The fact that hypoxia is associated with A3A may

or may not reflect the same phenomenon as the previous studies suggest. Nevertheless, the correlation is significant in this case and we should examine the relationship between A3A, A3B, and hypoxia further.

As for A3B, there is no direct previous evidence for its activation in hypoxia. From an a priori point of view, it is difficult to predict how A3B expression would be affected by hypoxia. However, we know that A3B expression is suppressed by TP53 (Menendez et al., 2017). On the other hand, TP53 has a complex relationship with hypoxia (Sermeus and Michiels, 2011). Originally, hypoxia is thought to lead to the accumulation of TP53 in cells (Graeber et al., 1994). However, it has been shown that mild hypoxia that is enough to stabilise HIF-1 α is not necessarily sufficient to cause TP53 accumulation; rather, sustained hypoxia (Haltermann and Federoff, 1999) and high levels of hypoxia (Hammond et al., 2002) can induce TP53 stabilisation and accumulation. On the other side, hypoxia selects for TP53 mutations by favouring cells with lower apoptotic potential (Graeber et al., 1996; Kim et al., 1997). Since TP53 can induce apoptosis, mutations in the gene can cause a cell to be favourably selected in hypoxia by evading cell death, leading to hypoxic tumours being more likely to contain TP53 mutations in general. Besides, chronic and acute hypoxia have different effects on not only TP53, but many other pathways (Pires et al., 2010). All these factors taken together makes the task of predicting the A3B expression response to hypoxia challenging. On top of all this, the study showing A3B repression by TP53 was conducted in cancer cell lines. Whether the finding translates to human cancers and TCGA data is debatable. Here, I discussed one aspect that might control the expression of A3B in different ways and the

incompleteness of the evidence, but it is not known whether A3B can also be activated or repressed by other pathways in cancer.

To conclude, we take the correlation between the expression of A3B and the hypoxia signature as a result on which we must build and that we must support with additional evidence since there is inconclusive evidence from the literature as to whether this is expected or not. It is also to note that both A3A and A3B are only able to be used as predictors of hypoxia in breast cancer according to the principal component regression model. This further reinforces the point that our results are not yet conclusive and there is still much to learn in order to deconvolute the complex association between APOBEC and hypoxia.

One general limitation of using the Buffa signature as a predictor of hypoxia that I must highlight is the fact that the signature was derived from transcriptome data. This implies that we are only directly observing changes at the transcriptional level due to hypoxia. We do not have any data concerning post-transcriptional events or cellular changes induced by hypoxia that do not act on the transcriptome. However, this does not completely discredit the current study, as the Buffa signature is validated and shown to be prognostic in multiple cancers, indicating that it is a potentially accurate predictor of overall hypoxia status. The development of newer, more accurate gene signature is ongoing work in my laboratory in collaboration with others. Work is currently under way in deriving cancer-specific gene signatures and signatures that are independent of intrinsic differences among cancer subtypes. The preliminary results from this work shows that in breast cancer, the expression of the new signature shows less variability

among samples of different subtypes. This confirms the knowledge that different subtypes have different hypoxia levels, triple negative being the most hypoxic. However, it also highlights the potential need to consider the cancer subtype into the analysis. As a next step, it would be interesting to apply the new signatures generated from the future study to the methods of this current study to compare and contrast the results.

Another member of the research group analysed the expression of APOBEC genes in cell lines exposed to 24h of hypoxia, with results that differ from what I present here. According to his analyses, A3B, A3C, and A3F are the only APOBEC genes that are significantly expressed. A3B is expressed significantly in most breast cell lines analysed and shows a negative differential expression between hypoxia and normoxia, as in A3B is expressed at a higher level in normoxia than in hypoxia. A3C is significantly expressed in BT20, DU4475, MDA231, and MDA468 cell lines but shows no differential expression between hypoxia and normoxia. A3F is expressed in BT474 and MDA468 and shows no differential expression between hypoxia and normoxia. Interestingly, MCF7, a widely-used breast cancer cell line, does not express any APOBEC gene at a significant level.

Although these results may seem unsettling compared to what we report here, it is important to keep in mind that a 24h hypoxia exposure in cell lines is not fully representative of what occurs in primary tumours such as the tissues sequenced by TCGA. First, by simplifying a model down to only one cell type takes away the intercellular chemical signalling that might occur in actual tumours. We know that

APOBEC genes can be controlled by cytokines (Covino et al., 2018). By using cell lines, we are only studying the cellular response to hypoxia. It is possible that the activation of APOBEC genes under hypoxia is done through external mediators from the microenvironment which are not present in cell culture. Second, an acute exposure to hypoxia does not elicit the same cellular response as chronic hypoxia does, both of which occur in tumours (Pires et al., 2010). Since the effects seen here can only be generalised to acute hypoxia, it is possible that the changes in APOBEC expression seen in TCGA data is due to chronic hypoxia and selection forces. Namely, chronic hypoxia selects for TP53 mutations (Graeber et al., 1994; Kim et al., 1997) and in turn removes the inhibition of A3B expression (Middlebrooks et al., 2016). A 24h exposure to hypoxia in cell lines would not be able to fully activate TP53 to begin with, as it requires prolonged exposure to hypoxia (Haltermann and Federoff, 1999) or high levels of hypoxia (Hammond et al., 2002), and the model at hand does not allow for the effects of selection to be studied. Therefore, although the cell line models are useful in reporting the relationship between APOBEC and hypoxia in a specific context, it cannot be generalised to TCGA tumour samples.

4.3 APOBEC and Hypoxia in Subtypes of Breast Cancer

Breast invasive carcinoma can be classified into sub-categories. Traditionally, histological subtypes have been used to visually distinguish and diagnose different cancers using various staining techniques and immunohistochemistry (Malhotra et al., 2010). With advents in microarray technology at the turn of the millennium, there was an interest in developing a molecular classification of breast cancer based on gene expression profiles in order to outright replace histological subtyping. With time, the new technology led to novel ways of classifying breast cancer and proved to be useful in conjunction with histology instead of replacing it (Weigelt et al., 2009). Here, we extend our analysis of the relationship between APOBEC and hypoxia to four molecular subtypes of breast cancer in order to study the heterogeneity that exists among them.

First, we observe that in the basal-like and HER2-enriched subtypes, there is a larger differential expression of A3A and A3B in tumour tissue compared to normal. We hypothesize that the mechanisms explaining increased expression in the two subgroups could be different. It is known even before molecular profiling was introduced in breast cancer that basal-like breast cancers are hypoxic (Rakha et al., 2009; Tan et al., 2009). From our analyses, it is also the subtype with the highest median expression of the Buffa signature in tumours. This supports the hypothesis that hypoxia might be driving APOBEC expression, as we see that the most hypoxic subtype also shows the largest differential expression of active APOBEC genes. As for the HER2-enriched subtype, the finding is not surprising. The original reports of A3B mutagenesis in cancers show

that the HER2-enriched subtype is enriched for mutations of APOBEC origin (Roberts et al., 2013). More recently, it has also been shown that HER2 amplification directly contributes to an increase in DNA damage and A3B activity (Kanu et al., 2016). Thus, we expect that the HER2-enriched subtype expresses A3 family genes preferentially, and perhaps more so than other subtypes.

Interestingly, when looking at the correlation between the expression of APOBEC genes and that of the Buffa signature, we see no significant correlation in the basal-like and HER2-enriched subtypes. Upon closer inspection of the data distribution in these subtypes, it is clear that this lack of correlation is due to a small variance in the data. The basal-like subtype samples all show high levels of hypoxia signature expression, and the HER2-enriched samples all show a high A3A and A3B expression. Without a range of samples from low to high expression, the Spearman's correlation score is bound to be low or insignificant.

The luminal subtypes both show a lower differential expression of the hypoxia signature compared to the first two subtypes. In contrast, the two subtypes show different patterns of A3B differential expression. The luminal B subtype has a large differential expression of A3B. A similar effect has previously been reported by a study examining the prognostic implications of A3B in luminal breast cancers (Cescon et al., 2015). The study shows that the luminal B subtype shows higher absolute expression of A3B in the luminal B subtype compared to luminal A, both in TCGA data and in microarray data. The authors note that A3B expression is linked to prognosis, and that highly proliferative cancers such as the luminal B subtype is associated with high A3B expression. In terms of correlation between

APOBEC and hypoxia signature expression, both luminal subtypes show significant correlation between the two, consistent with the analysis on BRCA as a whole.

Overall, we see that the subtypes of breast cancer show a varying degree of APOBEC gene expression and differential expression. Although the correlation between APOBEC and hypoxia signature expression disappears in two subtypes, it is not a sign that there is not association between the two, as it may be a result of the low power to associate the two in small subsets of homogeneous data.

The next step in this direction would be to analyse the prevalence of APOBEC mutational signatures in breast cancer subtypes. Differences among subtypes in this regard would provide insight into mechanisms controlling APOBEC activity, as the breast cancer subtypes have key distinguishing features. Starting from those, we can narrow down the factors that might influence APOBEC expression and mutational activity such as hypoxia, TP53 status, oestrogen receptor status, and more.

4.4 Mutational Signatures to Study the Relationship Between APOBEC and Hypoxia

Expanding on the correlational evidence for the association of APOBEC gene expression with hypoxia, I decided to investigate the APOBEC mutational signature. While correlation in gene expression is evidence of interaction, whether direct or indirect, between two sets of genes, it does not show a functional relationship nor provide evidence that the expressed genes lead to functional protein products. Analysing the prevalence of the APOBEC mutational signature allows us to assess the extent of APOBEC activity on genomic DNA on a functional level. Although different APOBEC proteins bind and mutate different sequences, we are interested in the signature of A3A and A3B.

From the results gathered, we see that BRCA, HNSC, and LUAD samples show significant correlation between A3A and A3B expression and the prevalence of the APOBEC mutational signature using all available analysis methods. The APOBEC mutational signature was first linked to breast cancer (Burns et al., 2013a), and a study has previously investigated the correlation between A3A/A3B expression and the count of APOBEC-induced mutations, although with different normalisation methods than what is presented here (Roberts et al., 2013). Here, we reaffirm the findings of these previous studies by showing that the association between expression and mutational signature is still present using four methods and that all have different strengths and weaknesses. HNSC and LUAD are also cancer types that have been reported to have a strong presence of the APOBEC mutational signature, however there is no report for the

correlation between expression and mutational signature (Burns et al., 2013b; Roberts et al., 2013). Here, we add to the previous studies by showing that there is strong evidence of a positive relationship between the expression of A3A/A3B and the prevalence of the APOBEC mutational signature in HNSC and LUAD.

Other cancer types analysed by Burns et al. and Roberts et al. show varying degrees of correlation between A3A/A3B expression and the presence of the mutational signature. LUSC shows significant correlation using three methods. The correlation calculated using the count of APOBEC-induced mutations is not significant. This can be due to the fact that the counts are not normalised. Nevertheless, a significant correlation using the three other methods is still a strong indicator of a positive relationship between A3A/A3B expression and the presence of the mutational signature. HNSC was not analysed using the enrichment score and the other methods report significant correlation, therefore all available evidence points towards a significant positive relationship between expression and mutational signature.

Interestingly, BLCA shows no evidence of correlation between A3A/A3B expression and the presence of the APOBEC mutational signature. We find that, in fact, our results do not go against the findings of the previous studies. The mutational signature is indeed present in BLCA, but simply not correlated with A3A/A3B expression. BLCA is the cancer type with the highest proportion of non-null deconstructSigs scores for mutational signatures 2 and 13, the ones attributed to APOBEC mutagenesis, using both deconstructSigs methods. For the expression of A3B, we see that most samples in BLCA have high expression,

thus the power to see a correlation is reduced, as the variance in the data is reduced. Therefore, the absence of a significant Spearman's correlation between A3A/A3B expression and the presence of the APOBEC mutational signature in BLCA does not necessarily indicate that the APOBEC proteins are not active in BLCA. Instead, the high overall expression level agree with high detection of a mutational signature. Also, A3A and A3B, although being the members with the most evidence towards involvement in cancer, are not the only proteins capable of generating mutations in the TCW consensus motif and it might be that other members of the family are contributing to this profile. More research into the role of other A3 family members in BLCA is required, but this is beyond the scope of the current study.

Another observation from the data collected is the fact that using both deconstructSigs methods, the cancer types that report a significant proportion of samples with the APOBEC mutational signature corresponds to exactly the cancer types in which the signature was initially seen in 2013. BLCA, BRCA, HNSC, LUAD, and LUSC all have more than 10% of samples exhibiting the presence of signatures 2 and 13. This reaffirms the fact that APOBEC activity is present in these cancer types using another method of analysis.

Turning our attention to the association between hypoxia and the APOBEC mutational signature, we calculate the correlation between the mean expression of the Buffa signature and the prevalence of the mutational signature using the four methods mentioned above. Many cancers show significant correlation using the count of TCW to TTW/TGW mutations, but only BRCA shows significant

correlation using deconstructSigs. One possible explanation for this is that highly hypoxic breast cancers, such as those of the basal-like subtype (Koboldt et al., 2012), would have mutated TP53 (Graeber et al., 1994; Kim et al., 1997). Since TP53 has been shown to suppress the expression of A3B (Menendez et al., 2017), a cancer with mutated TP53 would show increased A3B expression (Middlebrooks et al., 2016), and in turn, APOBEC-induced mutations. In addition, there is evidence that A3B selectively mutates TP53 in breast cancer (Burns et al., 2013a). One can even say that this creates a positive feedback loop by inhibiting the suppression of A3B expression. However, this is unlikely to be widespread, as if the feedback loop were to commonly exist, we would expect a higher overall presence of the APOBEC mutational signature and a higher proportion of samples with significant levels of the signature in breast cancer.

The finding regarding the association between hypoxia and the APOBEC mutational signature in BRCA is novel, as no study has investigated this relationship yet. However, this also means that the lack of available evidence and the lack of other types of studies on this topic makes our finding unsupported by independent evidence. Two main concerns stem from this. Is the correlation between the hypoxia signature expression and the APOBEC mutational signature presence truly representative of an association between hypoxia and APOBEC activity in BRCA? And is the absence of such correlation in other cancer types truly a sign that the association is not present? These are questions that are best answered by more studies using different methodology to validate our findings. One way to start looking further at this question is by studying the phenomenon in cell culture. This way, we can artificially induce hypoxia in controlled conditions

to observe the change in APOBEC-induced mutations. However, cell line models have their own limitations, as outlined above. We must be careful when drawing conclusions from them as the results may not be fully translatable to in vivo environments.

One limitation of this study is that we base our conclusions drawn from the results on the assumption that A3A and A3B are the proteins carrying out the TCW to TTV/TGW mutations observed. Although there is a great amount of evidence on A3A and A3B being the primary active APOBEC proteins in cancer that induce TCW to TTV/TGW mutations, there exists other APOBEC proteins that can carry out these mutations, namely the rest of the A3 family except A3G. We have already established that they are relatively highly expressed compared to other APOBEC genes. However, we have not assessed the level of involvement of these proteins in inducing mutations based on their cellular localisation. It is possible that whenever we do not see correlation between A3A/A3B expression and the prevalence of the APOBEC mutational signature when the signature is present, it is because another A3 protein is being activated in the nucleus, namely A3C and A3H. As a corollary, it is possible that the expression of another A3 gene correlates to a higher extent with the mutational signature, thus being a sign that A3A/A3B might not be the main active APOBEC proteins in that cancer.

In addition, the relationship inferred here between APOBEC mutations and hypoxia does not give clues as to the mechanism underlying it. It would be necessary to investigate whether hypoxia is inducing APOBEC activity or whether samples with high APOBEC activity are selected for in hypoxia. This can be done

by analysing the relationship between APOBEC gene amplification and the expression of a hypoxia signature, or by assessing amplification in cell lines exposed to hypoxia. If hypoxia is driving the expression of APOBEC genes through cellular pathways or transcription factors, it would be necessary to investigate the exact mechanism through which it acts.

Finally, a change in the prevalence of the APOBEC mutational signature can occur without any functional change to APOBEC gene expression and regulation. Other factors that can influence the extent to which the APOBEC signature is present are namely the availability of ssDNA in the nucleus, the import rate of APOBEC proteins into the nucleus, and the presence of factors inhibiting the action of APOBEC proteins. In the first case, increased double-stranded breaks and increased DNA repair by homologous recombination increases the abundance of ssDNA in the nucleus. In the second case, cellular mechanisms can promote or inhibit the import of APOBEC proteins into the nucleus. Namely, A3A import is inhibited in monocytes under normal conditions (Land et al., 2013). Overall, we must make assumptions about the state of these possible cellular mechanisms in order to draw conclusions from our data.

4.5 Pathway Analysis Shows a Functional Relationship between APOBEC and Hypoxia in Breast Cancer

The gene co-expression and pathway analysis shows that A3B is functionally closer to hypoxia than the other A3 genes. We see that among the genes co-expressed with A3B, the vast majority are also among the top 500 genes co-expressed with the Buffa hypoxia signature. On the contrary, only two genes are also co-expressed in any five A3 genes, and there is no overlap between the genes co-expressed with any five A3 genes and the top 500 genes co-expressed with the Buffa signature. The fact that A3B is functionally dissociated with the other APOBEC genes is concordant with the self-correlation analysis and principal component analysis conducted previously. It shows that the expression of A3B in cancer might be controlled through mechanisms that are different from the regulation of other A3 family genes. This is supported by evidence in the literature. Notably, TP53 acts differently on the expression of A3A and A3B (Menendez et al., 2017; Middlebrooks et al., 2016). In addition, A3 family genes are regulated through different mechanisms in innate immunity as well (Covino et al., 2018). Notably, A3A and A3G can be induced by interferons, whereas A3B can be induced by the NF- κ B signalling pathway. All of this indicates that it is not unexpected that A3B's expression in cancer is controlled in a manner that is dissociated from the rest of the A3 family. In turn, the evidence here suggests that A3B might be the main or even the only A3 family gene that is upregulated in hypoxia.

This effect is even more apparent when looking at the pathways enriched in the genes co-expressed with A3B, hypoxia, and any five A3 genes. Among the top 10 pathways for each group, the pathways enriched for in genes co-expressed with any five A3 genes is completely dissociated from the two other groups, which have three overlapping pathways.

4.6 Significance of the Work

The significance of the work done here is two-fold. First, it provides evidence for the regulation of APOBEC gene expression in cancer, with hypoxia being a potential driver. Second, it provides a commentary and comparison for the different methods of analysing the APOBEC mutational signature and mutational signatures in general.

Understanding the regulation of APOBEC genes can eventually lead to advances in therapies targeting APOBEC proteins, with the potential of reducing the mutation load in tumours with high native APOBEC expression. However, therapies revolving around APOBEC proteins still have a long way to go before reaching maturity and clinical applicability since their involvement and regulation in cancer is still poorly understood. In addition, recent evidence suggests that A3B overexpression does not cause adverse clinical phenomena, but rather that it may be an effect of them (Cescon et al., 2015). Nevertheless, APOBEC proteins are a cause of concern in cancer as it increases the mutation load, which can lead to genomic instability and cancer progression. Understanding its regulation and involvement in hypoxia can eventually lead to classification models and guide therapies for better effectiveness.

On a different front, there is currently no consensus as to how the APOBEC mutational signature should be quantified. This leads us to seek a comparison of the methods available, mainly between methods based on an empirical definition of the signature and those based on a mathematical definition. It is possible to say that since the mathematical model is generated in an a priori and unbiased

manner, it may better reflect the true mutational signature by including unspecific binding and secondary binding. It also has the power to deconvolute complex mixes of signatures that might be present in cancers with different aetiologies, such as lung cancer. However, the same methods introduce more noise in the data by nature of their regression algorithms and these signatures are not all empirically validated. In addition, the use of these signatures is not portable to other data types, as the signatures are derived based on a specific set of samples. These signatures may not accurately portray the mutational patterns in samples analysed by different sequencing methods or samples that might have a genomic composition that is vastly different from the samples used in the original study (Alexandrov et al., 2013). A re-calculation of mutational signatures using the original methods would be required for proper analysis. In fact, in ideal situations, we would always analyse the mutational signatures generated from the samples that we want to analyse. However, this process is time- and computational power-demanding, and not always viable. The empirical definition of the APOBEC mutational signature does not face this problem, as the simple mutational pattern used can be translated to any type of data reporting single nucleotides.

Overall, the work here serves as a primer for further investigation of the involvement of APOBEC family proteins in cancer and hypoxia. It will lead to more advanced studies investigating specific mechanisms of induction for APOBEC genes and their association with other hypoxia-related pathways. It opens the door for new studies to follow the model of using two main types of methods to look at the APOBEC mutational signature, and to adopt it as a way to get more consistency in the results presented.

4.7 General Remarks

Here, we present evidence for the association between hypoxia, APOBEC expression, and APOBEC-mediated mutagenesis in cancer. APOBEC has been known to be involved in cancer mutagenesis, but its regulation is still poorly understood, especially in the context of cancer. Preliminary evidence from cell line models show the activation of APOBEC genes in immune cells exposed to hypoxic conditions, driving us to investigate the same phenomenon in cancer.

Although the evidence presented here is novel and significant, it carries its intrinsic limitations. Our analysis methods do not investigate any cellular or systemic mechanism behind what we propose as an induction of APOBEC gene expression via hypoxia. Without this level of evidence, we cannot say with absolute confidence that the correlations seen between APOBEC expression and hypoxia, using any analysis performed here, is truly indicative of hypoxia driving APOBEC expression and activity. It may be an upstream mechanism leading to both hypoxia and APOBEC expression to be increased. Future work would be needed to address the questions that arise from this study, namely studies investigating a mechanism of induction for APOBEC genes under hypoxia.

In addition, the conclusions drawn from our analysis of the APOBEC mutational signature carry significant assumptions about factors that we cannot study at the moment, such as the inhibition of APOBEC function by other proteins, the import rate of APOBEC proteins into the nucleus, and the availability of ssDNA in the genome. These are all valid questions that need to be addressed in future studies.

It is also important to note that the association that other factors may be in play at controlling the expression and activity of APOBEC genes and proteins, and they need to be studied in order to deconvolute the factors affecting APOBEC in cancer. The nature of cancers also makes the complete deconvolution of factors leading to a specific phenomenon in cancer hard to achieve, as the heterogeneity of cancers is large. Environmental, genetic, and lifestyle factors, combined with the possibility of sporadic mutations all lead to noise when analysing the upstream causes of specific cancers.

4.8 Conclusion

In conclusion, my work here provides evidence for the association between APOBEC and hypoxia in cancer and serves as a primer for future studies to investigate the mechanisms of this association in addition to providing a comparison of the ways that we can measure and quantify the APOBEC mutational signature. The evidence here is preliminary and inconclusive as to the general question at hand and raises many more questions about APOBEC and hypoxia in cancer. Much future work is required to fully answer these questions and to obtain a complete picture of APOBEC and hypoxia in cancer.

Bibliography

Alexandrov, L.B., Nik-Zainal, S., Wedge, D.C., Aparicio, S.A.J.R., Behjati, S., Biankin, A. V., Bignell, G.R., Bolli, N., Borg, A., Børresen-Dale, A.-L., et al. (2013). Signatures of mutational processes in human cancer. *Nature* 500, 415–421.

Aran, D., Hu, Z., and Butte, A.J. (2017). xCell: Digitally portraying the tissue cellular heterogeneity landscape. *Genome Biol.* 18.

Baysal, B.E., De Jong, K., Liu, B., Wang, J., Patnaik, S.K., Wallace, P.K., and Taggart, R.T. (2013). Hypoxia-inducible C-to-U coding RNA editing downregulates SDHB in monocytes. *PeerJ* 1, e152.

Bennett, R.P., Presnyak, V., Wedekind, J.E., and Smith, H.C. (2008). Nuclear exclusion of the HIV-1 host defense factor APOBEC3G requires a novel cytoplasmic retention signal and is not dependent on RNA binding. *J. Biol. Chem.* 283, 7320–7327.

Berry, M.W., Browne, M., Langville, A.N., Pauca, V.P., and Plemmons, R.J. (2007). Algorithms and applications for approximate nonnegative matrix factorization. *Comput. Stat. Data Anal.* 52, 155–173.

Bonvin, M., Achermann, F., Greeve, I., Stroka, D., Keogh, A., Inderbitzin, D., Candinas, D., Sommer, P., Wain-Hobson, S., Vartanian, J.P., et al. (2006). Interferon-inducible expression of APOBEC3 editing enzymes in human hepatocytes and inhibition of hepatitis B virus replication. *Hepatology* 43, 1364–1374.

Brash, D.E., Rudolph, J.A., Simon, J.A., Lin, A., McKenna, G.J., Baden, H.P., Halperin, A.J., and Ponten, J. (1991). A role for sunlight in skin cancer: UV-induced p53 mutations in squamous cell carcinoma. *Proc. Natl. Acad. Sci.* 88, 10124–10128.

Bressac, B., Puisieux, A., Kew, M., Volkmann, M., Bozcall, S., Bella Mura, J., de la Monte, S., Carlson, R., Blum, H., Wands, J., et al. (1991). p53 mutation in hepatocellular carcinoma after aflatoxin exposure. *Lancet* 338, 1356–1359.

Buffa, F.M., Harris, A.L., West, C.M., and Miller, C.J. (2010). Large meta-analysis of multiple cancers reveals a common, compact and highly prognostic hypoxia metagene. *Br. J. Cancer* 102, 428–435.

Burns, M.B., Lackey, L., Carpenter, M.A., Rathore, A., Land, A.M., Leonard, B., Refsland, E.W., Kotandeniya, D., Tretyakova, N., Nikas, J.B., et al. (2013a). APOBEC3B is an enzymatic source of mutation in breast cancer. *Nature* 494, 366–370.

Burns, M.B., Temiz, N.A., and Harris, R.S. (2013b). Evidence for APOBEC3B mutagenesis in multiple human cancers. *Nat. Genet.* 45, 977–983.

Carmona-Saez, P., Chagoyen, M., Tirado, F., Carazo, J.M., and Pascual-Montano, A. (2007). GENECODIS: a web-based tool for finding significant concurrent annotations in gene lists. *Genome Biol.* 8, R3.

Cerami, E., Gao, J., Dogrusoz, U., Gross, B.E., Sumer, S.O., Aksoy, B.A., Jacobsen, A., Byrne, C.J., Heuer, M.L., Larsson, E., et al. (2012). The cBio Cancer Genomics Portal: An open platform for exploring multidimensional cancer genomics data. *Cancer Discov.* 2, 401–404.

Cescon, D.W., Haibe-Kains, B., and Mak, T.W. (2015). APOBEC3B expression in breast cancer reflects cellular proliferation, while a deletion polymorphism is associated with immune activation. *Proc. Natl. Acad. Sci. U. S. A.* 112, 2841–2846.

Chaipan, C., Smith, J.L., Hu, W.-S., and Pathak, V.K. (2013). APOBEC3G Restricts HIV-1 to a Greater Extent than APOBEC3F and APOBEC3DE in Human Primary CD4+ T Cells and Macrophages. *J. Virol.* 87, 444–453.

Chan, K., Roberts, S.A., Klimczak, L.J., Sterling, J.F., Saini, N., Malc, E.P., Kim, J., Kwiatkowski, D.J., Fargo, D.C., Mieczkowski, P.A., et al. (2015). An

APOBEC3A hypermutation signature is distinguishable from the signature of background mutagenesis by APOBEC3B in human cancers. *Nat. Genet.* **47**, 1067–1072.

Covino, D.A., Gauzzi, M.C., and Fantuzzi, L. (2018). Understanding the regulation of APOBEC3 expression: Current evidence and much to learn. *J. Leukoc. Biol.* **103**, 433–444.

Davidson, N.O., Innerarity, T.L., Scott, J., Smith, H., Driscoll, D.M., Teng, B., and Chan, L. (1995). Proposed nomenclature for the catalytic subunit of the mammalian apolipoprotein B mRNA editing enzyme: APOBEC-1. *RNA* **1**, 3.

Dhawan, A., Barberis, A., Cheng, W., Domingo, E., Maughan, T., Scott, J.G., Harris, A.L., Buffa, F.M., Kingdom, U., Clinic, C., et al. (2017). sigQC : A procedural approach evaluation of gene signatures for standardising the Correspondence : francesca.buffa@oncology.ox.ac.uk Flowchart of steps involved in the proposed sigQC protocol and sample output plots produced by the sigQC R package . Abstr. *BioRxiv*.

Forbes, S.A., Beare, D., Boutselakis, H., Bamford, S., Bindal, N., Tate, J., Cole, C.G., Ward, S., Dawson, E., Ponting, L., et al. (2017). COSMIC: somatic cancer genetics at high-resolution. *Nucleic Acids Res.* **45**, D777–D783.

Gao, J., Aksoy, B.A., Dogrusoz, U., Dresdner, G., Gross, B., Sumer, S.O., Sun, Y., Jacobsen, A., Sinha, R., Larsson, E., et al. (2013). Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci. Signal.* **6**, pl1-pl1.

Gibbs, P.E.M., and Lawrence, C.W. (1995). Novel mutagenic properties of abasic sites in *saccharomyces cerevisiae*. *J. Mol. Biol.* **251**, 229–236.

Graeber, T.G., Peterson, J.F., Tsai, M., Monica, K., Fornace, A.J., and Giaccia, A.J. (1994). Hypoxia induces accumulation of p53 protein, but activation of a G1-phase checkpoint by low-oxygen conditions is independent of p53 status. *Mol. Cell. Biol.* **14**, 6264–6277.

Graeber, T.G., Osmanian, C., Jacks, T., Housman, D.E., Koch, C.J., Lowe, S.W., and Giaccia, A.J. (1996). Hypoxia-mediated selection of cells with diminished apoptotic potential in solid tumours. *Nature* 379, 88–91.

Halterman, M.W., and Federoff, H.J. (1999). HIF-1 α and p53 Promote Hypoxia-Induced Delayed Neuronal Death in Models of CNS Ischemia. *Exp. Neurol.* 159, 65–72.

Hammond, E.M., Denko, N.C., Dorie, M.J., Abraham, R.T., and Giaccia, A.J. (2002). Hypoxia links ATR and p53 through replication arrest. *Mol. Cell. Biol.* 22, 1834–1843.

Hanahan, D., and Weinberg, R.A. (2011). Hallmarks of cancer: The next generation. *Cell* 144, 646–674.

Harris, R.S., Petersen-Mahrt, S.K., and Neuberger, M.S. (2002). RNA editing enzyme APOBEC1 and some of its homologs can act as DNA mutators. *Mol. Cell* 10, 1247–1253.

Jarmuz, A., Chester, A., Bayliss, J., Gisbourne, J., Dunham, I., Scott, J., and Navaratnam, N. (2002). An anthropoid-specific locus of orphan C to U RNA-editing enzymes on chromosome 22. *Genomics* 79, 285–296.

Kanu, N., Cerone, M.A., Goh, G., Zalmas, L.P., Bartkova, J., Dietzen, M., McGranahan, N., Rogers, R., Law, E.K., Gromova, I., et al. (2016). DNA replication stress mediates APOBEC3 family mutagenesis in breast cancer. *Genome Biol.* 17, 185.

Kim, C.Y., Tsai, M.H., Osmanian, C., Graeber, T.G., Lee, J.E., Giffard, R.G., DiPaolo, J.A., Peehl, D.M., and Giaccia, A.J. (1997). Selection of human cervical epithelial cells that possess reduced apoptotic potential to low-oxygen conditions. *Cancer Res.* 57, 4200–4204.

Koboldt, D.C., Fulton, R.S., McLellan, M.D., Schmidt, H., Kalicki-Veizer, J., McMichael, J.F., Fulton, L.L., Dooling, D.J., Ding, L., Mardis, E.R., et al. (2012). Comprehensive molecular portraits of human breast tumours. *Nature* 490, 61–

70.

Krokan, H.E., and Bjørås, M. (2013). Base excision repair. *Cold Spring Harb. Perspect. Biol.* 5, a012583.

Lackey, L., Demorest, Z.L., Land, A.M., Hultquist, J.F., Brown, W.L., and Harris, R.S. (2012). APOBEC3B and AID have similar nuclear import mechanisms. *J. Mol. Biol.* 419, 301–314.

Land, A.M., Law, E.K., Carpenter, M.A., Lackey, L., Brown, W.L., and Harris, R.S. (2013). Endogenous APOBEC3A DNA cytosine deaminase is cytoplasmic and nongenotoxic. *J. Biol. Chem.* 288, 17253–17260.

Leonard, B., McCann, J.L., Starrett, G.J., Kosyakovsky, L., Luengas, E.M., Molan, A.M., Burns, M.B., McDougale, R.M., Parker, P.J., Brown, W.L., et al. (2015). The PKC/NF- κ B signaling pathway induces APOBEC3B expression in multiple human cancers. *Cancer Res.* 75, 4538–4547.

Li, B., and Dewey, C.N. (2011). RSEM: Accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* 12.

Malhotra, G.K., Zhao, X., Band, H., and Band, V. (2010). Histological, molecular and functional subtypes of breast cancers. *Cancer Biol. Ther.* 10, 955–960.

Mangeat, B., Turelli, P., Caron, G., Friedli, M., Perrin, L., and Trono, D. (2003). Broad antiretroviral defence by human APOBEC3G through lethal editing of nascent reverse transcripts. *Nature* 424, 99–103.

Marino, D., Perković, M., Hain, A., Vasudevan, A.A.J., Hofmann, H., Hanschmann, K.M., Mühlebach, M.D., Schumann, G.G., König, R., Cichutek, K., et al. (2016). APOBEC4 Enhances the Replication of HIV-1. *PLoS One* 11, e0155422.

Maruyama, W., Shirakawa, K., Matsui, H., Matsumoto, T., Yamazaki, H., Sarca, A.D., Kazuma, Y., Kobayashi, M., Shindo, K., and Takaori-Kondo, A. (2016). Classical NF- κ B pathway is responsible for APOBEC3B expression in cancer

cells. *Biochem. Biophys. Res. Commun.* 478, 1466–1471.

Matsumoto, T., Marusawa, H., Endo, Y., Ueda, Y., Matsumoto, Y., and Chiba, T. (2006). Expression of APOBEC2 is transcriptionally regulated by NF- κ B in human hepatocytes. *FEBS Lett.* 580, 731–735.

Menendez, D., Nguyen, T.-A., Snipe, J., and Resnick, M.A. (2017). The Cytidine Deaminase APOBEC3 Family Is Subject to Transcriptional Regulation by p53. *Mol. Cancer Res.* 15, 735–743.

Middlebrooks, C.D., Banday, A.R., Matsuda, K., Udquim, K.I., Onabajo, O.O., Paquin, A., Figueroa, J.D., Zhu, B., Koutros, S., Kubo, M., et al. (2016). Association of germline variants in the APOBEC3 region with cancer risk and enrichment with APOBEC-signature mutations in tumors. *Nat. Genet.* 48, 1330–1338.

Muramatsu, M., Sankaranand, V.S., Anant, S., Sugai, M., Kinoshita, K., Davidson, N.O., and Honjo, T. (1999). Specific expression of activation-induced cytidine deaminase (AID), a novel member of the RNA-editing deaminase family in germinal center B cells. *J. Biol. Chem.* 274, 18470–18476.

Muramatsu, M., Kinoshita, K., Fagarasan, S., Yamada, S., Shinkai, Y., and Honjo, T. (2000). Class switch recombination and hypermutation require activation-induced cytidine deaminase (AID), a potential RNA editing enzyme. *Cell* 102, 553–563.

Navaratnam, N., Patel, D., Shah, R., Greeve, J.C., Powell, L.M., Knott, T.J., and Scott, J. (1991). An additional editing site is present in apolipoprotein B mRNA. *Nucleic Acids Res.* 19, 1741–1744.

Navaratnam, N., Morrison, J.R., Bhattacharya, S., Patel, D., Funahashi, T., Giannoni, F., Teng, B.B., Davidson, N.O., and Scott, J. (1993). The p27 catalytic subunit of the apolipoprotein B mRNA editing enzyme is a cytidine deaminase. *J. Biol. Chem.* 268, 20709–20712.

Nik-Zainal, S., Alexandrov, L.B., Wedge, D.C., Van Loo, P., Greenman, C.D.,

Raine, K., Jones, D., Hinton, J., Marshall, J., Stebbings, L.A., et al. (2012). Mutational processes molding the genomes of 21 breast cancers. *Cell* 149, 979–993.

Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H., and Kanehisa, M. (1999). KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 27, 29–34.

Okuyama, S., Marusawa, H., Matsumoto, T., Ueda, Y., Matsumoto, Y., Endo, Y., Takai, A., and Chiba, T. (2012). Excessive activity of apolipoprotein B mRNA editing enzyme catalytic polypeptide 2 (APOBEC2) contributes to liver and lung tumorigenesis. *Int. J. Cancer* 130, 1294–1301.

Perou, C.M., Sørile, T., Eisen, M.B., Van De Rijn, M., Jeffrey, S.S., Renshaw, C.A., Pollack, J.R., Ross, D.T., Johnsen, H., Akslen, L.A., et al. (2000). Molecular portraits of human breast tumours. *Nature* 406, 747–752.

Pfeifer, G.P., Denissenko, M.F., Olivier, M., Tretyakova, N., Hecht, S.S., and Hainaut, P. (2002). Tobacco smoke carcinogens, DNA damage and p53 mutations in smoking-associated cancers. *Oncogene* 21–48, 7435–7451.

Pires, I.M., Bencokova, Z., Milani, M., Folkes, L.K., Li, J.-L., Stratford, M.R., Harris, A.L., and Hammond, E.M. (2010). Effects of Acute versus Chronic Hypoxia on DNA Damage Responses and Genomic Instability. *Cancer Res.* 70, 925–935.

Pleasance, E.D., Cheetham, R.K., Stephens, P.J., McBride, D.J., Humphray, S.J., Greenman, C.D., Varela, I., Lin, M.L., Ordóñez, G.R., Bignell, G.R., et al. (2010a). A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature* 463, 191–196.

Pleasance, E.D., Stephens, P.J., O’Meara, S., McBride, D.J., Meynert, A., Jones, D., Lin, M.L., Beare, D., Lau, K.W., Greenman, C., et al. (2010b). A small-cell lung cancer genome with complex signatures of tobacco exposure. *Nature* 463, 184–190.

- Rakha, E.A., Elsheikh, S.E., Aleskandarany, M.A., Habashi, H.O., Green, A.R., Powe, D.G., El-Sayed, M.E., Benhasouna, A., Brunet, J.-S., Akslen, L.A., et al. (2009). Triple-Negative Breast Cancer: Distinguishing between Basal and Nonbasal Subtypes. *Clin. Cancer Res.* 15, 2302–2310.
- Roberts, S.A., Lawrence, M.S., Klimczak, L.J., Grimm, S.A., Fargo, D., Stojanov, P., Kiezun, A., Kryukov, G. V, Carter, S.L., Saksena, G., et al. (2013). An APOBEC cytidine deaminase mutagenesis pattern is widespread in human cancers. *Nat. Genet.* 45, 970–976.
- Rogozin, I.B., Basu, M.K., Jordan, I.K., Pavlov, Y.I., and Koonin, E. V. (2005). APOBEC4, a new member of the AID/APOBEC family of polynucleotide (deoxy)cytidine deaminases predicted by computational analysis. *Cell Cycle* 4, 1281–1285.
- Rosenthal, R., McGranahan, N., Herrero, J., Taylor, B.S., and Swanton, C. (2016). deconstructSigs: Delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution. *Genome Biol.* 17.
- Salter, J.D., Bennett, R.P., and Smith, H.C. (2016). The APOBEC Protein Family: United by Structure, Divergent in Function. *Trends Biochem. Sci.* 41, 578–594.
- Sanger, F., Nicklen, S., and Coulson, A.R. (1977). DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci.* 74, 5463–5467.
- Sato, Y., Probst, H.C., Tatsumi, R., Ikeuchi, Y., Neuberger, M.S., and Rada, C. (2010). Deficiency in APOBEC2 leads to a shift in muscle fiber type, diminished body mass, and myopathy. *J. Biol. Chem.* 285, 7111–7118.
- Sermeus, A., and Michiels, C. (2011). Reciprocal influence of the p53 and the hypoxic pathways. *Cell Death Dis.* 2, e164.
- Sharma, S., Patnaik, S.K., Thomas Taggart, R., Kannisto, E.D., Enriquez, S.M., Gollnick, P., and Baysal, B.E. (2015). APOBEC3A cytidine deaminase induces

RNA editing in monocytes and macrophages. *Nat. Commun.* 6.

Sharma, S., Patnaik, S.K., Taggart, R.T., and Baysal, B.E. (2016). The double-domain cytidine deaminase APOBEC3G is a cellular site-specific RNA editing enzyme. *Sci. Rep.* 6.

Sharma, S., Patnaik, S.K., Kemer, Z., and Baysal, B.E. (2017). Transient overexpression of exogenous APOBEC3A causes C-to-U RNA editing of thousands of genes. *RNA Biol.* 14, 603–610.

Sheehy, A.M., Gaddis, N.C., Choi, J.D., and Malim, M.H. (2002). Isolation of a human gene that inhibits HIV-1 infection and is suppressed by the viral Vif protein. *Nature* 418, 646–650.

Smith, H.C., Bennett, R.P., Kizilyer, A., McDougall, W.M., and Prohaska, K.M. (2012). Functions and regulation of the APOBEC family of proteins. *Semin. Cell Dev. Biol.* 23, 258–268.

Sonna, L.A., Cullivan, M.L., Sheldon, H.K., Pratt, R.E., and Lilly, C.M. (2003). Effect of hypoxia on gene expression by human hepatocytes (HepG2). *Physiol. Genomics* 12, 195–207.

Sorlie, T., Perou, C.M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., Hastie, T., Eisen, M.B., van de Rijn, M., Jeffrey, S.S., et al. (2001). Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc. Natl. Acad. Sci.* 98, 10869–10874.

Tan, E.Y., Yan, M., Campo, L., Han, C., Takano, E., Turley, H., Candiloro, I., Pezzella, F., Gatter, K.C., Millar, E.K.A., et al. (2009). The key hypoxia regulated gene CAIX is upregulated in basal-like breast tumours and is associated with resistance to chemotherapy. *Br. J. Cancer* 100, 405–411.

Teng, B., Burant, C., and Davidson, N. (1993). Molecular cloning of an apolipoprotein B messenger RNA editing protein. *Science* (80-.). 260, 1816–1819.

Thielen, B.K., McNevin, J.P., McElrath, M.J., Hunt, B.V.S., Klein, K.C., and Lingappa, J.R. (2010). Innate immune signaling induces high levels of TC-specific deaminase activity in primary monocyte-derived cells through expression of APOBEC3A isoforms. *J. Biol. Chem.* 285, 27753–27766.

Vaupel, P., and Mayer, A. (2007). Hypoxia in cancer: significance and impact on clinical outcome. *Cancer Metastasis Rev.* 26, 225–239.

Weigelt, B., Baehner, F.L., and Reis-Filho, J.S. (2009). The contribution of gene expression profiling to breast cancer classification, prognostication and prediction: a retrospective of the last decade. *J. Pathol.* 220, n/a-n/a.

Witkin, E.M. (1969). Ultraviolet-Induced Mutation and DNA Repair. *Annu. Rev. Microbiol.* 23, 487–514.

Supplementary Tables

Table S1: Genes in the Buffa Hypoxia Signature

Gene name	Gene description
ACOT7	acyl-CoA thioesterase 7 [Source:HGNC Symbol;Acc:HGNC:24157]
ADM	adrenomedullin [Source:HGNC Symbol;Acc:HGNC:259]
AK4	adenylate kinase 4 [Source:HGNC Symbol;Acc:HGNC:363]
ANKRD37	ankyrin repeat domain 37 [Source:HGNC Symbol;Acc:HGNC:29593]
ANLN	anillin actin binding protein [Source:HGNC Symbol;Acc:HGNC:14082]
BNIP3	BCL2 interacting protein 3 [Source:HGNC Symbol;Acc:HGNC:1084]
MRGBP	MRG domain binding protein [Source:HGNC Symbol;Acc:HGNC:15866]
HILPDA	hypoxia inducible lipid droplet associated [Source:HGNC Symbol;Acc:HGNC:28859]
CA9	carbonic anhydrase 9 [Source:HGNC Symbol;Acc:HGNC:1383]
CDKN3	cyclin dependent kinase inhibitor 3 [Source:HGNC Symbol;Acc:HGNC:1791]
CHCHD2	coiled-coil-helix-coiled-coil-helix domain containing 2 [Source:HGNC Symbol;Acc:HGNC:21645]
CORO1C	coronin 1C [Source:HGNC Symbol;Acc:HGNC:2254]
CTSL2	cathepsin V [Source:HGNC Symbol;Acc:HGNC:2538]
DDIT4	DNA damage inducible transcript 4 [Source:HGNC Symbol;Acc:HGNC:24944]
ENO1	enolase 1 [Source:HGNC Symbol;Acc:HGNC:3350]
ESRP1	epithelial splicing regulatory protein 1 [Source:HGNC Symbol;Acc:HGNC:25966]
GAPDH	glyceraldehyde-3-phosphate dehydrogenase [Source:HGNC Symbol;Acc:HGNC:4141]

GPI	glucose-6-phosphate isomerase [Source:HGNC Symbol;Acc:HGNC:4458]
HK2	hexokinase 2 [Source:HGNC Symbol;Acc:HGNC:4923]
KIF20A	kinesin family member 20A [Source:HGNC Symbol;Acc:HGNC:9787]
KIF4A	kinesin family member 4A [Source:HGNC Symbol;Acc:HGNC:13339]
LDHA	lactate dehydrogenase A [Source:HGNC Symbol;Acc:HGNC:6535]
LRRC42	leucine rich repeat containing 42 [Source:HGNC Symbol;Acc:HGNC:28792]
MAD2L2	mitotic arrest deficient 2 like 2 [Source:HGNC Symbol;Acc:HGNC:6764]
MAP7D1	MAP7 domain containing 1 [Source:HGNC Symbol;Acc:HGNC:25514]
MCTS1	MCTS1, re-initiation and release factor [Source:HGNC Symbol;Acc:HGNC:23357]
MIF	macrophage migration inhibitory factor [Source:HGNC Symbol;Acc:HGNC:7097]
MRPL13	mitochondrial ribosomal protein L13 [Source:HGNC Symbol;Acc:HGNC:14278]
MRPL15	mitochondrial ribosomal protein L15 [Source:HGNC Symbol;Acc:HGNC:14054]
MRPS17	mitochondrial ribosomal protein S17 [Source:HGNC Symbol;Acc:HGNC:14047]
NDRG1	N-myc downstream regulated 1 [Source:HGNC Symbol;Acc:HGNC:7679]
P4HA1	prolyl 4-hydroxylase subunit alpha 1 [Source:HGNC Symbol;Acc:HGNC:8546]
PFKP	phosphofructokinase, platelet [Source:HGNC Symbol;Acc:HGNC:8878]
PGAM1	phosphoglycerate mutase 1 [Source:HGNC Symbol;Acc:HGNC:8888]
PGAM4	phosphoglycerate mutase family member 4 [Source:HGNC Symbol;Acc:HGNC:21731]
PGK1	phosphoglycerate kinase 1 [Source:HGNC Symbol;Acc:HGNC:8896]
PNP	purine nucleoside phosphorylase [Source:HGNC Symbol;Acc:HGNC:7892]
PSMA7	proteasome subunit alpha 7 [Source:HGNC Symbol;Acc:HGNC:9536]

PSRC1	proline and serine rich coiled-coil 1 [Source:HGNC Symbol;Acc:HGNC:24472]
SEC61G	Sec61 translocon gamma subunit [Source:HGNC Symbol;Acc:HGNC:18277]
SHCBP1	SHC binding and spindle associated 1 [Source:HGNC Symbol;Acc:HGNC:29547]
SLC16A1	solute carrier family 16 member 1 [Source:HGNC Symbol;Acc:HGNC:10922]
SLC25A32	solute carrier family 25 member 32 [Source:HGNC Symbol;Acc:HGNC:29683]
SLC2A1	solute carrier family 2 member 1 [Source:HGNC Symbol;Acc:HGNC:11005]
TPI1	triosephosphate isomerase 1 [Source:HGNC Symbol;Acc:HGNC:12009]
TUBA1A	tubulin alpha 1a [Source:HGNC Symbol;Acc:HGNC:20766]
TUBA1B	tubulin alpha 1b [Source:HGNC Symbol;Acc:HGNC:18809]
TUBA1C	tubulin alpha 1c [Source:HGNC Symbol;Acc:HGNC:20768]
TUBB6	tubulin beta 6 class V [Source:HGNC Symbol;Acc:HGNC:20776]
UTP11	UTP11, small subunit processome component [Source:HGNC Symbol;Acc:HGNC:24329]
VEGFA	vascular endothelial growth factor A [Source:HGNC Symbol;Acc:HGNC:12680]
YKT6	YKT6 v-SNARE homolog [Source:HGNC Symbol;Acc:HGNC:16959]

Genes contained in the Buffa hypoxia signature (Buffa et al., 2010). These are the top ranked genes by connectivity in a gene co-expression network analysis starting from 10 curated seed genes. Gene descriptions generated through g:Convert, a tool to convert gene names and annotations:

<https://biit.cs.ut.ee/gprofiler/gconvert.cgi> (Reimand et al., 2016).

Table S2: Genes Co-Expressed with APOBEC3 Genes in Breast Cancer

Genes co-expressed with a minimum of five A3 genes

SH3BP1	WARS	AIF1	AKNA	APOBEC3G	APOL1
APOL2	APOL3	APOL6	ARHGAP25	ARHGAP27	ARHGAP30
B2M	BIRC3	BTN2A2	BTN3A1	BTN3A2	BTN3A3
C10orf54	C19orf66	C1QA	C1QB	C1QC	C1R
C1orf38	CASP1	CASP4	CCL5	CD37	CD40
CD48	CD4	CD52	CD53	CD74	CD86
CD97	CIITA	CMPK2	CORO1A	CSF2RB	CTSS
CXCL10	CXCL9	CYBA	CYTH4	DAPP1	DEF6
DENND3	EPSTI1	EVI2B	FERMT3	FGL2	FMNL1
FYB	GBP1	GBP2	GBP4	GBP5	GIMAP4
GIMAP5	GMFG	GMIP	GPSM3	HCK	HCLS1
HCP5	HERC5	HERC6	HLA-A	HLA-B	HLA-C
HLA-DMA	HLA-DMB	HLA-DPA1	HLA-DPB1	HLA-DQB1	HLA-DRA
HLA-DRB1	HLA-E	HLA-F	HLA-H	ICAM1	ICAM3
IFI16	IFI30	IFI44L	IFI44	IFIH1	IFIT2
IFIT3	IKZF1	IL10RA	IL18BP	IL2RB	IL2RG
IL32	INPP5D	IRF1	IRF8	ITGB2	JAK3
LAP3	LAPTM5	LAT2	LCP2	LGALS9	LIMD2
LOC96610	LPXN	LYN	MPEG1	MX2	MYO1F
NCKAP1L	NFKBIE	NLRC5	NMI	NR1H3	OAS2
OPTN	PARP14	PARVG	PIK3CD	PIM2	PLCB2
PLCG2	PLEKH01	PLEK	PLSCR1	PML	PSMB10
PSMB8	PSMB9	RAC2	RASSF4	RCSD1	RELB
RHOG	RNASE6	RNF166	RSAD2	RUNX3	SAMD9L
SAMD9	SASH3	SELL	SELPLG	SEMA4D	SERPINB9
SERPING1	SH3KBP1	SIGLEC1	SLAMF7	SLA	SLC15A3
SP100	SPI1	SPOCK2	SRGN	STAT1	STAT2
STK10	TAP1	TAP2	TCN2	TMEM140	TNFAIP3
TNFRSF1B	TRAF1	TRIM21	TRIM22	TRPV2	UBD
XAF1	ADAM8	ADAP2	ADCY7	APBB1IP	ARRB2
BIN1	BST2	C2	C3AR1	CCDC69	CCL2
CCR1	CD14	CD163	CD68	CD83	CECR1

CFLAR	CLEC2D	CMKLR1	COTL1	CPVL	CTSC
CXCL16	CYBB	DDX58	DDX60L	DDX60	DGKA
DOCK2	EFHD2	EMILIN2	EMP3	EVI2A	FBX06
FCER1G	FCGR2A	FPR3	GLIPR2	GM2A	HAPLN3
HAVCR2	HLA-DQA1	HLA-DQA2	HMHA1	ICAM2	IFI27
IFI35	IFITM1	IL16	IL27RA	IRF5	IRF7
IRF9	ISG20	ITGAL	ITGAX	KCNAB2	KIAA1949
LAIR1	LCP1	LGMN	LILRB4	LYZ	MAN2B1
MAP3K14	MGAT1	MICAL1	MOBK12C	MPP1	MS4A6A
MX1	MYD88	NCF2	NECAP2	NFKB2	NLRP1
NPL	OAS1	OASL	PAG1	PARP12	PARP9
PIK3AP1	PIM1	PLEKH02	PPM1M	PSME2	PTK2B
PTPRC	RASSF5	RIN3	RNF19B	S1PR2	SAMHD1
SEL1L3	SEPT6	SH3TC1	SIRPA	SLAMF8	SLC7A7
SP110	SP140L	TAPBPL	TAPBP	TFEB	TGM2
TLR2	TNFAIP8	TNIP1	TRANK1	TRIM38	TTC7A
TYMP	TYROBP	UBA7	UBE2L6	UNC13D	VAMP5
VCAM1	WIPF1	ZNF438			

Genes co-expressed with A3B

SH3BP1	WARS	ANLN	ARHGAP11A	ASF1B	ASPM
ATP5C1	AURKA	BIRC5	BUB1B	BUB1	C16orf61
C20orf24	CBX2	CCNA2	CCNB1	CCNB2	CDC123
CDC20	CDC25B	CDC7	CDCA3	CDCA5	CDCA8
CDK1	CDT1	CENPE	CENPN	CEP55	CHAF1B
CHEK1	CHEK2	CKS1B	CKS2	COQ2	CTPS
DBF4	DDX39	DEK	DEPDC1	DLGAP5	DNAJB11
DNAJC9	DONSON	DSN1	E2F1	ECE2	EIF2C2
EIF2S2	EPR1	ESPL1	EXO1	EZH2	FAM136A
FAM83D	FANCA	FEN1	FOXM1	GIN51	GMNN
GPR172A	GTPBP4	H2AFZ	HDGF	HJURP	HMGA1
HMMR	HSPA14	IMPA2	IQGAP3	IRAK1	KIF11
KIF20A	KIF23	KIF2C	KIF4A	KIFC1	KPNA2
LMNB1	LMNB2	MAD2L1	MASTL	MCM2	MCM5
MCM6	MELK	MKI67	MRPL15	MTHFD2	MYBL2
NCAPG	NCAPH	NDC80	NEK2	NIP7	NUDT5
NUF2	NUP93	NUSAP1	PCNA	PHF19	PHF5A

PKMYT1	PLK1	PNP	POC1A	POLR2D	POP1
PPIL5	PPPDE2	PRC1	PSMB2	PSMD14	PTTG1
RANBP1	RBM17	RCC1	RECQL4	RFC4	RIPK2
RPN1	RRM2	SLC25A19	SLC25A5	SLC7A5	SNRPA1
SPAG5	STIL	STMN1	SUV39H2	TACC3	TCF19
THOC4	TK1	TPX2	TRIP13	TROAP	TTK
TUBA1B	TUBA1C	TUBB	TYMS	UBE2C	UBE2S
UBE2T	VRK1	WDR67	YARS	ZDHHC18	AARS
ADM2	ADSL	ATAD2	C15orf23	C16orf75	C4orf46
C9orf140	CCDC99	CCNE2	CCT5	CDC6	CENPF
CENPL	CIT	DKC1	DRG1	DTL	ECT2
FAM111B	FANCD2	FANCI	FBX05	GPSM2	HMGB2
ILF2	INCENP	JOSD1	KIAA0101	MCM3	MCM4
MCM7	MLF1IP	MYBL1	NCAPD2	NCAPG2	NFKBIL2
PBK	PDXP	PRIM2	RACGAP1	RANGAP1	RAN
SEPHS1	SEPT3	SGOL2	SMC4	TIMELESS	TMPO
TOMM22	TOP2A	UHRF1	WDR76	XPO5	XRCC6
ZWINT					

Genes co-expressed with a minimum of five A3 genes, partial to hypoxia

APOL1	APOL3	APOL6	ARHGAP25	B2M	BIRC3
BTN3A1	BTN3A3	C1orf38	CASP1	CCL5	CCR1
CD48	CD74	CIITA	CORO1A	CTSS	CXCL10
CYBB	CYTH4	EPSTI1	FMNL1	FYB	GBP1
GBP4	GBP5	GIMAP4	GIMAP5	HCP5	HLA-A
HLA-B	HLA-C	HLA-DMA	HLA-DMB	HLA-DRA	HLA-E
HLA-F	HLA-H	IFI16	IFI44L	IFIH1	IFIT3
IL18BP	IL2RB	IL2RG	IRF1	JAK3	LAP3
LCP2	LGALS9	MPEG1	MX2	NLRC5	NMI
OAS2	OASL	PARP14	PARVG	PML	PSMB9
PSME2	PTPRC	SAMD9L	SAMHD1	SASH3	SELL
SEMA4D	SERPINB9	SH3BP1	SIGLEC1	SLAMF7	SLC15A3
SP100	STAT1	STAT2	TAP1	TAP2	TNFRSF1B
TRAF1	TRANK1	TRIM21	TRIM22	UBA7	UBE2L6
WARS	XAF1	AIF1	APBB1IP	APOBEC3G	APOL2
ARHGAP30	BST2	C10orf54	C19orf66	C1QA	C1QB
C1QC	C2	C3AR1	CCL2	CD163	CD37

CD4	CD52	CD53	CD68	CD83	CD86
CFLAR	CMKLR1	CMPK2	CSF2RB	CXCL9	DDX58
DDX60L	DDX60	DHX58	DOCK2	DTX3L	EVI2A
EVI2B	FCER1G	FERMT3	FGL2	FPR3	GBP2
GMFG	GPSM3	HAVCR2	HCK	HCLS1	HERC5
HERC6	HLA-DPA1	IFI27	IFI30	IFI35	IFI44
IFI6	IFIT2	IFIT5	IFITM1	IKZF1	IL10RA
INPP5D	IRF8	IRF9	ISG15	ISG20	ITGAL
ITGAX	ITGB2	LAIR1	LAPTM5	LAT2	LILRB4
LPXN	LYN	LYZ	MS4A6A	MX1	MYO1F
NCF2	NCKAP1L	OAS1	OAS3	OPTN	PARP12
PARP9	PIK3AP1	PIK3CD	PIK3IP1	PLCB2	PLEKH02
PLEK	PLSCR1	PSMB10	PSMB8	RAC2	RASSF4
RNASE6	RNF19B	RNF213	RSAD2	SAMD9	SERPING1
SLAMF8	SLA	SLC7A7	SPI1	SPOCK2	SRGN
TCN2	TMEM140	TNFAIP3	TNFRSF14	TRIM14	TRPV2
TYMP	TYROBP	USP18	WIPF1	ZNFX1	

Genes co-expressed with A3B, partial to hypoxia

<none>

We calculate the Spearman's correlation coefficient between every APOBEC3 gene and every gene reported by TCGA in BRCA. Genes with lower expression than the median expression of all genes across all samples in BRCA are excluded. Gene pairs with a correlation coefficient of 0.3 or greater and an FDR-adjusted p value < 0.05 are considered co-expressed. We list genes that are co-expressed with five or more APOBEC3 genes as well as all genes co-expressed with A3B. The same analysis is performed using Spearman's correlation partial to the mean expression of the Buffa hypoxia signature. There is no gene co-expressed with A3B using this method.

Table S3: Top Genes Co-Expressed with the Buffa Hypoxia Signature in Breast Cancer

TUBA1C	CDC20	ENO1	TPI1	GAPDH	KIF2C
PTTG1	PLK1	MYBL2	UBE2C	CCNB2	PSMB2
CDCA8	CENPN	CDCA5	ECE2	BIRC5	CDCA3
YBX1	C16orf61	RANBP1	TUBA1B	CTPS	FAM83D
TPX2	TRIP13	CEP55	FOXO1	MELK	SHMT2
GPI	SNRPG	TUBB	AURKA	CDT1	UQCRH
GARS	NDC80	HJURP	PSMG1	UBE2S	CCNA2
HN1	MRPL37	PGK1	SNRPA1	CKS2	PSMA7
CKS1B	RAN	PRDX4	STMN1	YARS	CCNB1
NUP93	CCT5	CBX2	CHEK1	UQCRHL	DLGAP5
TTK	NCAPG	NUDT5	NUTF2	H2AFX	FAM136A
NCAPH	KPNA2	KIFC1	CDC123	KCMF1	PDCD5
C20orf24	RCC1	RFC4	PNP	GTPBP4	MRPS5
EIF2S2	KIF20A	TROAP	COX4NB	PGAM1	RFC2
DNAJB11	FEN1	SLC7A5	CSTB	KIF4A	EPR1
PSMD7	DKC1	MRPL15	MCM7	UBE2T	BUB1
EXO1	ANLN	KARS	PPM1G	PDIA6	DDX39
DEPDC1	RRM2	MCM2	THOC4	SUV39H2	SNRPD1
UTP11L	PPP1R14B	MAD2L1	ENY2	E2F1	ATP5C1
TACC3	TKT	EZH2	CCT6A	DONSON	GMPS
CDC25B	GLRX3	PFKP	GLT25D1	CDK1	FANCA
MCM6	SLC25A5	MEMO1	PDHA1	KIF23	PKMYT1
TK1	MRPL47	ALG3	BOLA3	C15orf23	MTHFD1L
GIN51	TBC1D7	HMGA1	RHEB	CCT4	UCK2
MAGOH	TYMS	PLOD1	PSMD14	RPN1	BYSL
SLC25A19	TMEM189	GPR172A	NIP7	MRPS17	C6orf153
C16orf57	CDCA4	CHEK2	H2AFX	NUF2	VRK1
EIF3B	PHF5A	RBM17	NDRG1	PRC1	MTHFD2
TOMM5	SEC61G	C20orf20	SF3B14	NCS1	DBF4
C9orf30	CIAPIN1	TBRG4	SNRPC	LYAR	LMNB2
POP1	C16orf80	ASF1B	PTDSS1	CYCS	LSM2
BUB1B	TOMM40	CCT7	ATP1B3	RIPK2	ACOT7
GPN1	C8orf76	EBP	NARF	HSPD1	NOP2

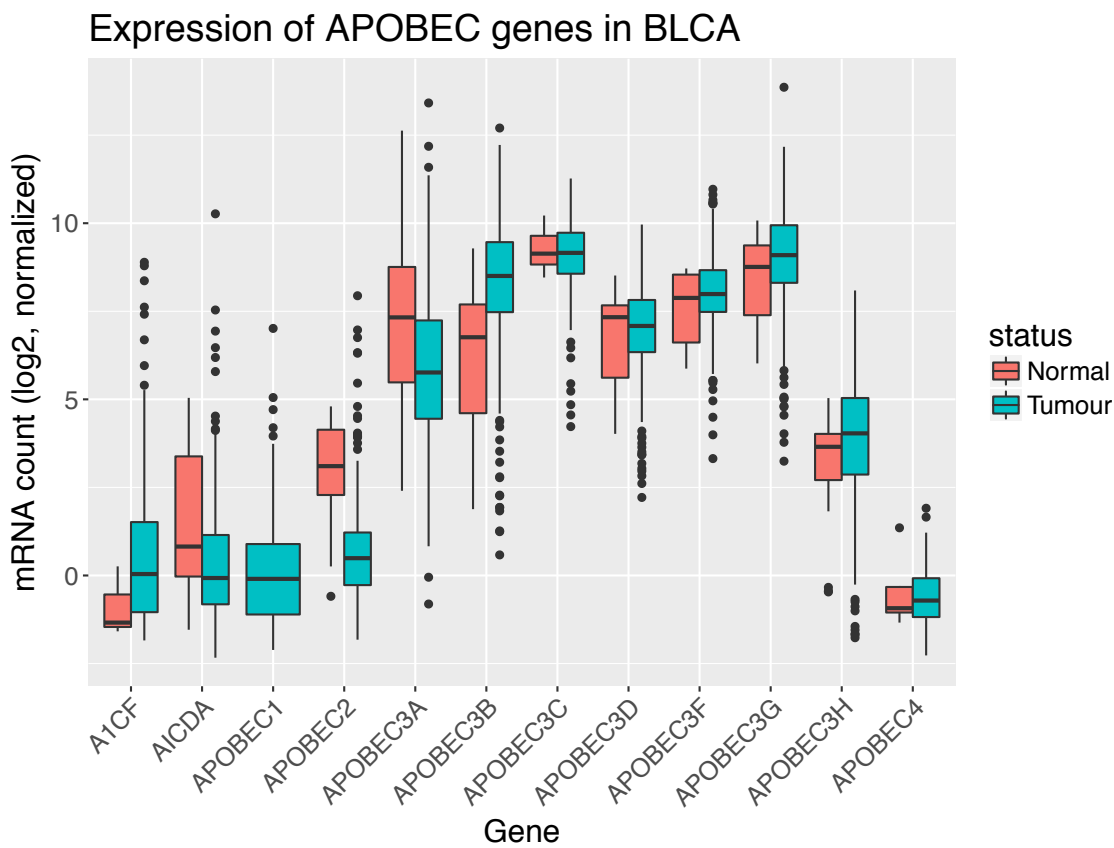
C9orf140	NDUFA9	NASP	STIL	CDK16	SHFM1
HDGF	POLR2D	CDK4	CDK2AP1	RBM38	MKI67
LMNB1	IMPA2	ADSL	SRD5A1	STIP1	NOL10
TMSB10	PAK1IP1	NT5DC2	TEAD4	DNAJC9	NCAPD2
RNASEH2A	NOP58	CIRH1A	PCNA	LDHA	PVR
NOP16	C3orf26	BRIX1	DCAF13	C6orf115	PSMD2
PPIA	TCP1	MED8	GNL2	ZWINT	NUSAP1
ATP5G3	WDR77	CDC6	NFKBIL2	SRM	RNASEH1
EIF5AL1	ZNF259	C21orf45	PSMA4	MCM3	NUDT1
ASNS	COX5A	HMGB3	FAM58A	MCM5	CALR
CHAF1B	EIF4EBP1	NDUFAF4	FAF1	ILF2	GGH
ELOVL1	PSMA5	HMMR	AIMP2	ARPC5L	MRT04
C6orf129	EIF5A	POC1A	PHGDH	SPAG5	TCF19
ANP32B	RPF2	CEBPB	PRDX1	TCEB1	AHSA1
RECQL4	PSMC2	MDH1	TUBB3	MTERFD1	EBNA1BP2
NOP56	MRPL13	RBM28	PFN1	NDUFB9	KIF11
S100A10	EIF2C2	IFRD1	PN01	TALD01	PSMA2
CCDC58	IRAK1	PKM2	PLAUR	HSPE1	SERBP1
MRPS22	CFL1	GMNN	CHCHD3	RPIA	SF3A3
HDAC2	PFDN2	MARS	SSRP1	PPPDE2	AK2
MANF	SEC13	BTG3	TAF11	PTMA	MCM4
SEPHS1	ATP5J2	TXN	POLR2F	NDUFS5	SUV39H1
TOMM22	CDC7	ERI3	PSMB5	MRPL11	AARS
UHRF1	C8orf59	HNRNPAB	RCC2	ADRM1	TARS
GPSM2	TMEM206	GPT2	SSBP1	MRPL12	BRI3
SLC2A1	PDXP	EIF2B3	CPSF3	MRPL36	FAM127B
PSMB4	MRPL2	ESPL1	SNRPA	POP7	DPH2
FAM96B	XP05	NAE1	RACGAP1	SLC16A3	MRPL51
FDPS	PPIL5	AMD1	FBXO5	TAGLN2	UBE2V2
AP2M1	RPUSD3	NAA10	ATAD3A	DRG1	CENPE
RUVBL1	CCT3	RAP2B	CCDC99	MAD2L2	YWHAQ
USP39	KIAA1609	FAM98A	LAPTM4B	MRPS15	NEK2
IQGAP3	TCOF1	ARHGAP11A	PPIL1	IDH2	DDOST
NCL	JMJD6	YKT6	CYC1	MDH2	S100A9
RRP1	VAC14	PDAP1	KIAA0101	PES1	SRPK1
POLR1C	BOP1	BID	PUF60	RPP25	TMEM201
HPRT1	ADPRHL2	SF3B3	CCT8	PWP2	TIMM50
STOML2	DHCR7	PSMC1	SRPRB	UNKNOWN_5	C16orf75

PBK	NUDCD1	POLD2	MAGOHB	LSM6	HSPA14
TMEM65	PSMG3	ASPM	ACTB	HMBS	C3orf1
DCUN1D5	UFD1L	SNRPF	PSMC3	PREB	CCDC90A
C11orf84	RANGAP1	MRPS12	YEATS2	ODC1	S100A11
BAK1	RPP38	MVD	PABPC4	GALNS	CENPF
FTSJ1	SNRPD3	PSMB7	PRIM2	C7orf68	CCDC86
PRELID1	C11orf75	NSDHL	SLC25A13	C3orf37	TBPL1
PPIH	CHCHD2	CEBPG	YRDC	SLC25A33	LOC152217
PDXK	SMARCB1	SH3BGRL3	PLOD3	COX7B	FANCI
DTYMK	DGUOK	PDCL3	LSM4	MRPL14	ACTL6A
CACYBP	DAZAP1	KCTD5	PLP2	PDE7A	SLC5A6
TPD52L2	B4GALT2	SFPQ	MRPS24	PSMA3	FARSB
MTCH2	U2AF1				

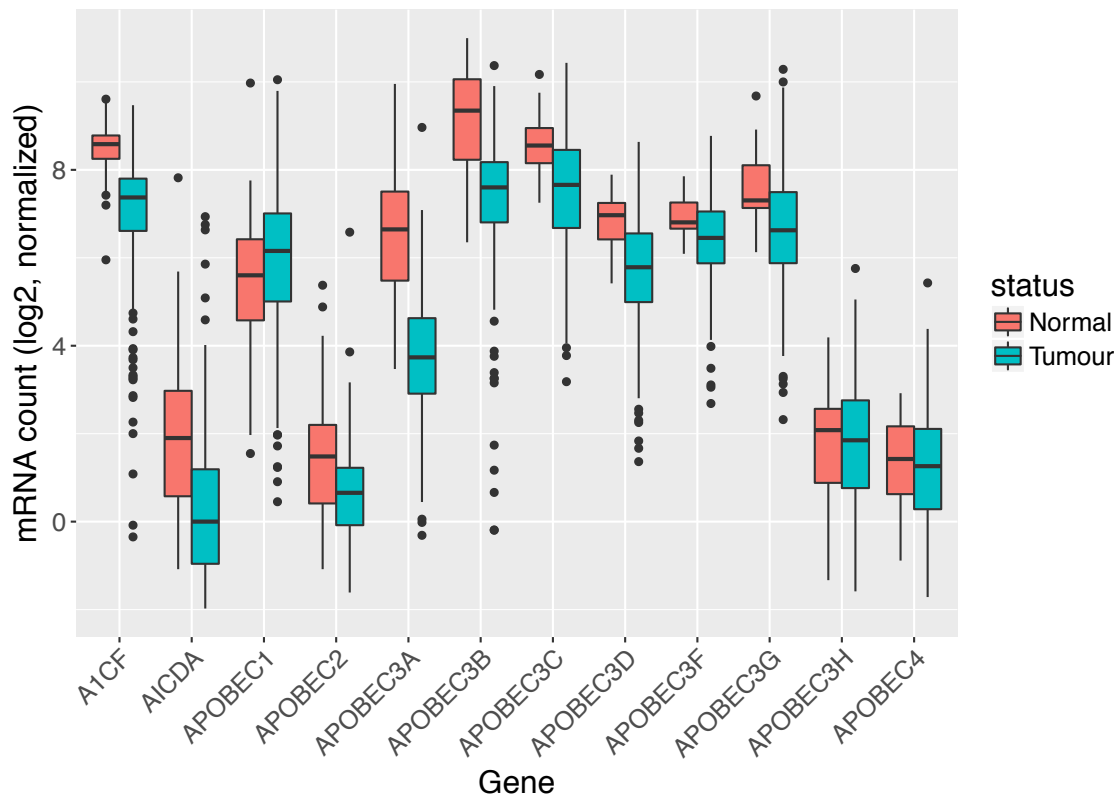
We calculate the Spearman's correlation coefficient between every gene in the Buffa hypoxia signature and every gene reported by TCGA in BRCA. Genes with lower expression than the median expression of all genes across all samples in BRCA are excluded. Only gene pairs with an FDR-adjusted p value < 0.05 are considered. For each gene in TCGA, we take the sum of its correlation coefficients with all genes in the Buffa hypoxia signature. We rank the genes by their summed coefficients and display the top 500 genes here. Genes are in descending order of summed coefficients, by row, then by columns.

Supplementary Figures

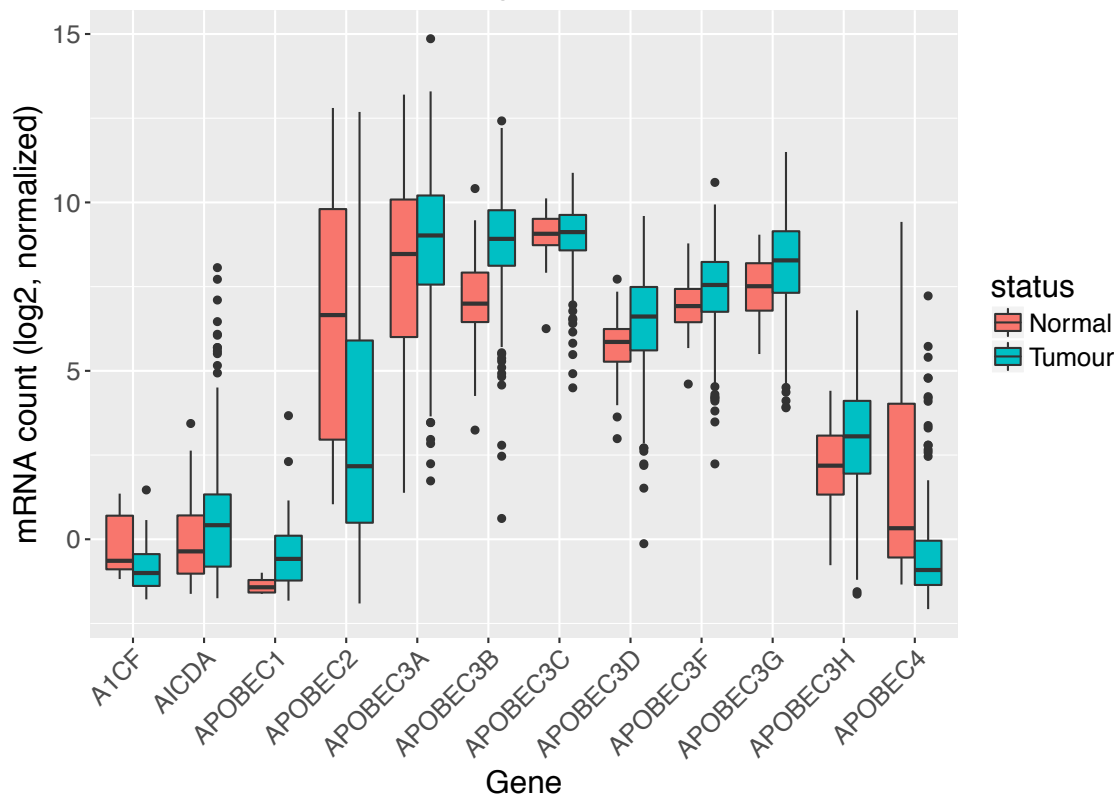
Figure S1: Expression of APOBEC Genes in 12 Cancer Types

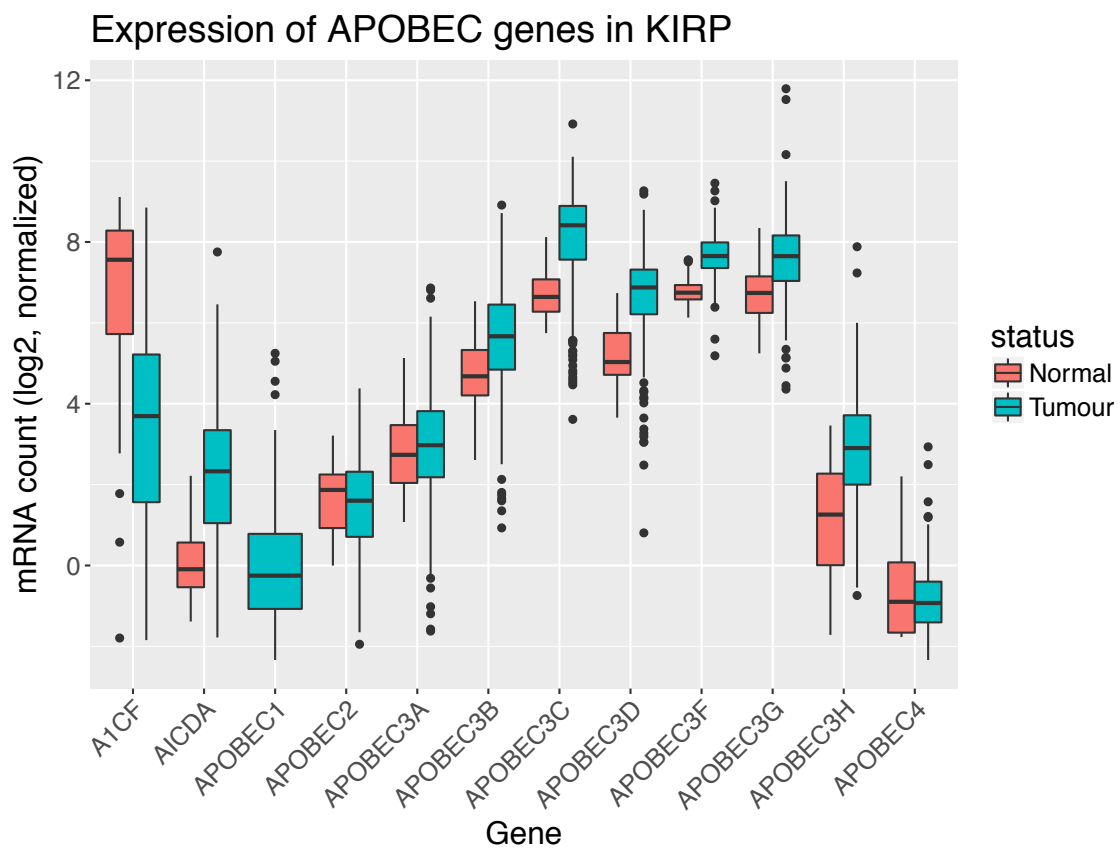
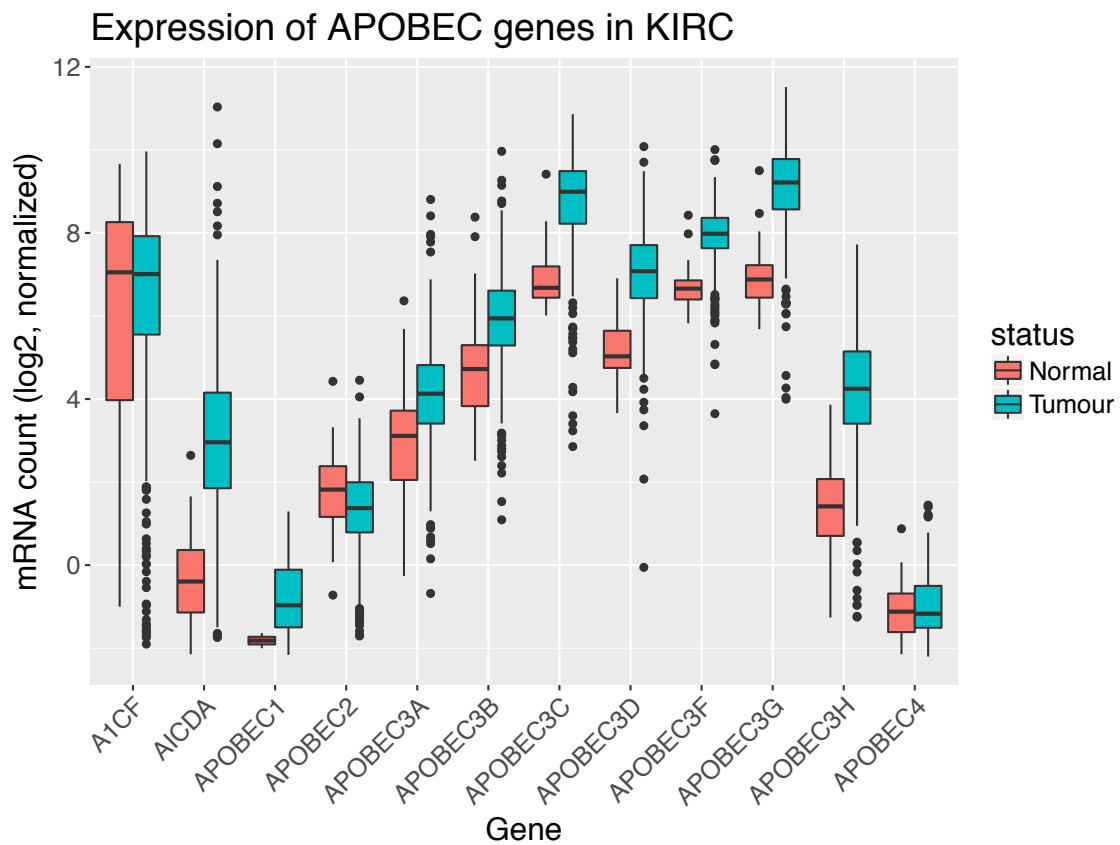


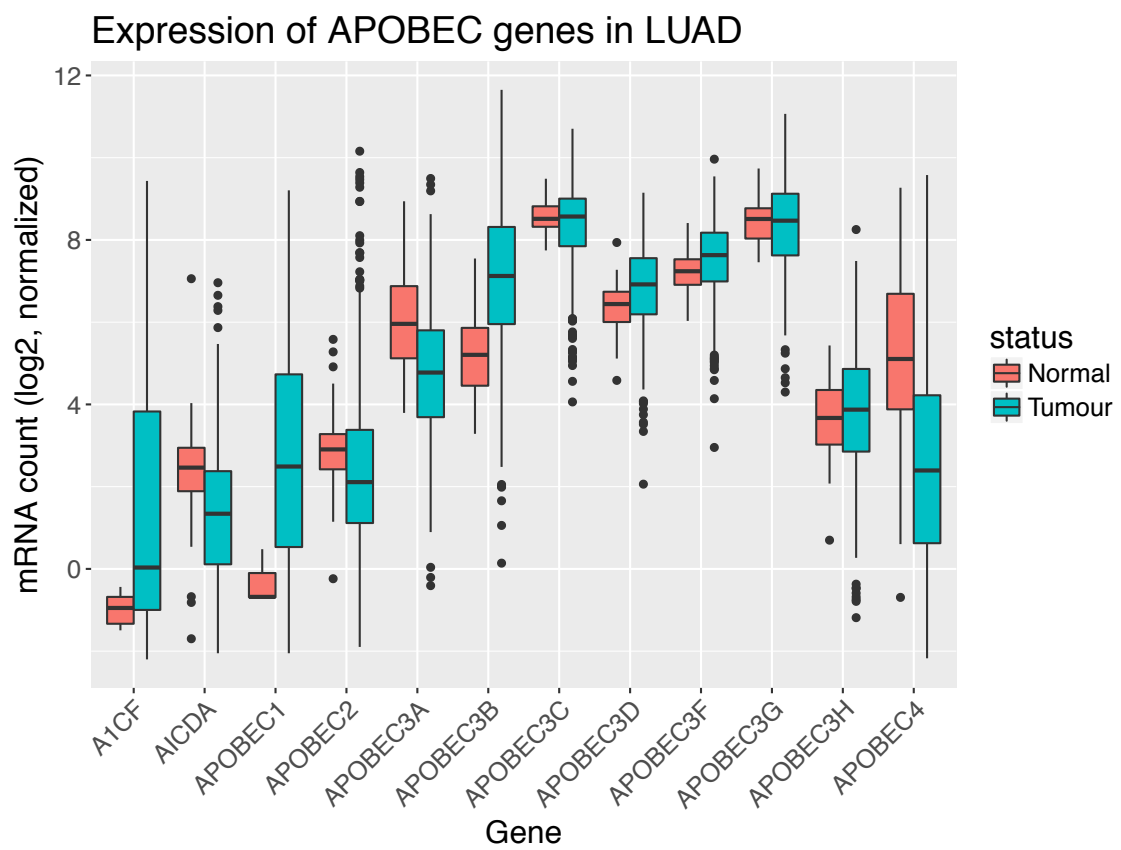
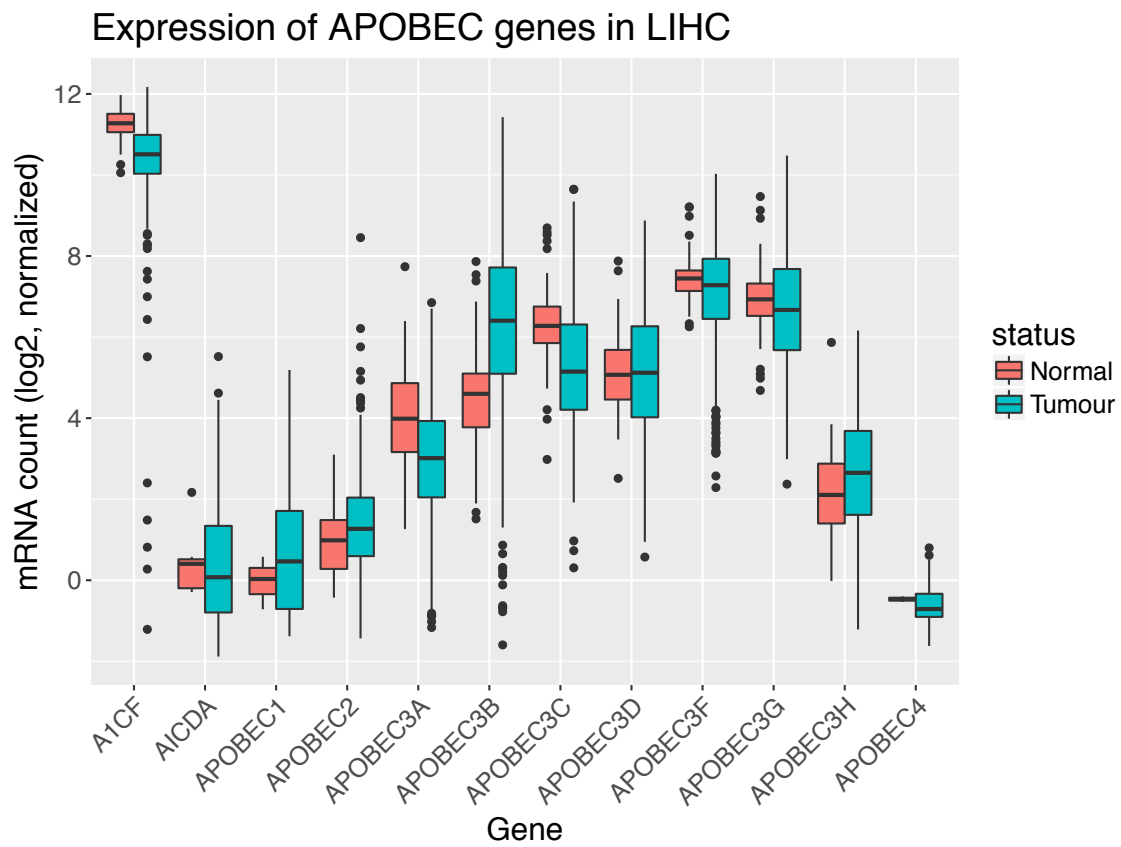
Expression of APOBEC genes in COAD



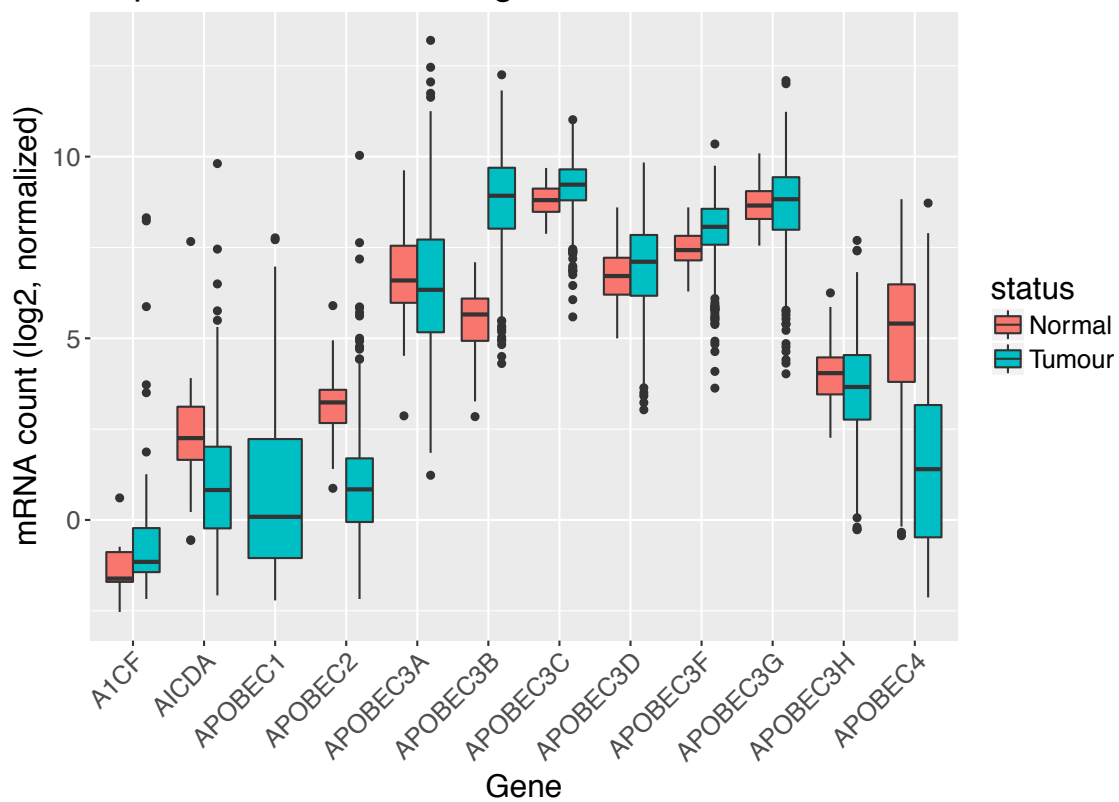
Expression of APOBEC genes in HNSC



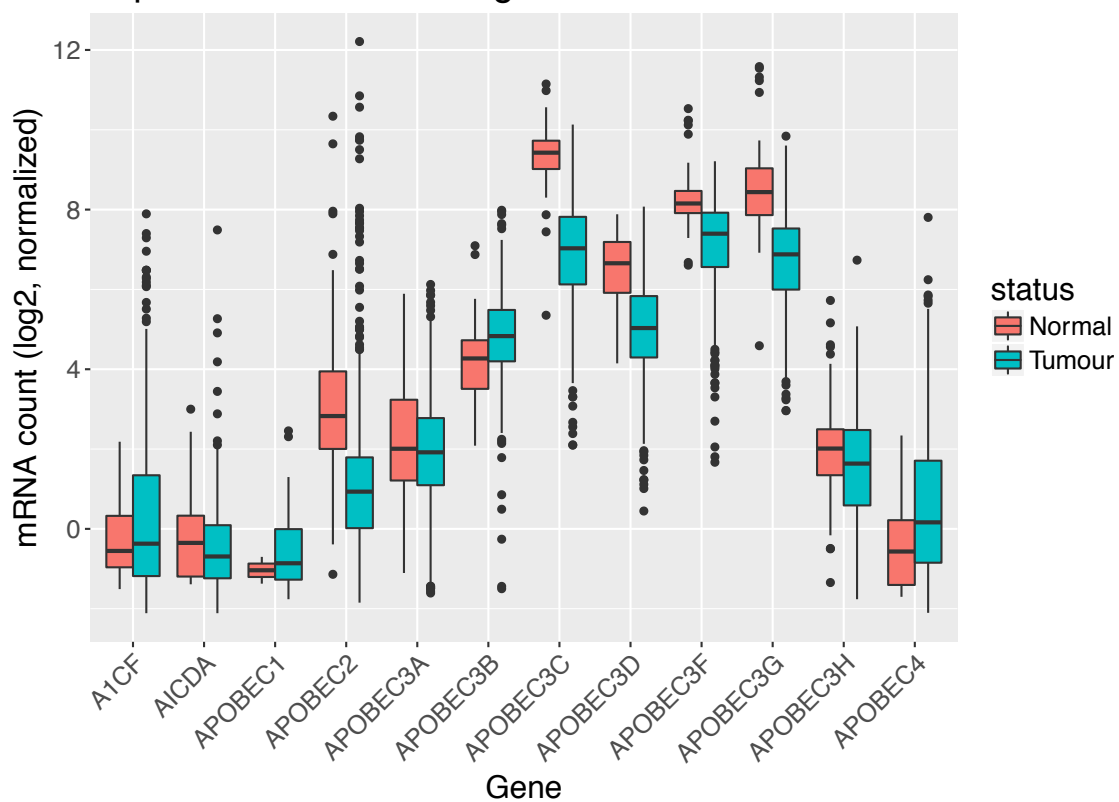


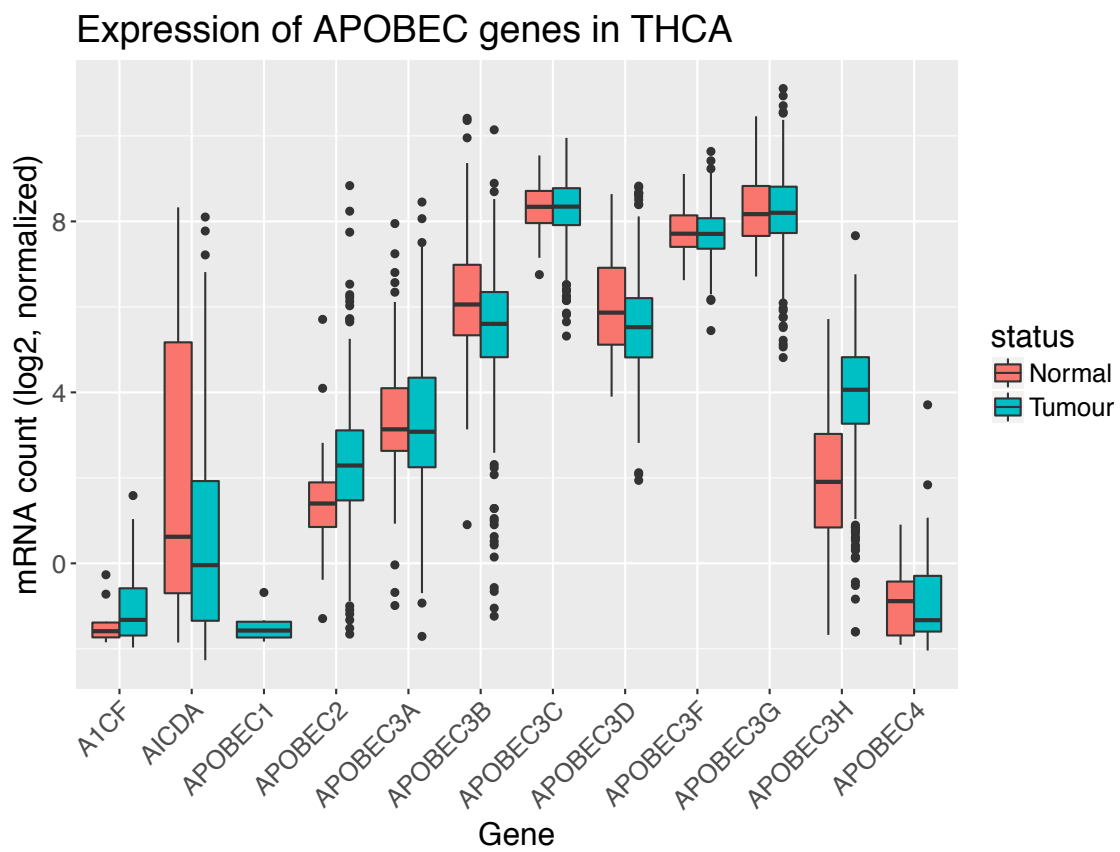
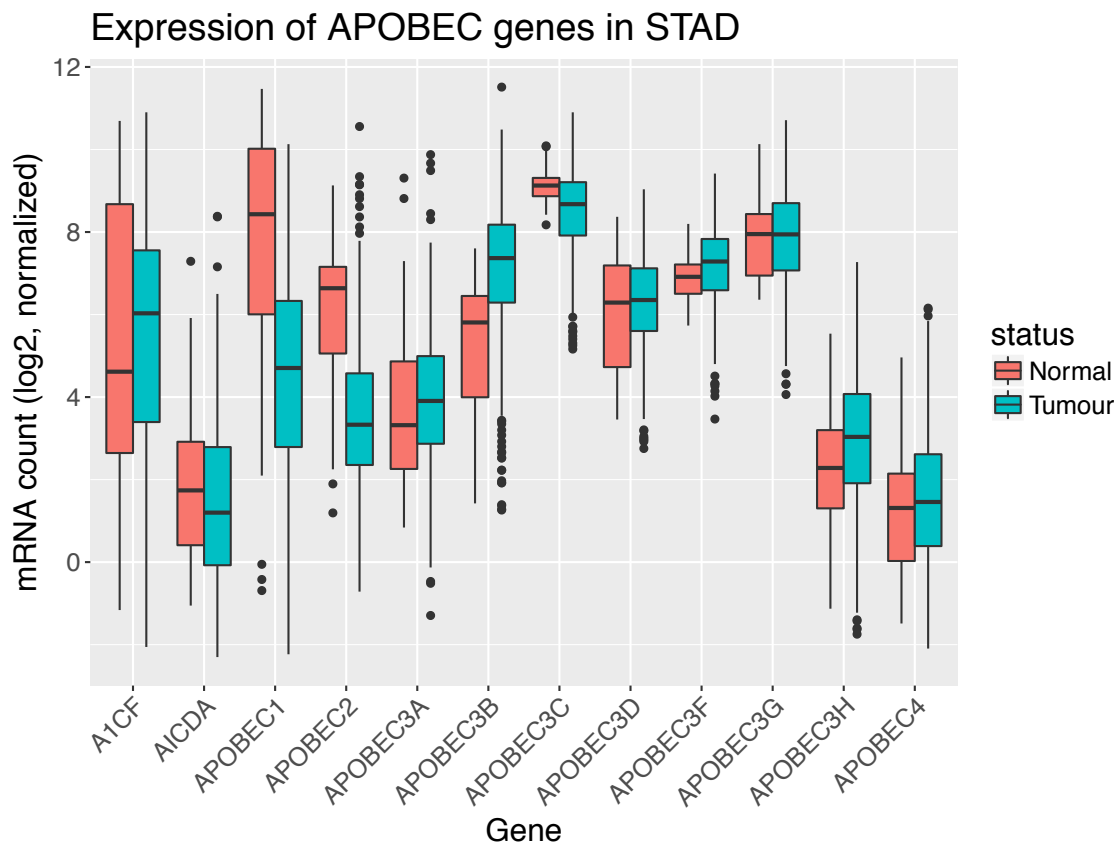


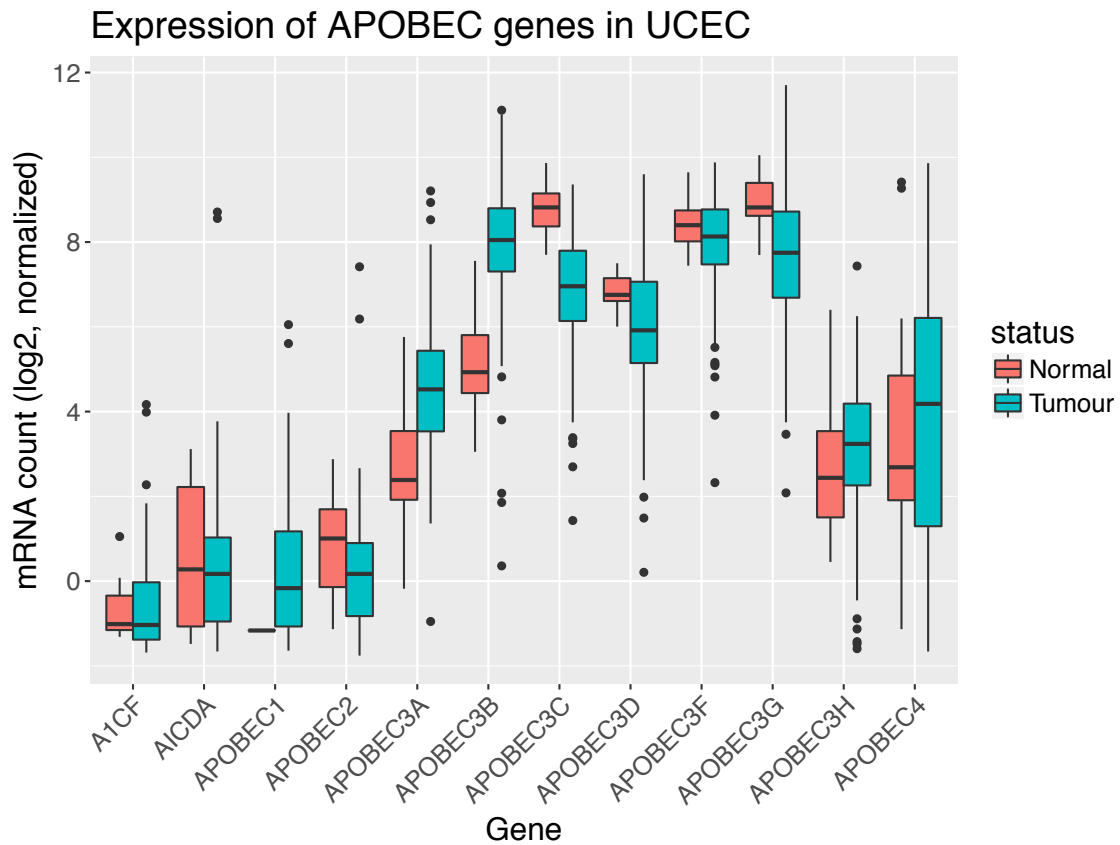
Expression of APOBEC genes in LUSC



Expression of APOBEC genes in PRAD

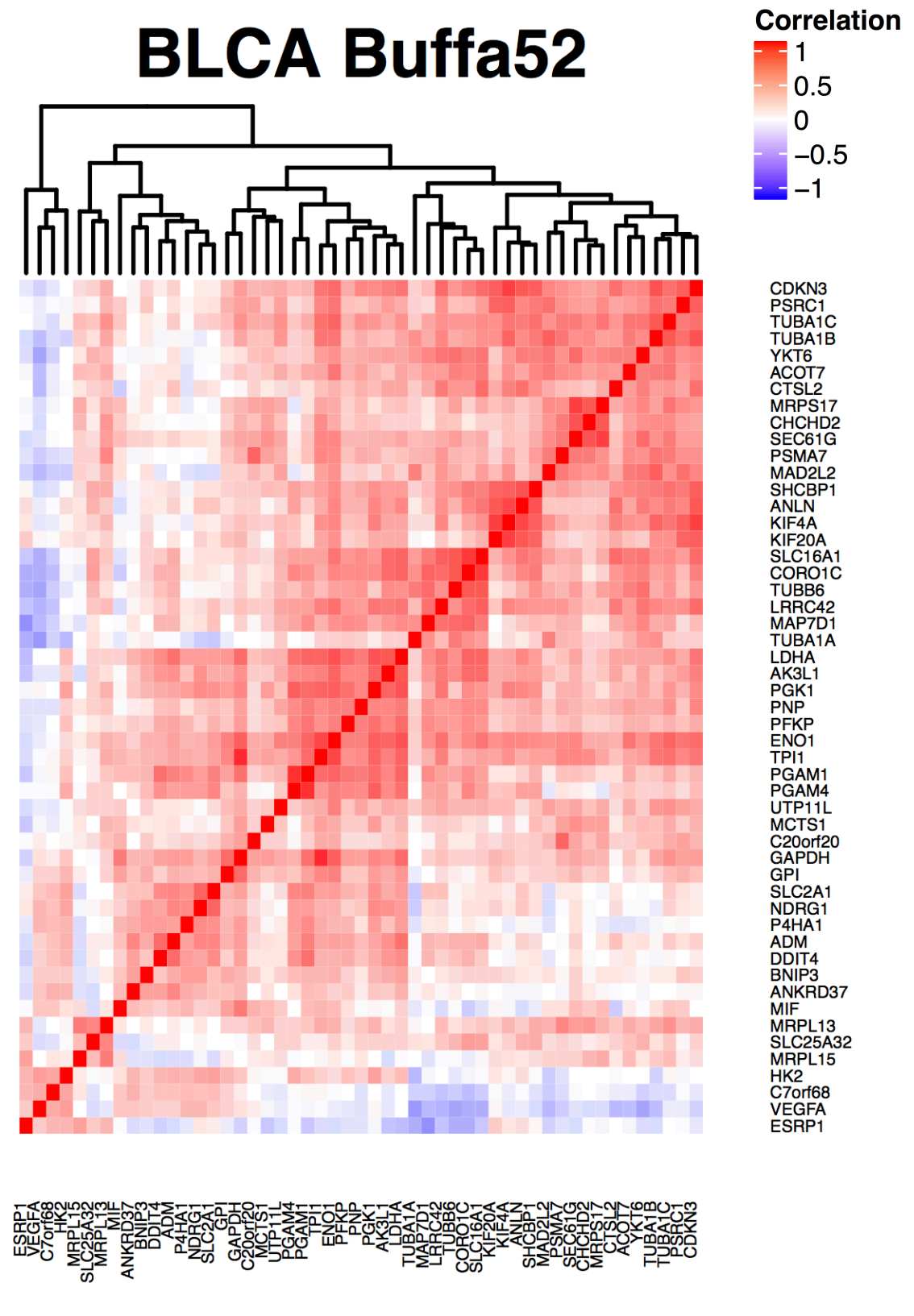




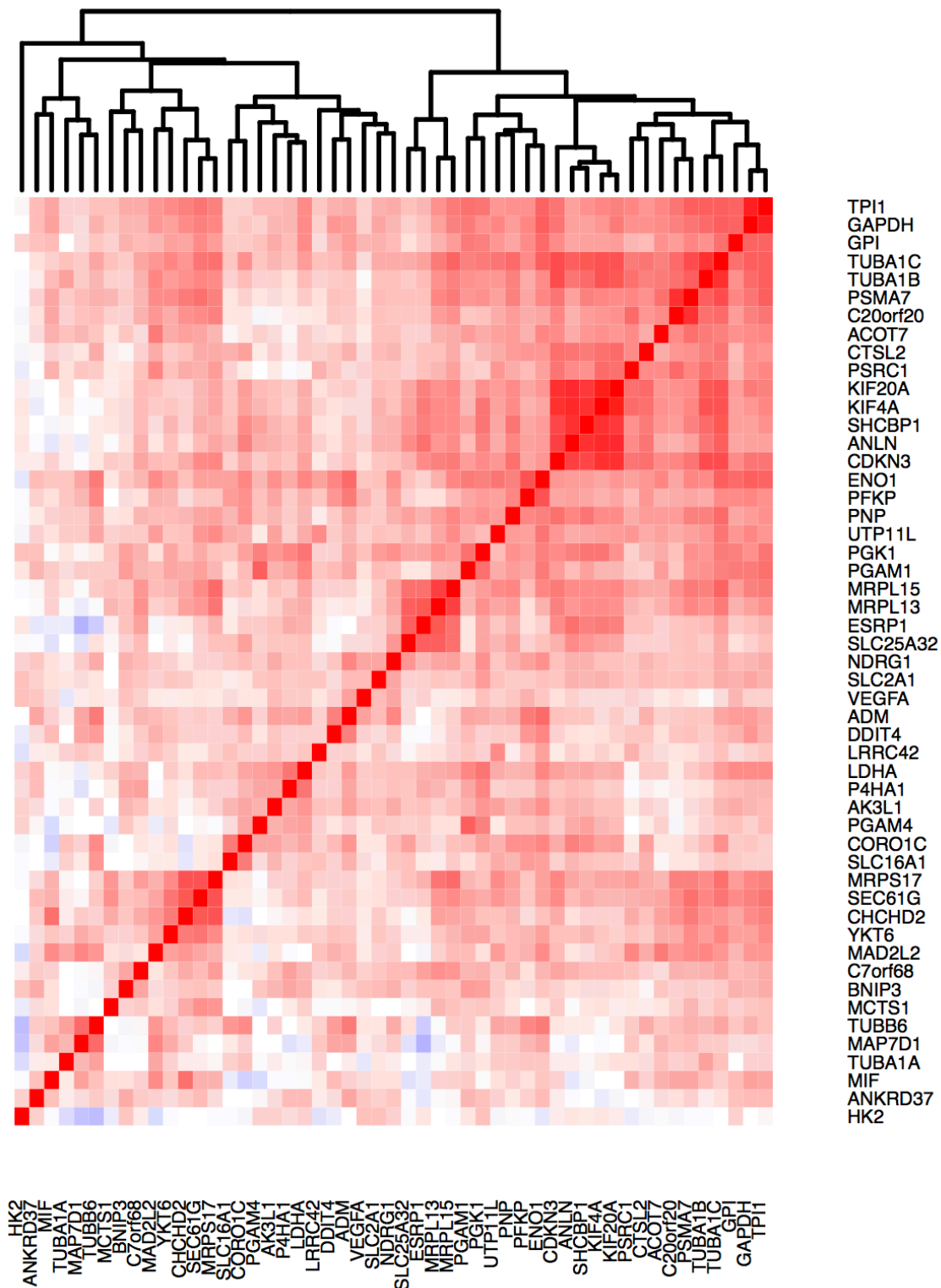


Expression of individual APOBEC genes in 13 cancer types. Data is log 2 transformed, normalised RNA-seq by expectation maximization (RSEM) values from TCGA. Values for each sample are displayed. Boxes represent the range from the first to the third quartiles. Centre line represents the median. Whiskers represent 3rd quartile plus 1.5x the interquartile range or the maximal value, whichever is smaller, and the 1st quartile minus 1.5x the interquartile range or the minimal value, whichever is larger. Dots represent values not contained within whiskers.

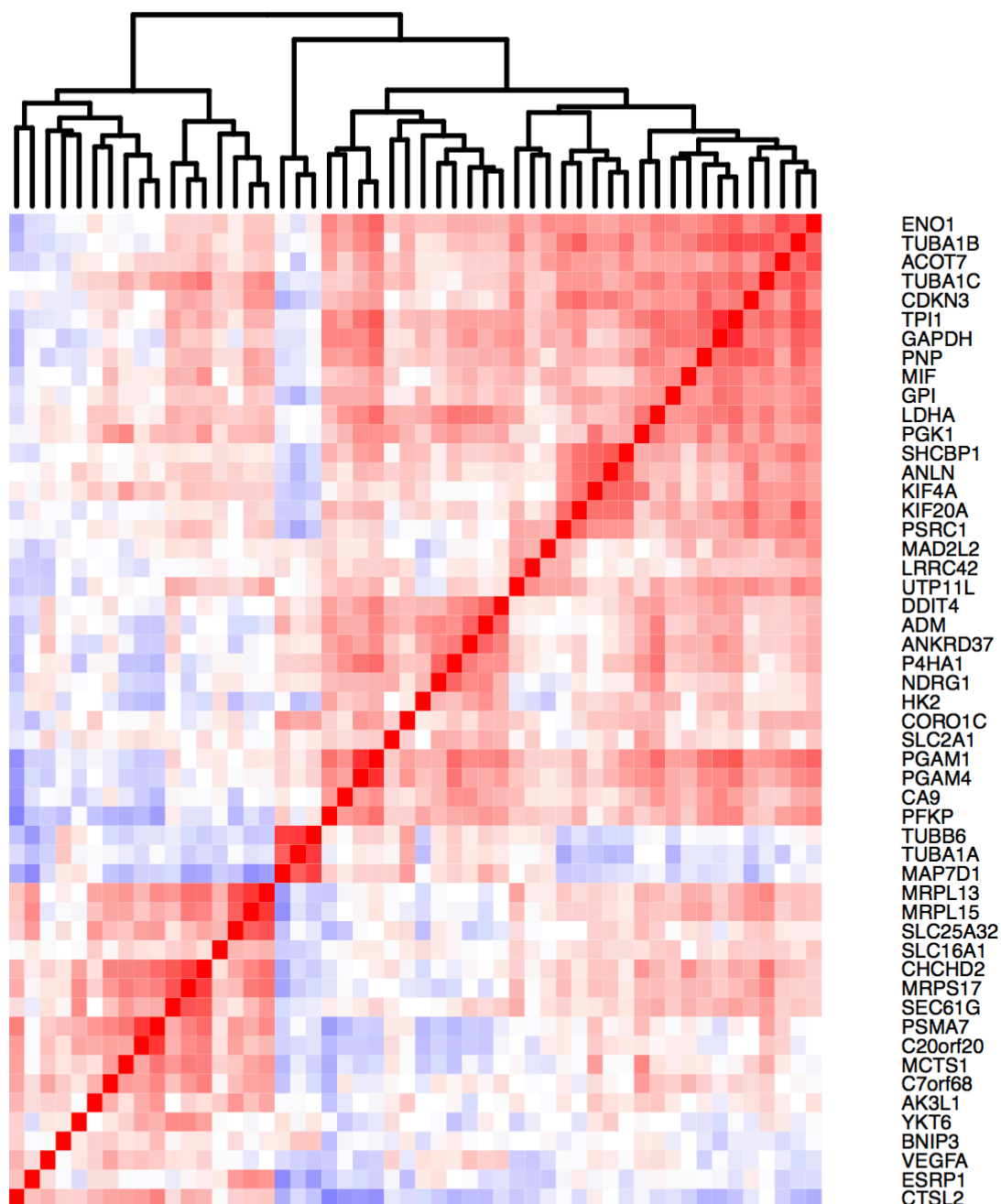
Figure S2: Self-Correlation of the Expression of the Buffa Hypoxia Signature in 13 Cancer Types



BRCA Buffa52

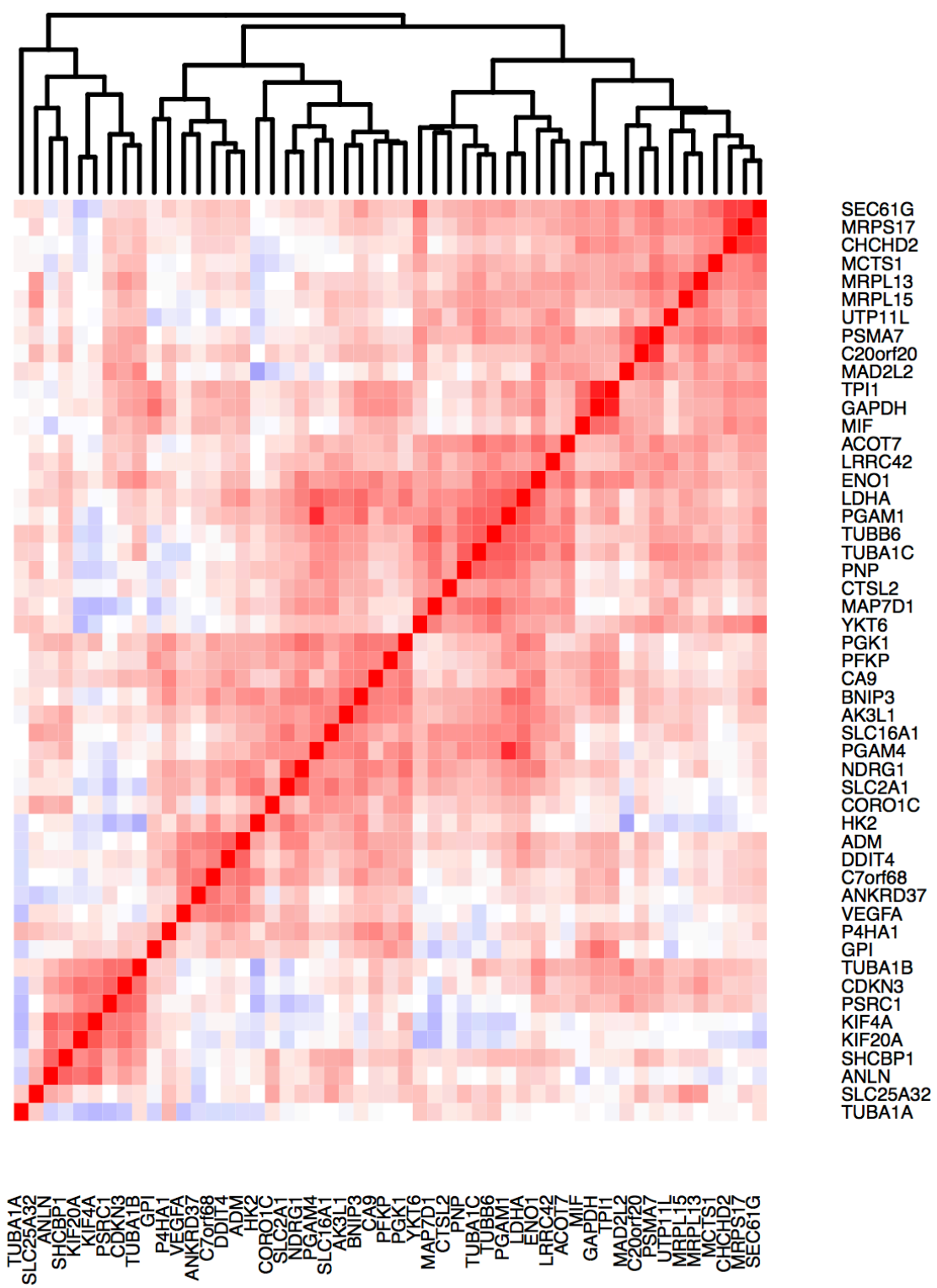


COAD Buffa52

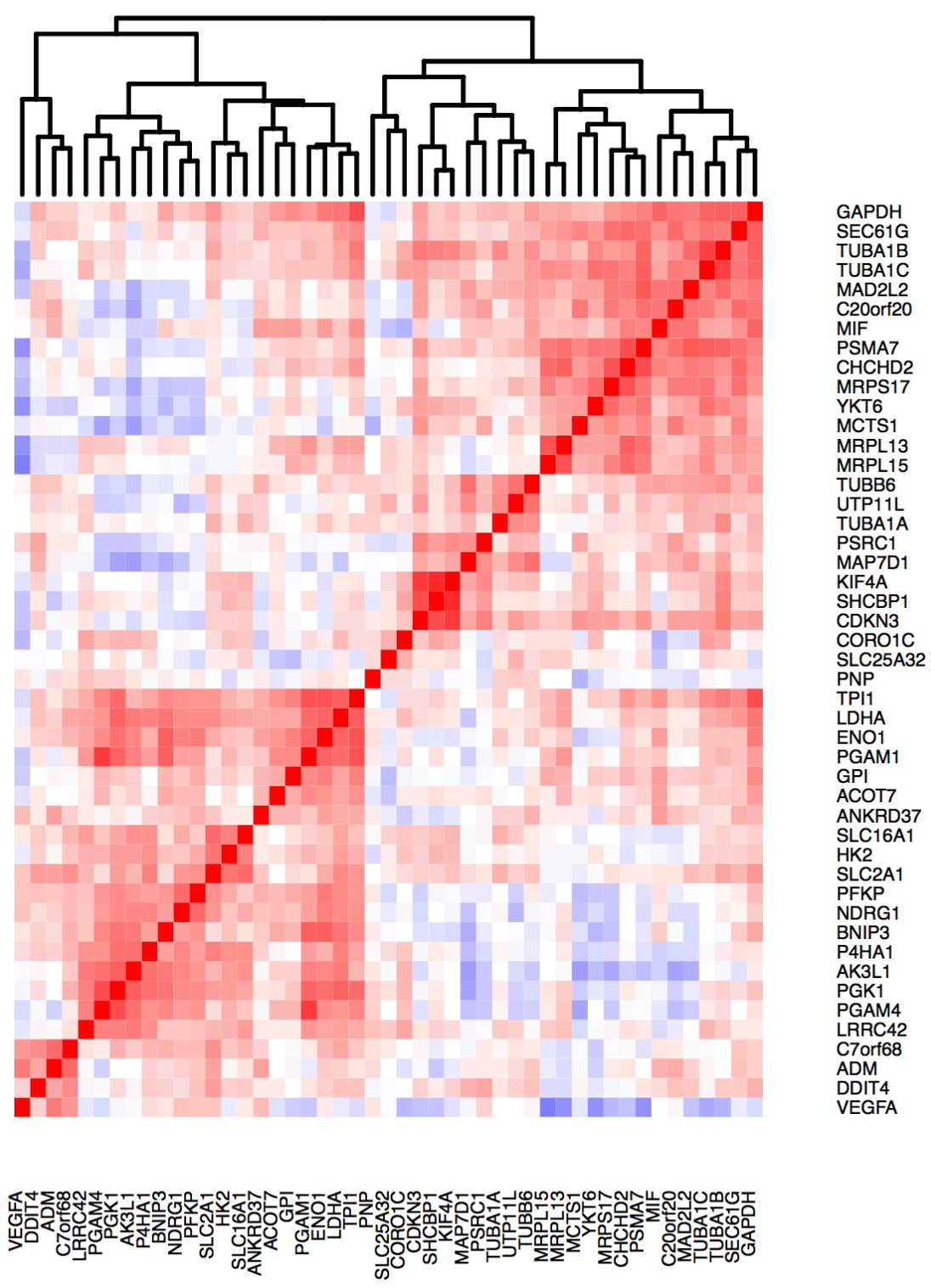


CTSL2
ESRP1
VEGFA
BNIP3
YKT6
AK3L1
C7orf68
MCTS1
C20orf20
PSMA7
SEC61G
MRPS17
CHCHD2
SLC16A1
SLC25A32
MRPL13
MRPL15
MAP7D1
TUBA1A
TUBB6
PFKP
CA9
PGAM4
PGAM1
SLC2A1
CORO1C
HK2
NDRG1
P4HA1
ANKRD37
ADM
DDIT4
UTP11L
LRRC42
MAD2L2
PSRC1
KIF20A
KIF4A
ANLN
SHCBP1
PGK1
LDHA
GPI
MIF
PNP
GAPDH
TPI1
CDKN3
TUBA1C
ACOT7
TUBA1B
ENO1

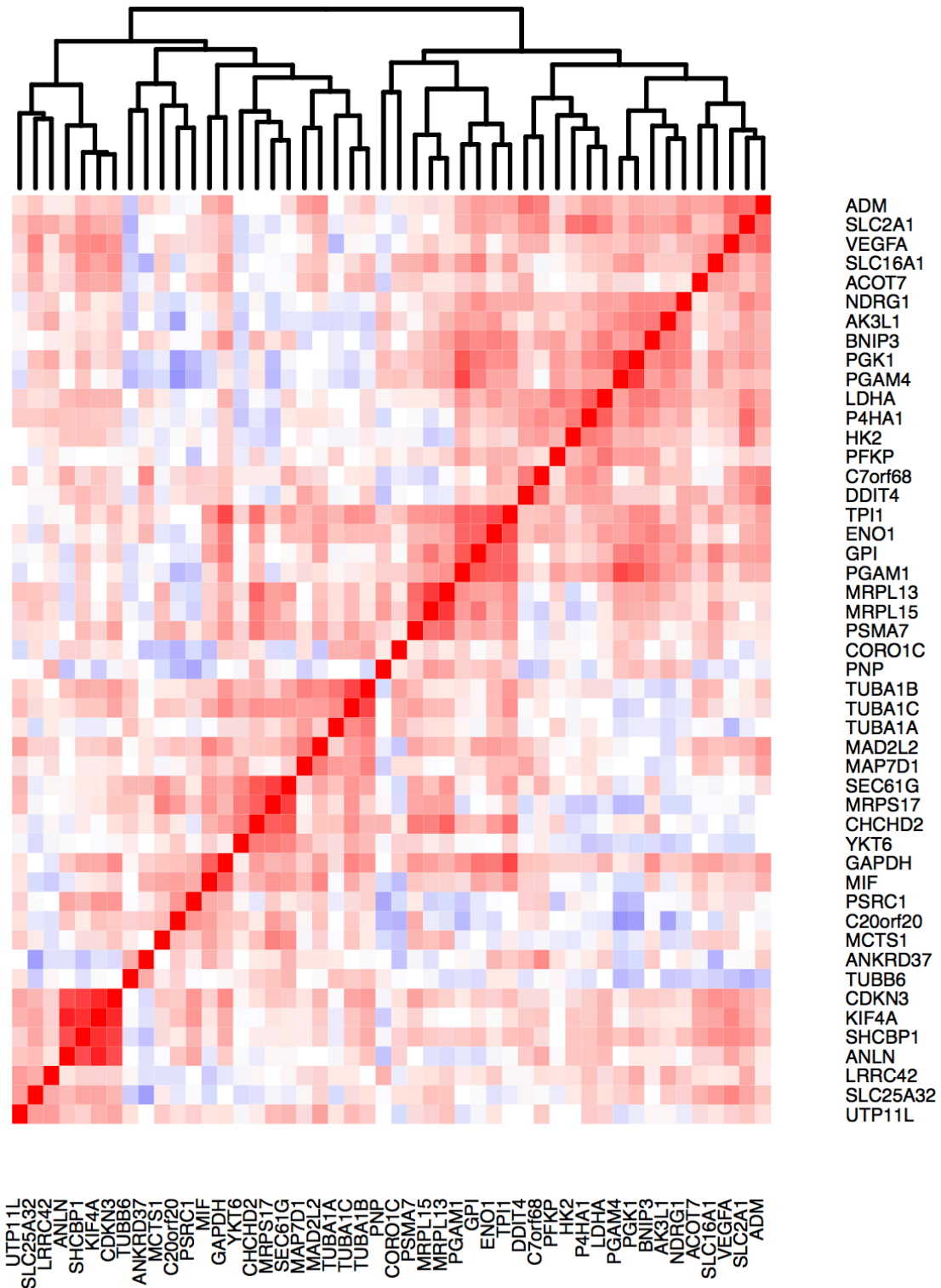
HNSC Buffa52



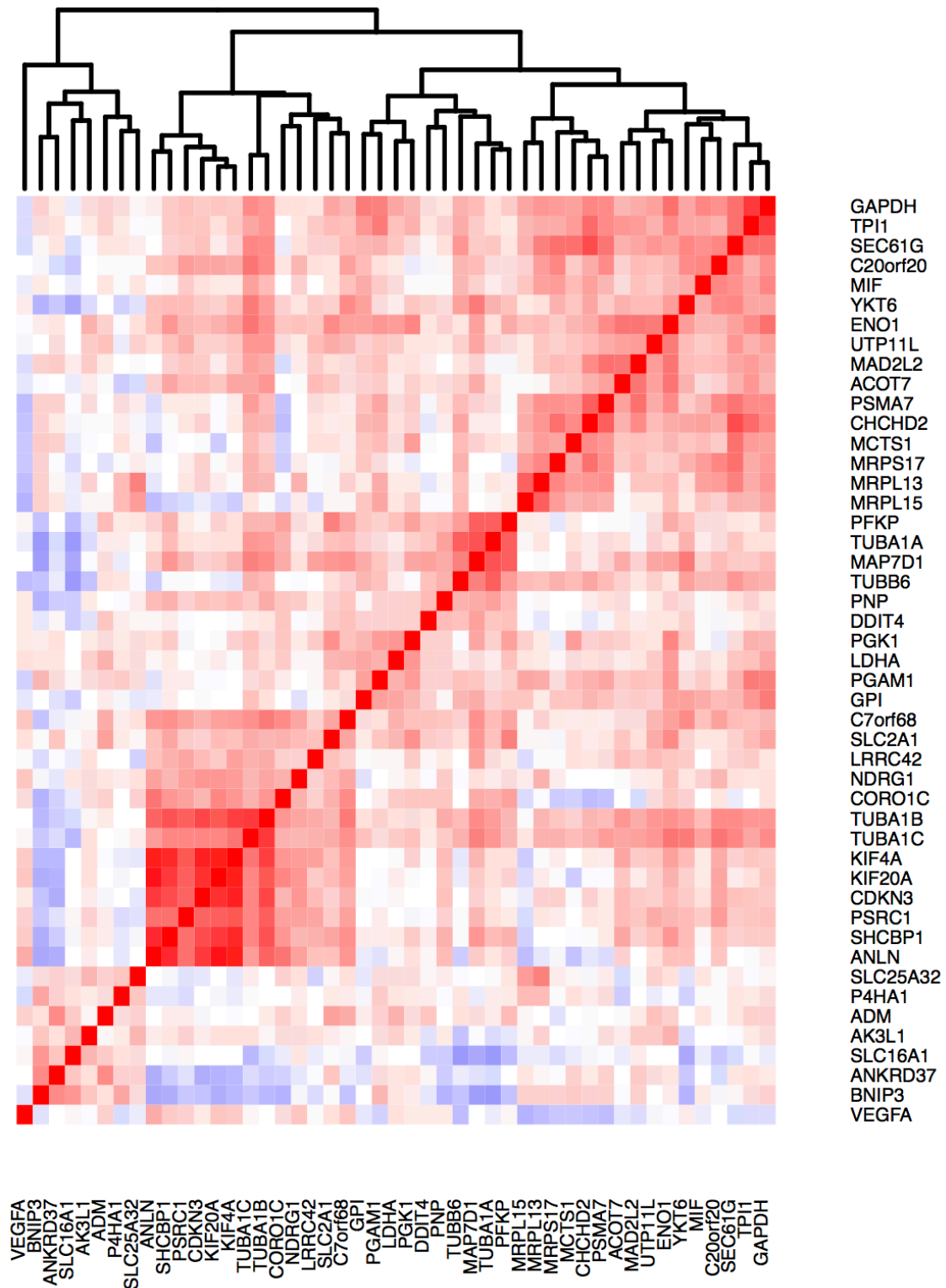
KIRC Buffa52



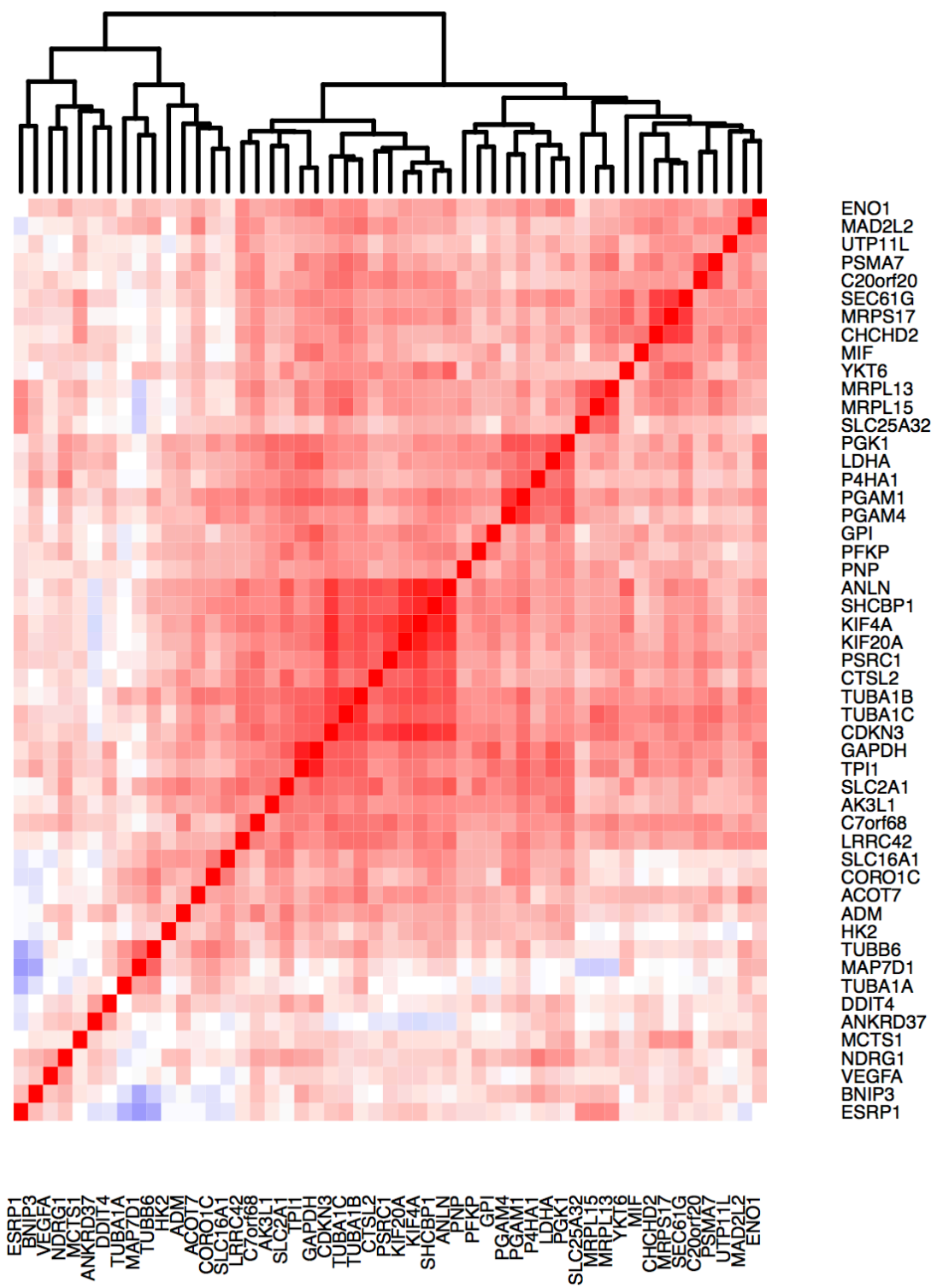
KIRP Buffa52



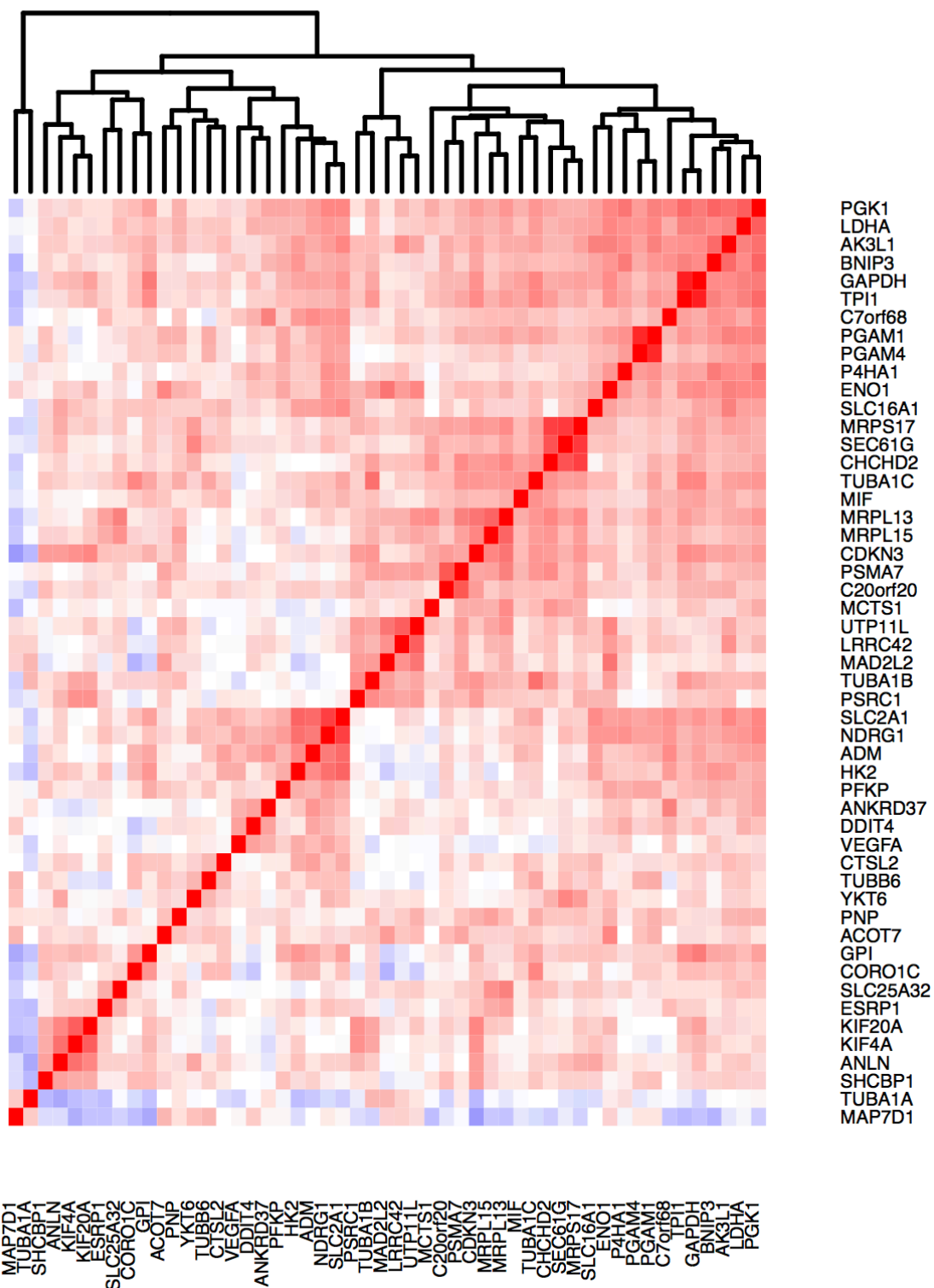
LIHC Buffa52



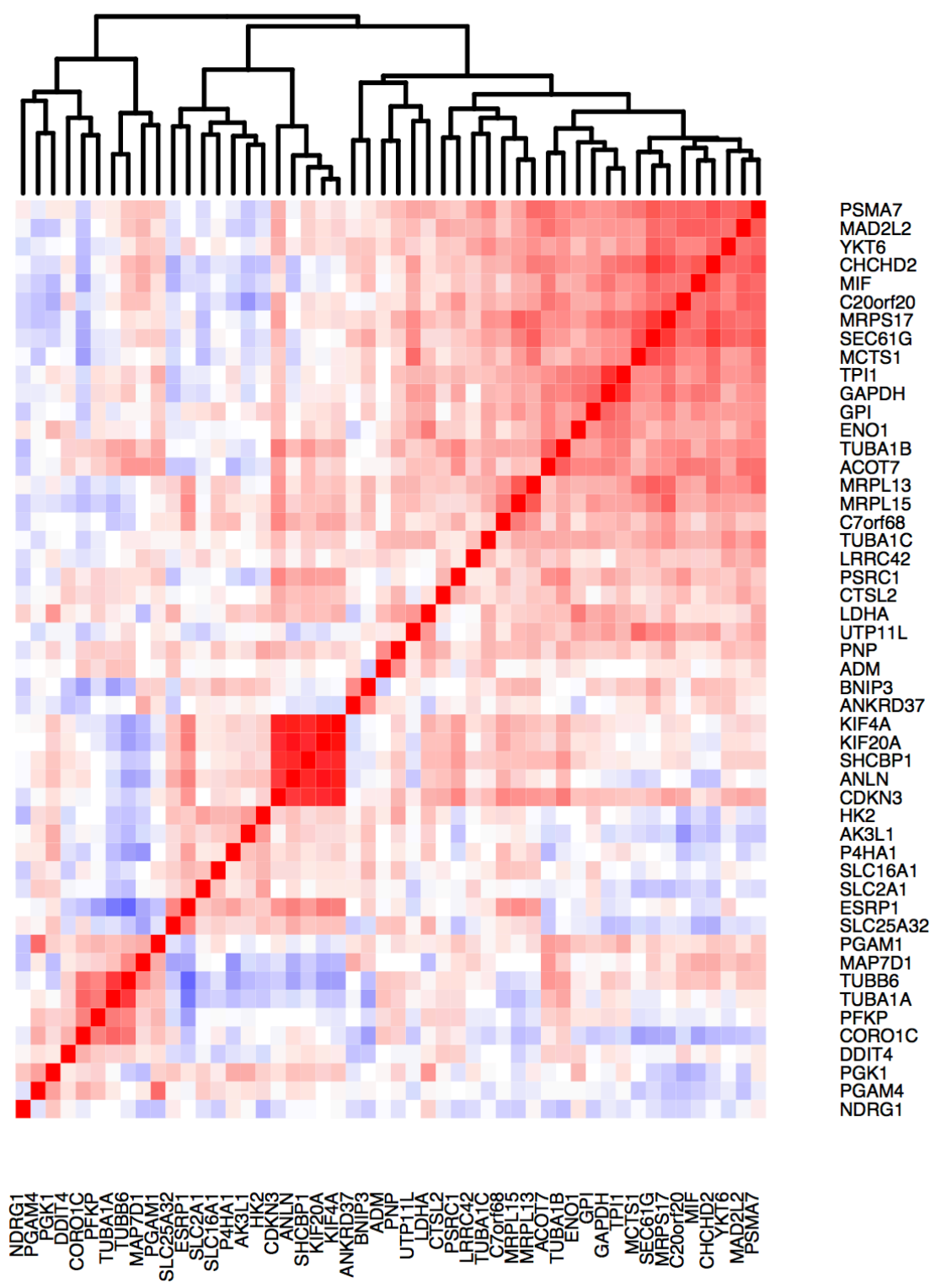
LUAD Buffa52



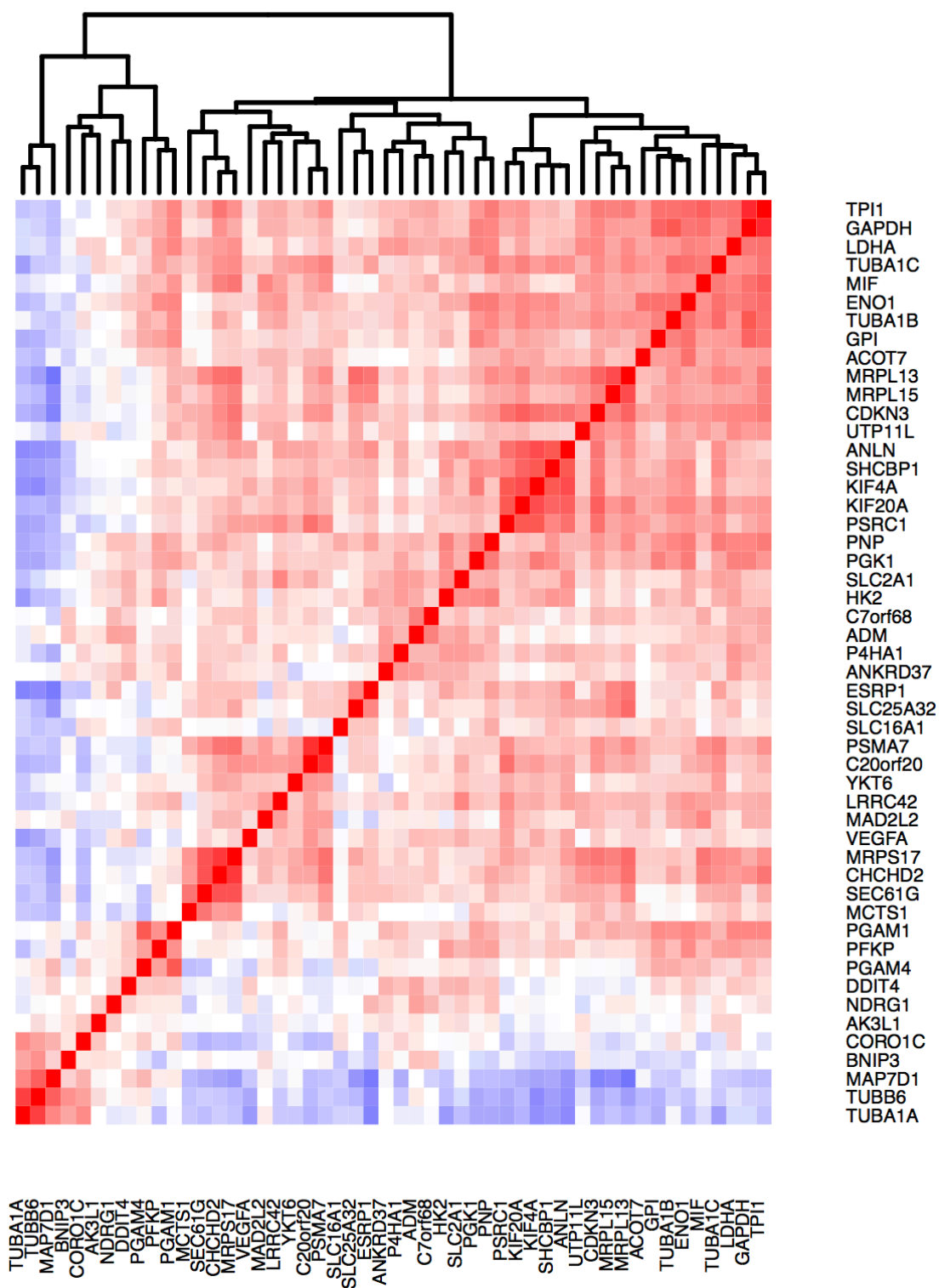
LUSC Buffa52



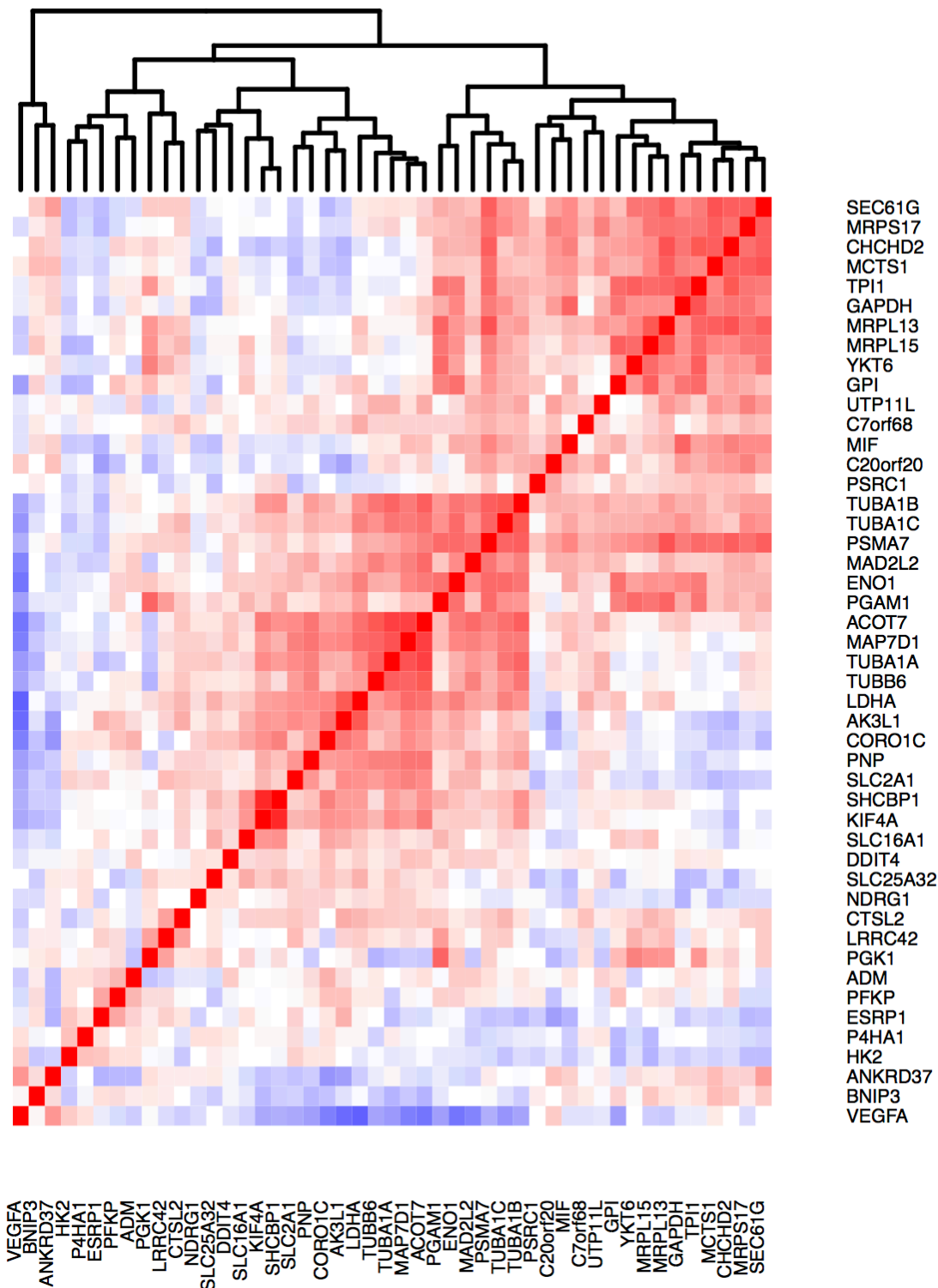
PRAD Buffa52



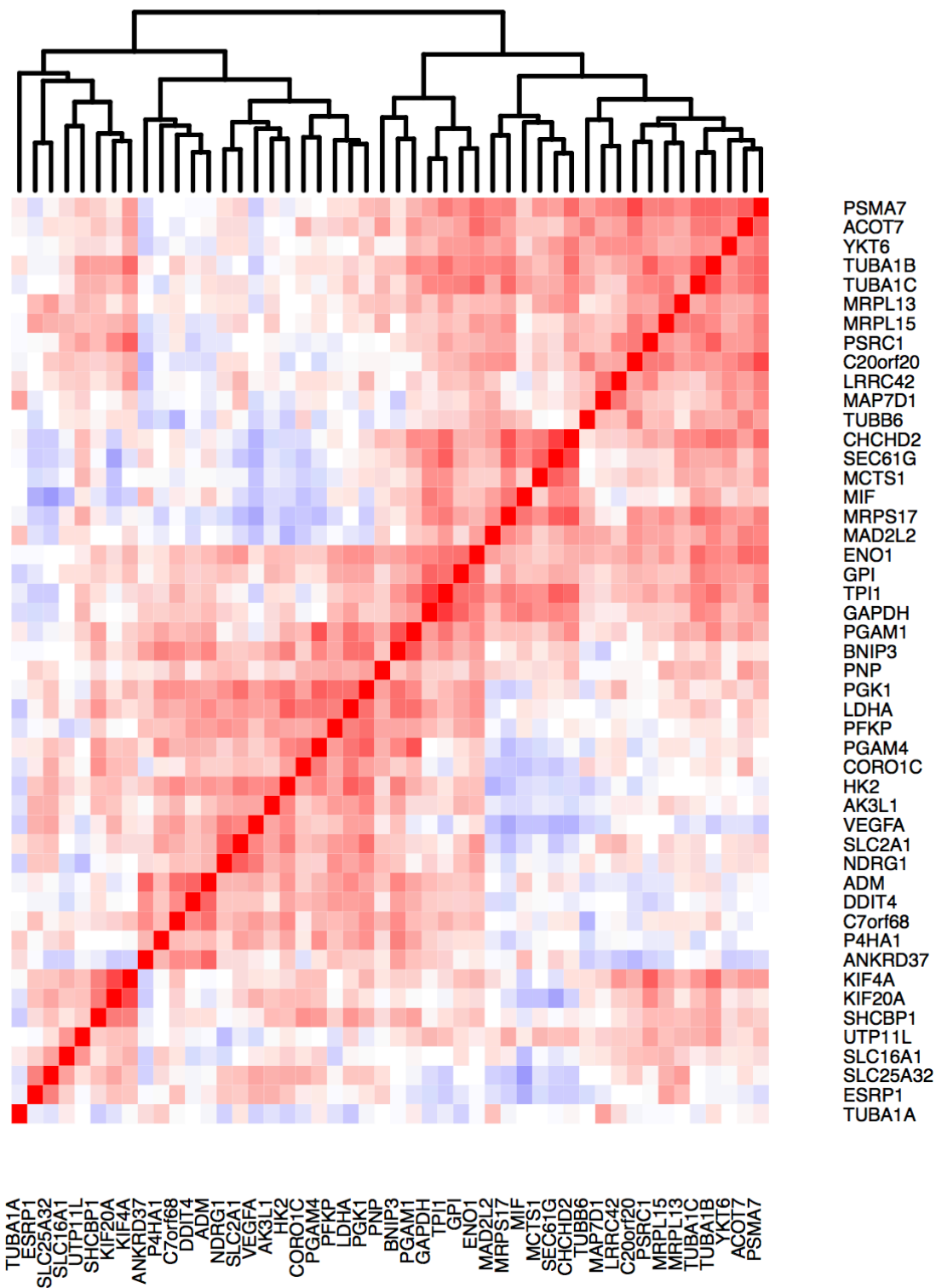
STAD Buffa52



THCA Buffa52

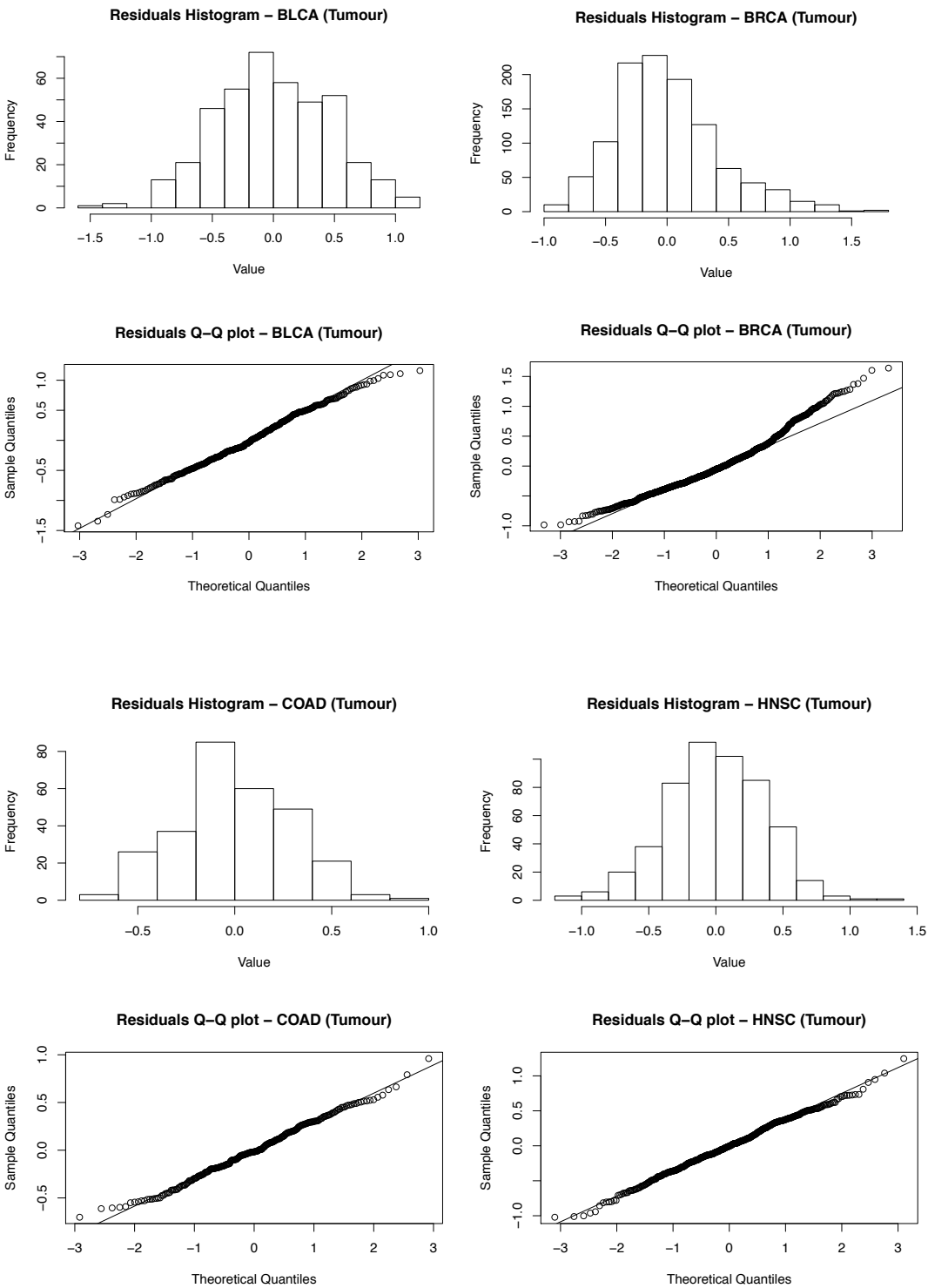


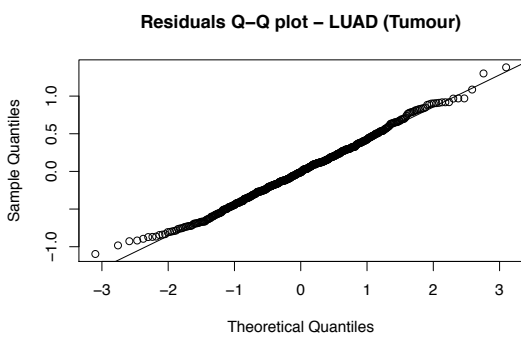
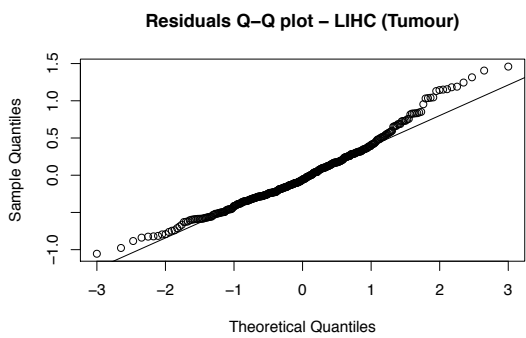
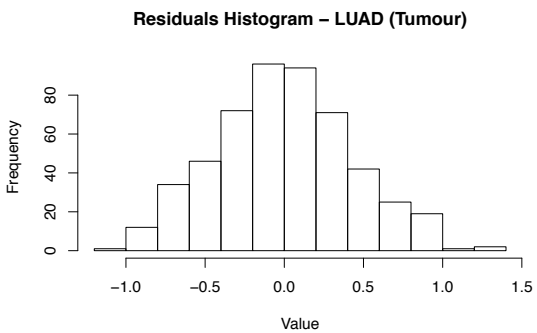
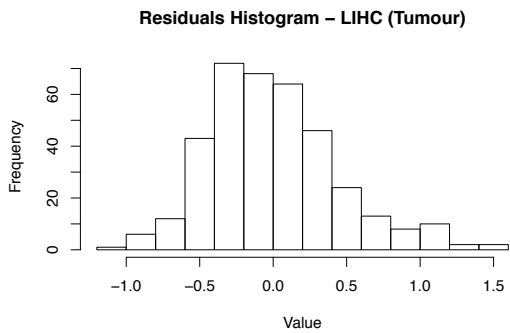
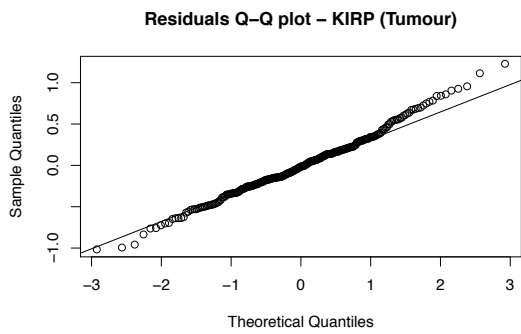
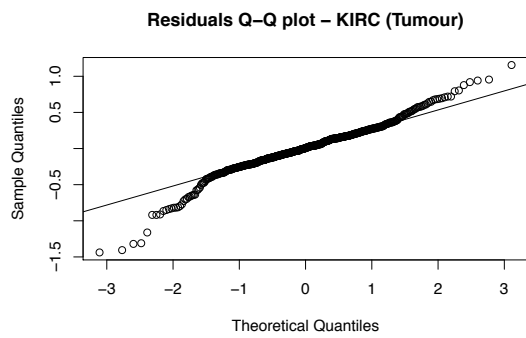
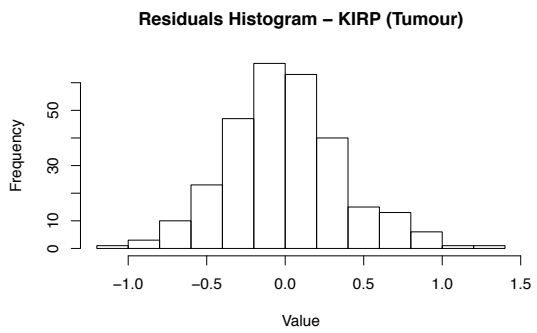
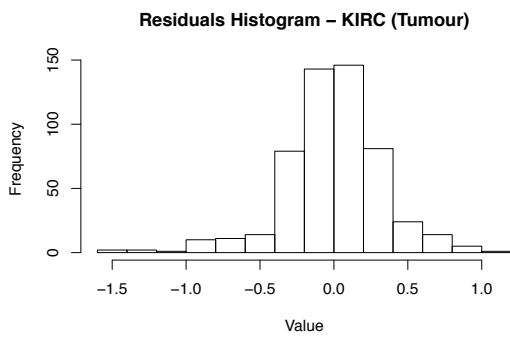
UCEC Buffa52



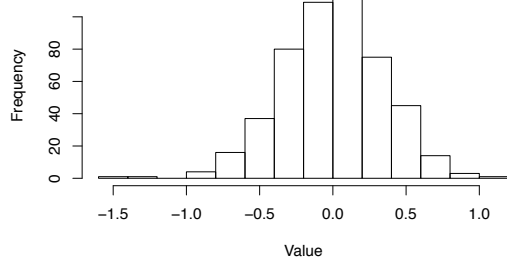
Self-correlation of the Buffa hypoxia signature in terms of mRNA expression across 13 cancer types. Correlation calculated as Spearman's rho. A highly self-correlated signature is a prerequisite for usability in a given data set. Figure generated by sigQC.

Figure S3: Residuals Histogram and Q-Q Plots for the Principal Component Regression of the Expression of APOBEC3 Genes in Relation to the Mean Hypoxia Signature

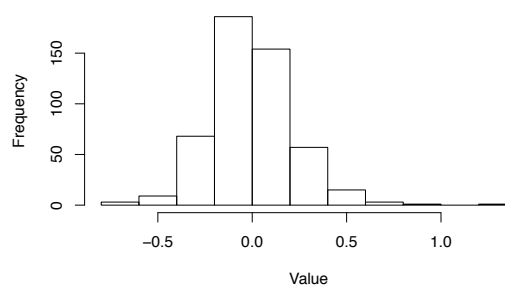




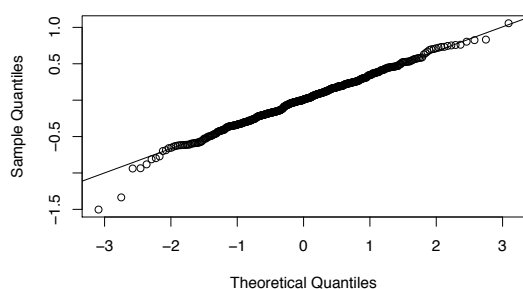
Residuals Histogram – LUSC (Tumour)



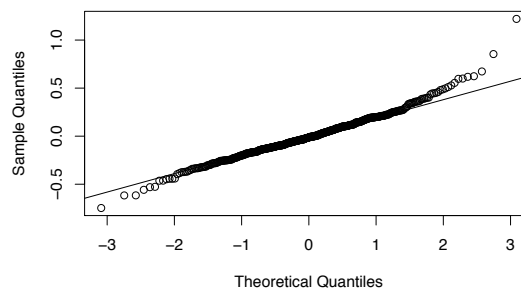
Residuals Histogram – PRAD (Tumour)



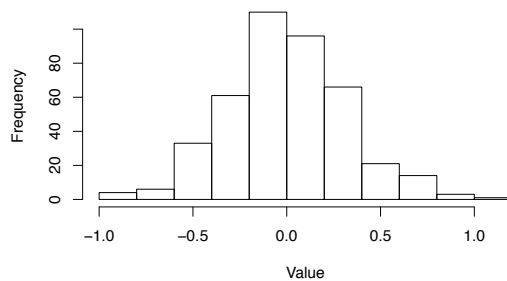
Residuals Q-Q plot – LUSC (Tumour)



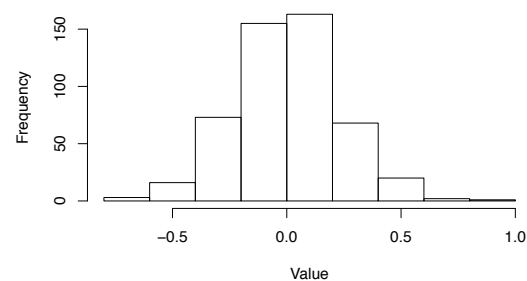
Residuals Q-Q plot – PRAD (Tumour)



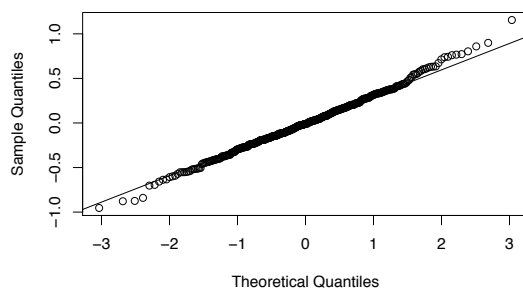
Residuals Histogram – STAD (Tumour)



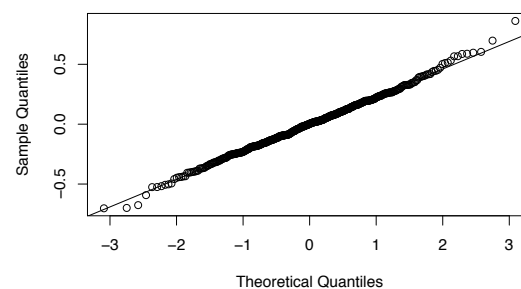
Residuals Histogram – THCA (Tumour)

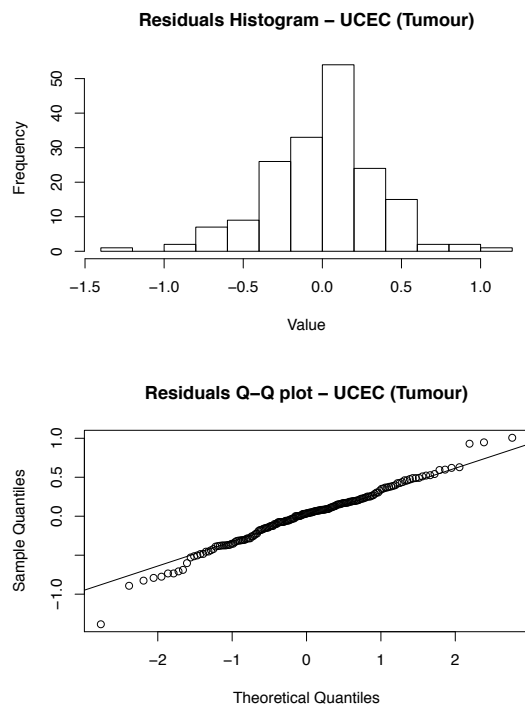


Residuals Q-Q plot – STAD (Tumour)



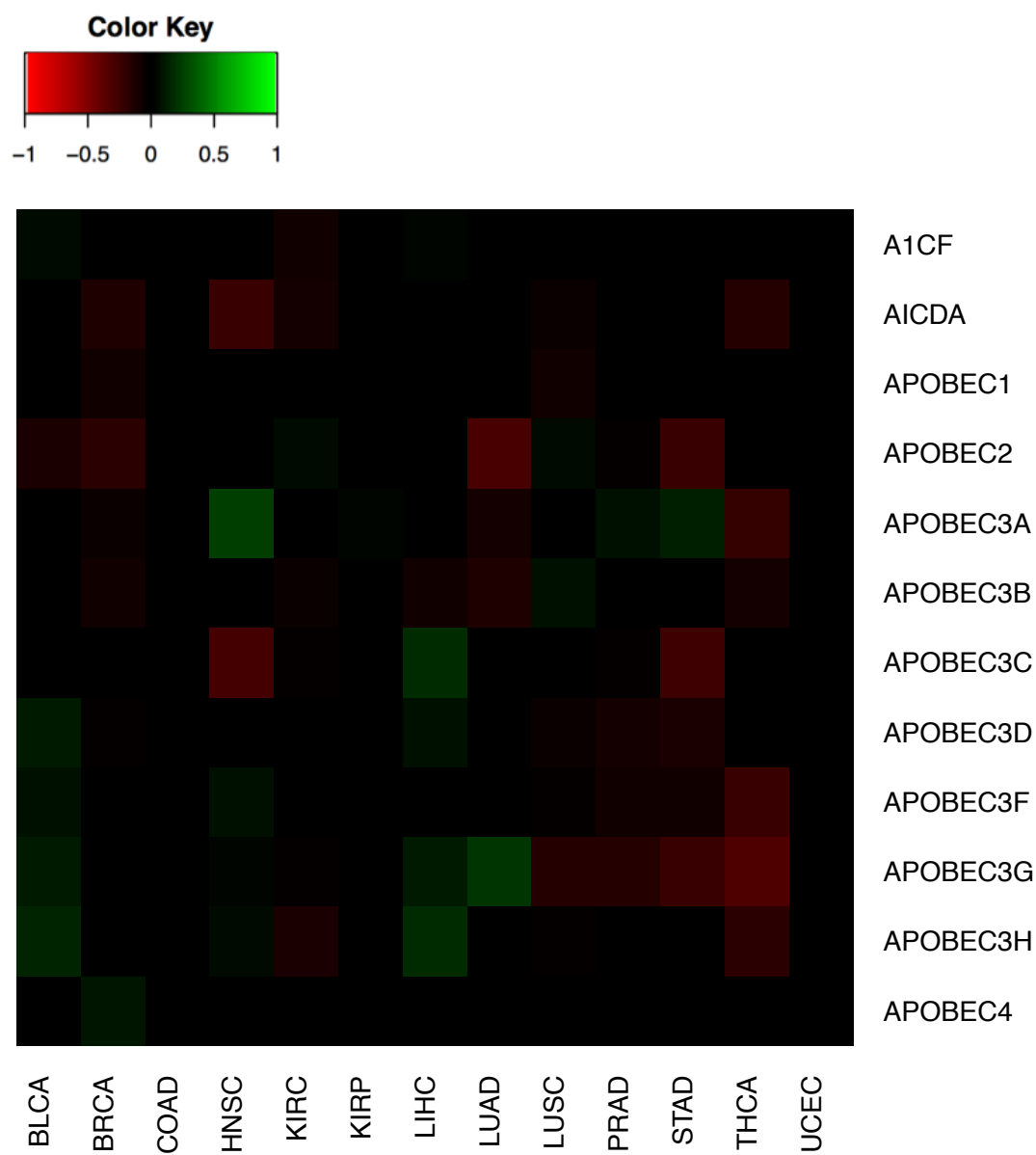
Residuals Q-Q plot – THCA (Tumour)





The residuals of the the principal component regression of the expression of APOBEC3 genes in relation to the mean Buffa hypoxia signature is plotted as a histogram. A normal distribution in the residuals histogram and a linear trend in the Q-Q plot are required for a linear regression model to be accurate.

Figure S4: Difference in Spearman’s Rho between Partial Correlation and Direct Correlation between APOBEC and Hypoxia



The heatmap illustrates the difference in Spearman’s rho between the correlation with correction for xCell signatures and without. Spearman’s rho for a direct correlation is subtracted from the rho for a correlation partial to significant xCell signatures.

