LIST OF TABLES

Table 1.1	Known mutational signatures due to DNA damage	8
Table 1.2	Known mutational signatures due to the activity of DNA repair mechanisms	24
Table 2.1	Similarities between simulated mutational signatures	47
Table 3.1	Summary of breast cancer samples and their data sources	72
Table 3.2	Validating consensus mutational signatures found in breast cancer	84
Table 4.1	Validating consensus mutational signatures found in human cancer	97
Table 5.1	Mutational signatures and age of diagnosis	119
Table 6.1	Summary of the deciphered signatures of mutational processes in human cancer	123

LIST OF FIGURES

Figure 1.1	Somatic mutations in cancer versus nucleotide polymorphisms in the germline	5
Figure 1.2	Illustration of mutational processes operative in a cancer	7
Figure 2.1	Simulated examples of mutational signatures defined over different mutational alphabets	46
Figure 2.2	Simulated example of a mutational catalogue of cancer genome	48
Figure 2.3	Simulated example of three mutational signatures active in a single cancer genome	51
Figure 2.4	Simulated example of mutational signatures deciphered from a set of mutational catalogues	54
Figure 2.5	Deciphering mutational signatures from a set of 100 simulated mutational catalogues.	61
Figure 2.6	Design for simulating four mutational signatures with different similarities between them	63
Figure 2.7	Deciphering mutational signatures with different similarities between them	63
Figure 2.8	Deciphering mutational signatures from different sets of cancer genomes	64
Figure 2.9	Dependencies between mutational signatures and mutational catalogues of cancer genomes	64
Figure 2.10	Dependencies between mutational signatures and numbers of somatic mutations	65
Figure 2.11	Deciphering mutational signatures with different contributions in mutational catalogues	66
Figure 2.12	Deciphering errors of exposures and accuracy of mutational signatures	67
Figure 2.13	Evaluating the error rate of identified contributions of mutations signatures	68
Figure 3.1	Mutational signatures extracted from 119 breast cancer genomes	73
Figure 3.2	Breast cancer whole-genome mutational signatures with indels and dinucleotides	75
Figure 3.3	Breast cancer whole-genome mutational signatures with strand-bias	76
Figure 3.4	Signature BC-WG-2 with additional sequence context	77
Figure 3.5	Signature BC-WG-6 with additional sequence context	78
Figure 3.6	Mutational signatures extracted from 884 breast cancer exomes	79
Figure 3.7	Breast cancer exome mutational signatures with indels and dinucleotides	80
Figure 3.8	Breast cancer exome mutational signatures with strand-bias	81
Figure 3.9	Clustering of breast cancer signatures derived from whole-genome and exome data	82
Figure 3.10	Contributions of mutational signatures in a selected set of 25 breast cancer samples	85
Figure 3.11	Summary of the contributions of the mutational signatures in breast cancer	85
Figure 3.12	Samples harbouring *BRCA1/2* mutations and contributions of mutational signatures	87
Figure 3.13	Estrogen receptor positive/negative samples and contributions of mutational signatures	88
Figure 3.14	Age of diagnosis and mutations due to different mutational signatures	88
Figure 4.1	Samples used for deciphering signatures of mutational processes in human cancer	91
Figure 4.2	Mutational burden in human cancer	92
Figure 4.3	Clustering of mutational signatures	94
Figure 4.4	Types of statuses for validating mutational signatures	95
Figure 4.5	Consensus validated mutational signatures in human cancer	98

Figure 4.6	Consensus mutational signatures that failed validation	99
Figure 4.7	Consensus mutational signatures for which it is not possible to perform validation	99
Figure 4.8	Consensus mutational signatures with strand-bias	101
Figure 4.9	Signatures of mutational processes and the cancer types in which they are found	104
Figure 4.10	Prevalence of validated mutational signatures across all cancer types	105
Figure 4.11	Contributions of mutational signatures in a selected set of cancer types	107
Figure 5.1	Samples harbouring *BRCA1/2* mutations and contributions of Signature 3	114
Figure 5.2	Associating exposures of mutational signatures to cigarette smoking	116
Figure 5.3	Associating molecular or clinical features with the activity of mutational signatures	118

LIST OF ABBREVIATIONS

8-oxoG	7,8-dihydro-8-oxoguanine
AID	Activation-induced cytosine deaminase
APEX1	Apurinic/apyrimidinic endonuclease
API	Application programming interface
APOBEC	Apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like
BAM	Binary sequence alignment and mapping
BER	Base excision repair
BIR	Break-induced replication
BSS	Blind source separation
CGP	Cancer Genome Project
CPD	Cyclobutane pyrimidine dimer
D-loop	Displacement-loop
DA-NER	Domain associated nucleotide excision repair
dbSNP	Single Nucleotide Polymorphism Database
DNA	Deoxyribonucleic acid
dNTP	Deoxynucleoside triphosphate
DSBR	Classical double-strand break repair
FDR	False discovery rate
GG-NER	Global genome-wide nucleotide excision repair
GLM	Generalized linear model
GRCh37	Genome Reference Consortium human genome (build 37)
IARC	International Agency for Research on Cancer
ICA	Independent component analysis
ICGC	International Cancer Genome Consortium
IGHV	Immunoglobulin gene hypermutation
Indel	Small insertion/deletion
LP-BER	Long patch base excision repair
MMEJ	Microhomology mediated end joining
MMR	DNA mismatch repair
NER	Nucleotide excision repair
NHEJ	Non-homologous end joining

NHLBI	National Heart, Lung, and Blood Institute
NMF	Nonnegative matrix factorization
PAH	Polycyclic aromatic hydrocarbons
PCR	Polymerase chain reaction
Pol	Polymerase
POL II	RNA polymerase II
POLE	DNA polymerase epsilon catalytic subunit A
RFC	Replication factor C
RNA	Ribonucleic acid
RNS	Reactive nitrogen species
ROS	Reactive oxygen species
SDSA	Synthesis-dependent strand annealing
SNP	Single nucleotide polymorphism
SP-BER	Short patch base excision repair
SSA	Single-strand annealing
TC-BER	Transcription coupled base excision repair
TC-NER	Transcription coupled nucleotide excision repair
TCGA	The Cancer Genome Atlas
TET	Ten-eleven translocation methylcytosine dioxygenase
TP53	Tumour protein p53
UCSC	University of California, Santa Cruz
UV	Ultraviolet
UV-A	Ultraviolet A
UV-B	Ultraviolet B
UV-C	Ultraviolet B

APPENDIX I: Alphabets of mutational types

This appendix contains information for the alphabets of mutation types used throughout the course of this thesis. These alphabets were termed Ξ_6, Ξ_{96}, Ξ_{99}, Ξ_{192}, and Ξ_{1536} in chapter 2 and each one will be discussed in more details in the next few sections.

Mutational alphabet Ξ_6

The Ξ_6 alphabet is perhaps the simplest possible alphabet as it considers only the six types of somatic substitutions: **C>A**, **C>G**, **C>T**, **T>A**, **T>C**, **T>G**. All mutations are denoted using the pyrimidine of the Watson-Crick base pair as the reference and, in this appendix, these substitutions are coloured consistently with the way they are plotted in the majority of figures throughout this thesis.

Mutational alphabet Ξ_{96}

The Ξ_{96} alphabet provides greater resolution for examining the six types of single nucleotide variants (*i.e.*, the Ξ_6 alphabet) by including the immediate sequence context of each mutated base. In this alphabet, a mutation type contains a somatic substitution and both the 5' and 3' base next to the somatic mutation. For example, a **C>T** mutation can be characterized as …**TpCpG**…>…**TpTpG**… (mutated base underlined and presented as the pyrimidine partner of the mutated base pair) generating 96 possible mutation types – (6 types of substitutions) * (4 types of 5' bases) * (4 types of 3' bases). Table listing each of the 96 substitution types, the reference trinucleotide, and the mutated trinucleotide is provided below.

Sub	Ref	Mut	Sub	Ref	Mut
C>A	ApCpA	ApApA	T>A	ApTpA	ApApA
C>A	ApCpC	ApApC	T>A	ApTpC	ApApC
C>A	ApCpG	ApApG	T>A	ApTpG	ApApG
C>A	ApCpT	ApApT	T>A	ApTpT	ApApT
C>A	CpCpA	CpApA	T>A	CpTpA	CpApA
C>A	CpCpC	CpApC	T>A	CpTpC	CpApC
C>A	CpCpG	CpApG	T>A	CpTpG	CpApG
C>A	CpCpT	CpApT	T>A	CpTpT	CpApT
C>A	GpCpA	GpApA	T>A	GpTpA	GpApA
C>A	GpCpC	GpApC	T>A	GpTpC	GpApC
C>A	GpCpG	GpApG	T>A	GpTpG	GpApG
C>A	GpCpT	GpApT	T>A	GpTpT	GpApT

C>A	TpCpA	TpApA	T>A	TpTpA	TpApA
C>A	TpCpC	TpApC	T>A	TpTpC	TpApC
C>A	TpCpG	TpApG	T>A	TpTpG	TpApG
C>A	TpCpT	TpApT	T>A	TpTpT	TpApT
C>G	ApCpA	ApGpA	T>C	ApTpA	ApCpA
C>G	ApCpC	ApGpC	T>C	ApTpC	ApCpC
C>G	ApCpG	ApGpG	T>C	ApTpG	ApCpG
C>G	ApCpT	ApGpT	T>C	ApTpT	ApCpT
C>G	CpCpA	CpGpA	T>C	CpTpA	CpCpA
C>G	CpCpC	CpGpC	T>C	CpTpC	CpCpC
C>G	CpCpG	CpGpG	T>C	CpTpG	CpCpG
C>G	CpCpT	CpGpT	T>C	CpTpT	CpCpT
C>G	GpCpA	GpGpA	T>C	GpTpA	GpCpA
C>G	GpCpC	GpGpC	T>C	GpTpC	GpCpC
C>G	GpCpG	GpGpG	T>C	GpTpG	GpCpG
C>G	GpCpT	GpGpT	T>C	GpTpT	GpCpT
C>G	TpCpA	TpGpA	T>C	TpTpA	TpCpA
C>G	TpCpC	TpGpC	T>C	TpTpC	TpCpC
C>G	TpCpG	TpGpG	T>C	TpTpG	TpCpG
C>G	TpCpT	TpGpT	T>C	TpTpT	TpCpT
C>T	ApCpA	ApTpA	T>G	ApTpA	ApGpA
C>T	ApCpC	ApTpC	T>G	ApTpC	ApGpC
C>T	ApCpG	ApTpG	T>G	ApTpG	ApGpG
C>T	ApCpT	ApTpT	T>G	ApTpT	ApGpT
C>T	CpCpA	CpTpA	T>G	CpTpA	CpGpA
C>T	CpCpC	CpTpC	T>G	CpTpC	CpGpC
C>T	CpCpG	CpTpG	T>G	CpTpG	CpGpG
C>T	CpCpT	CpTpT	T>G	CpTpT	CpGpT
C>T	GpCpA	GpTpA	T>G	GpTpA	GpGpA
C>T	GpCpC	GpTpC	T>G	GpTpC	GpGpC
C>T	GpCpG	GpTpG	T>G	GpTpG	GpGpG
C>T	GpCpT	GpTpT	T>G	GpTpT	GpGpT
C>T	TpCpA	TpTpA	T>G	TpTpA	TpGpA
C>T	TpCpC	TpTpC	T>G	TpTpC	TpGpC
C>T	TpCpG	TpTpG	T>G	TpTpG	TpGpG
C>T	TpCpT	TpTpT	T>G	TpTpT	TpGpT

Mutational alphabet Ξ_{99}

The Ξ_{99} alphabet extends Ξ_{96} by including three additional mutation types, *viz.*, (i) double nucleotide substitutions, (ii) small insertions or deletions at short tandem repeats, and (iii) small insertions or deletions overlapping with microhomologies at breakpoints.

Mutational alphabet Ξ_{192}

The Ξ_{192} alphabet elaborates Ξ_{96} by considering the transcriptional strand on which a substitution resides. In contrast to all other alphabets, Ξ_{192} is defined only in the regions of the genome where transcription occurs, which in these analyses has been limited to the genomic footprints of protein coding genes. For example, the **C>T** mutations at **TpCpA** are split into two categories: the **C>T** mutations at **TpCpA** occurring on the untranscribed strand of a gene and the **C>T** mutations at **TpCpA** occurring on the transcribed strand. Similarly, all 96 mutations types from Ξ_{96} are extended to form the Ξ_{192} alphabet.

Mutational alphabet Ξ_{1536}

The Ξ_{1536} further extends Ξ_{96} by including two bases 5' and 3' to the mutated base resulting in 1,536 possible mutated pentanucleotides - (6 types of substitutions) * (16 types of the two immediate 5' bases) * (16 types of the two immediate 3' bases). For example, using the Ξ_{1536} alphabet, one of the 256 subclasses of a **C>T** mutation is …**ApTpCpGpC**… > …**ApTpTpGpC**… For brevity, the complete list of mutation types included in Ξ_{1536} is not provided here.

APPENDIX II: List of analysed samples

This appendix contains a summary list of all samples analysed throughout the course of this thesis. Summarized information is provided for all 7,042 separated by sequencing types (exome sequencing versus whole-genome sequencing), cancer types, and respective data sources. It should be noted that the pilocytic astrocytomas dataset contains a small number of other paediatric low-grade gliomas and paediatric low-grade glioneuronal tumours. Information for each individual sample including its mutational catalogues and somatic mutations (both before and after filtering) could be found at ftp://ftp.sanger.ac.uk/pub/cancer/AlexandrovEtAl.

Exome sample types and data sources	Total
ALL	**140**
doi:10.1038/nature10725	15
doi:10.1038/ng.2508	29
doi:10.1038/ng.2532	42
New unpublished samples	54
AML	**147**
TCGA data portal	147
Bladder	**136**
TCGA data portal	136
Breast	**844**
doi:10.1038/nature10933	63
doi:10.1038/nature11017	9
New unpublished samples	5
TCGA data portal	767
Cervix	**38**
TCGA data portal	38
CLL	**103**
doi:10.1038/ng.1032	80
ICGC data portal	23
Colorectum	**559**
doi:10.1038/nature11282	70
TCGA data portal	489
Oesophageal	**146**
doi:10.1038/ng.2591	146
Glioblastoma	**98**
ICGC data portal	50
TCGA data portal	48
Glioma Low Grade	**217**
TCGA data portal	217

Whole-genome sample types and data sources	Total
ALL	**1**
New unpublished samples	1
AML	**7**
doi:10.1038/nature10738	7
Breast	**119**
doi:10.1016/j.cell.2012.04.024	21
New unpublished samples	98
CLL	**28**
doi:10.1038/nature10113	4
New unpublished samples	24
Liver	**88**
ICGC data portal	66
New unpublished samples	22
Lung Adenocarcinoma	**24**
doi:10.1016/j.cell.2012.08.029	24
Lymphoma B-cell	**24**
doi:10.1038/ng.2468	1
New unpublished samples	23
Medulloblastoma	**100**
New unpublished samples	100
Pancreas	**15**
New unpublished samples	15
Pilocytic Astrocytoma	**101**
doi:10.1038/ng.2611	38
New unpublished samples	63
Grand Total	**507**

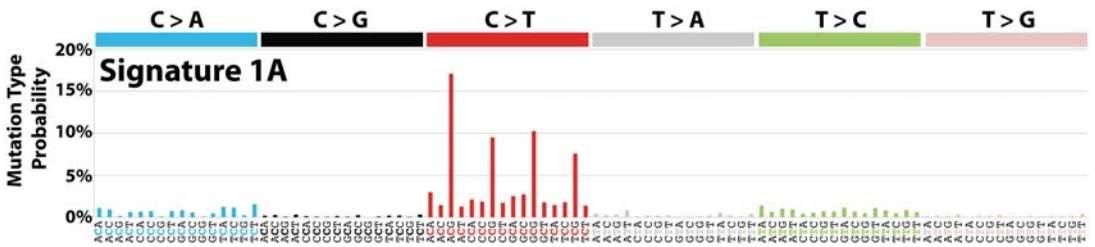
Head and Neck	**380**
doi:10.1126/science.1206923	12
doi:10.1126/science.1208130	68
TCGA data portal	300
Kidney Chromophobe	**65**
TCGA data portal	65
Kidney Clear Cell	**325**
doi:10.1038/ng.1014	10
doi:10.1038/ng.2323	7
TCGA data portal	308
Kidney Papillary	**100**
TCGA data portal	100
Lung Adenocarcinoma	**636**
doi:10.1016/j.cell.2012.08.029	150
doi:10.1038/nature07423	30
doi:10.1101/gr.145144.112	75
TCGA data portal	381
Lung Small Cell	**70**
doi:10.1038/ng.2396	29
doi:10.1038/ng.2405	40
ICGC data portal	1
Lung Squamous	**176**
TCGA data portal	176
Lymphoma B-cell	**24**
doi:10.1038/nature10351	16
doi:10.1038/ng.2468	8
Melanoma	**396**
doi:10.1016/j.cell.2012.06.024	92
doi:10.1038/nature11071	28
doi:10.1038/ng.1041	8
ICGC data portal	1
New unpublished samples	17
TCGA data portal	250
Myeloma	**69**
New unpublished samples	69
Neuroblastoma	**210**
doi:10.1038/ng.2493	13
doi:10.1038/ng.2529	197
Ovary	**471**
doi:10.1126/science.1196333	8
TCGA data portal	463
Pancreas	**98**
doi:10.1073/pnas.1118046108	22
doi:10.1126/science.120060	10
ICGC data portal	37

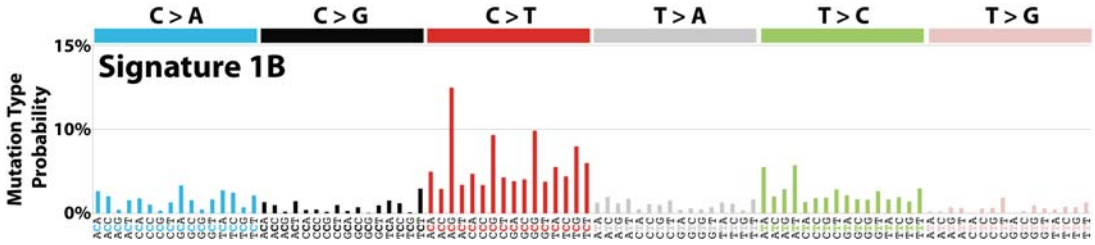
TCGA data portal	29
Prostate	**330**
doi:10.1038/nature09744	7
doi:10.1038/nature11125	61
doi:10.1038/ng.2279	112
TCGA data portal	150
Stomach	**212**
doi:10.1038/ng.2246	14
doi:10.1038/ng.982	22
ICGC data portal	10
TCGA data portal	166
Thyroid	**304**
TCGA data portal	304
Uterus	**241**
TCGA data portal	241
Grand Total	**6,535**

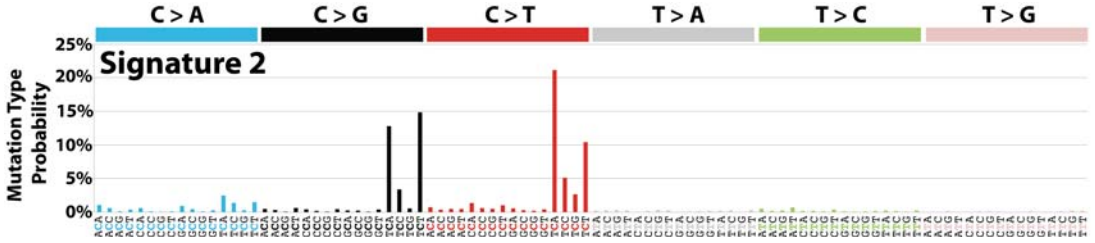
APPENDIX III: Mutational signatures in human cancer

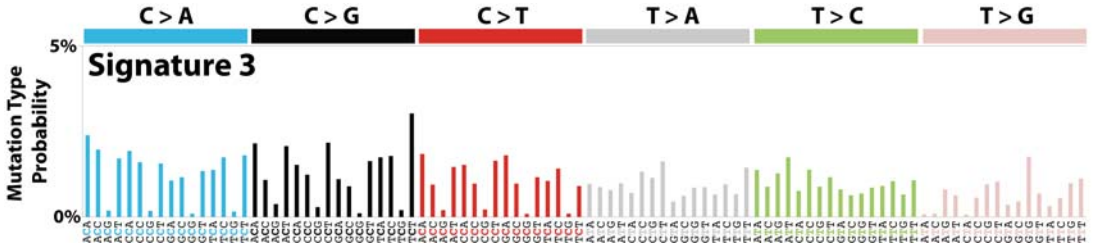
This appendix contains high-resolution figures for the twenty-seven consensus mutational signatures that were deciphered by applying the developed computational approach across the spectrum of human cancer (chapter 4). Each mutational signature is shown using the same plot. Signatures are displayed based on the trinucleotide frequency of the human genome. The probability bars for each of the six types of substitutions as well as the mutated bases are displayed in different colours. The mutation types are displayed on the horizontal axes, while vertical axes depict the percentages of mutations attributed to specific mutation types. The plots are ordered by signature validation types: validated mutational signatures (Signatures 1A, 1B, 2 through 21), mutational signatures that failed validation (Signatures R1, R2, and R3), and mutational signatures for which it was not possible to perform validation (Signatures U1 and U2).

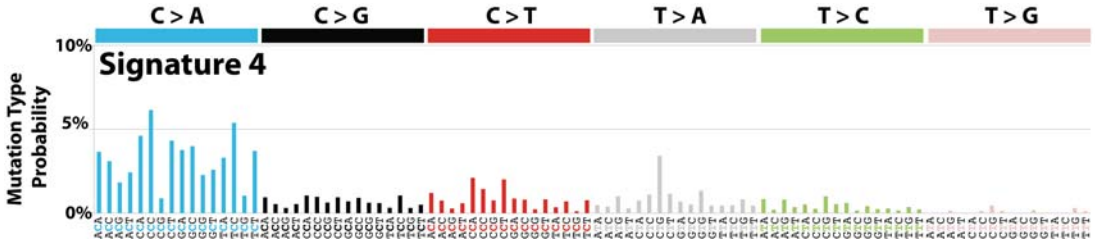
Validated mutational signatures

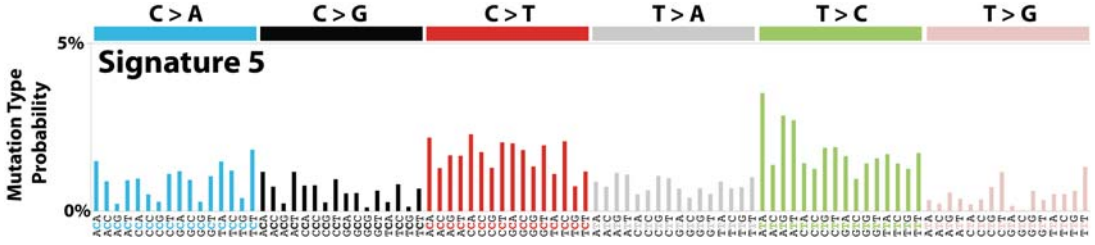


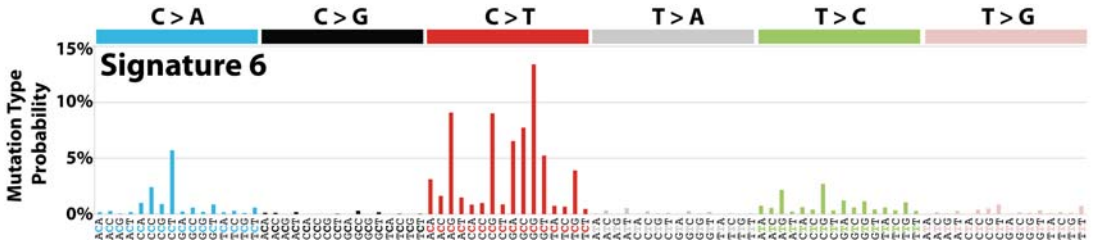


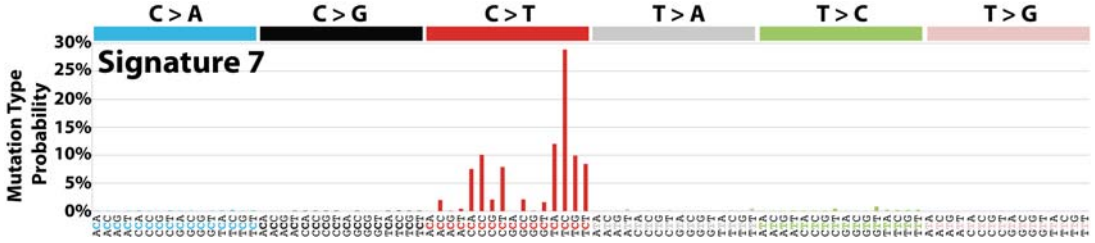


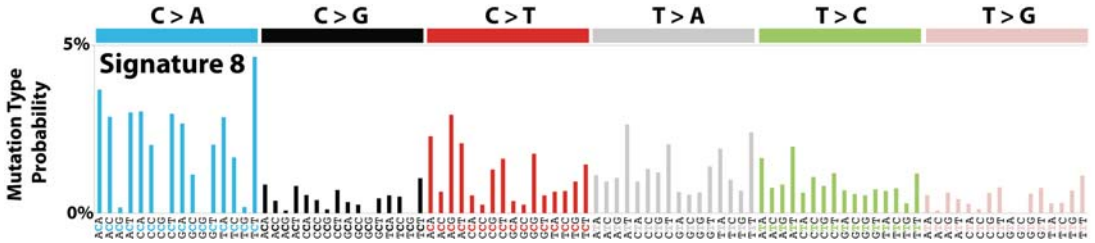


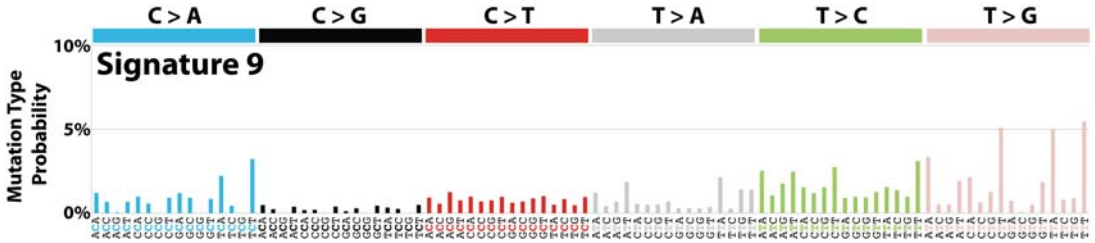


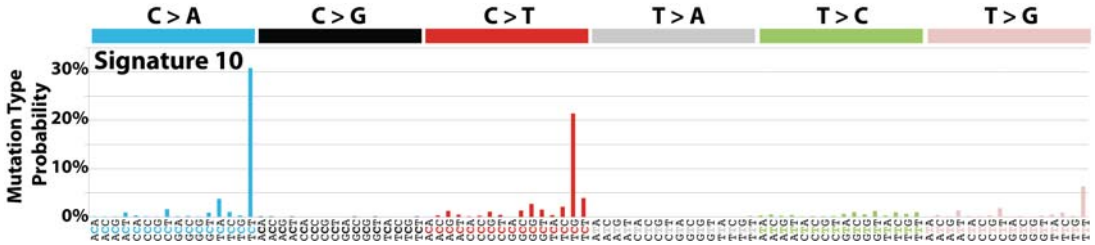


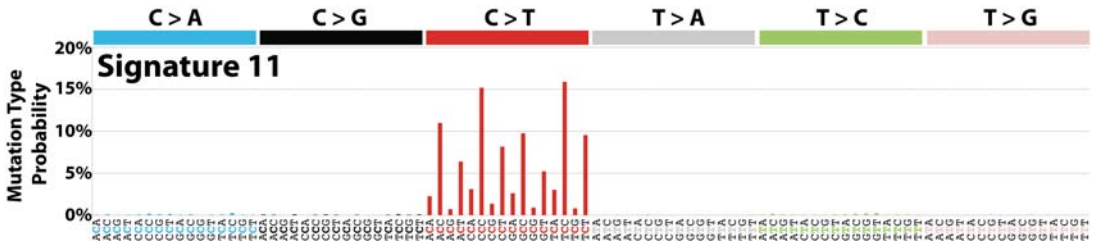

C > A
C > G
C > T
T > A
T > C
T > G
Signature 6
Mutation Type Probability
15%
10%
5%
0%

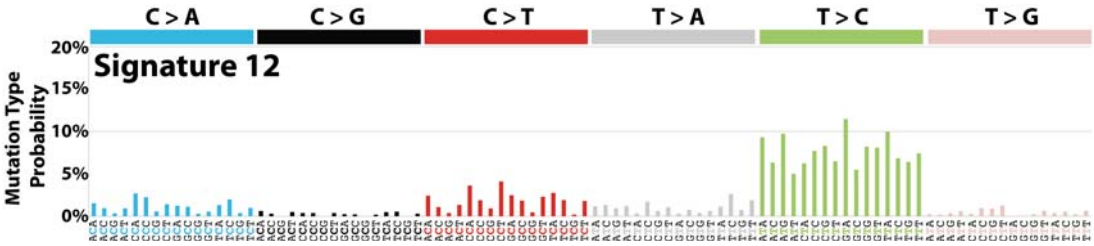

C > A
C > G
C > T
T > A
T > C
T > G
Signature 7
Mutation Type Probability
30%
25%
20%
15%
10%
5%
0%

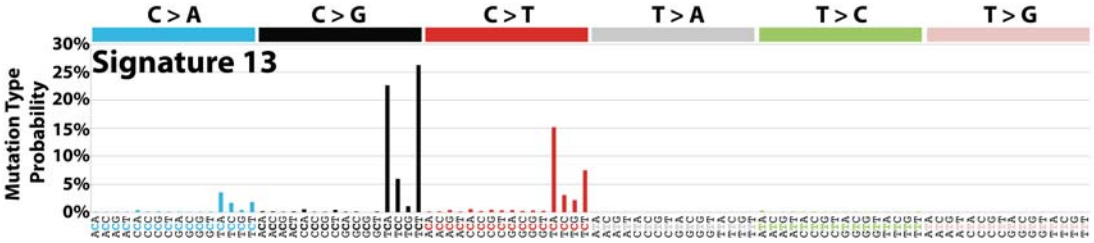

C > A
C > G
C > T
T > A
T > C
T > G
Signature 8
Mutation Type Probability
5%
0%

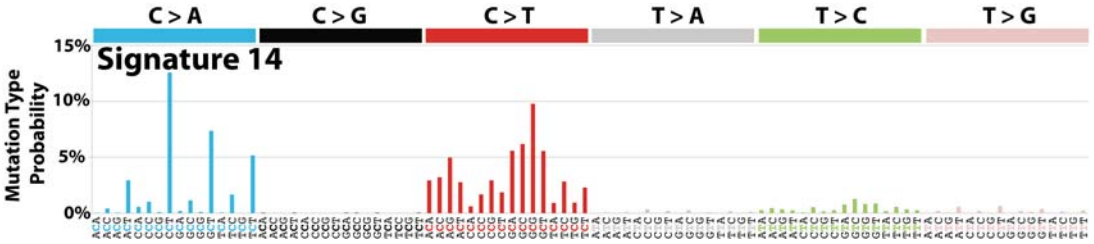

C > A
C > G
C > T
T > A
T > C
T > G
Signature 9
Mutation Type Probability
10%
5%
0%

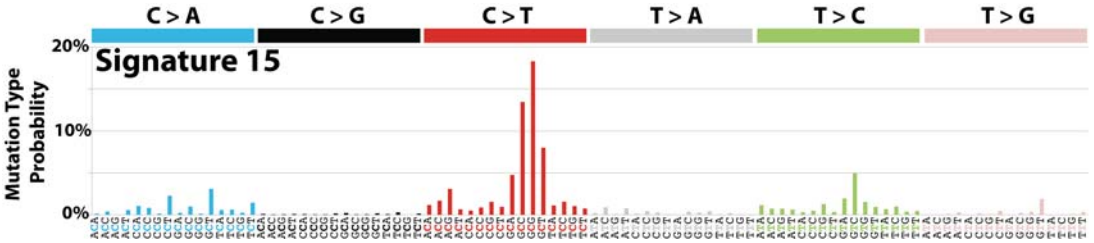

C > A
C > G
C > T
T > A
T > C
T > G
Signature 10
Mutation Type Probability
30%
20%
10%
0%

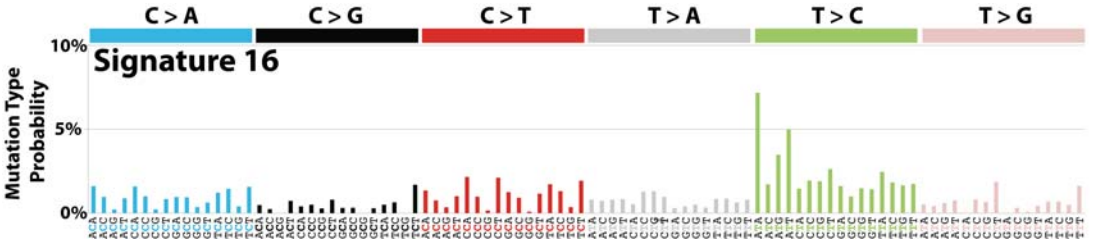

C > A
C > G
C > T
T > A
T > C
T > G
Signature 11
Mutation Type Probability
20%
15%
10%
5%
0%

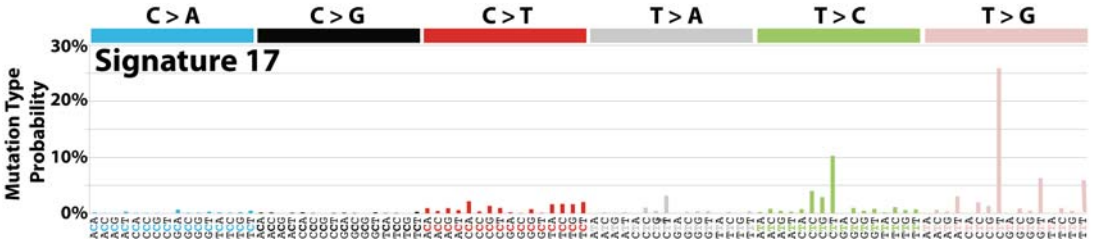

C > A
C > G
C > T
T > A
T > C
T > G
Signature 12
Mutation Type Probability
20%
15%
10%
5%
0%

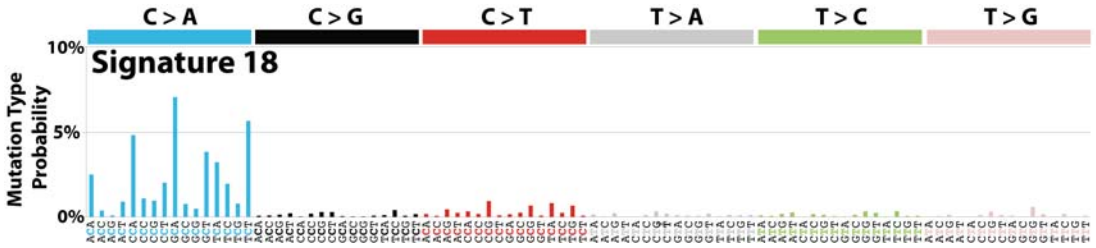

C > A
C > G
C > T
T > A
T > C
T > G
Signature 13
Mutation Type Probability
30%
25%
20%
15%
10%
5%
0%

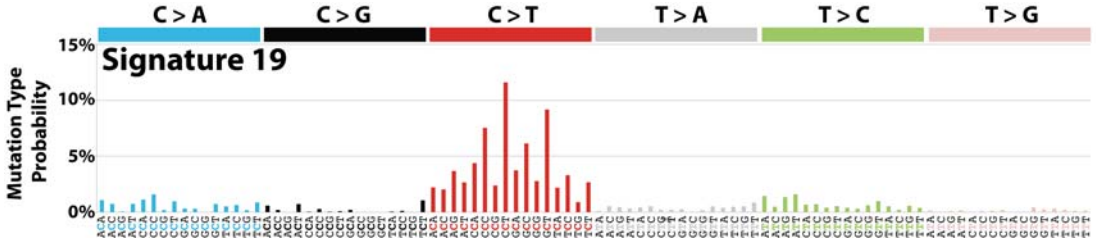

C > A
C > G
C > T
T > A
T > C
T > G
Signature 14
Mutation Type Probability
15%
10%
5%
0%

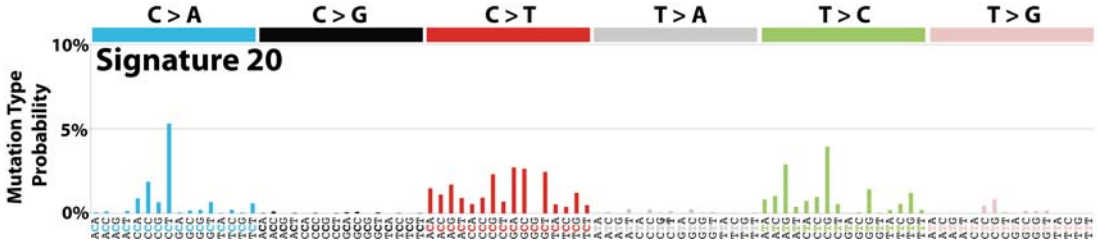

C > A
C > G
C > T
T > A
T > C
T > G
Signature 15
Mutation Type Probability
20%
10%
0%

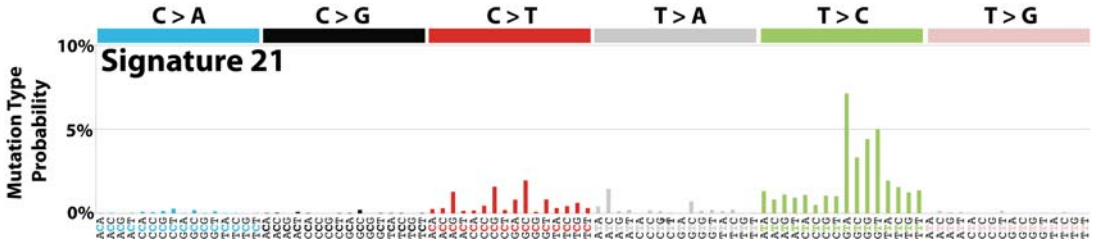

C > A
C > G
C > T
T > A
T > C
T > G
Signature 16
Mutation Type Probability
10%
5%
0%


C > A
C > G
C > T
T > A
T > C
T > G
Signature 17
Mutation Type Probability
30%
20%
10%
0%

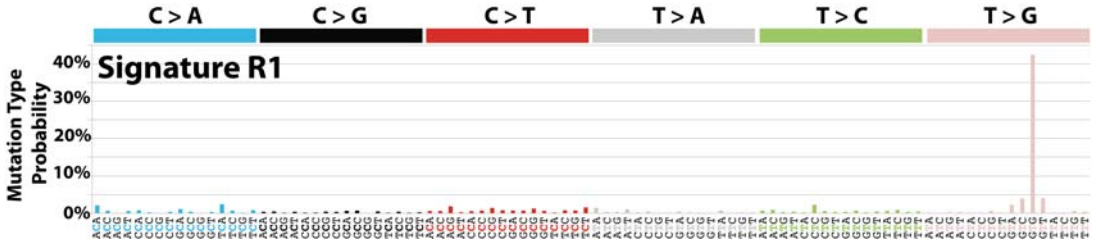

C > A
C > G
C > T
T > A
T > C
T > G
Signature 18
Mutation Type Probability
10%
5%
0%

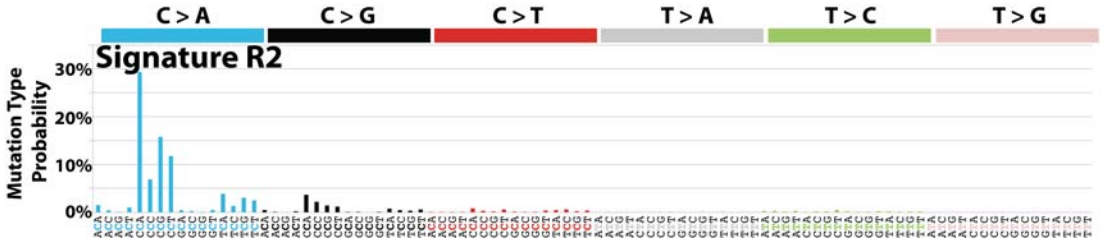

C > A
C > G
C > T
T > A
T > C
T > G
Signature 19
Mutation Type Probability
15%
10%
5%
0%

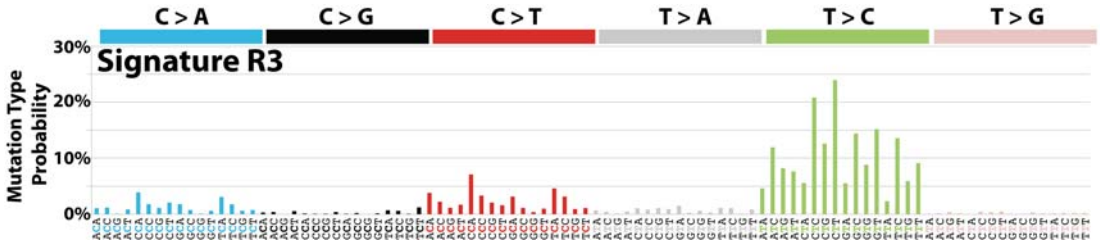

C > A
C > G
C > T
T > A
T > C
T > G
Signature 20
Mutation Type Probability
10%
5%
0%


C > A
C > G
C > T
T > A
T > C
T > G
Signature 21
Mutation Type Probability
10%
5%
0%

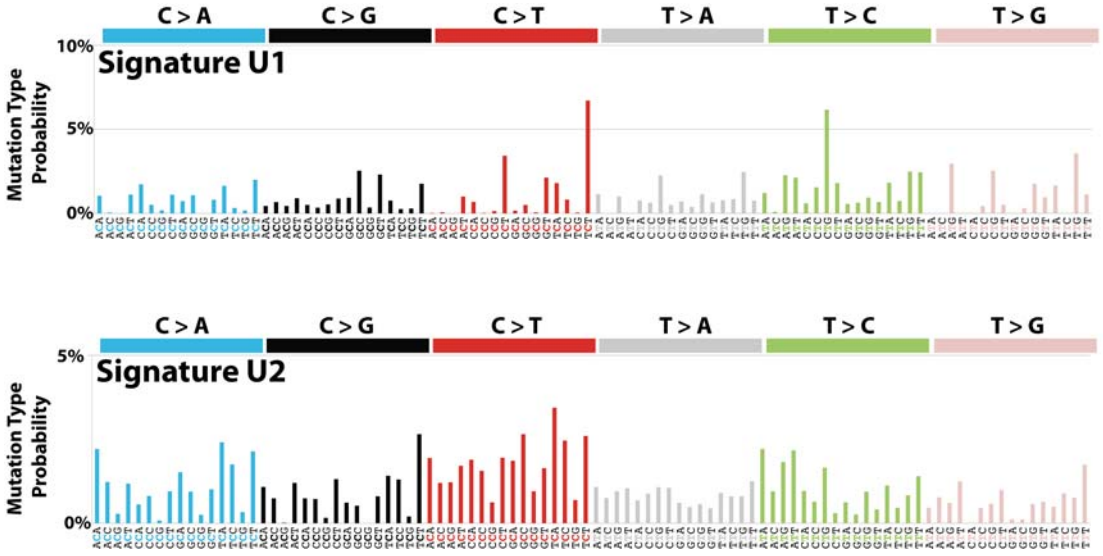
Mutational signatures that failed validation





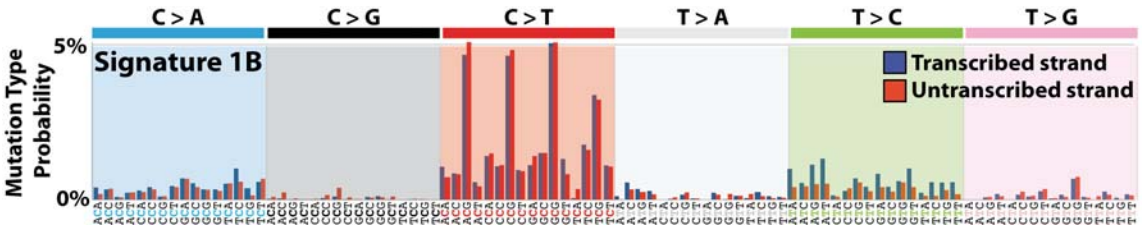


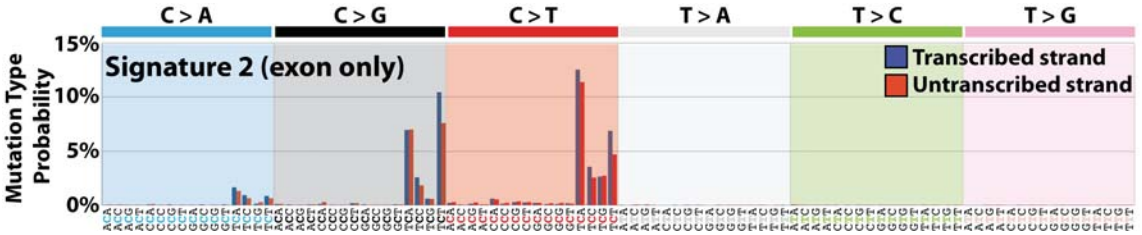
Mutational signatures for which validation was not performed

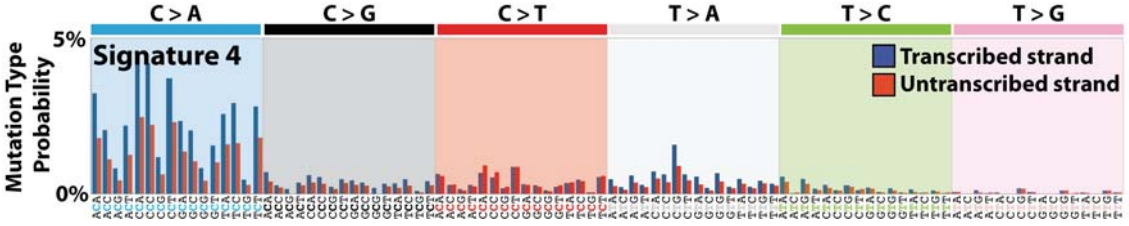


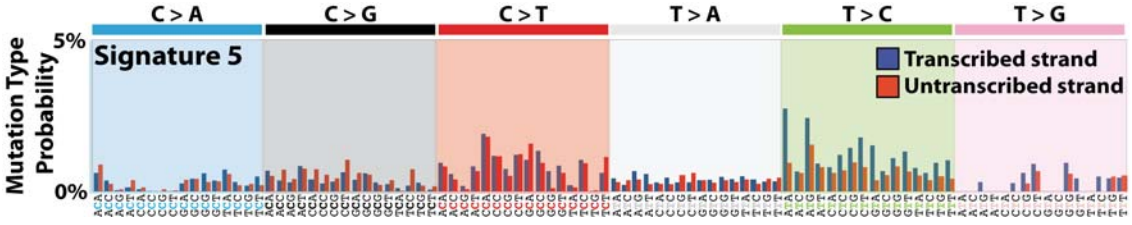
APPENDIX IV: Mutational signatures with transcriptional strand-bias

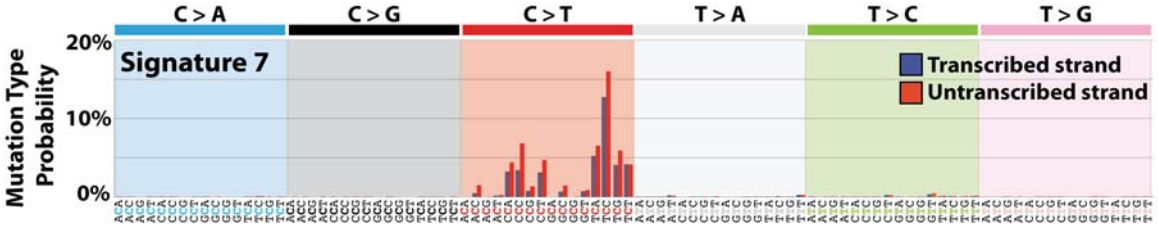
This appendix contains high-resolution figures for the nine consensus mutational signatures that exhibit transcriptional strand-bias. Each mutational signature is shown using the same figure format based on a 192 substitution classification incorporating the substitution type, the sequence context immediately 5' and 3' to the mutated base and whether the mutated base (in pyrimidine context) is on the transcribed or untranscribed strand. The panels for each of the six types of substitutions as well as the mutated bases are displayed in different colours. Mutations on the transcribed pyrimidine strand are displayed in blue while mutations on the untranscribed strand are displayed in red.

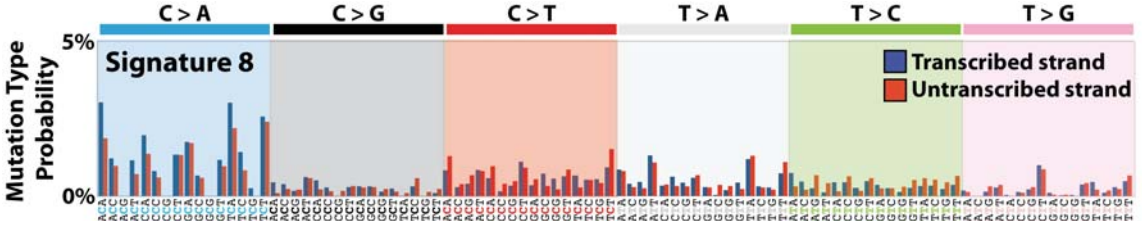

C > A
C > G
C > T
T > A
T > C
T > G
Signature 1B
Transcribed strand
Untranscribed strand
Mutation Type Probability
5%
0%

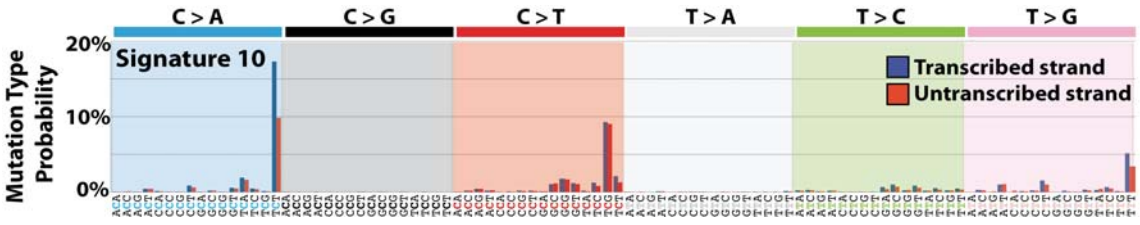

C > A
C > G
C > T
T > A
T > C
T > G
Signature 2 (exon only)
Transcribed strand
Untranscribed strand
Mutation Type Probability
15%
10%
5%
0%

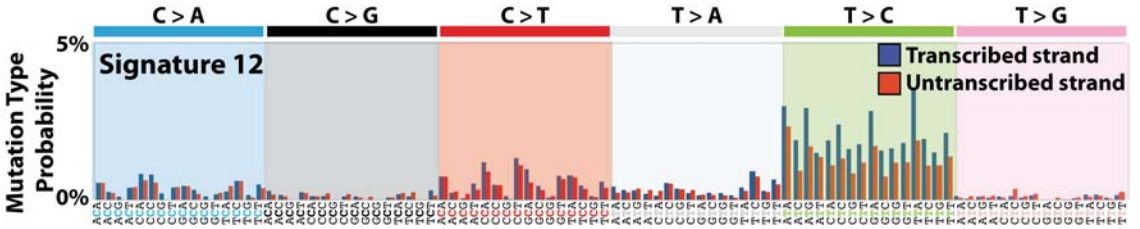

C > A
C > G
C > T
T > A
T > C
T > G
Signature 4
Transcribed strand
Untranscribed strand
Mutation Type Probability
5%
0%

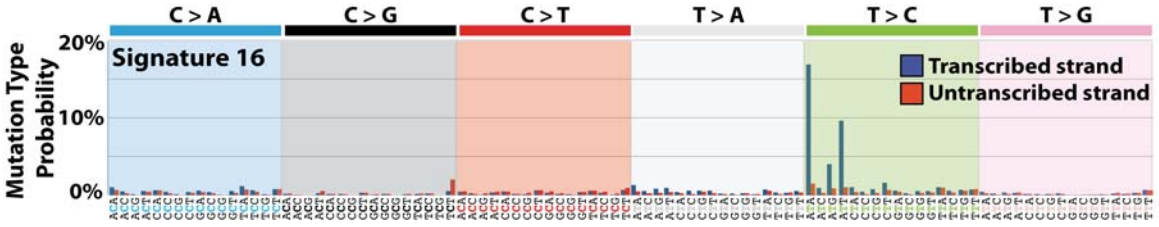

C > A
C > G
C > T
T > A
T > C
T > G
Signature 5
Transcribed strand
Untranscribed strand
Mutation Type Probability
5%
0%


C > A
C > G
C > T
T > A
T > C
T > G
Signature 7
Transcribed strand
Untranscribed strand
Mutation Type Probability
20%
10%
0%


C > A
C > G
C > T
T > A
T > C
T > G
Signature 8
Transcribed strand
Untranscribed strand
Mutation Type Probability
5%
0%


C > A
C > G
C > T
T > A
T > C
T > G
Signature 10
Mutation Type Probability
20%
10%
0%
Transcribed strand
Untranscribed strand

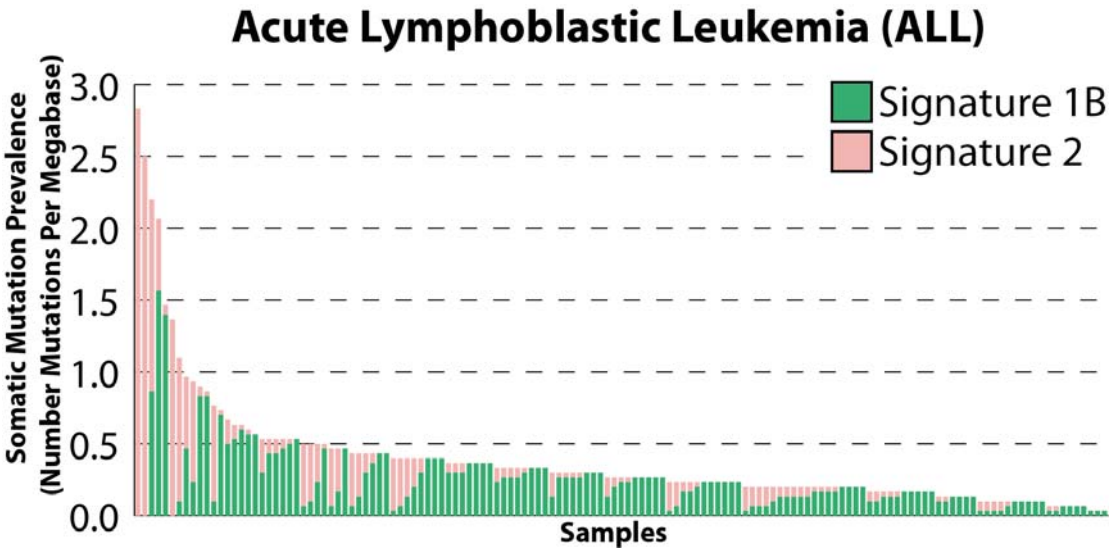

C > A
C > G
C > T
T > A
T > C
T > G
Signature 12
Mutation Type Probability
5%
0%
Transcribed strand
Untranscribed strand

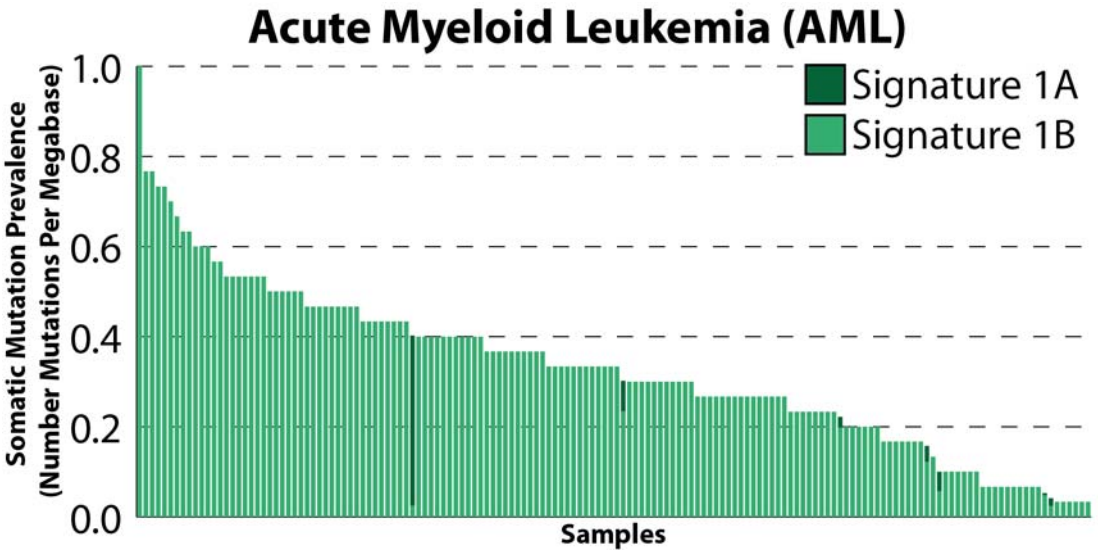

C > A
C > G
C > T
T > A
T > C
T > G
Signature 16
Mutation Type Probability
20%
10%
0%
Transcribed strand
Untranscribed strand

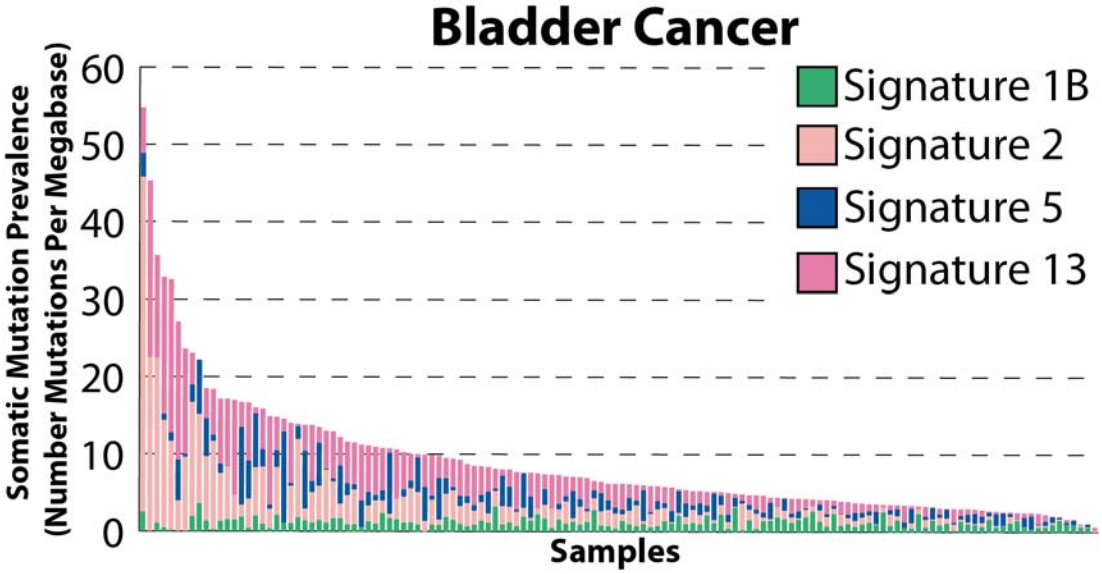
APPENDIX V: Contributions of mutational signatures in individual samples

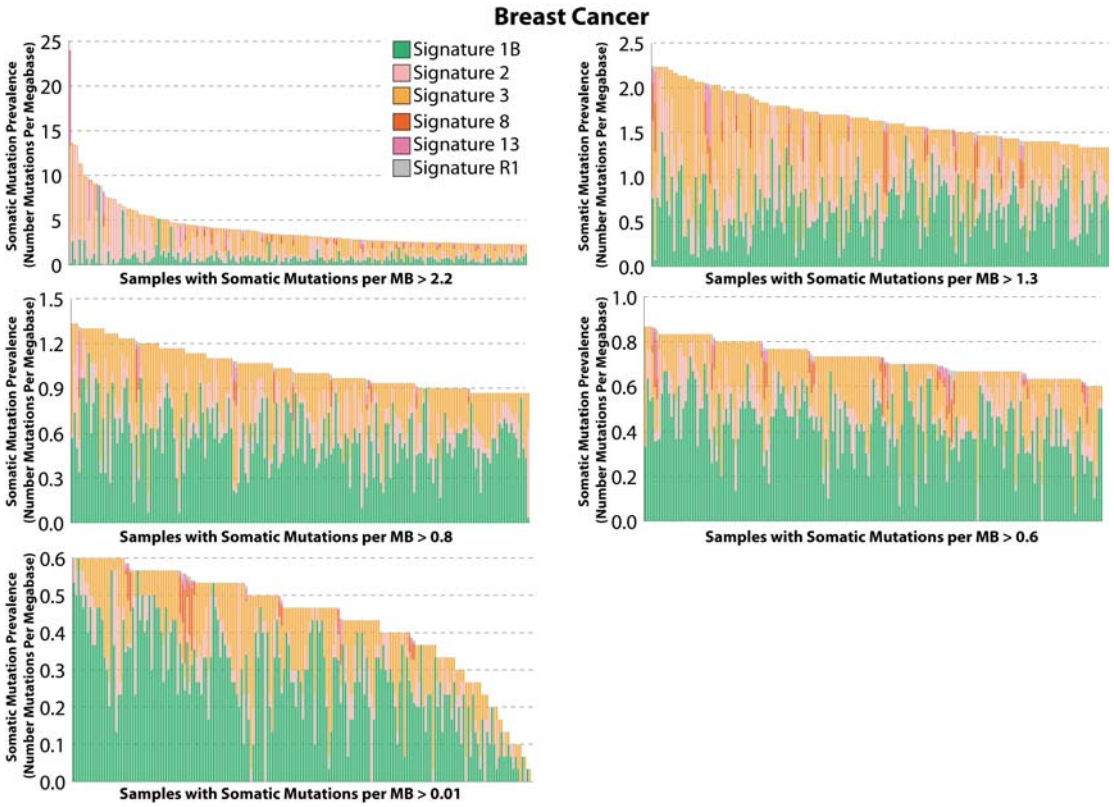
This appendix contains a high-resolution figure for each of the 30 examined cancer types (chapter 4). Each figure depicts all of the samples in a single cancer type and shows the contributions of the consensus mutational signatures (found in that cancer type) for each sample. All figures use the same format: samples are displayed on the horizontal axis, sorted in descending order based on the numbers of somatic mutations per megabase found in each sample, and the somatic mutation prevalence is displayed on the vertical axis. Mutational signatures are displayed in distinct colours, consistent in all figures. For clarity, several panels are provided (and clearly labelled) when the number of samples is too high or the somatic prevalence differs significantly between samples. Figures are displayed on individual pages, labelled to clearly show the names of the cancer types, and they are ordered alphabetically based on the names of these cancer types. In general, all samples are displayed in each cancer type and the two exceptions are denoted with an asterisk in the appropriate figures and listed below:

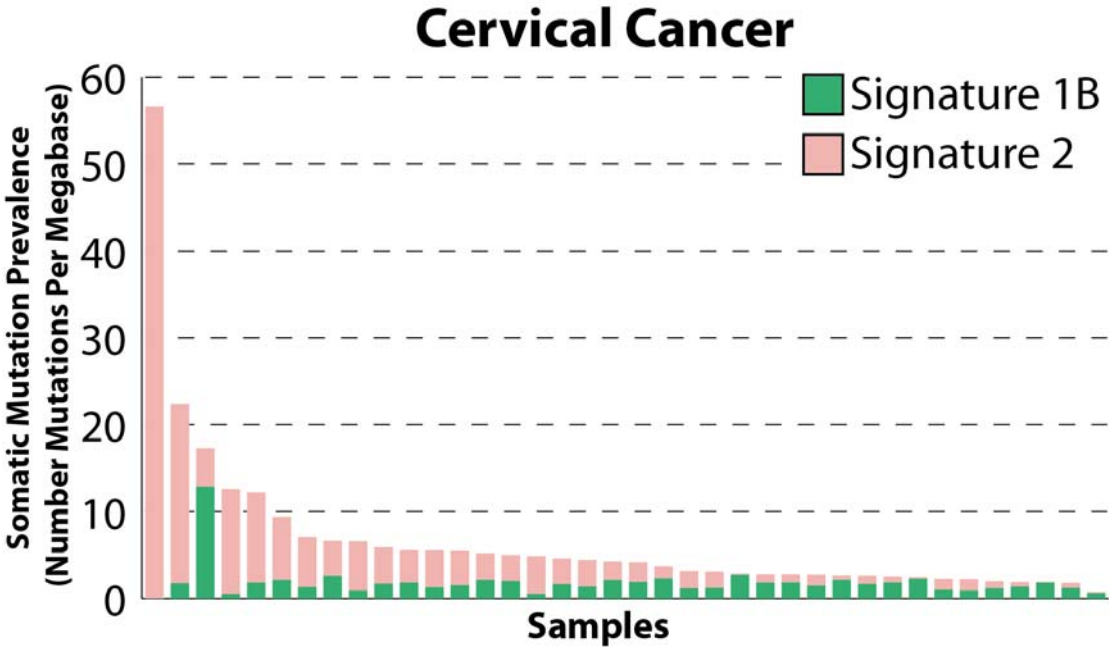
- For clarity, in glioma low grade, one hypermutator sample purely of Signature 14 (254 mutations per MB) is not displayed.
- In lung squamous, one hypermutator sample purely of Signature 7 (72 mutations per MB) is not displayed. Signature 7 is associated with exposure to ultraviolet light, an unlikely carcinogen for lung cancer. As such, this TCGA sample is most likely either a melanoma metastasis or a misannotated sample. Thus, the association between Signature 7 and lung squamous has not been discussed in chapter 4 and this association has not been displayed in Figure 4.9.

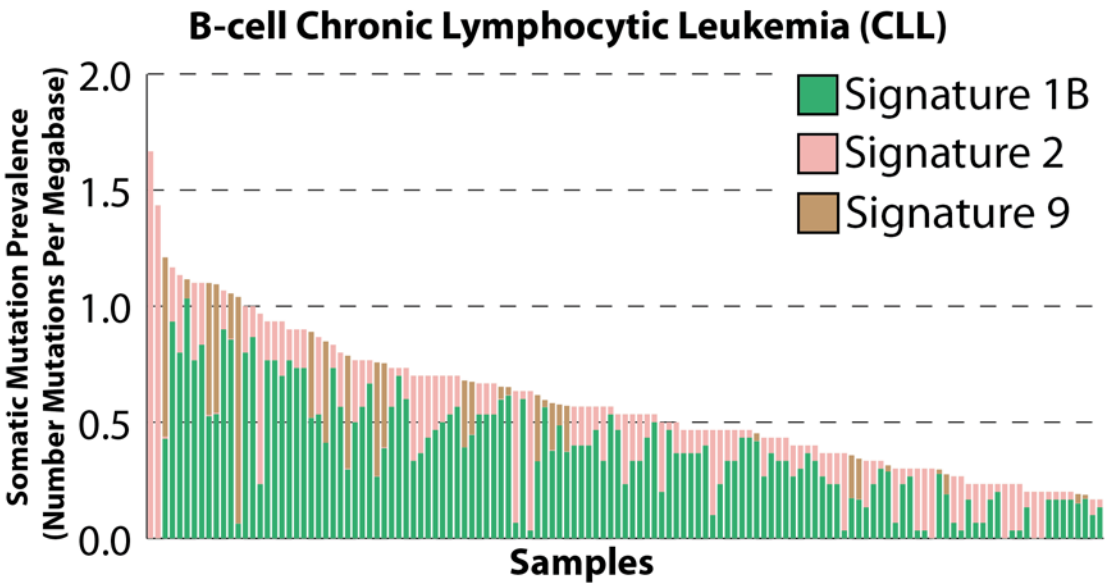

Acute Lymphoblastic Leukemia (ALL)
Somatic Mutation Prevalence
(Number Mutations Per Megabase)
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Signature 1B
Signature 2
Samples

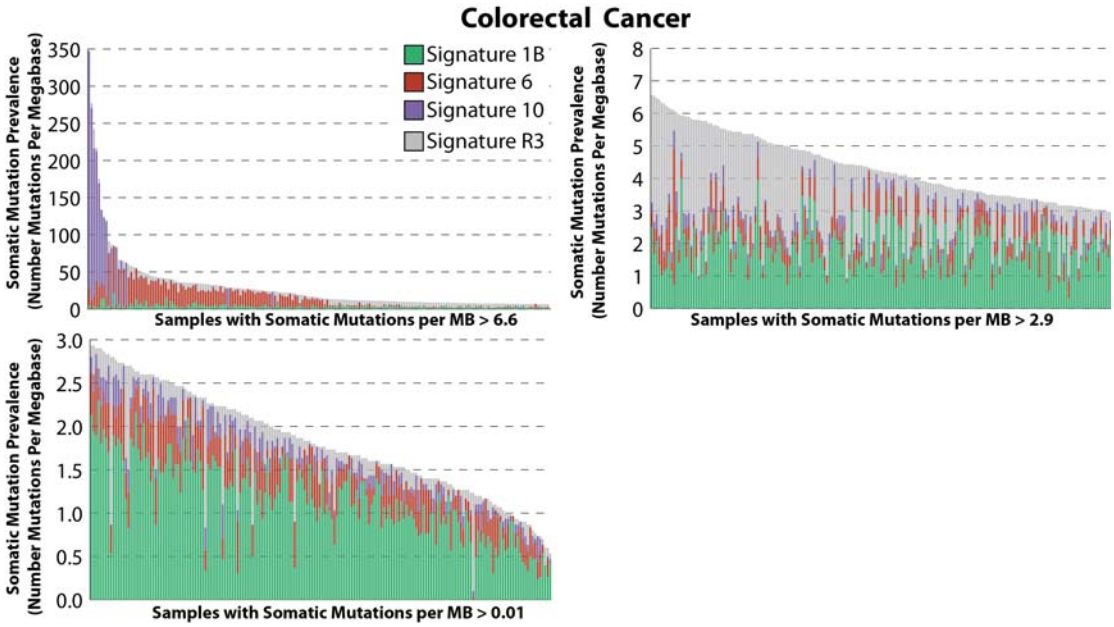

Acute Myeloid Leukemia (AML)
Somatic Mutation Prevalence
(Number Mutations Per Megabase)
1.0
0.8
0.6
0.4
0.2
0.0
Signature 1A
Signature 1B
Samples

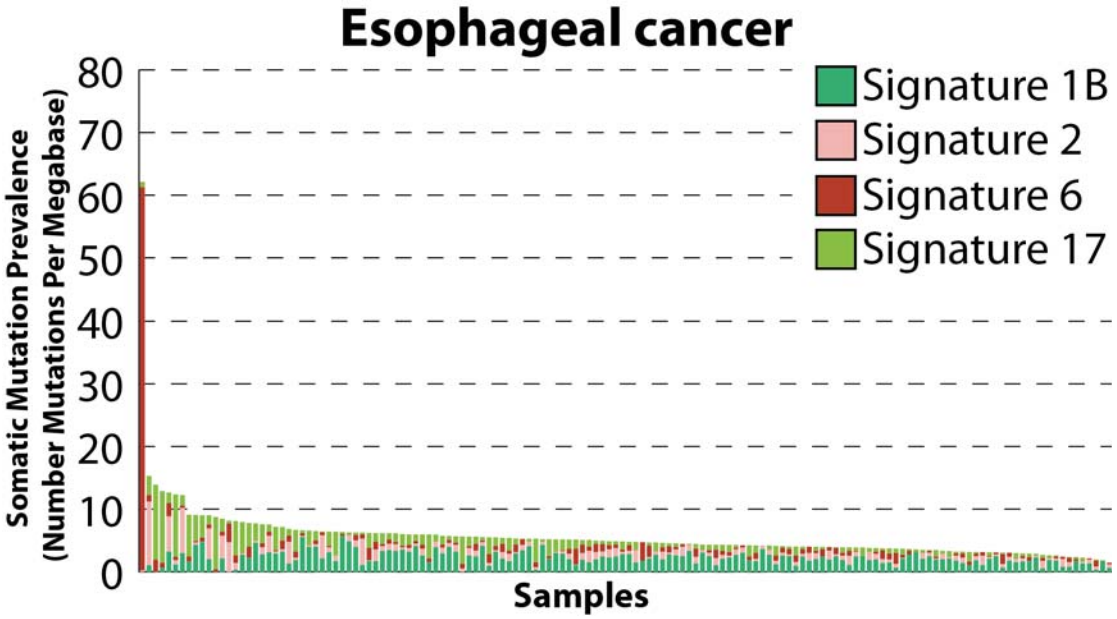

Bladder Cancer
Somatic Mutation Prevalence
(Number Mutations Per Megabase)
60
50
40
30
20
10
0
Samples
Signature 1B
Signature 2
Signature 5
Signature 13

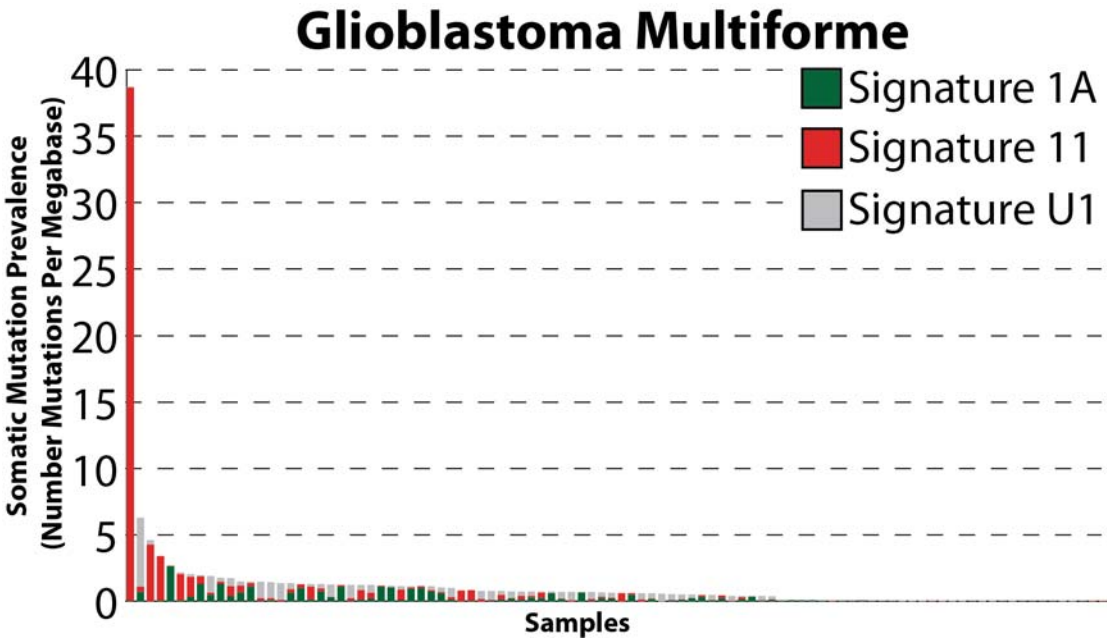

Breast Cancer
Signature 1B
Signature 2
Signature 3
Signature 8
Signature 13
Signature R1
Somatic Mutation Prevalence (Number Mutations Per Megabase)
Samples with Somatic Mutations per MB > 2.2
Samples with Somatic Mutations per MB > 1.3
Samples with Somatic Mutations per MB > 0.8
Samples with Somatic Mutations per MB > 0.6
Samples with Somatic Mutations per MB > 0.01

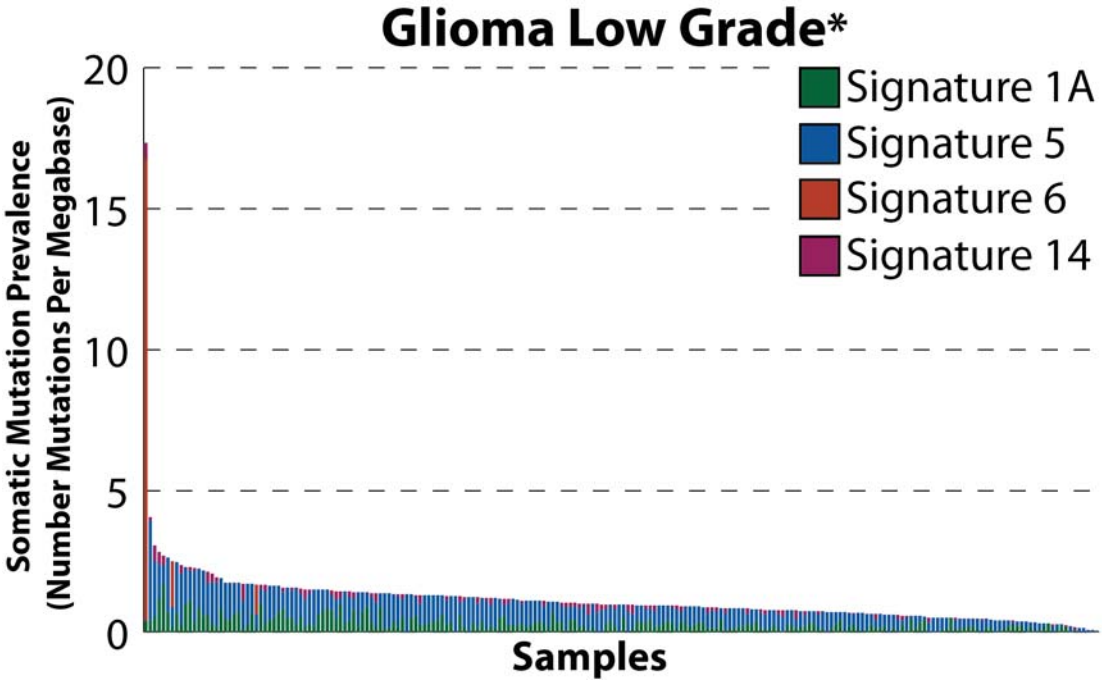

Cervical Cancer
Signature 1B
Signature 2
Somatic Mutation Prevalence
(Number Mutations Per Megabase)
60
50
40
30
20
10
0
Samples

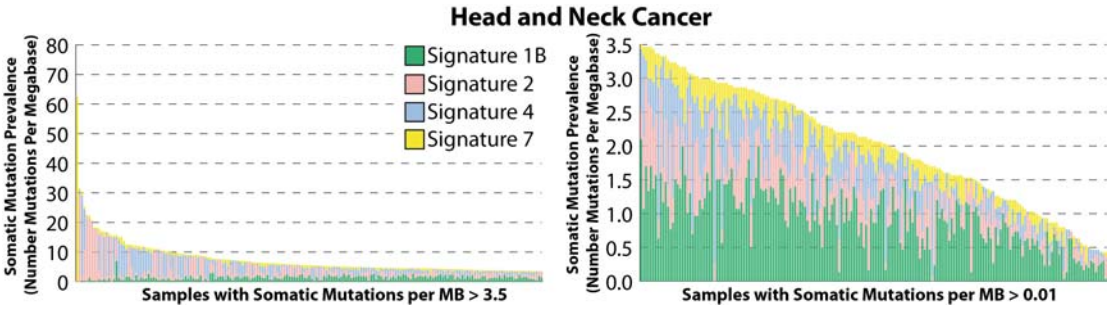

B-cell Chronic Lymphocytic Leukemia (CLL)
Somatic Mutation Prevalence
(Number Mutations Per Megabase)
2.0
1.5
1.0
0.5
0.0
Signature 1B
Signature 2
Signature 9
Samples

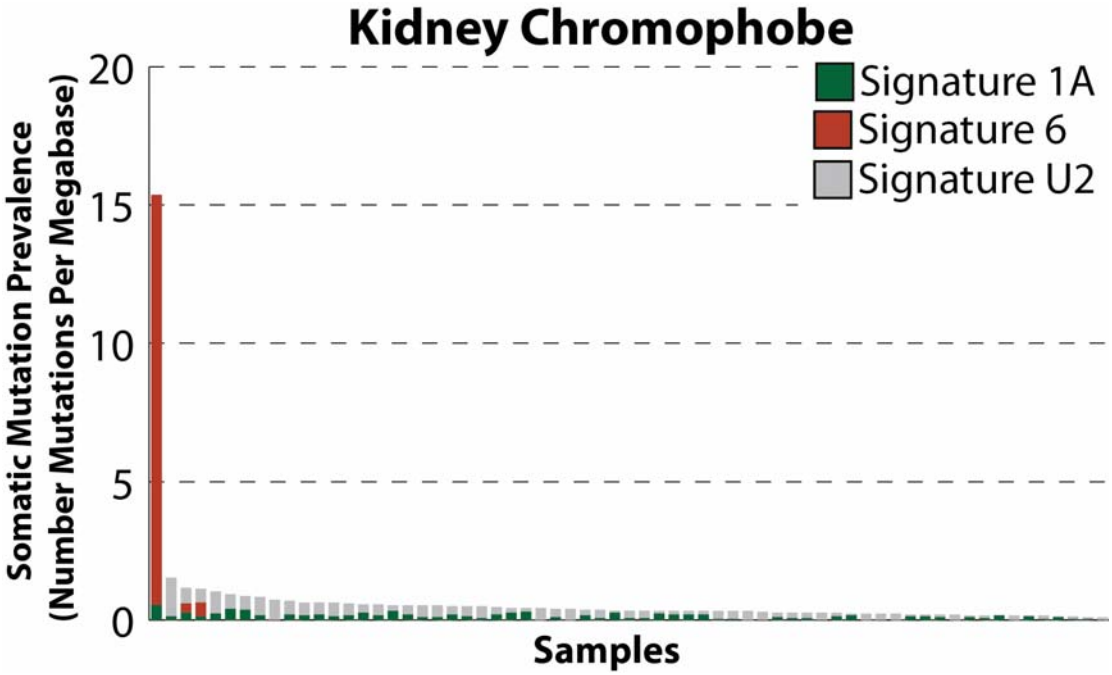

Colorectal Cancer
Signature 1B
Signature 6
Signature 10
Signature R3
Somatic Mutation Prevalence (Number Mutations Per Megabase)
Samples with Somatic Mutations per MB > 6.6
Samples with Somatic Mutations per MB > 2.9
Samples with Somatic Mutations per MB > 0.01

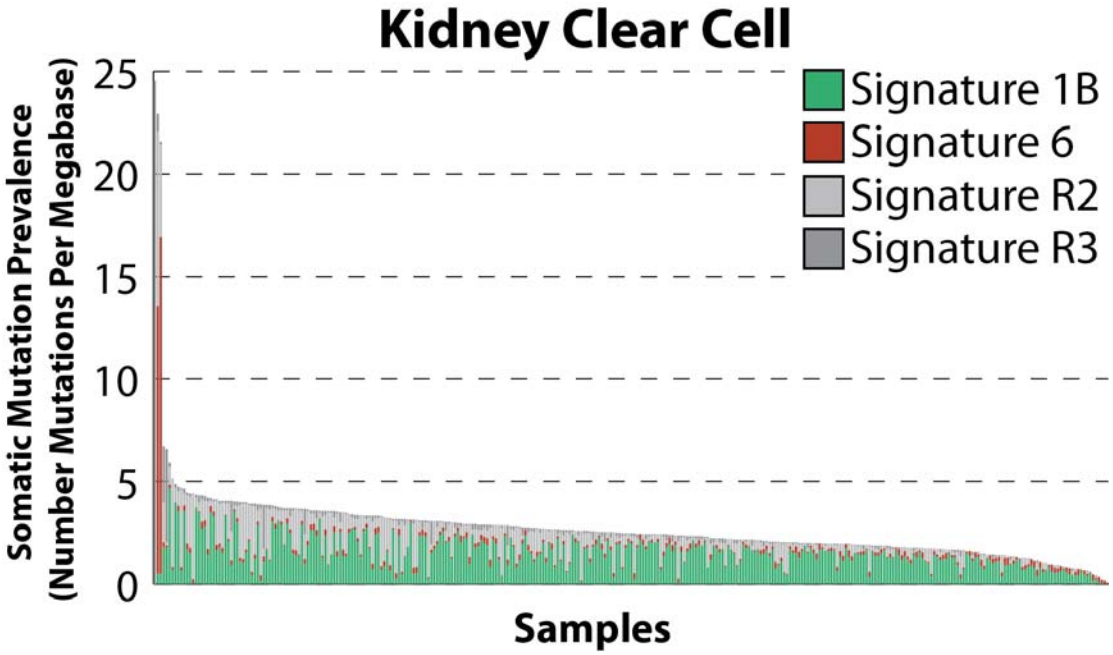

Esophageal cancer
Somatic Mutation Prevalence
(Number Mutations Per Megabase)
80
70
60
50
40
30
20
10
0
Samples
Signature 1B
Signature 2
Signature 6
Signature 17

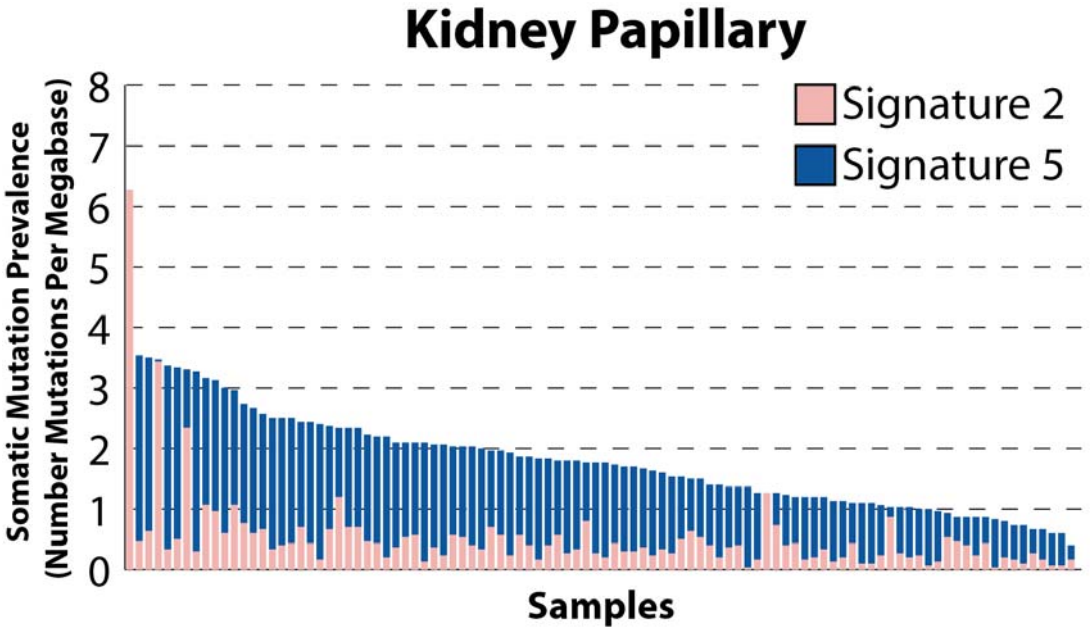

Glioblastoma Multiforme
Somatic Mutation Prevalence
(Number Mutations Per Megabase)
40
35
30
25
20
15
10
5
0
Samples
Signature 1A
Signature 11
Signature U1

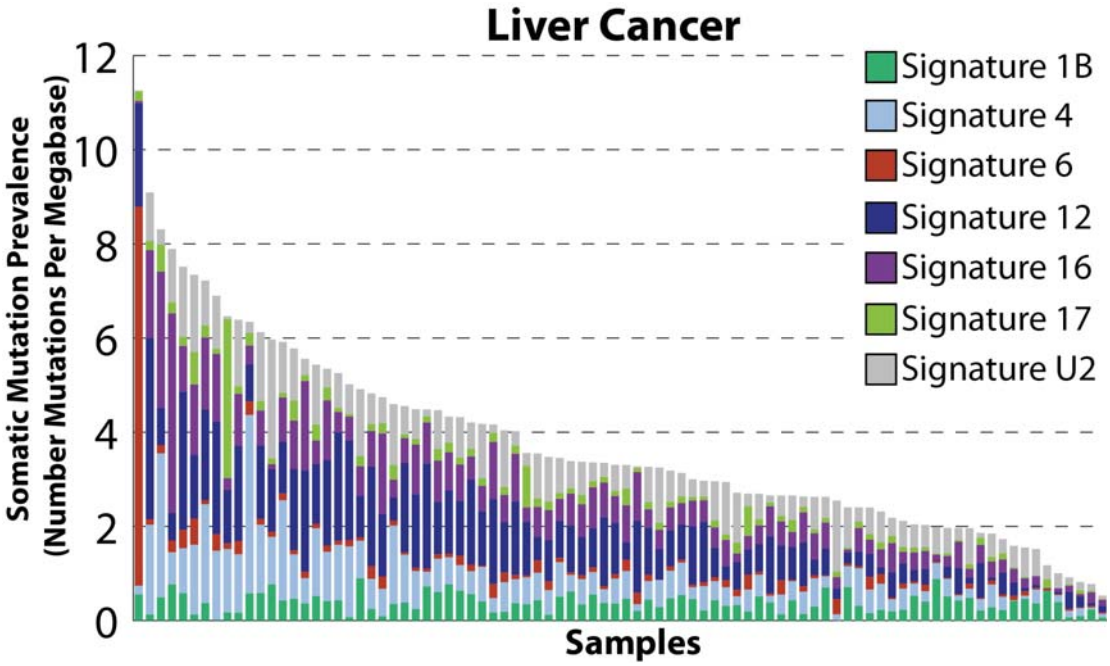

Glioma Low Grade*
Somatic Mutation Prevalence
(Number Mutations Per Megabase)
20
15
10
5
0
Samples
Signature 1A
Signature 5
Signature 6
Signature 14

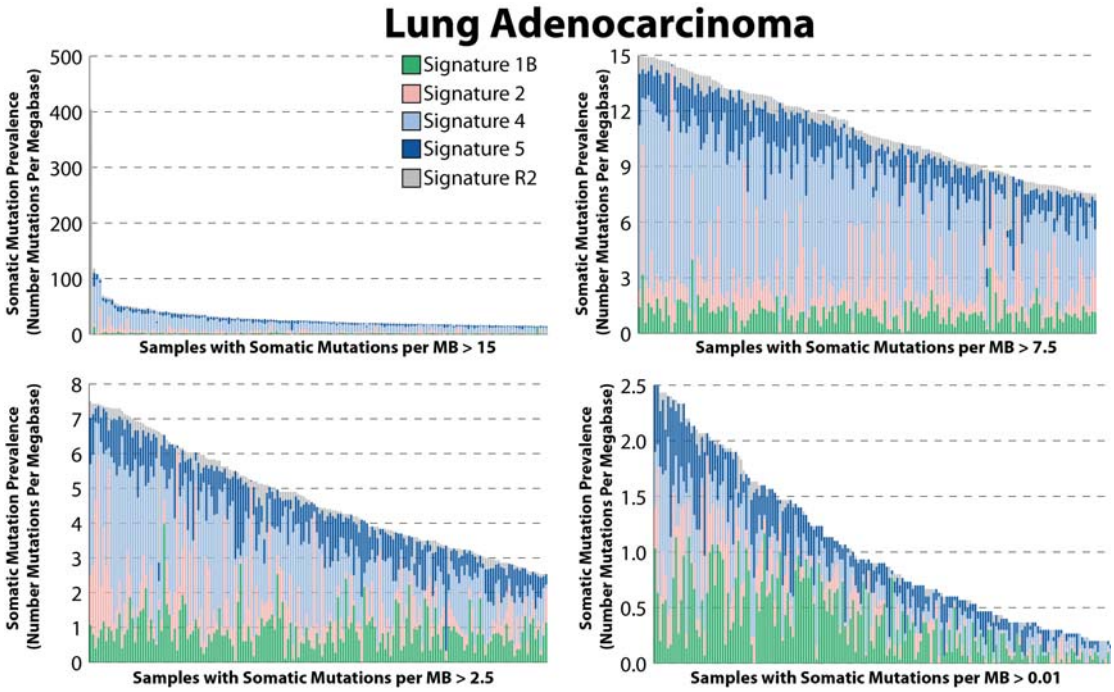

Head and Neck Cancer
Somatic Mutation Prevalence (Number Mutations Per Megabase)
80
70
60
50
40
30
20
10
0
Signature 1B
Signature 2
Signature 4
Signature 7
Samples with Somatic Mutations per MB > 3.5
Somatic Mutation Prevalence (Number Mutations Per Megabase)
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Samples with Somatic Mutations per MB > 0.01

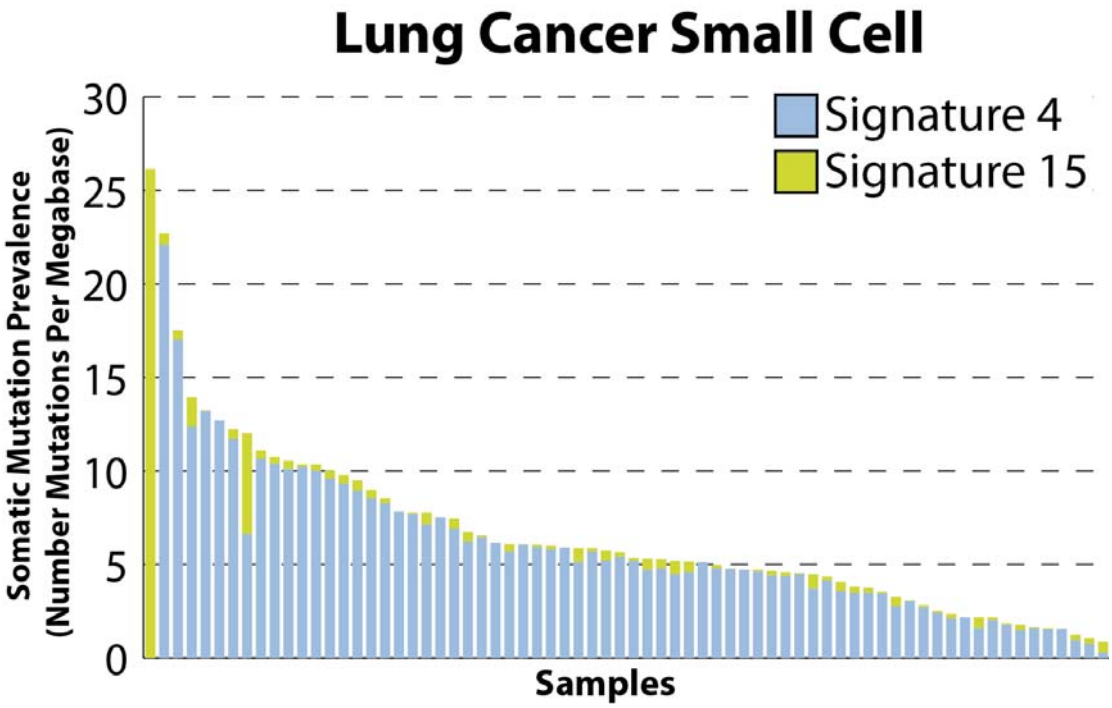

Kidney Chromophobe
Somatic Mutation Prevalence
(Number Mutations Per Megabase)
20
15
10
5
0
Samples
Signature 1A
Signature 6
Signature U2

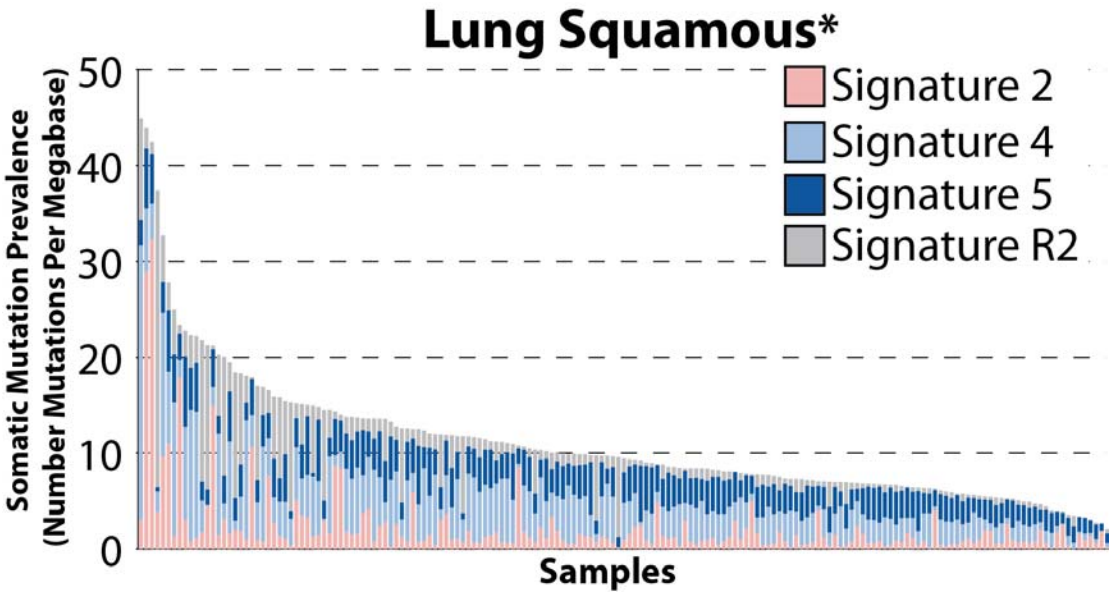

Kidney Clear Cell
Somatic Mutation Prevalence
(Number Mutations Per Megabase)
25
20
15
10
5
0
Signature 1B
Signature 6
Signature R2
Signature R3
Samples

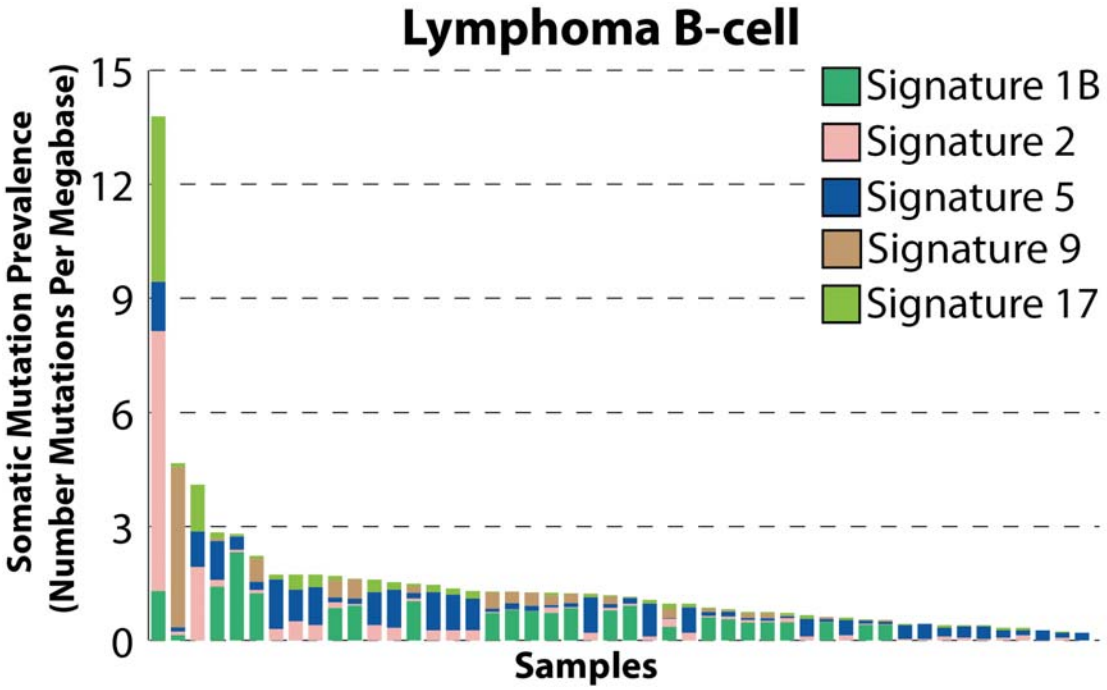

Kidney Papillary
Somatic Mutation Prevalence
(Number Mutations Per Megabase)
8
7
6
5
4
3
2
1
0
Signature 2
Signature 5
Samples

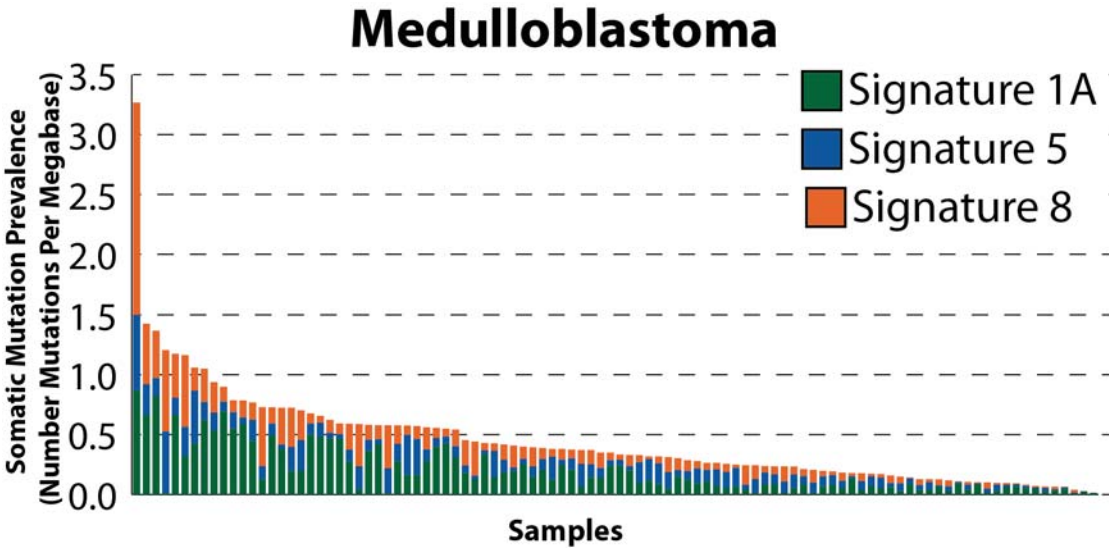

Liver Cancer
Somatic Mutation Prevalence
(Number Mutations Per Megabase)
12
10
8
6
4
2
0
Samples
Signature 1B
Signature 4
Signature 6
Signature 12
Signature 16
Signature 17
Signature U2

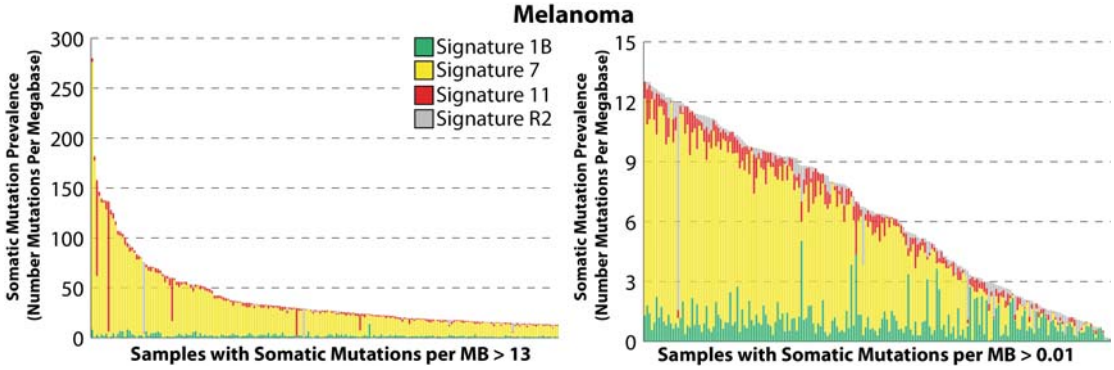

Lung Adenocarcinoma
Signature 1B
Signature 2
Signature 4
Signature 5
Signature R2
Somatic Mutation Prevalence (Number Mutations Per Megabase)
Samples with Somatic Mutations per MB > 15
Samples with Somatic Mutations per MB > 7.5
Samples with Somatic Mutations per MB > 2.5
Samples with Somatic Mutations per MB > 0.01

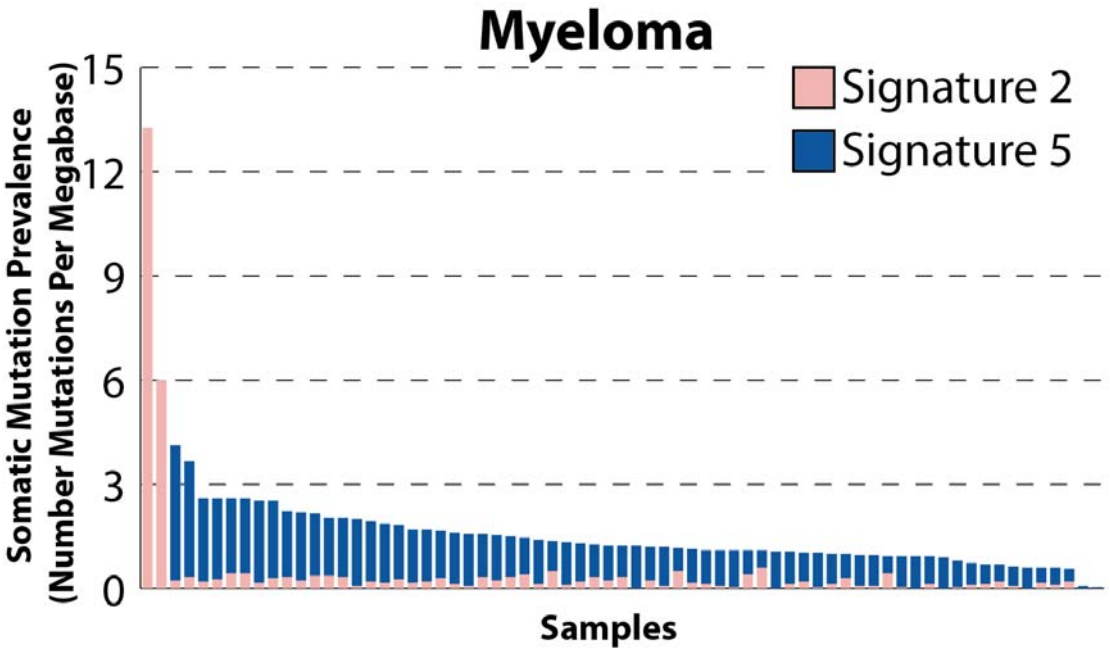

Lung Cancer Small Cell
Signature 4
Signature 15
Somatic Mutation Prevalence
(Number Mutations Per Megabase)
30
25
20
15
10
5
0
Samples

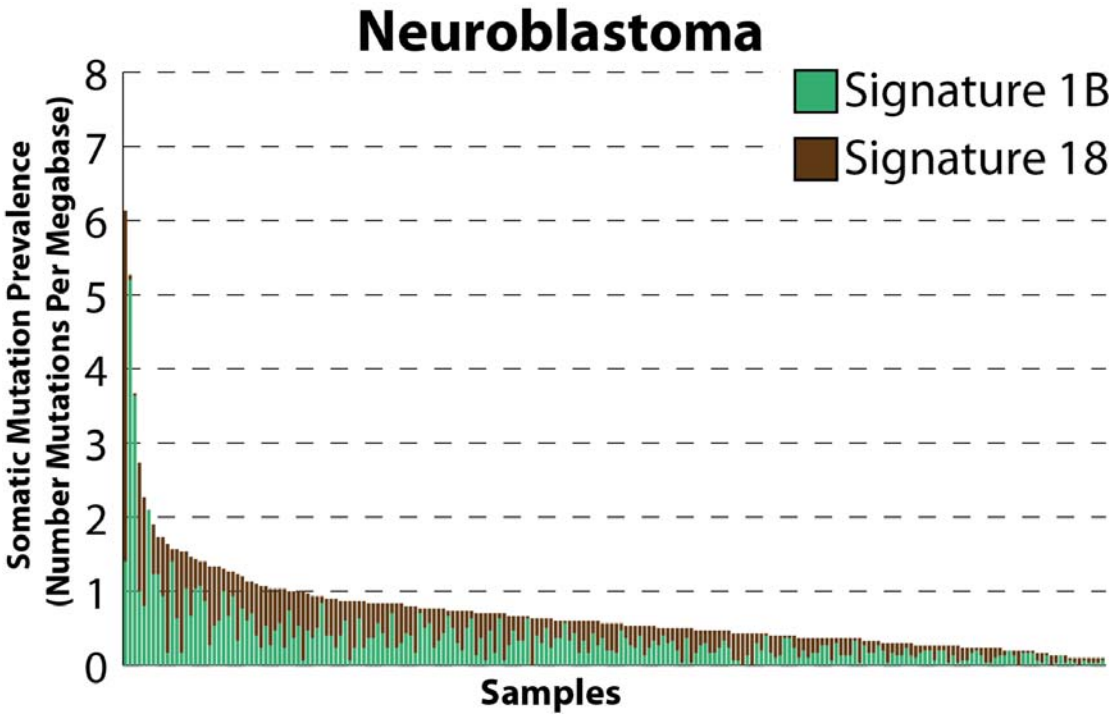

Lung Squamous*
Somatic Mutation Prevalence
(Number Mutations Per Megabase)
50
40
30
20
10
0
Samples
Signature 2
Signature 4
Signature 5
Signature R2


Lymphoma B-cell
Somatic Mutation Prevalence
(Number Mutations Per Megabase)
15
12
9
6
3
0
Samples
Signature 1B
Signature 2
Signature 5
Signature 9
Signature 17

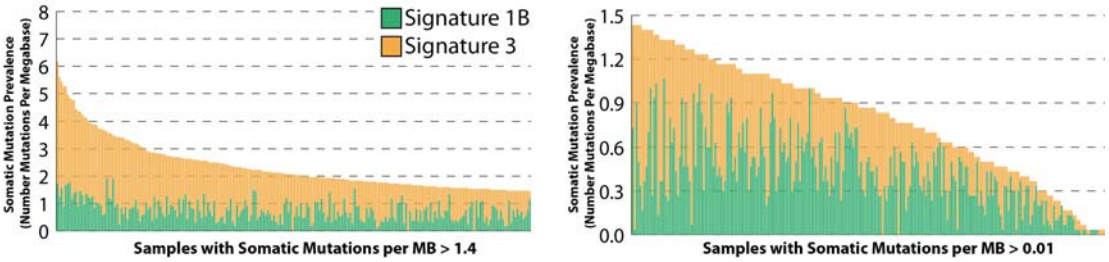

Medulloblastoma
Somatic Mutation Prevalence
(Number Mutations Per Megabase)
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Signature 1A
Signature 5
Signature 8
Samples

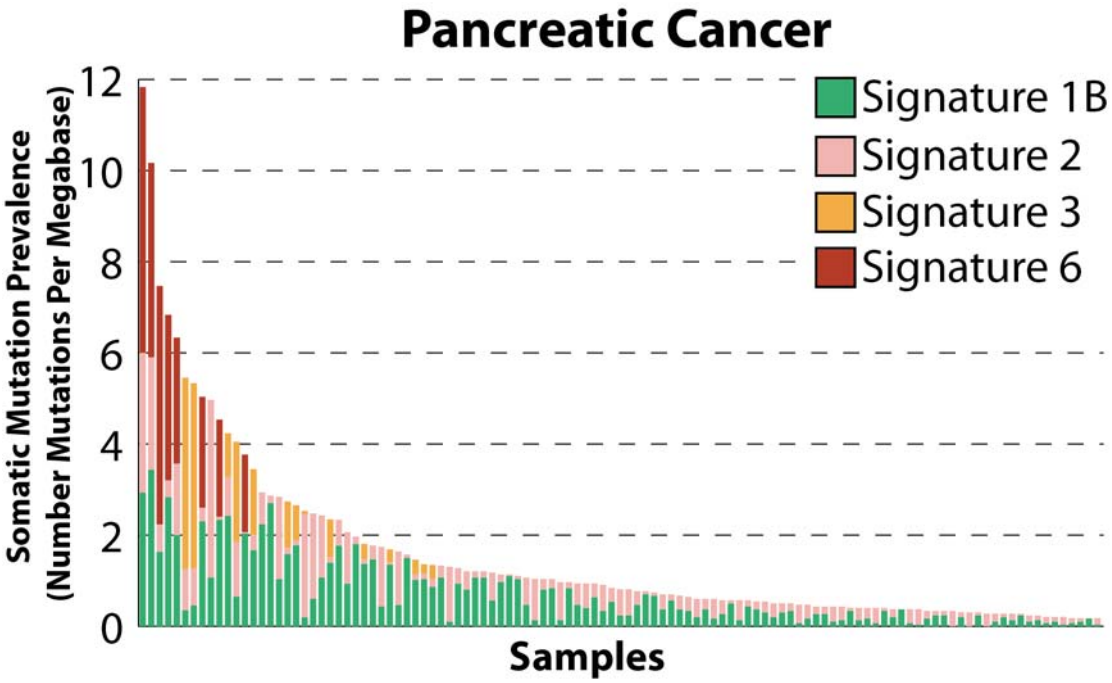

Melanoma
Signature 1B
Signature 7
Signature 11
Signature R2
Somatic Mutation Prevalence (Number Mutations Per Megabase)
300
250
200
150
100
50
0
Samples with Somatic Mutations per MB > 13
Somatic Mutation Prevalence (Number Mutations Per Megabase)
15
12
9
6
3
0
Samples with Somatic Mutations per MB > 0.01

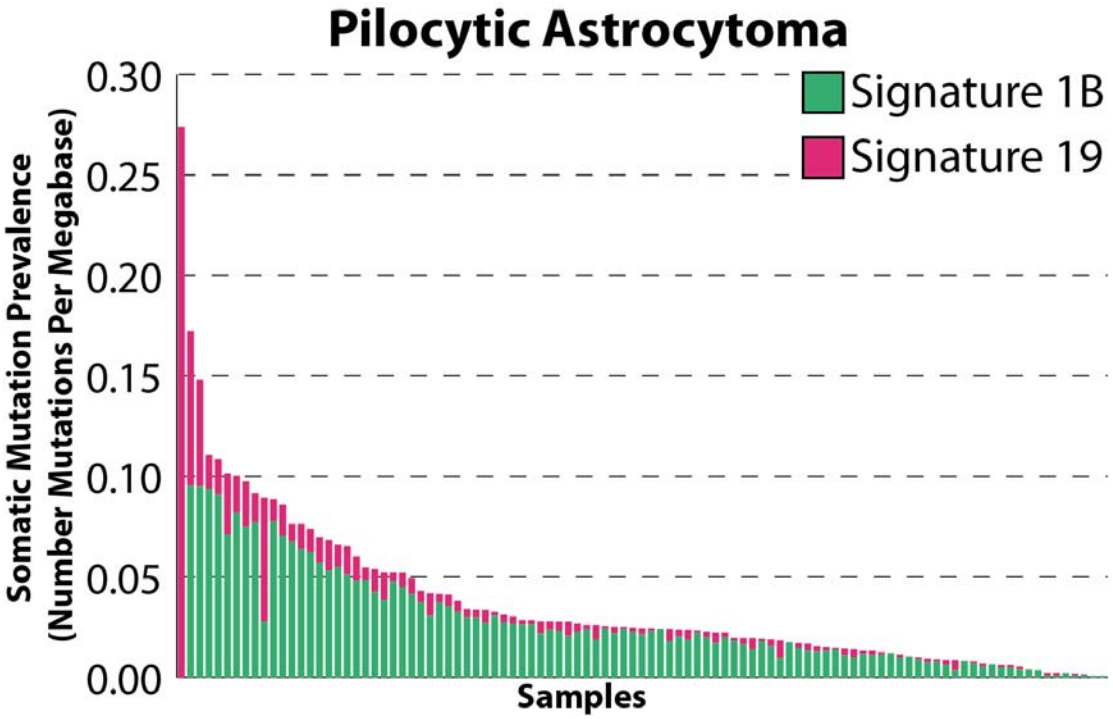

Myeloma
Signature 2
Signature 5
Somatic Mutation Prevalence
(Number Mutations Per Megabase)
15
12
9
6
3
0
Samples

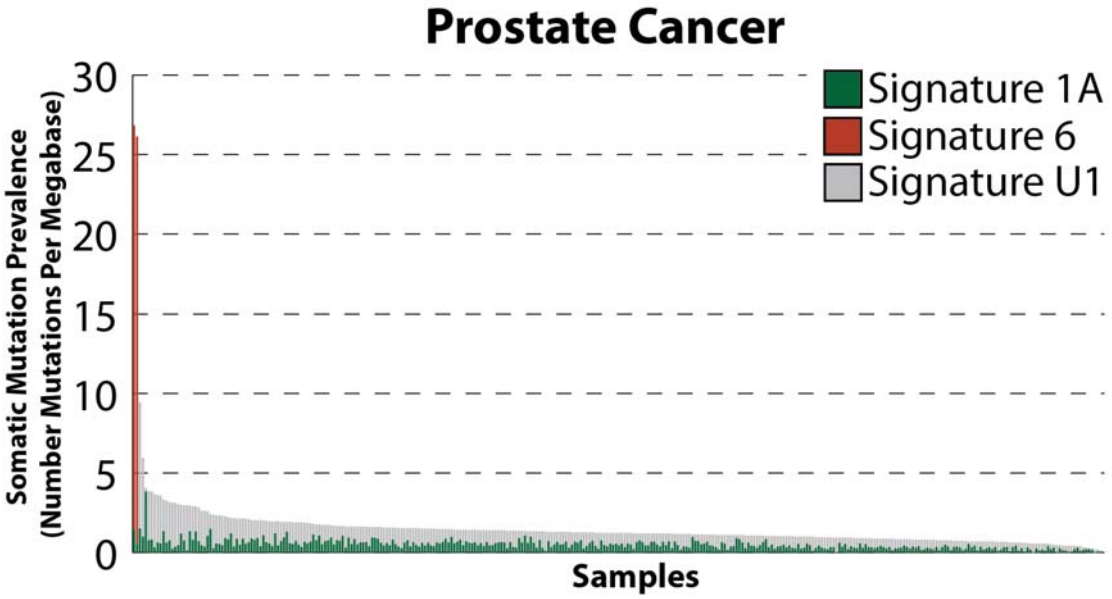

Neuroblastoma
Signature 1B
Signature 18
Somatic Mutation Prevalence
(Number Mutations Per Megabase)
8
7
6
5
4
3
2
1
0
Samples

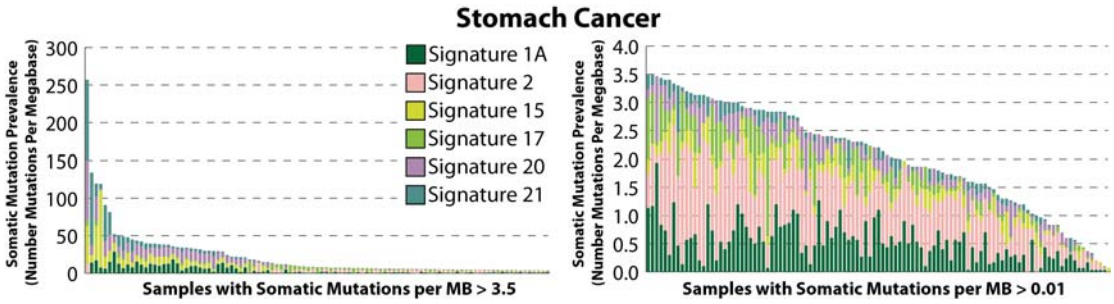
Ovarian Cancer

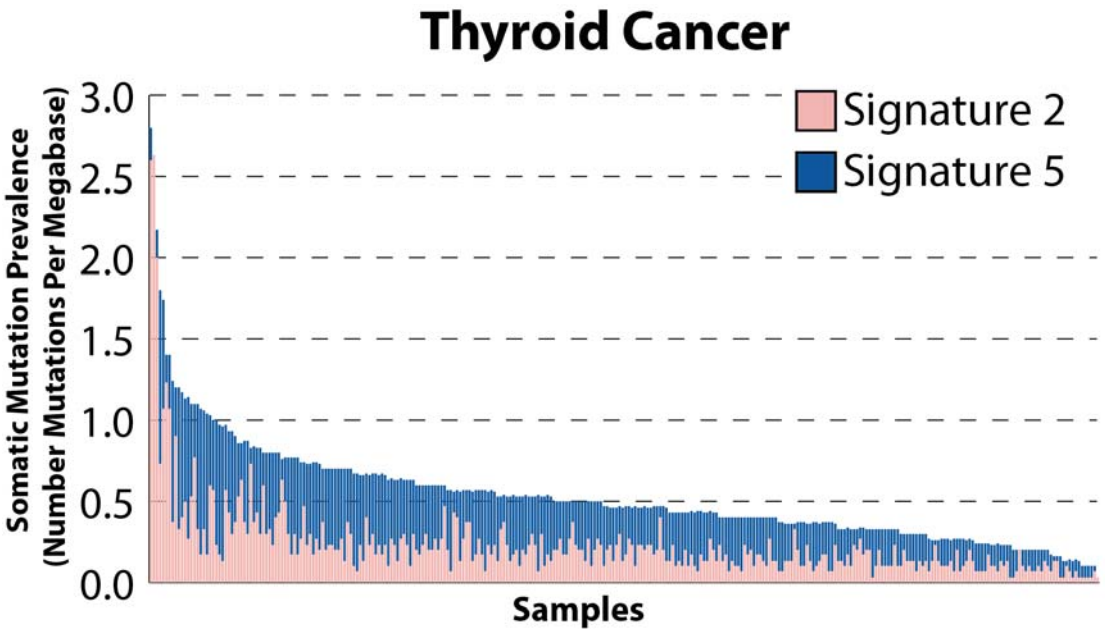


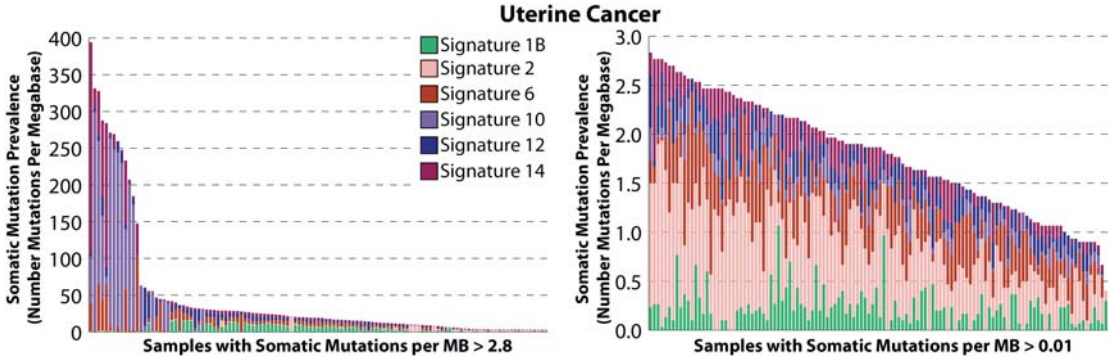

Pancreatic Cancer
Somatic Mutation Prevalence
(Number Mutations Per Megabase)
12
10
8
6
4
2
0
Samples
Signature 1B
Signature 2
Signature 3
Signature 6


Pilocytic Astrocytoma
Signature 1B
Signature 19
Somatic Mutation Prevalence
(Number Mutations Per Megabase)
0.30
0.25
0.20
0.15
0.10
0.05
0.00
Samples


Prostate Cancer
Somatic Mutation Prevalence
(Number Mutations Per Megabase)
30
25
20
15
10
5
0
Signature 1A
Signature 6
Signature U1
Samples

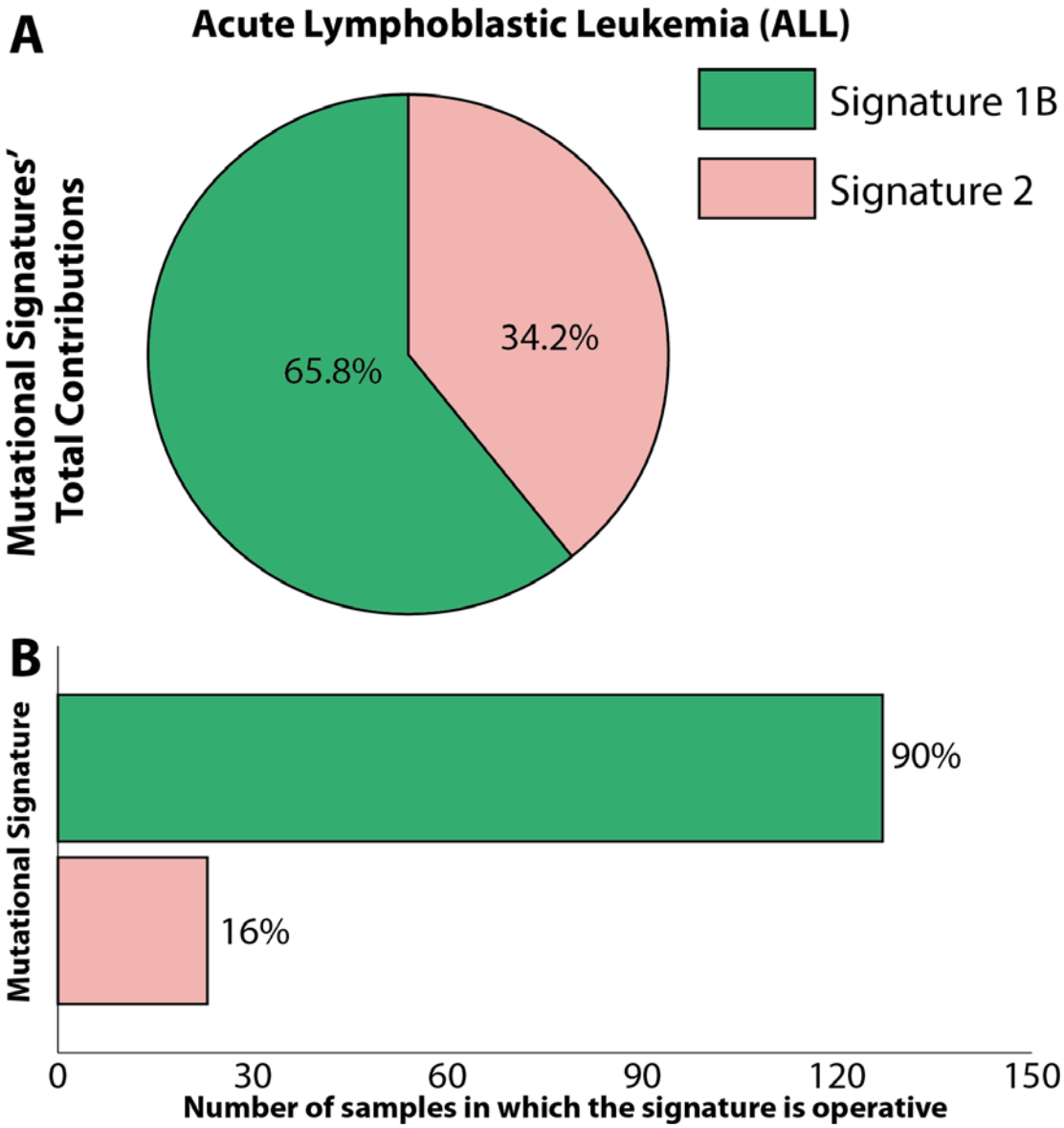

Stomach Cancer
Somatic Mutation Prevalence (Number Mutations Per Megabase)
300
250
200
150
100
50
0
Signature 1A
Signature 2
Signature 15
Signature 17
Signature 20
Signature 21
Samples with Somatic Mutations per MB > 3.5
Somatic Mutation Prevalence (Number Mutations Per Megabase)
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Samples with Somatic Mutations per MB > 0.01

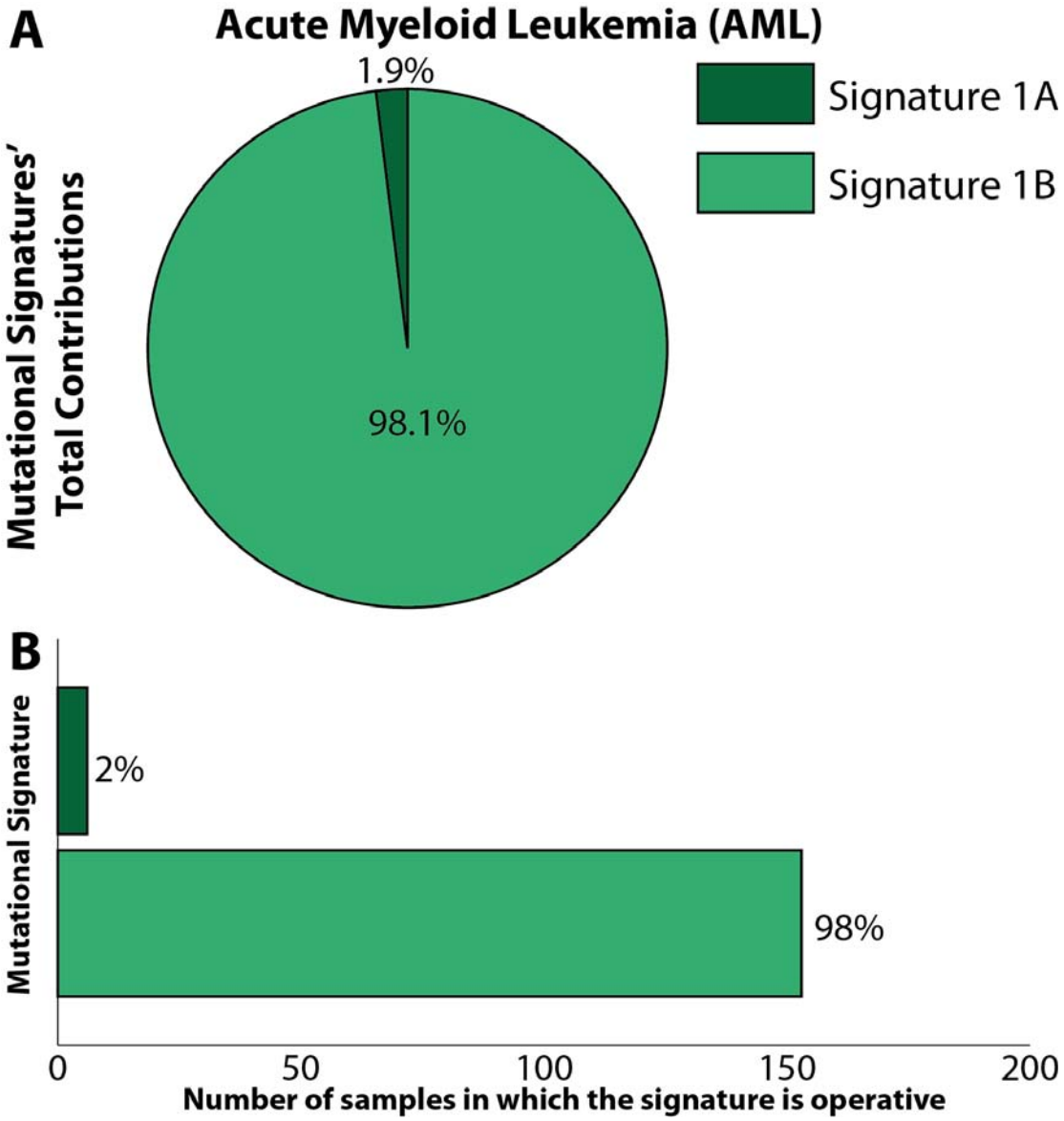

Thyroid Cancer
Somatic Mutation Prevalence
(Number Mutations Per Megabase)
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Signature 2
Signature 5
Samples

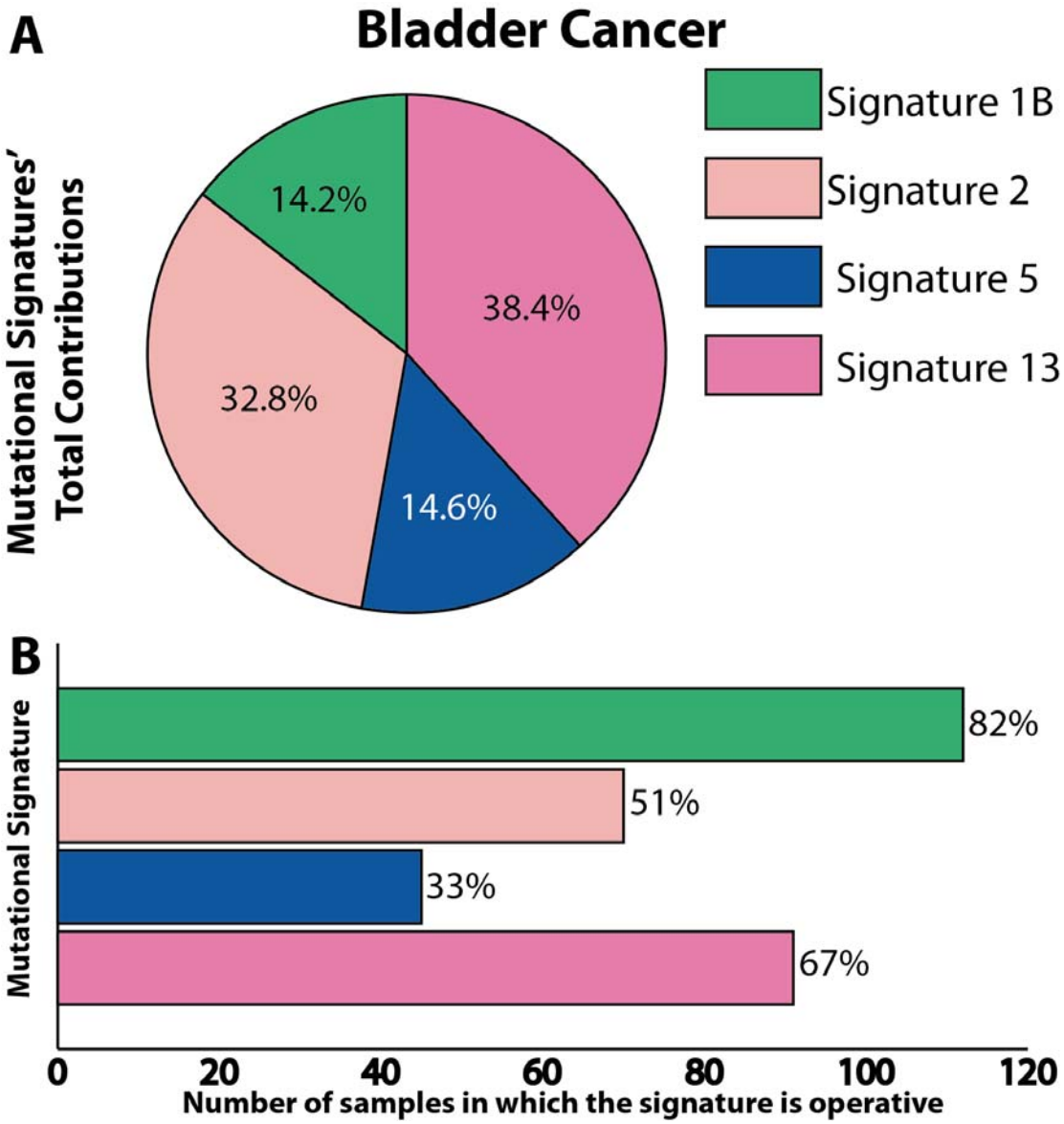

Uterine Cancer
Somatic Mutation Prevalence (Number Mutations Per Megabase)
400
350
300
250
200
150
100
50
0
Signature 1B
Signature 2
Signature 6
Signature 10
Signature 12
Signature 14
Samples with Somatic Mutations per MB > 2.8
Somatic Mutation Prevalence (Number Mutations Per Megabase)
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Samples with Somatic Mutations per MB > 0.01

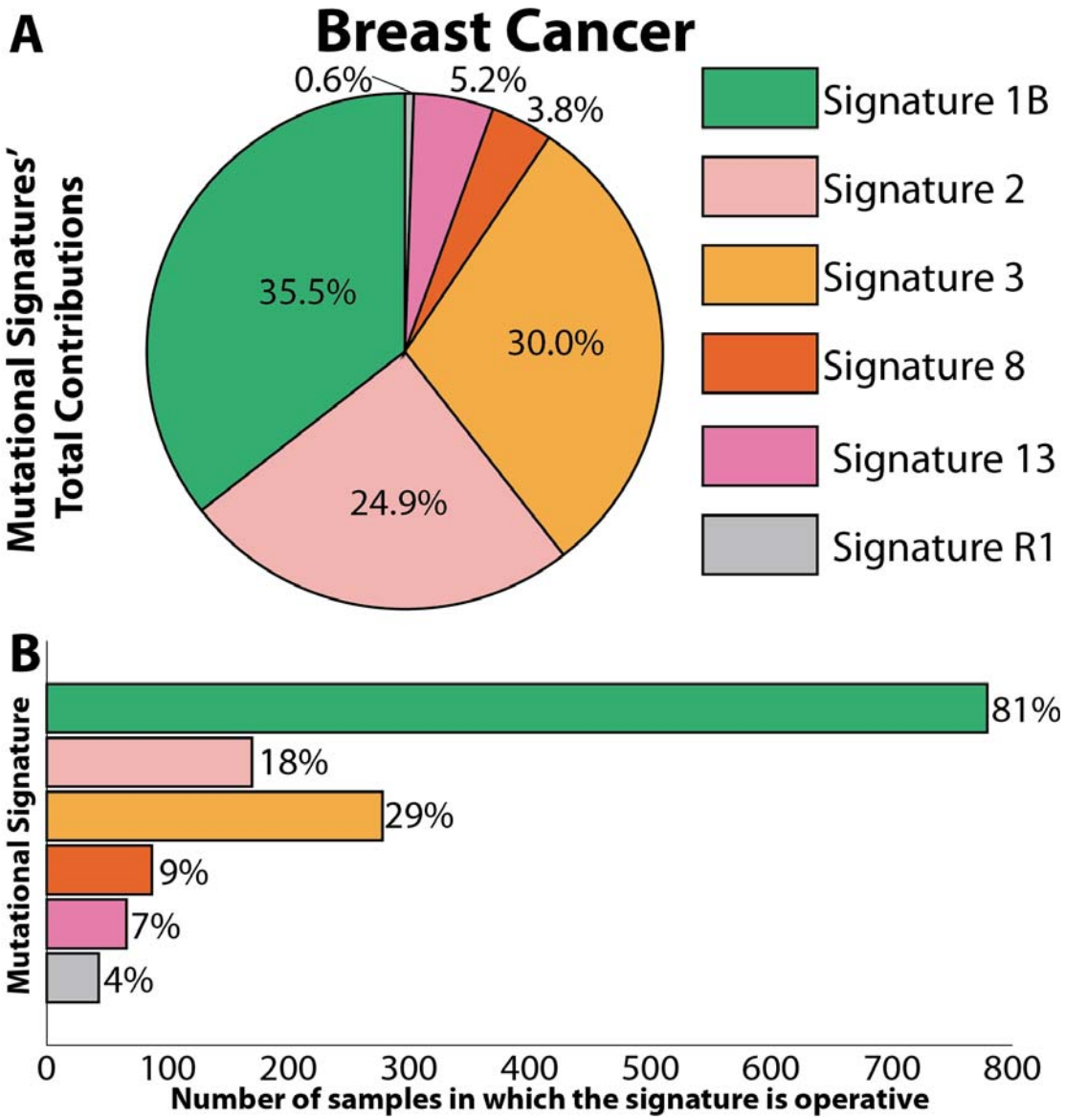
APPENDIX VI: Summary of signatures' contributions in cancer types

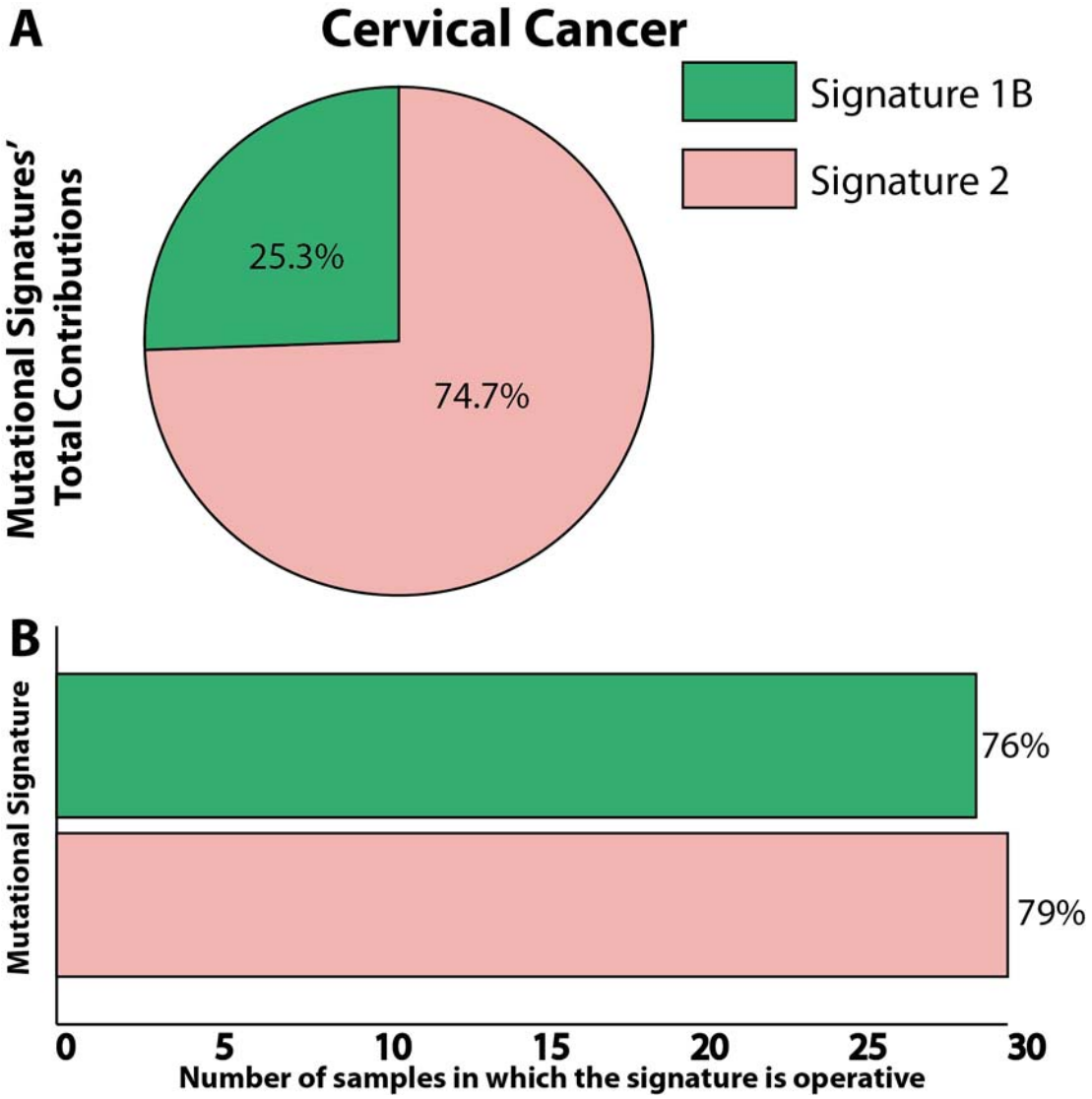
This appendix contains a high-resolution figure for each of the 30 examined cancer types (chapter 4). Each figure depicts a single cancer type and provides a summary of the contributions of the mutational signatures found in this cancer type. All figures have two panels: panel ***A*** depicting the percentage of total mutations contributed by each of the operative mutational signatures in that cancer type and panel ***B*** depicting the percentage and number of samples in which each mutational signature contributes significant number of somatic mutations. For most signatures, significant number of mutations in a sample is defined as more than 100 substitutions or more than 25% of all mutations in that sample. Mutational signatures are displayed in distinct colours, consistent in both panels of each figure as well as in all figures in Appendices V and VI. Figures are displayed on individual pages, labelled to clearly show the names of the cancer types, and they are ordered alphabetically based on the names of these cancer types. In general, all samples are included in the summary of each cancer type. The only exception (denoted with an asterisk in the appropriate figure) is one lung squamous hypermutator sample purely of Signature 7 (72 mutations per MB). Signature 7 is associated with exposure to ultraviolet light, an unlikely carcinogen for lung cancer. As such, this TCGA sample is most likely either a melanoma metastasis or a misannotated sample. Thus, the association between Signature 7 and lung squamous has not been discussed in chapter 4 and this association has not been displayed in Figure 4.9.

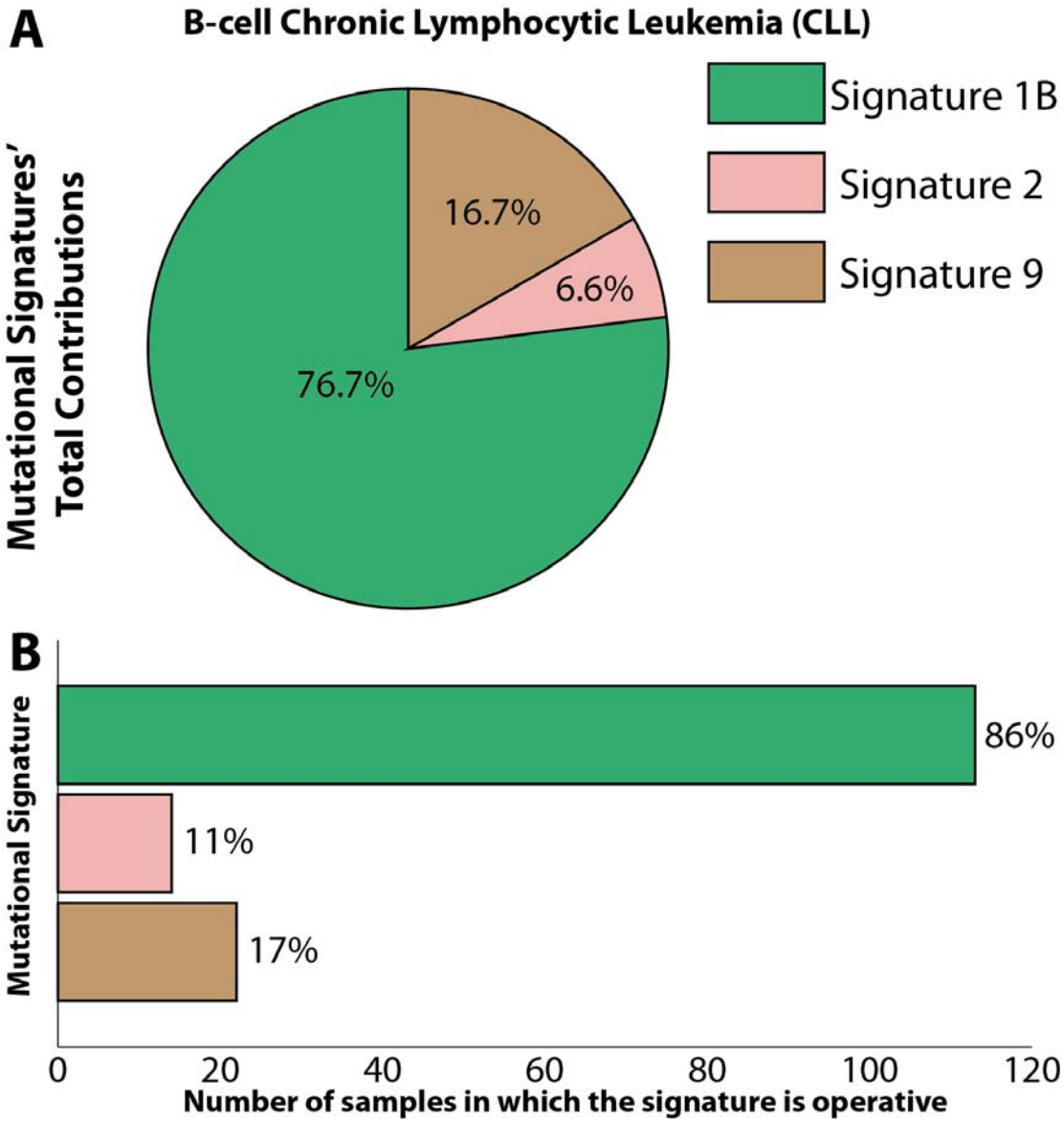

A
Acute Lymphoblastic Leukemia (ALL)
Mutational Signatures' Total Contributions
Signature 1B
Signature 2
65.8%
34.2%
B
Mutational Signature
90%
16%
0
30
60
90
120
150
Number of samples in which the signature is operative

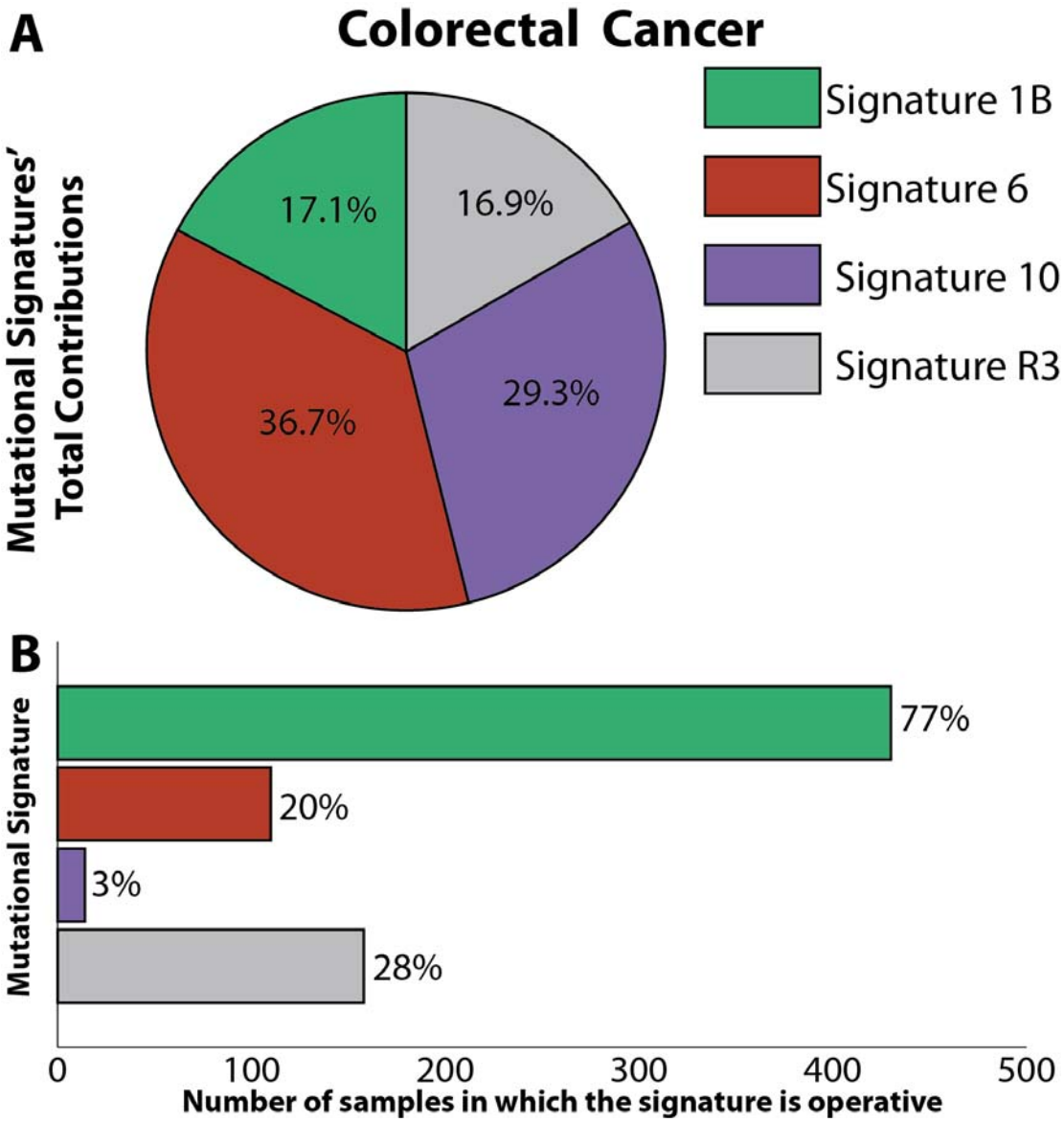

A
Acute Myeloid Leukemia (AML)
1.9%
Signature 1A
Signature 1B
Mutational Signatures' Total Contributions
98.1%
B
Mutational Signature
2%
98%
0
50
100
150
200
Number of samples in which the signature is operative

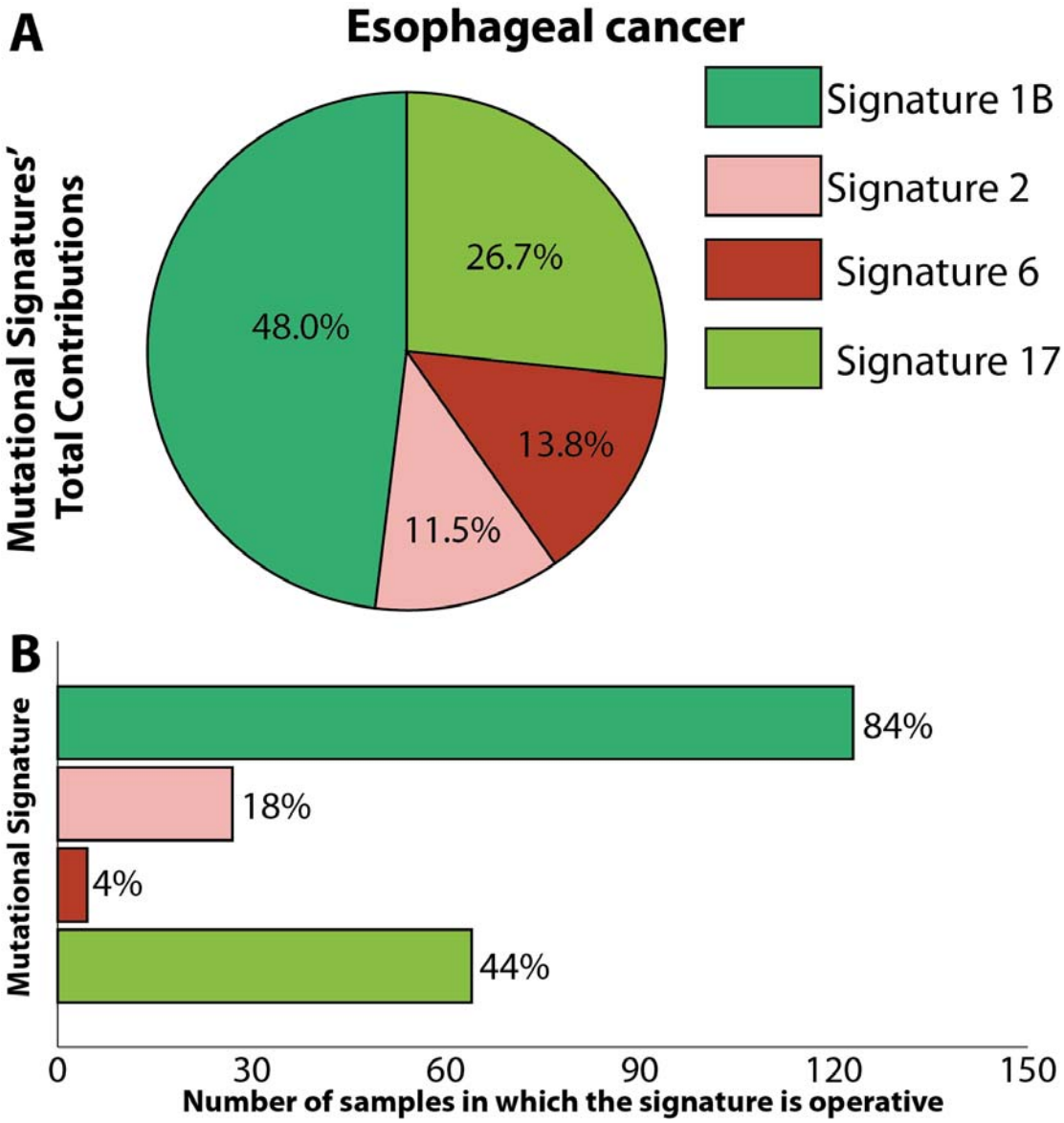

A
Bladder Cancer
Mutational Signatures' Total Contributions
14.2%
38.4%
32.8%
14.6%
Signature 1B
Signature 2
Signature 5
Signature 13
B
Mutational Signature
82%
51%
33%
67%
0
20
40
60
80
100
120
Number of samples in which the signature is operative

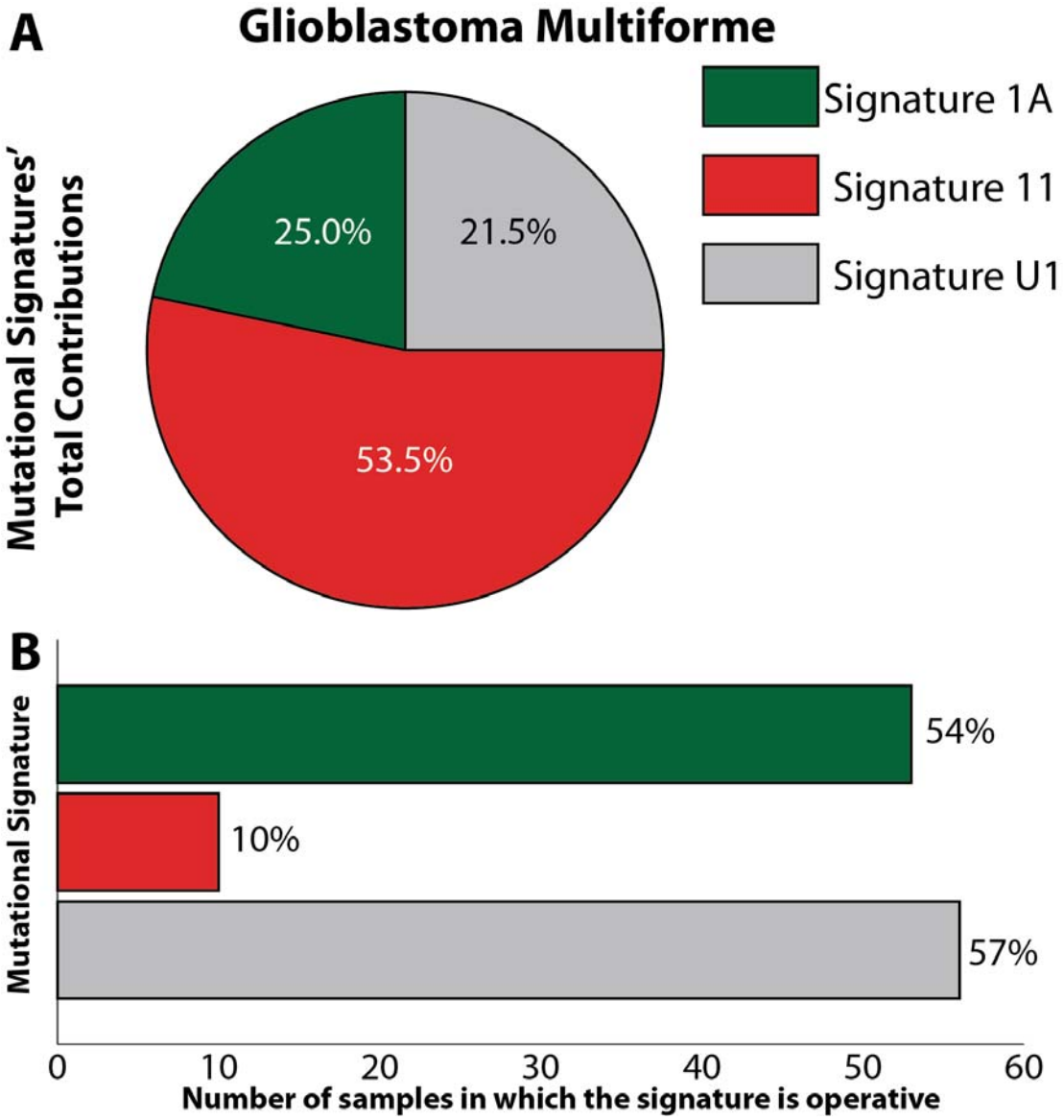

A
Breast Cancer
Mutational Signatures' Total Contributions
0.6%
5.2%
3.8%
35.5%
30.0%
24.9%
Signature 1B
Signature 2
Signature 3
Signature 8
Signature 13
Signature R1
B
Mutational Signature
81%
18%
29%
9%
7%
4%
0
100
200
300
400
500
600
700
800
Number of samples in which the signature is operative

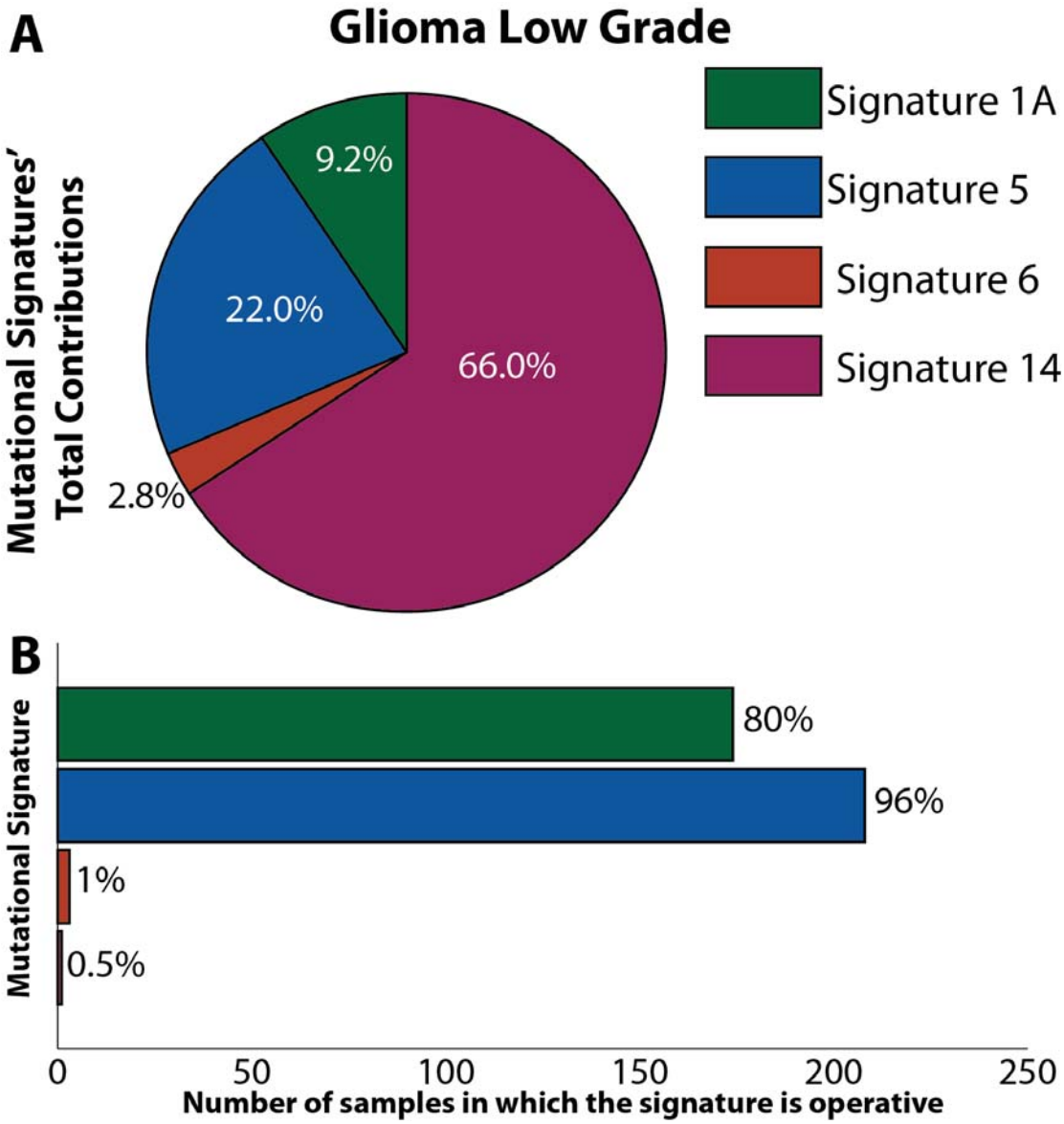

A
Cervical Cancer
Signature 1B
Signature 2
Mutational Signatures' Total Contributions
25.3%
74.7%
B
Mutational Signature
76%
79%
0
5
10
15
20
25
30
Number of samples in which the signature is operative

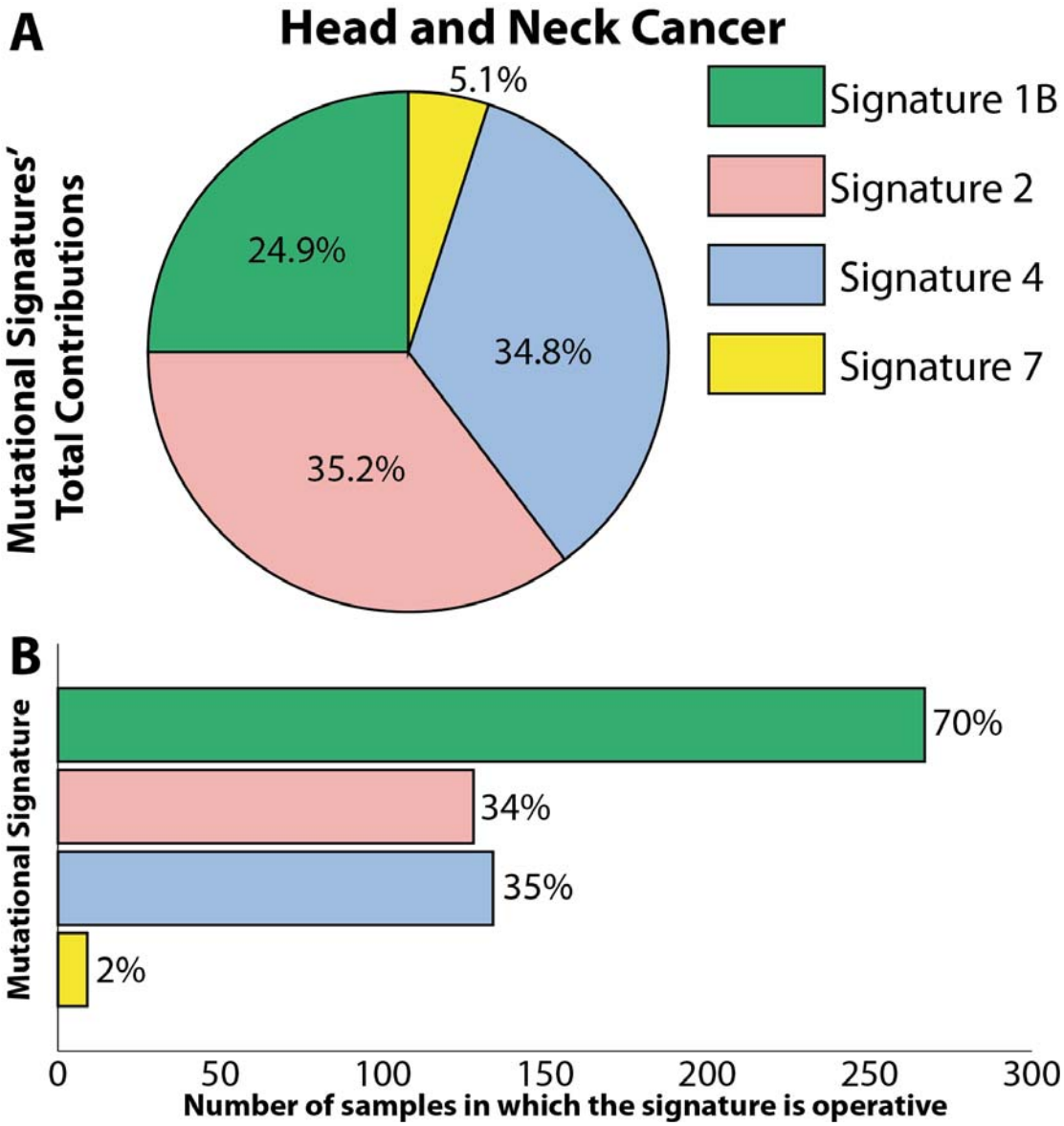

A
B-cell Chronic Lymphocytic Leukemia (CLL)
Mutational Signatures' Total Contributions
Signature 1B
Signature 2
Signature 9
16.7%
6.6%
76.7%
B
Mutational Signature
86%
11%
17%
0
20
40
60
80
100
120
Number of samples in which the signature is operative

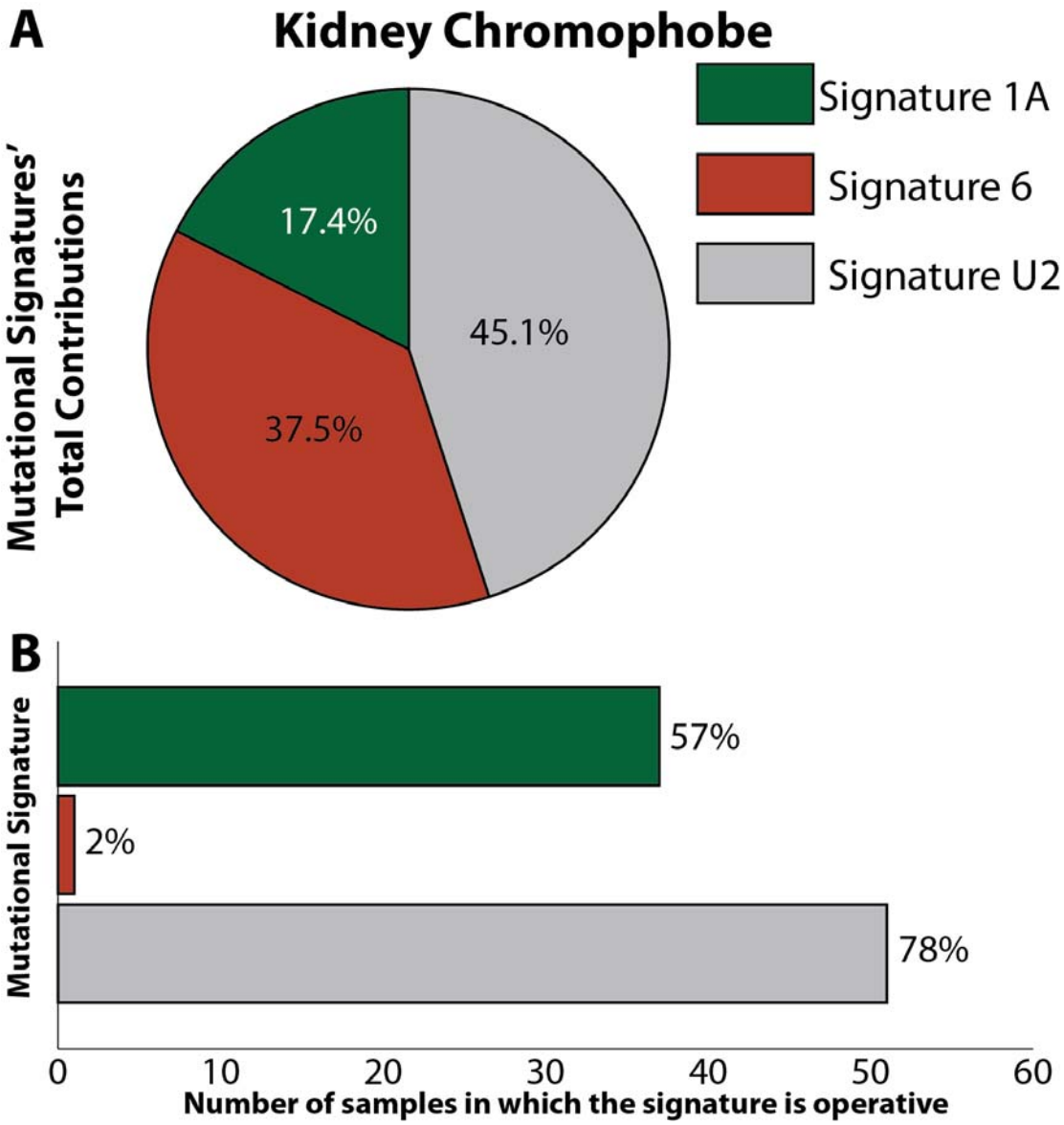

A
Colorectal Cancer
Mutational Signatures' Total Contributions
17.1%
16.9%
36.7%
29.3%
Signature 1B
Signature 6
Signature 10
Signature R3
B
Mutational Signature
77%
20%
3%
28%
0
100
200
300
400
500
Number of samples in which the signature is operative

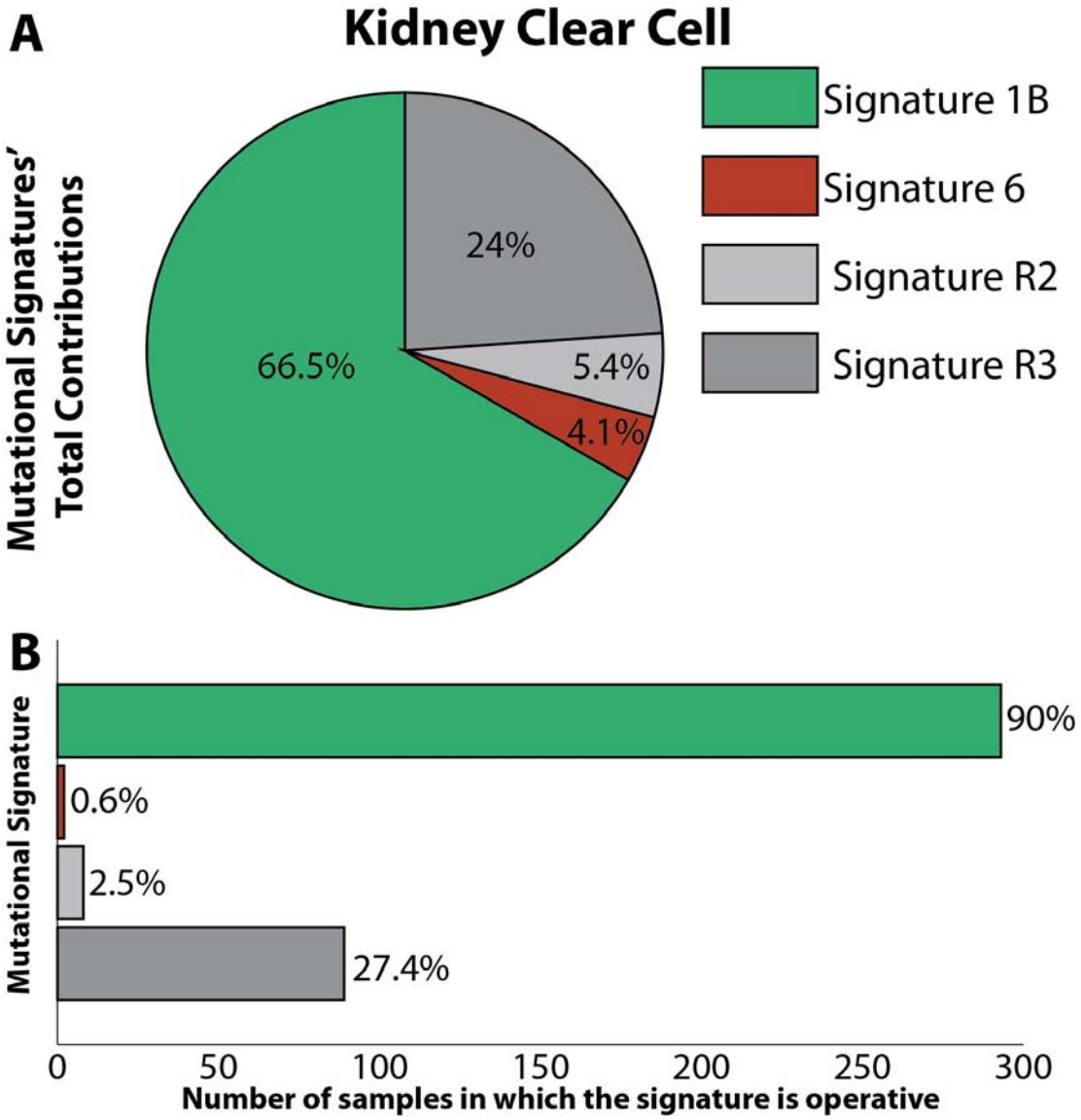

A
Esophageal cancer
Mutational Signatures' Total Contributions
48.0%
26.7%
13.8%
11.5%
Signature 1B
Signature 2
Signature 6
Signature 17
B
Mutational Signature
84%
18%
4%
44%
0
30
60
90
120
150
Number of samples in which the signature is operative

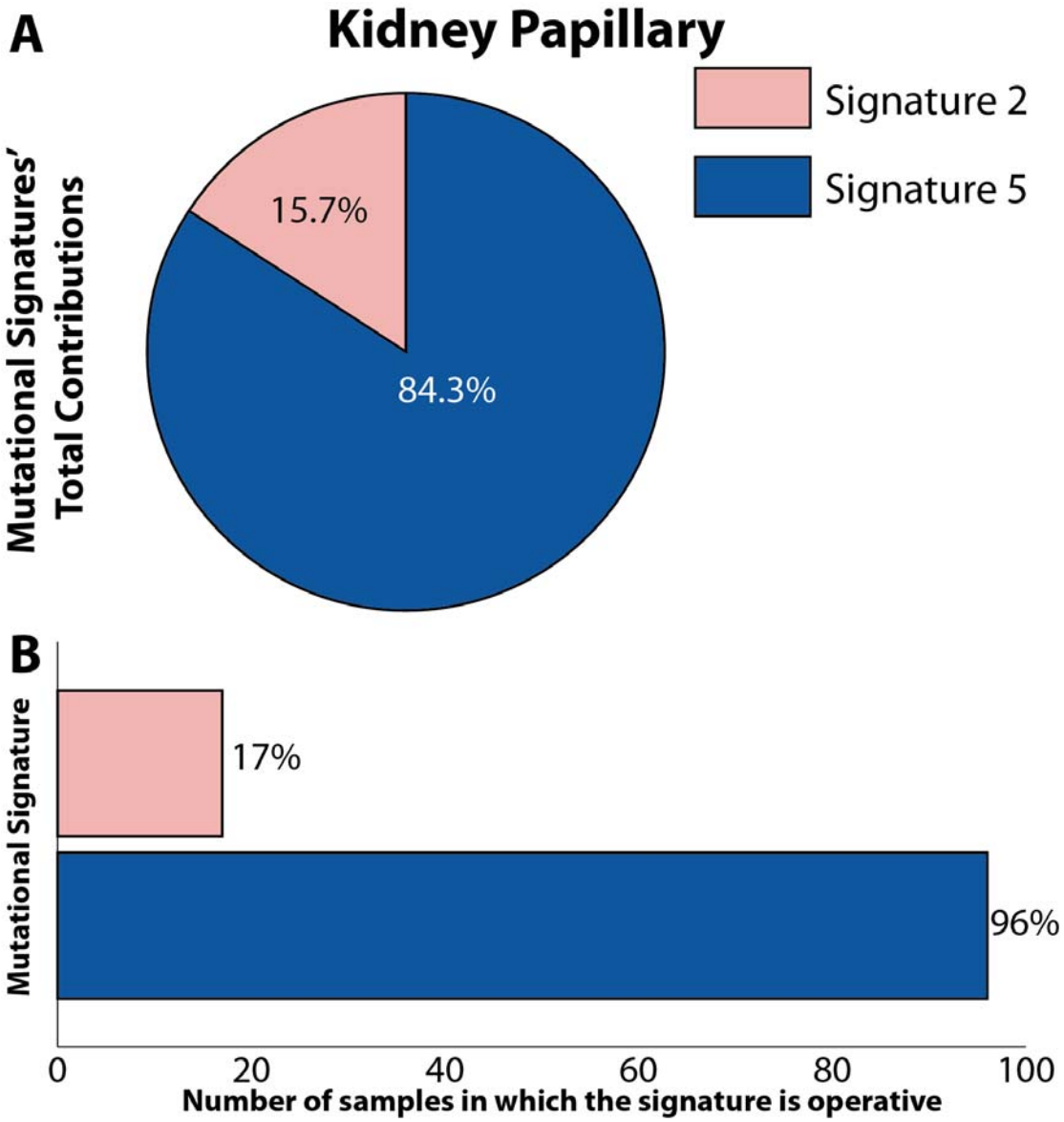

A
Glioblastoma Multiforme
Mutational Signatures' Total Contributions
25.0%
21.5%
53.5%
Signature 1A
Signature 11
Signature U1
B
Mutational Signature
54%
10%
57%
0
10
20
30
40
50
60
Number of samples in which the signature is operative

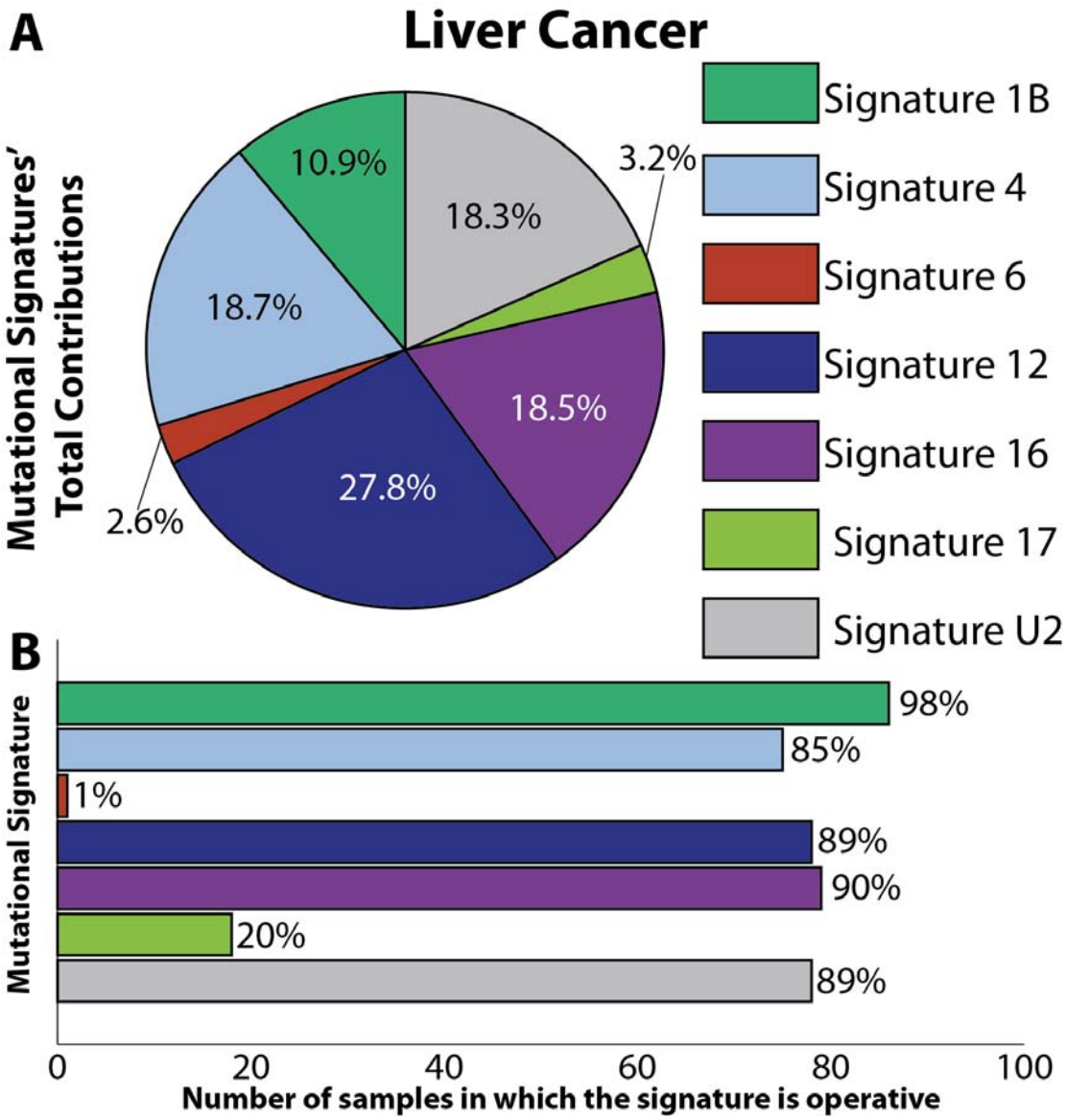

A
Glioma Low Grade
Mutational Signatures' Total Contributions
9.2%
22.0%
66.0%
2.8%
Signature 1A
Signature 5
Signature 6
Signature 14
B
Mutational Signature
80%
96%
1%
0.5%
0
50
100
150
200
250
Number of samples in which the signature is operative

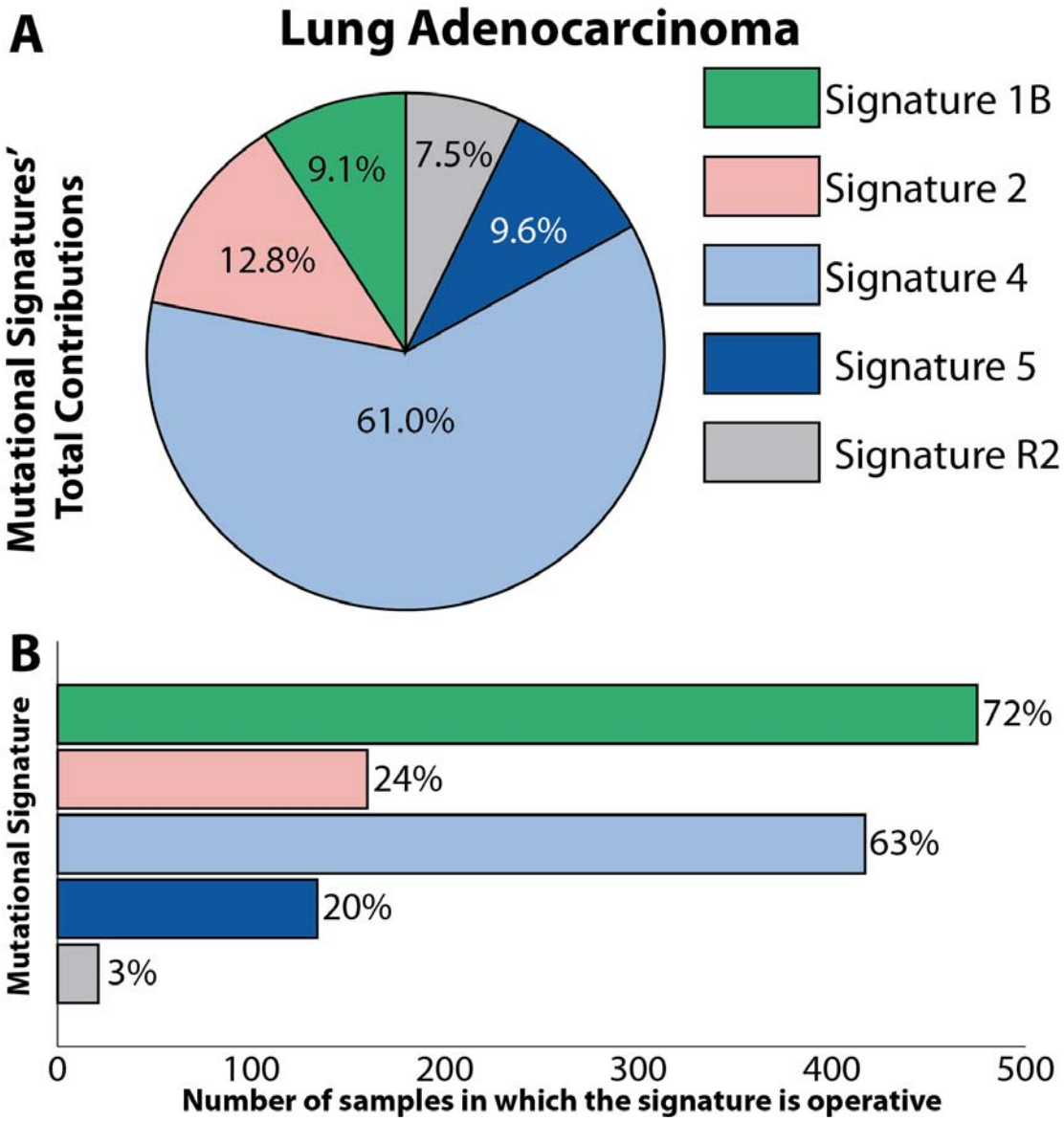

A
Head and Neck Cancer
Mutational Signatures' Total Contributions
5.1%
24.9%
34.8%
35.2%
Signature 1B
Signature 2
Signature 4
Signature 7
B
Mutational Signature
70%
34%
35%
2%
0
50
100
150
200
250
300
Number of samples in which the signature is operative

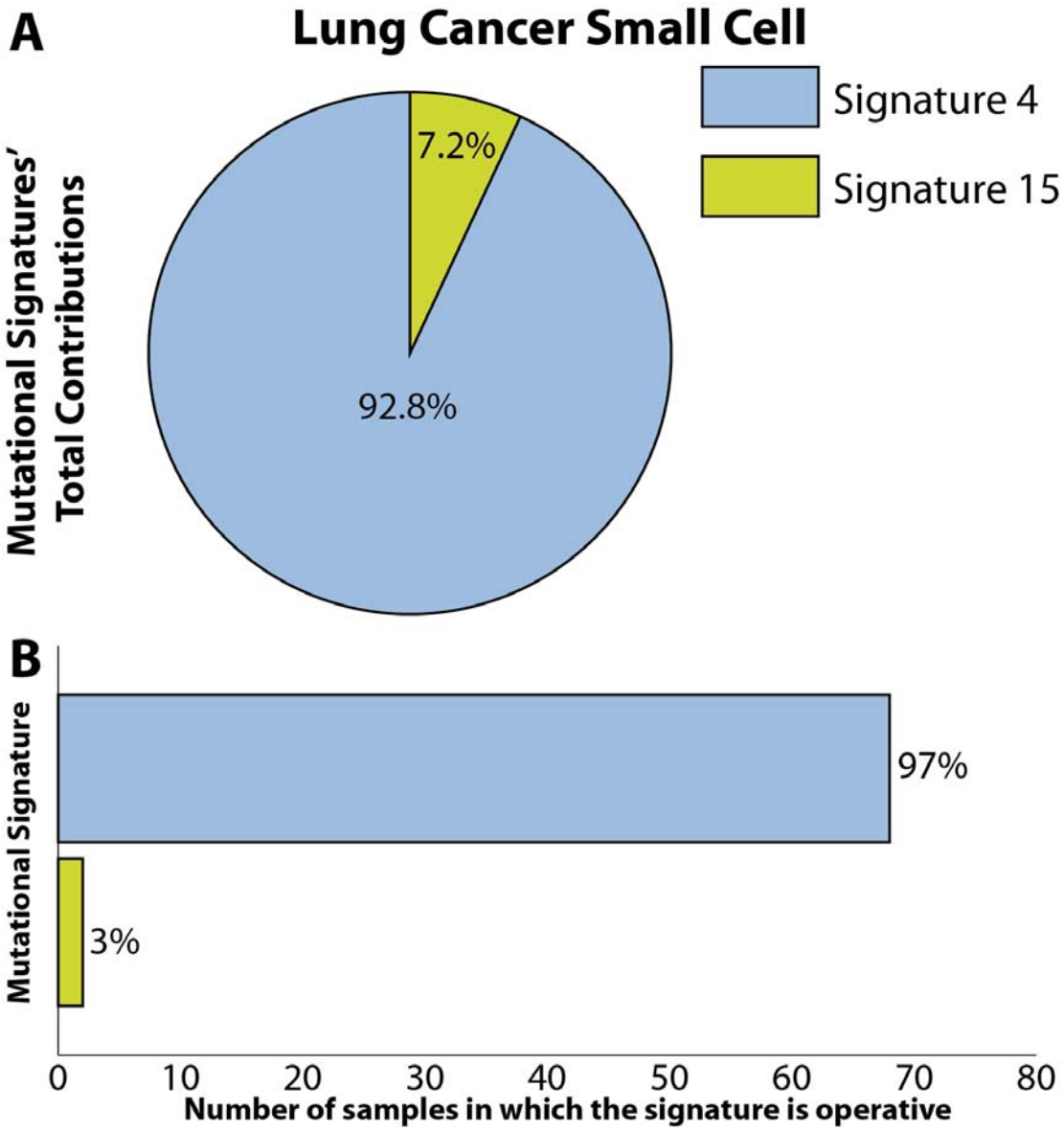

A
Kidney Chromophobe
Mutational Signatures' Total Contributions
17.4%
45.1%
37.5%
Signature 1A
Signature 6
Signature U2
B
Mutational Signature
57%
2%
78%
0
10
20
30
40
50
60
Number of samples in which the signature is operative

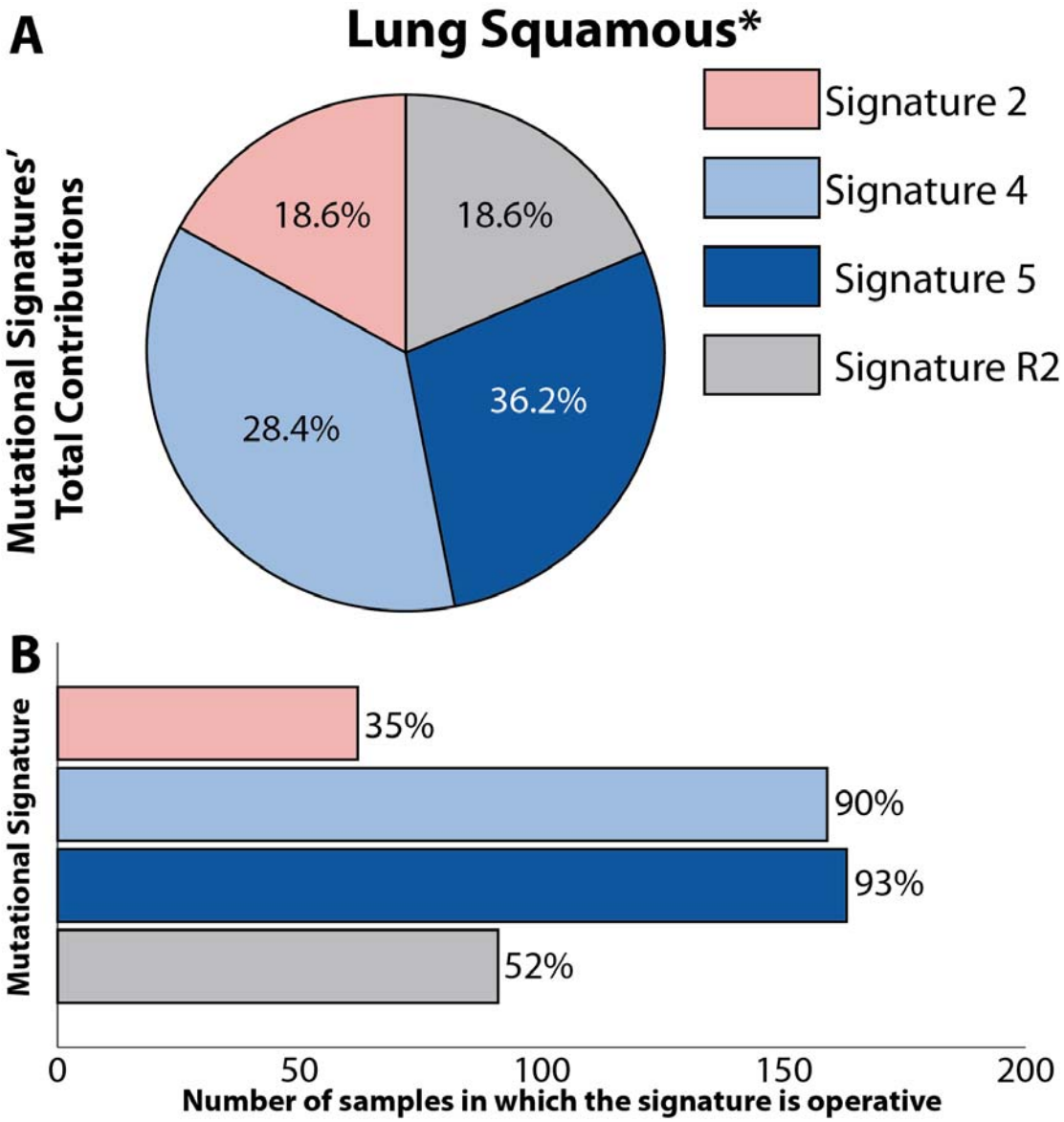

A
Kidney Clear Cell
Mutational Signatures' Total Contributions
66.5%
24%
5.4%
4.1%
Signature 1B
Signature 6
Signature R2
Signature R3
B
Mutational Signature
90%
0.6%
2.5%
27.4%
0
50
100
150
200
250
300
Number of samples in which the signature is operative

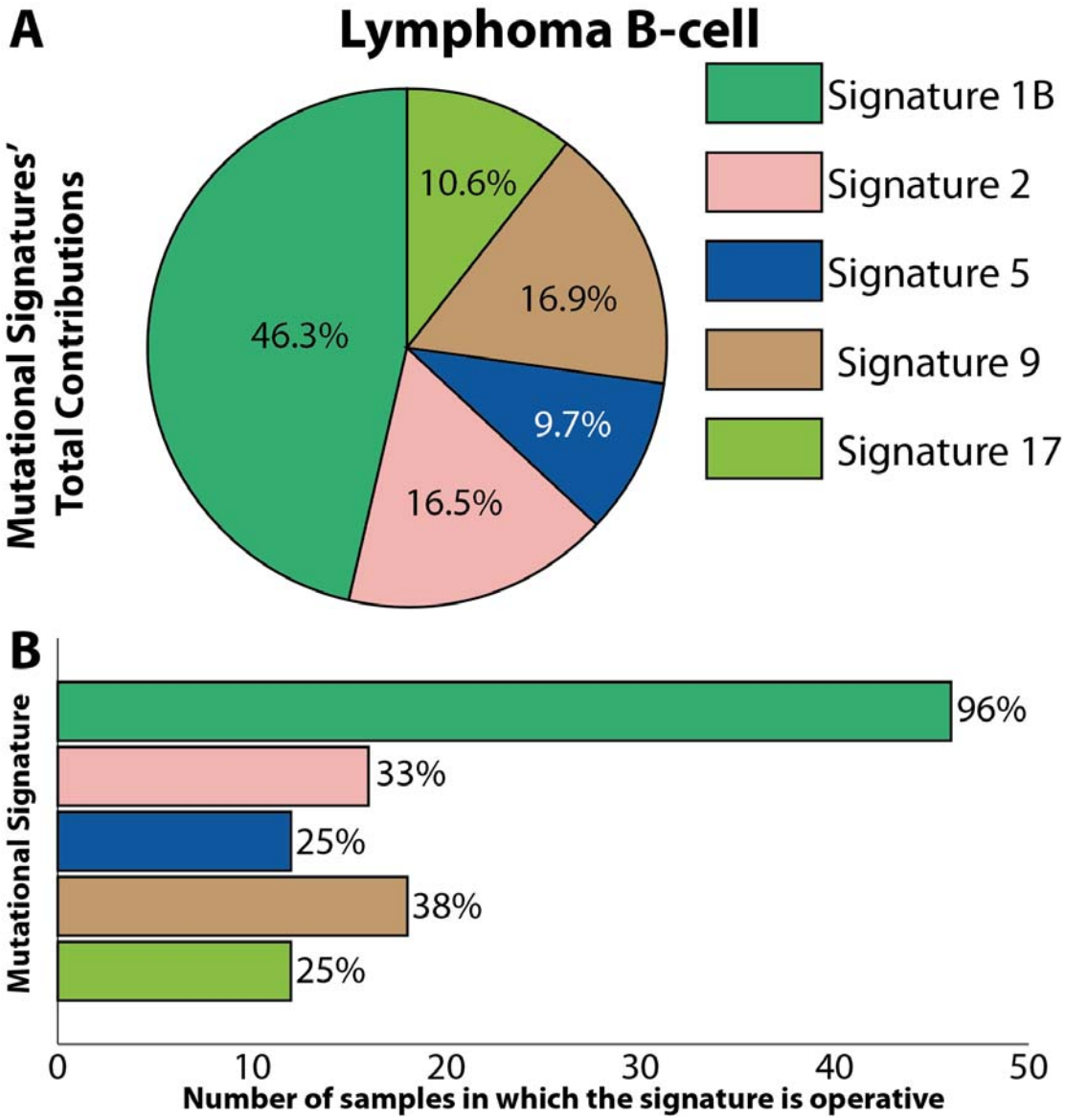

A
Kidney Papillary
Signature 2
Signature 5
Mutational Signatures' Total Contributions
15.7%
84.3%
B
Mutational Signature
17%
96%
0
20
40
60
80
100
Number of samples in which the signature is operative

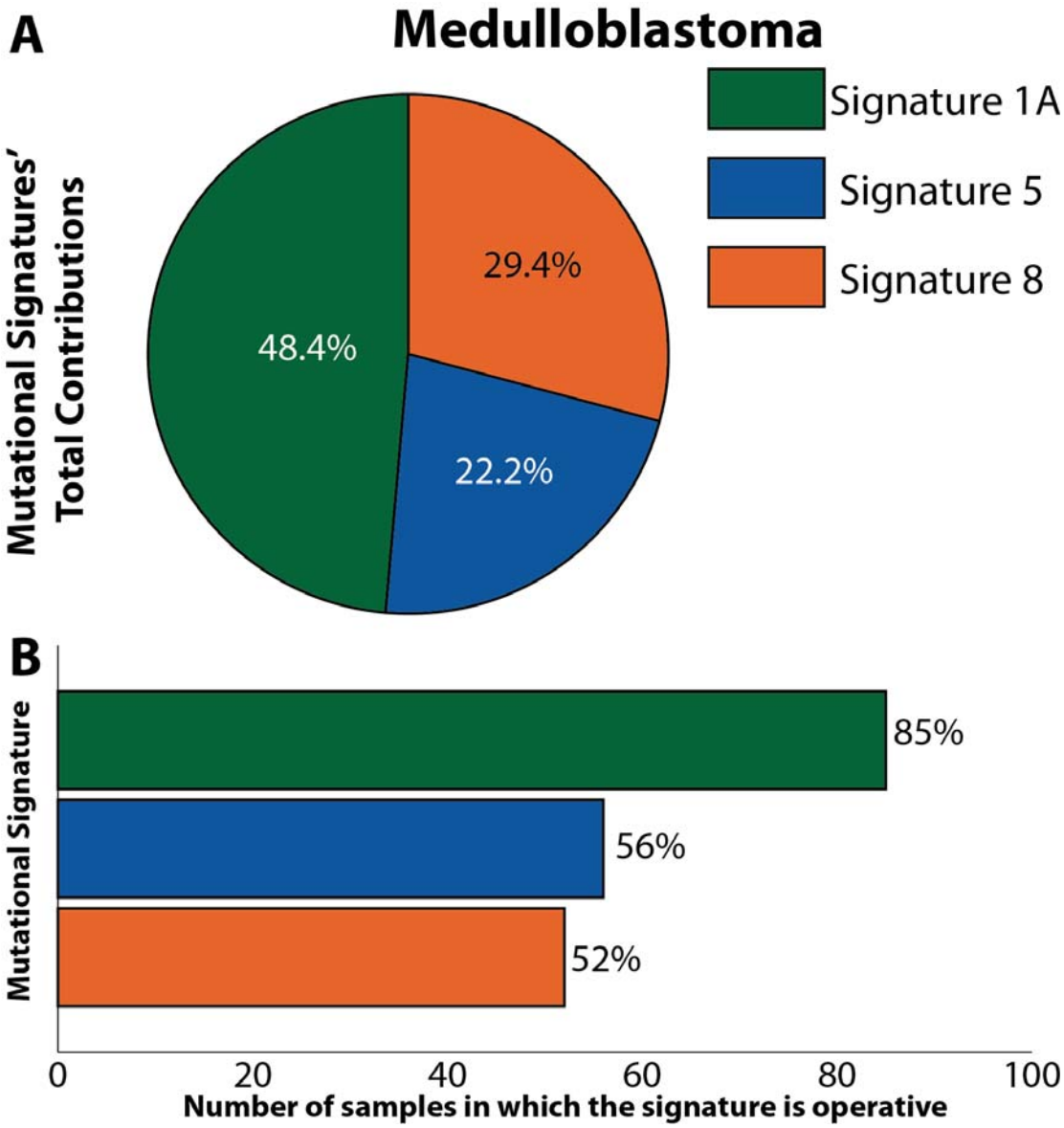

A
Liver Cancer
Mutational Signatures' Total Contributions
10.9%
18.3%
3.2%
18.7%
18.5%
27.8%
2.6%
Signature 1B
Signature 4
Signature 6
Signature 12
Signature 16
Signature 17
Signature U2
B
Mutational Signature
98%
85%
1%
89%
90%
20%
89%
0
20
40
60
80
100
Number of samples in which the signature is operative

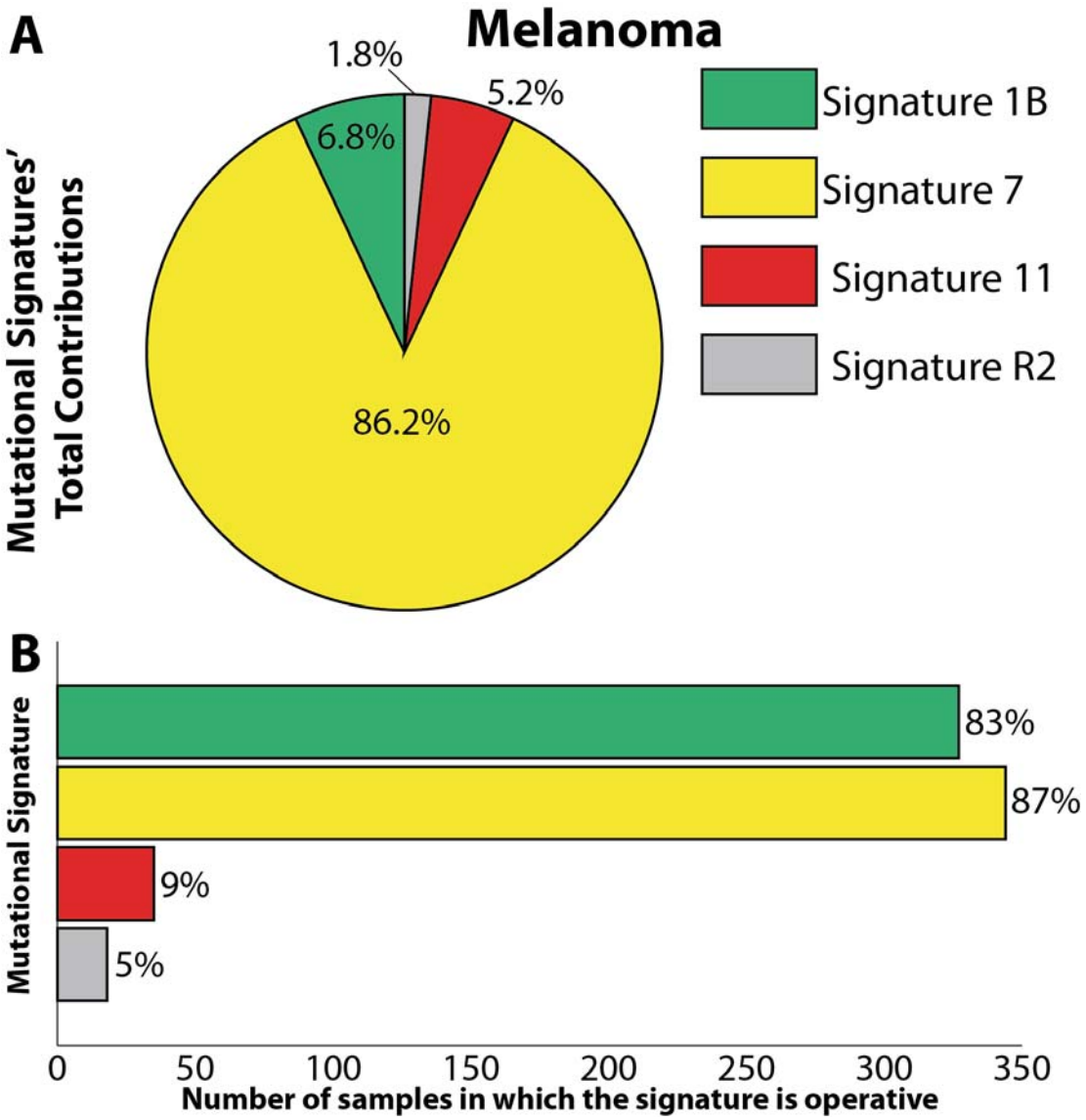

A
Lung Adenocarcinoma
Mutational Signatures' Total Contributions
9.1%
7.5%
9.6%
12.8%
61.0%
Signature 1B
Signature 2
Signature 4
Signature 5
Signature R2
B
Mutational Signature
72%
24%
63%
20%
3%
0
100
200
300
400
500
Number of samples in which the signature is operative

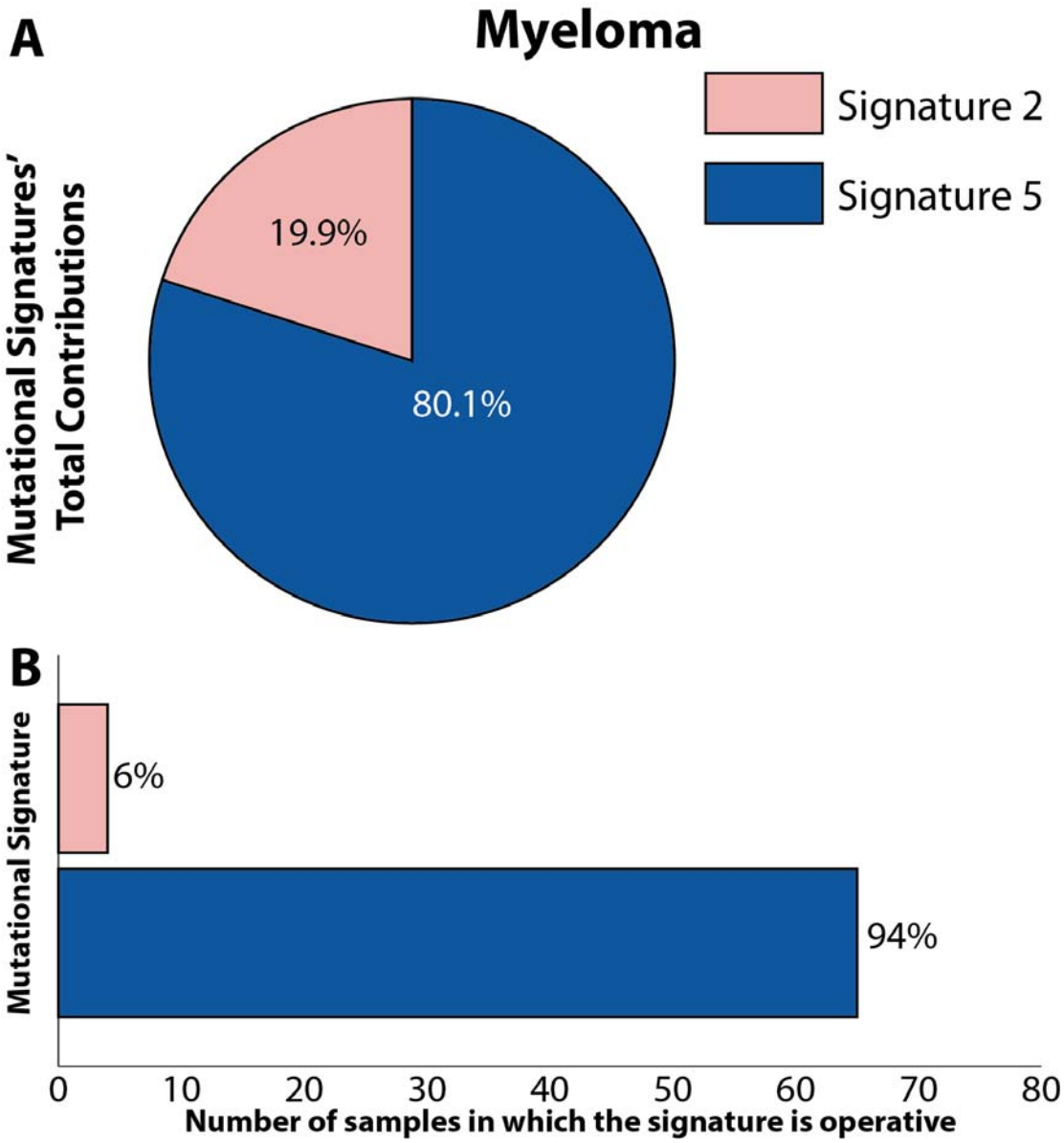

A
Lung Cancer Small Cell
Mutational Signatures' Total Contributions
Signature 4
Signature 15
7.2%
92.8%
B
Mutational Signature
97%
3%
0
10
20
30
40
50
60
70
80
Number of samples in which the signature is operative

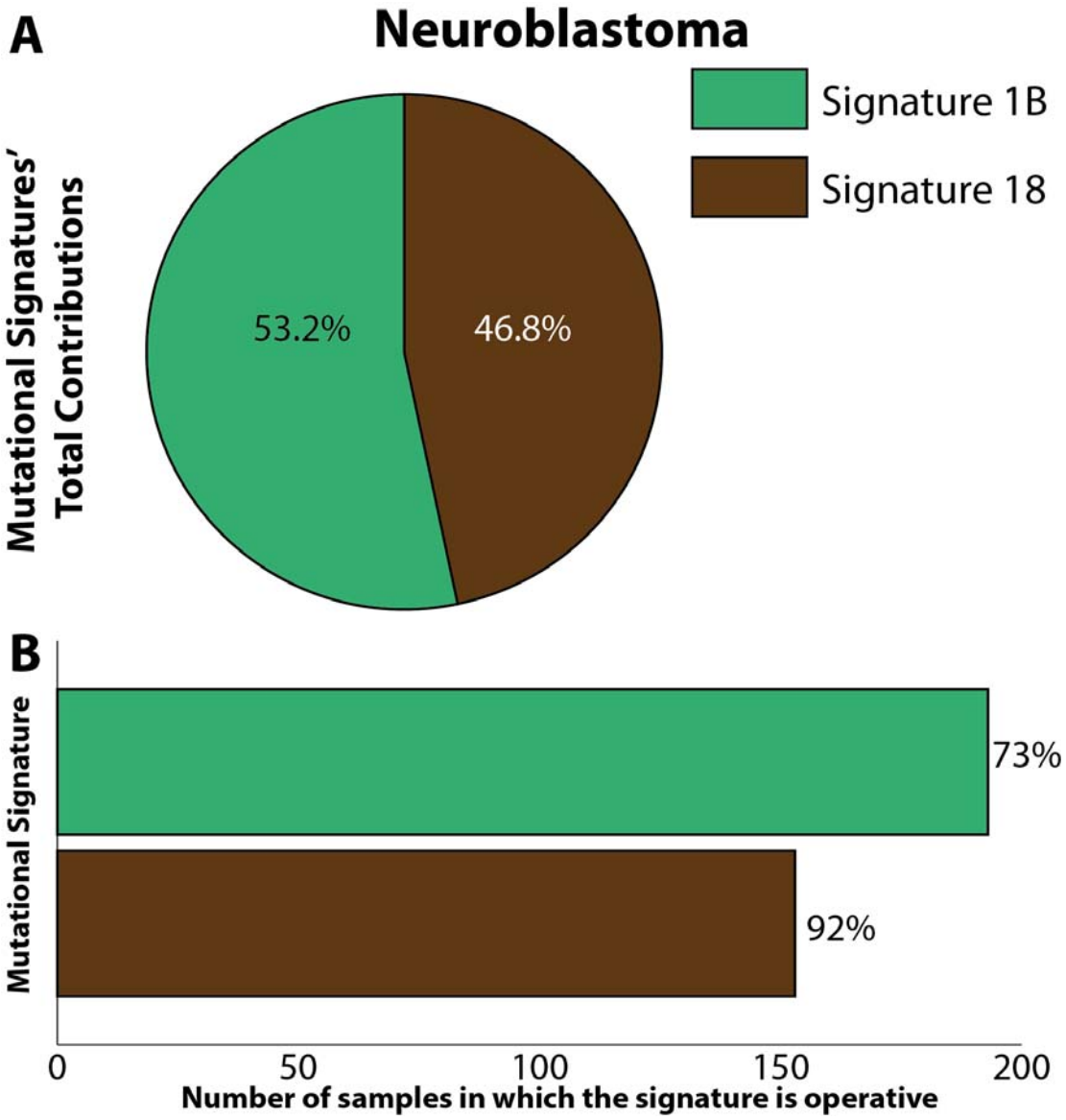

A
Lung Squamous*
Mutational Signatures' Total Contributions
18.6%
18.6%
36.2%
28.4%
Signature 2
Signature 4
Signature 5
Signature R2
B
Mutational Signature
35%
90%
93%
52%
0
50
100
150
200
Number of samples in which the signature is operative

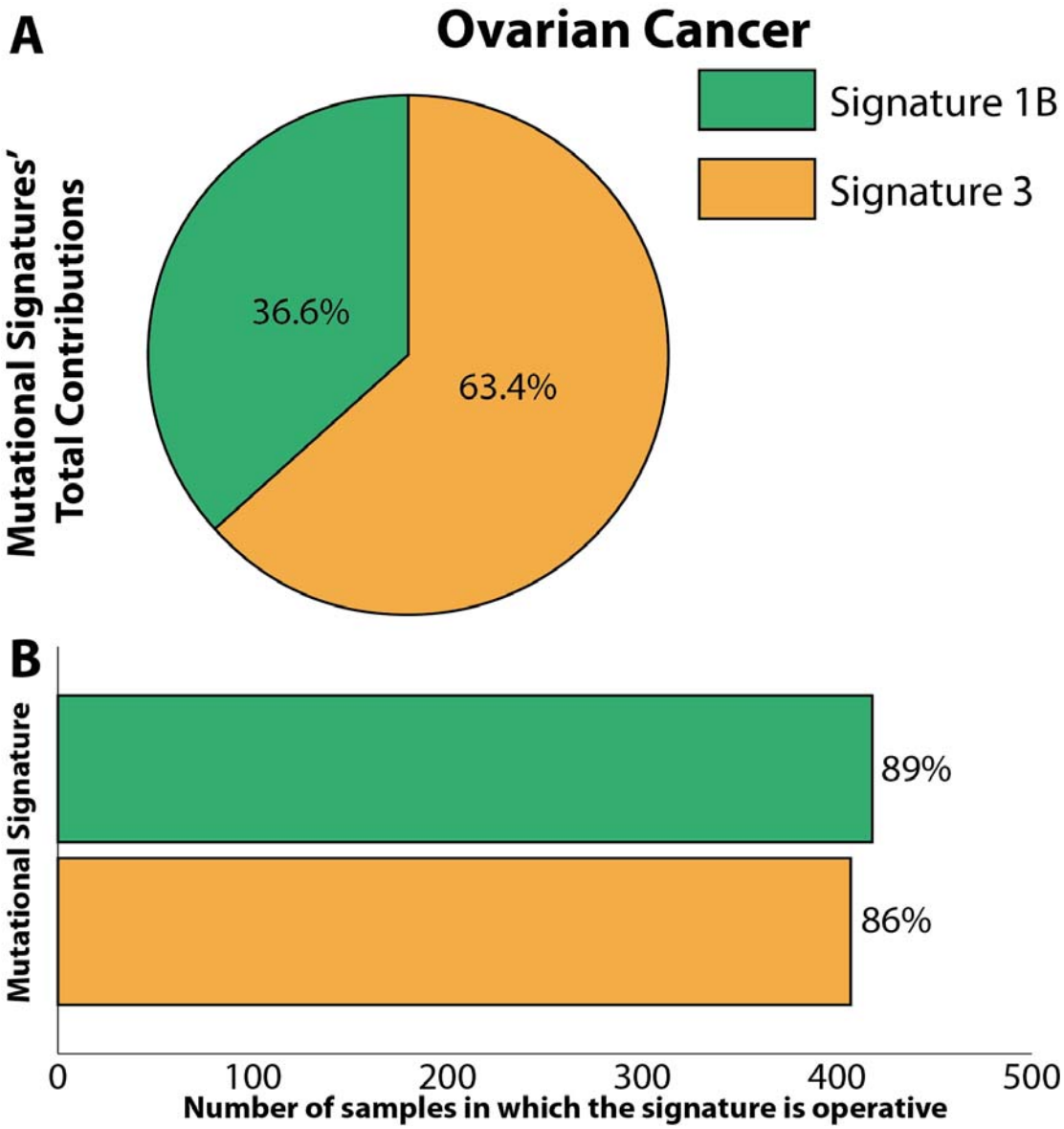

A
Lymphoma B-cell
Mutational Signatures' Total Contributions
46.3%
10.6%
16.9%
9.7%
16.5%
Signature 1B
Signature 2
Signature 5
Signature 9
Signature 17
B
Mutational Signature
96%
33%
25%
38%
25%
0
10
20
30
40
50
Number of samples in which the signature is operative

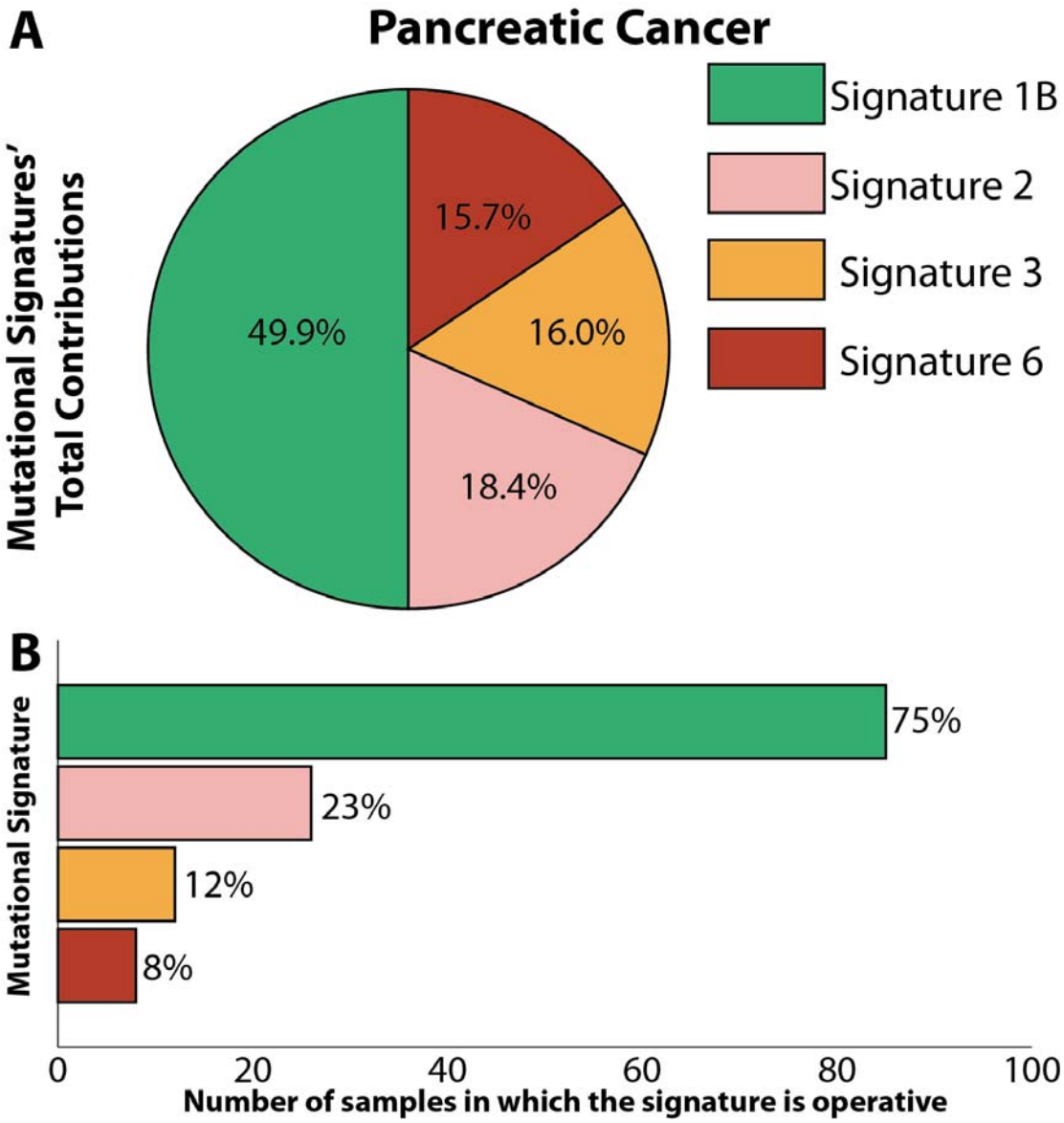

A
Medulloblastoma
Mutational Signatures' Total Contributions
Signature 1A
Signature 5
Signature 8
48.4%
29.4%
22.2%
B
Mutational Signature
85%
56%
52%
0
20
40
60
80
100
Number of samples in which the signature is operative

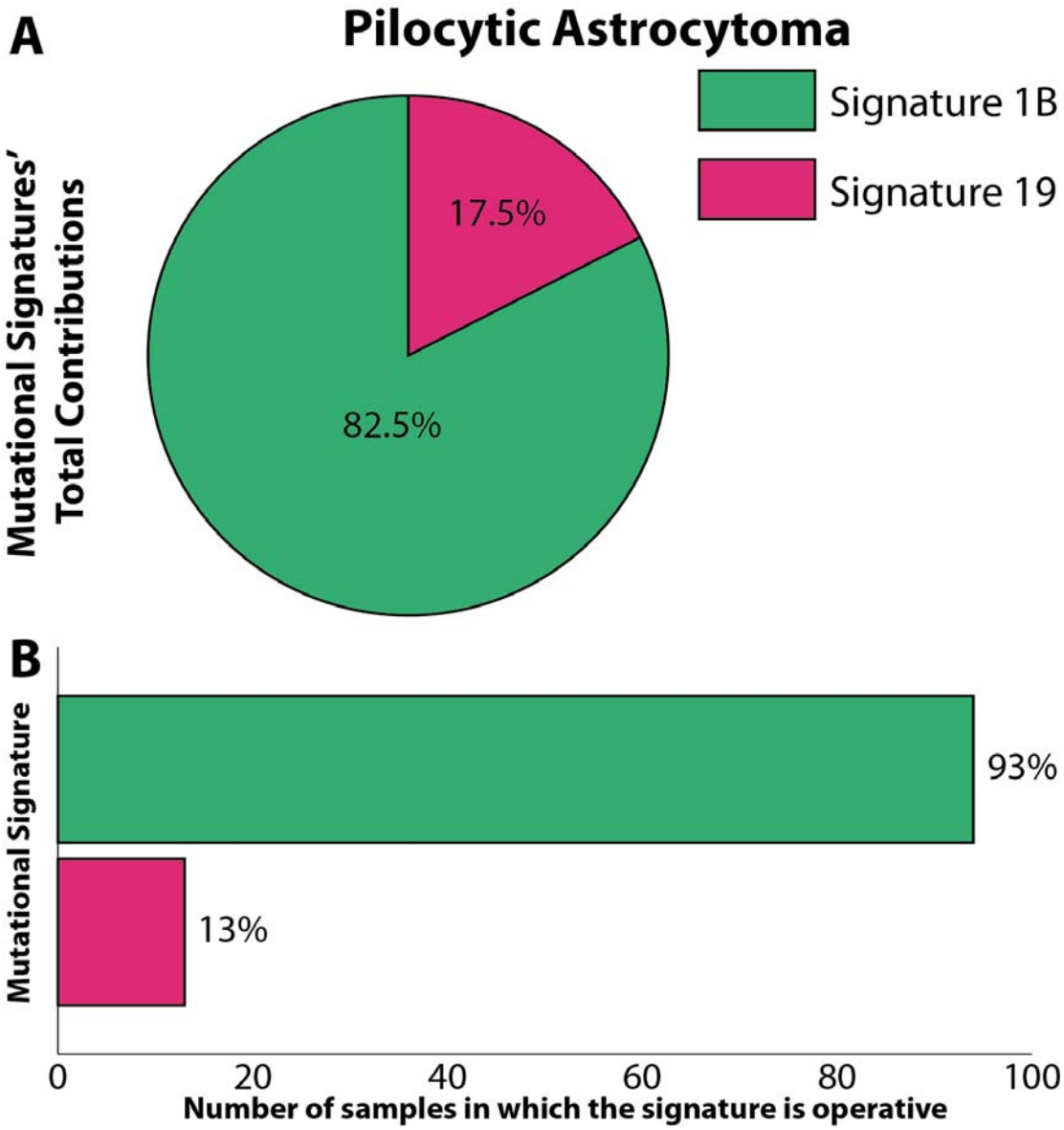

A
Melanoma
Mutational Signatures' Total Contributions
1.8%
5.2%
6.8%
86.2%
Signature 1B
Signature 7
Signature 11
Signature R2
B
Mutational Signature
83%
87%
9%
5%
0
50
100
150
200
250
300
350
Number of samples in which the signature is operative

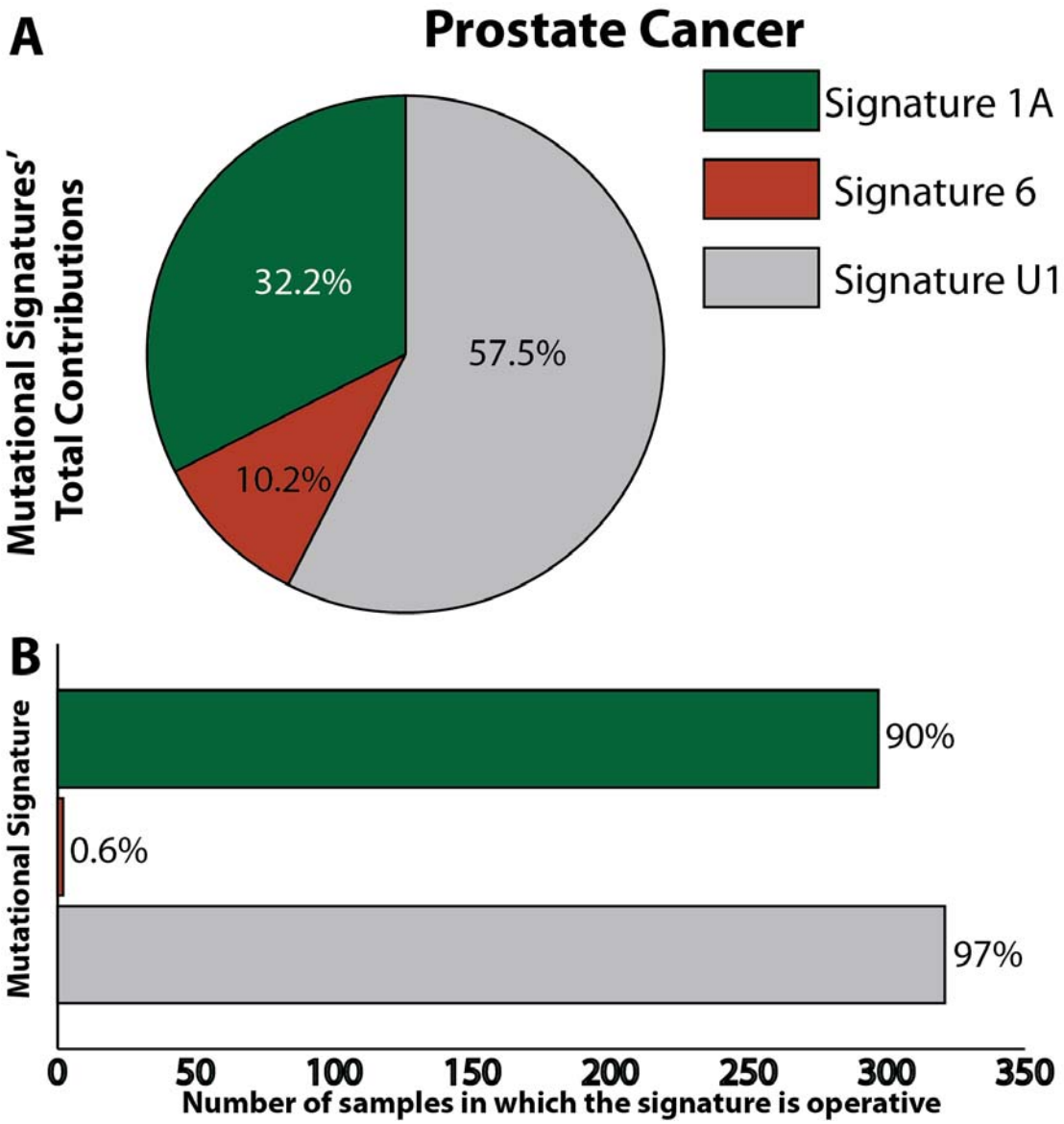

A
Myeloma
Mutational Signatures' Total Contributions
Signature 2
Signature 5
19.9%
80.1%
B
Mutational Signature
6%
94%
0
10
20
30
40
50
60
70
80
Number of samples in which the signature is operative

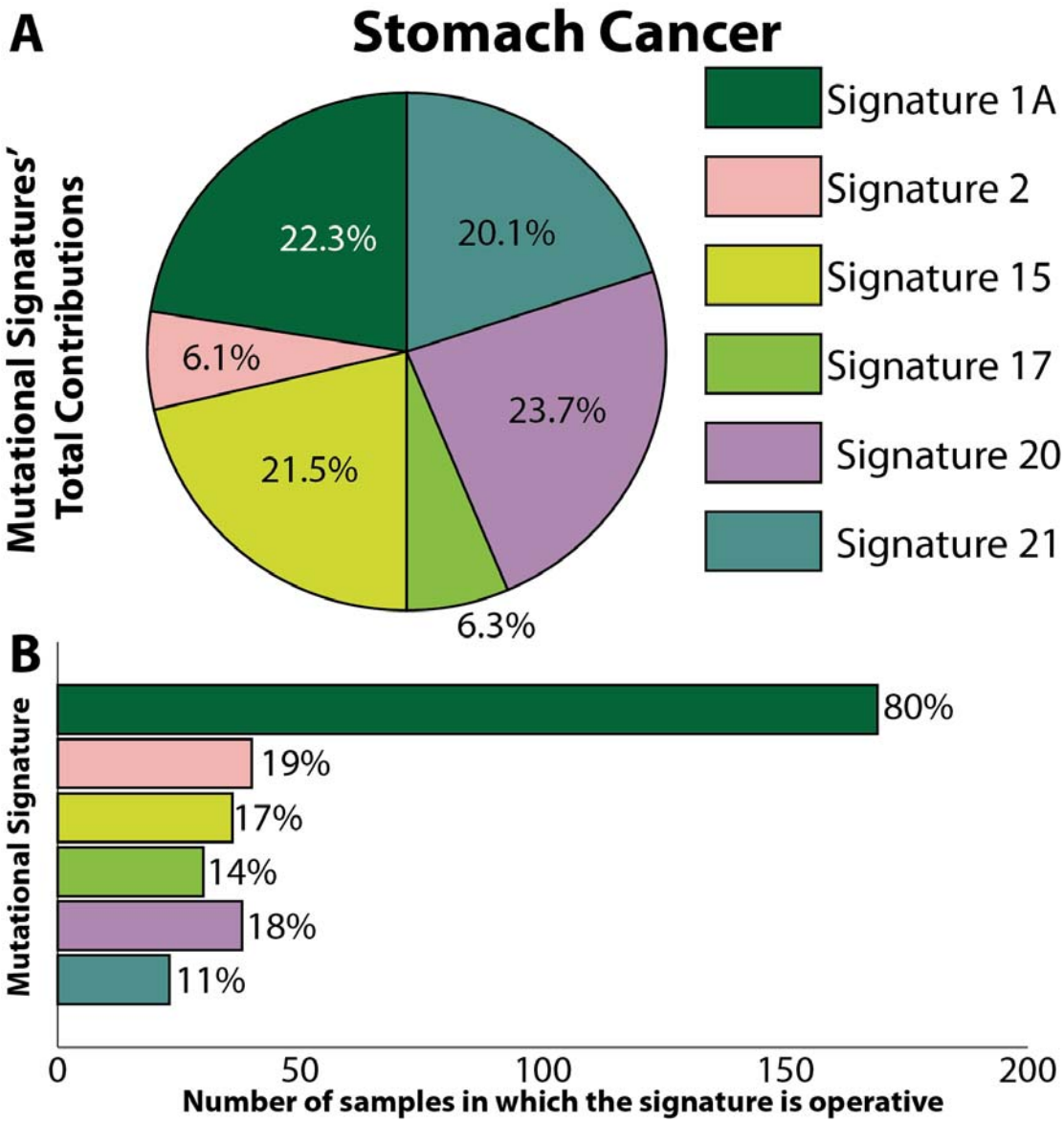

A
Neuroblastoma
Signature 1B
Signature 18
Mutational Signatures' Total Contributions
53.2%
46.8%
B
Mutational Signature
73%
92%
0
50
100
150
200
Number of samples in which the signature is operative

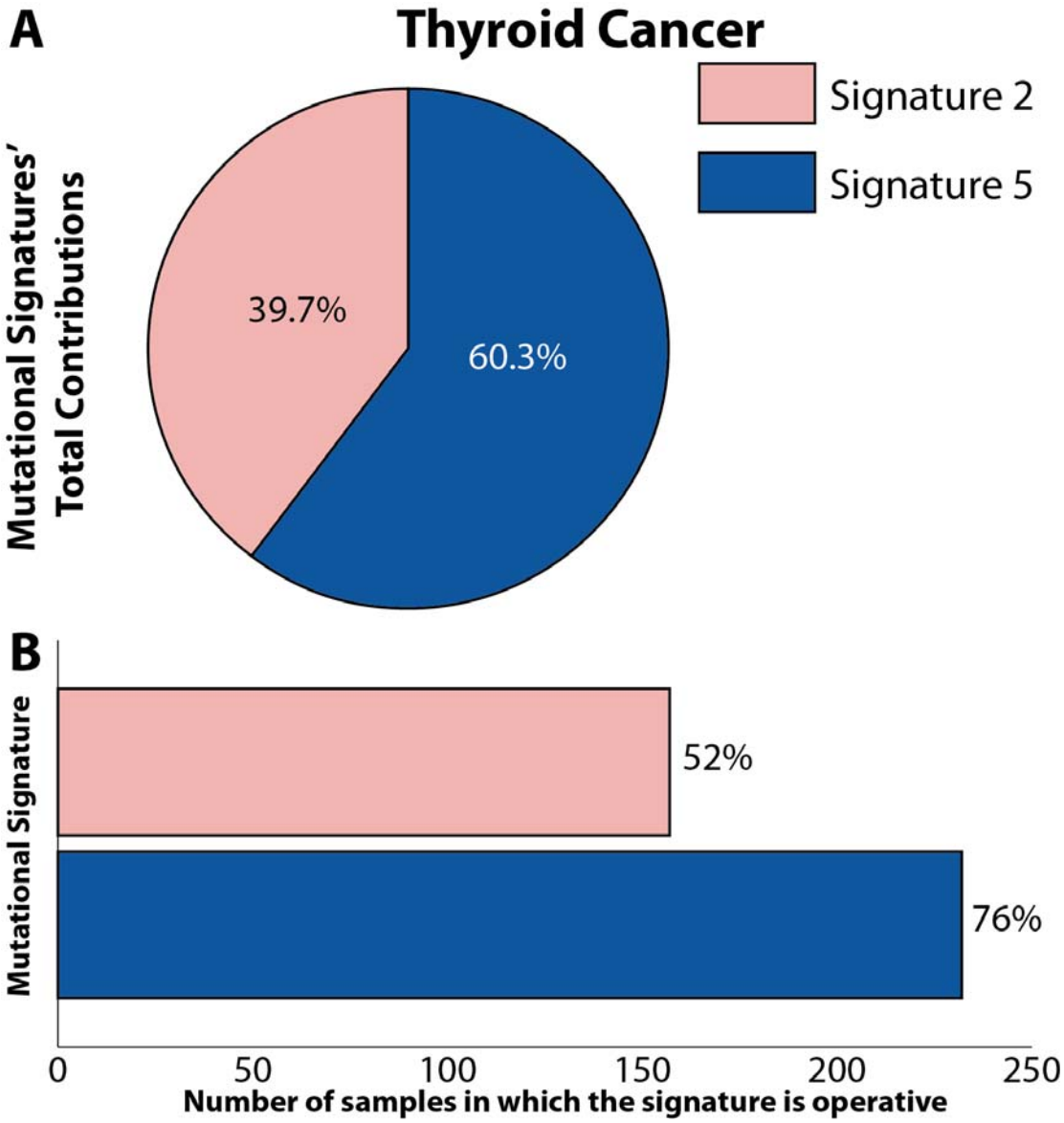

A
Ovarian Cancer
Mutational Signatures' Total Contributions
Signature 1B
Signature 3
36.6%
63.4%
B
Mutational Signature
89%
86%
0
100
200
300
400
500
Number of samples in which the signature is operative

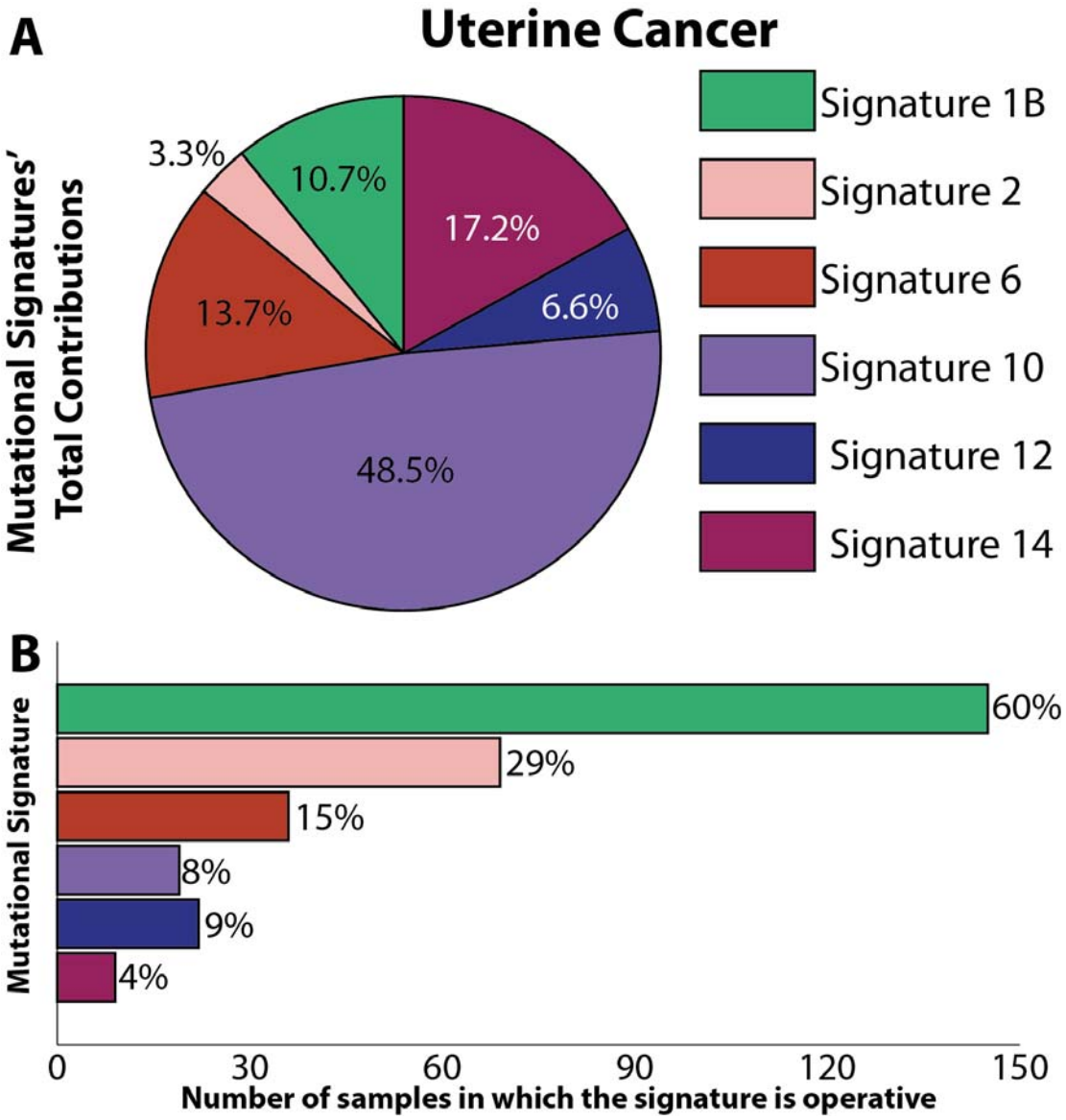

A
Pancreatic Cancer
Mutational Signatures' Total Contributions
49.9%
15.7%
16.0%
18.4%
Signature 1B
Signature 2
Signature 3
Signature 6
B
Mutational Signature
75%
23%
12%
8%
0
20
40
60
80
100
Number of samples in which the signature is operative


A
Pilocytic Astrocytoma
Signature 1B
Signature 19
Mutational Signatures' Total Contributions
17.5%
82.5%
B
Mutational Signature
93%
13%
0
20
40
60
80
100
Number of samples in which the signature is operative


A
Prostate Cancer
Mutational Signatures' Total Contributions
32.2%
10.2%
57.5%
Signature 1A
Signature 6
Signature U1
B
Mutational Signature
90%
0.6%
97%
0
50
100
150
200
250
300
350
Number of samples in which the signature is operative


A
Stomach Cancer
Mutational Signatures' Total Contributions
22.3%
20.1%
6.1%
23.7%
21.5%
6.3%
Signature 1A
Signature 2
Signature 15
Signature 17
Signature 20
Signature 21
B
Mutational Signature
80%
19%
17%
14%
18%
11%
0
50
100
150
200
Number of samples in which the signature is operative


A
Thyroid Cancer
Signature 2
Signature 5
Mutational Signatures' Total Contributions
39.7%
60.3%
B
Mutational Signature
52%
76%
0
50
100
150
200
250
Number of samples in which the signature is operative


A
Uterine Cancer
Mutational Signatures' Total Contributions
3.3%
10.7%
17.2%
13.7%
6.6%
48.5%
Signature 1B
Signature 2
Signature 6
Signature 10
Signature 12
Signature 14
B
Mutational Signature
60%
29%
15%
8%
9%
4%
0
30
60
90
120
150
Number of samples in which the signature is operative

APPENDIX VII: Publications associated with this thesis

This appendix contains the references of the articles that have been written and published as part of this thesis. The articles are separated into two categories: (i) main articles – four manuscripts directly related to developing and presenting the approach for deciphering mutational signatures and applying this approach to a large scale of whole-genome and whole-exome sequencing data, and (ii) supporting articles – seven manuscripts in which mutational signatures (and/or patterns of somatic mutations) have been examined. It is worth noting that this list of manuscripts does not include another six published articles unrelated to mutational signatures and/or cancer nor does it include another seven articles currently under review with which I have been involved during the course of my doctoral studies. Lastly, it should be noted that this thesis is almost entirely written based on the four main mutational signatures articles.

Main articles

Alexandrov LB and Stratton MR (2014) Mutational Signatures: The Patterns of Somatic Mutations Hidden in Cancer Genomes. **Current Opinion in Genetics & Development** 24, 52-60 (invited review article/corresponding author).

Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SA, Behjati S, Biankin AV, Bignell GR, Bolli N, Borg A, Borresen-Dale AL, Boyault S, Burkhardt B, Butler AP, Caldas C, Davies HR, Desmedt C, Eils R, Eyfjord JE, Foekens JA, Greaves M, Hosoda F, Hutter B, Ilicic T, Imbeaud S, Imielinsk M, Jager N, Jones DT, Jones D, Knappskog S, Kool M, Lakhani SR, Lopez-Otin C, Martin S, Munshi NC, Nakamura H, Northcott PA, Pajic M, Papaemmanuil E, Paradiso A, Pearson JV, Puente XS, Raine K, Ramakrishna M, Richardson AL, Richter J, Rosenstiel P, Schlesner M, Schumacher TN, Span PN, Teague JW, Totoki Y, Tutt AN, Valdes-Mas R, van Buuren MM, van 't Veer L, Vincent-Salomon A, Waddell N, Yates LR, Zucman-Rossi J, Futreal PA, McDermott U, Lichter P, Meyerson M, Grimmond SM, Siebert R, Campo E, Shibata T, Pfister SM, Campbell PJ, and Stratton MR (2013) Signatures of mutational processes in human cancer. **Nature** 500:415-421.

Alexandrov LB, Nik-Zainal S, Wedge DC, Campbell PJ, and Stratton MR (2013) Deciphering signatures of mutational processes operative in human cancer. **Cell Reports** 3:246-259.

Nik-Zainal S, **Alexandrov LB,** Wedge DC, Van Loo P, Greenman CD, Raine K, Jones D, Hinton J, Marshall J, Stebbings LA, Menzies A, Martin S, Leung K, Chen L, Leroy C, Ramakrishna M, Rance R, Lau KW, Mudie LJ, Varela I, McBride DJ, Bignell GR, Cooke SL, Shlien A, Gamble J, Whitmore I, Maddison M, Tarpey PS, Davies HR, Papaemmanuil E, Stephens PJ, McLaren S, Butler AP, Teague JW, Jonsson G, Garber JE, Silver D, Miron P, Fatima A, Boyault S, Langerod A, Tutt A, Martens JW, Aparicio SA, Borg A, Salomon AV, Thomas G, Borresen-Dale AL, Richardson AL, Neuberger MS, Futreal PA, Campbell PJ, Stratton MR (2012) Mutational processes molding the genomes of 21 breast cancers. **Cell** 149:979-993.

Supporting articles

Behjati S, Huch M, van Boxtel R, Karthaus W, Wedge DC, Tamuri AU, Martincorena I, Petljak M, **Alexandrov LB**, Gundem G, Tarpey PS, Roerink S, Blokker J, Maddison M, Mudie L, Robinson B, Nik-Zainal S, Campbell P, Goldman N, van de Wetering M, Cuppen E, Clevers H, and Stratton MR (2014) Genome sequencing of normal cells reveals developmental lineages and mutational processes. **Nature (AOP)** doi:10.1038/nature13448

Murchison EP, Wedge DC, **Alexandrov LB**, Fu B, Martincorena I, Ning Z, Tubio J, Werner EI, Allen J, Barboza di Nardi A, Donelan EM, Marino G, Fassati A, Campbell PJ, Yang F, Burt A, Weiss RA, and Stratton MR (2014) Transmissible dog cancer genome reveals the origin and history of an ancient cell lineage. **Science** 343:437-440.

Bolli B, Avet-Loiseau H, Wedge D, Van Loo P, **Alexandrov LB**, Martincorena I, Dawson K, Iorio F, Nik-Zainal S, Bignell G, Hinton H, Li Y, Tubio J, McLaren S, O'Meara S, Butler AS, Teague J, Mudie L, Anderson E, Rashid N, Tai YT, Shammas M, Sperling A, Fulciniti M, Richardson P, Parmigiani G, Magrangeas F, Minvielle S, Moreau P, Attal M, Facon T, Futreal A, Anderson K, and Campbell PJ (2014) Heterogeneity of genomic architecture and evolution in multiple myeloma. **Nature Communications** 5 (2997).

Papaemmanuil E, Rapado I, Li Y, Potter NE, Wedge DC, Tubio J, **Alexandrov LB**, Loo PV, Cooke SL, Marshall J, Martincorena I, Hinton J, Gundem G, van Delft FW, Nik-Zainal S, Jones DR, Ramakrishna M, Titley I, Stebbings L, Leroy C, Menzies A, Gamble K, Robinson B, Mudie L, Raine K, O'Meara S, Teague JW, Butler AP, Cazzaniga G, Biondi A, Zuna J, Kempski H, Muschen M, Ford AM, Stratton MR, Greaves M, and Campbell PJ (2014) RAG-mediated recombination is the predominant driver of oncogenic rearrangement in ETV6-RUNX1 acute lymphoblastic leukemia. **Nature Genetics** 46 (2): 116-125.

Wong CC, Martincorena I, Rust AG, Rashid M, Alifrangis C, **Alexandrov LB**,

Tiffen JC, Kober C, Green AR, Massie CE, Nangalia J, Lempidaki S, Döhner H, Döhner K, Bray SJ, McDermott U, Papaemmanuil E, Campbell PJ, and Adams DJ (2013) Inactivating *CUX1* mutations promote tumorigenesis. **Nature Genetics** 46 (1): 33-38.

Nik-Zainal S, Van Loo P, Wedge DC, **Alexandrov LB**, Greenman CD, Lau KW, Raine K, Jones D, Marshall J, Ramakrishna M, Shlien A, Cooke SL, Hinton J, Menzies A, Stebbings LA, Leroy C, Jia M, Rance R, Mudie LJ, Gamble SJ, Stephens PJ, McLaren S, Tarpey PS, Papaemmanuil E, Davies HR, Varela I, McBride DJ, Bignell GR, Leung K, Butler AP, Teague JW, Martin S, Jonsson G, Mariani O, Boyault S, Miron P, Fatima A, Langerod A, Aparicio SA, Tutt A, Sieuwerts AM, Borg A, Thomas G, Salomon AV, Richardson AL, Borresen-Dale AL, Futreal PA, Stratton MR, Campbell PJ, and Breast Cancer Working Group of the International Cancer Genome C (2012) The life history of 21 breast cancers. **Cell** 149:994-1007

Murchison EP, Schulz-Trieglaff OB, Ning Z, **Alexandrov LB**, Bauer MJ, Fu B, Hims M, Ding Z, Ivakhno S, Stewart C, Ng BL, Wong W, Aken B, White S, Alsop A, Becq J, Bignell GR, Cheetham RK, Cheng W, Connor TR, Cox AJ, Feng ZP, Gu Y, Grocock RJ, Harris SR, Khrebtukova I, Kingsbury Z, Kowarsky M, Kreiss A, Luo S, Marshall J, McBride DJ, Murray L, Pearse AM, Raine K, Rasolonjatovo I, Shaw R, Tedder P, Tregidgo C, Vilella AJ, Wedge DC, Woods GM, Gormley N, Humphray S, Schroth G, Smith G, Hall K, Searle SM, Carter NP, Papenfuss AT, Futreal PA, Campbell PJ, Yang F, Bentley DR, Evers DJ, and Stratton MR (2012) Genome sequencing and analysis of the Tasmanian devil and its transmissible cancer. **Cell** 148:780-791