Somatic Mutations Analysis for IBC WGS data

March 07, 2016

Xiaotong Li

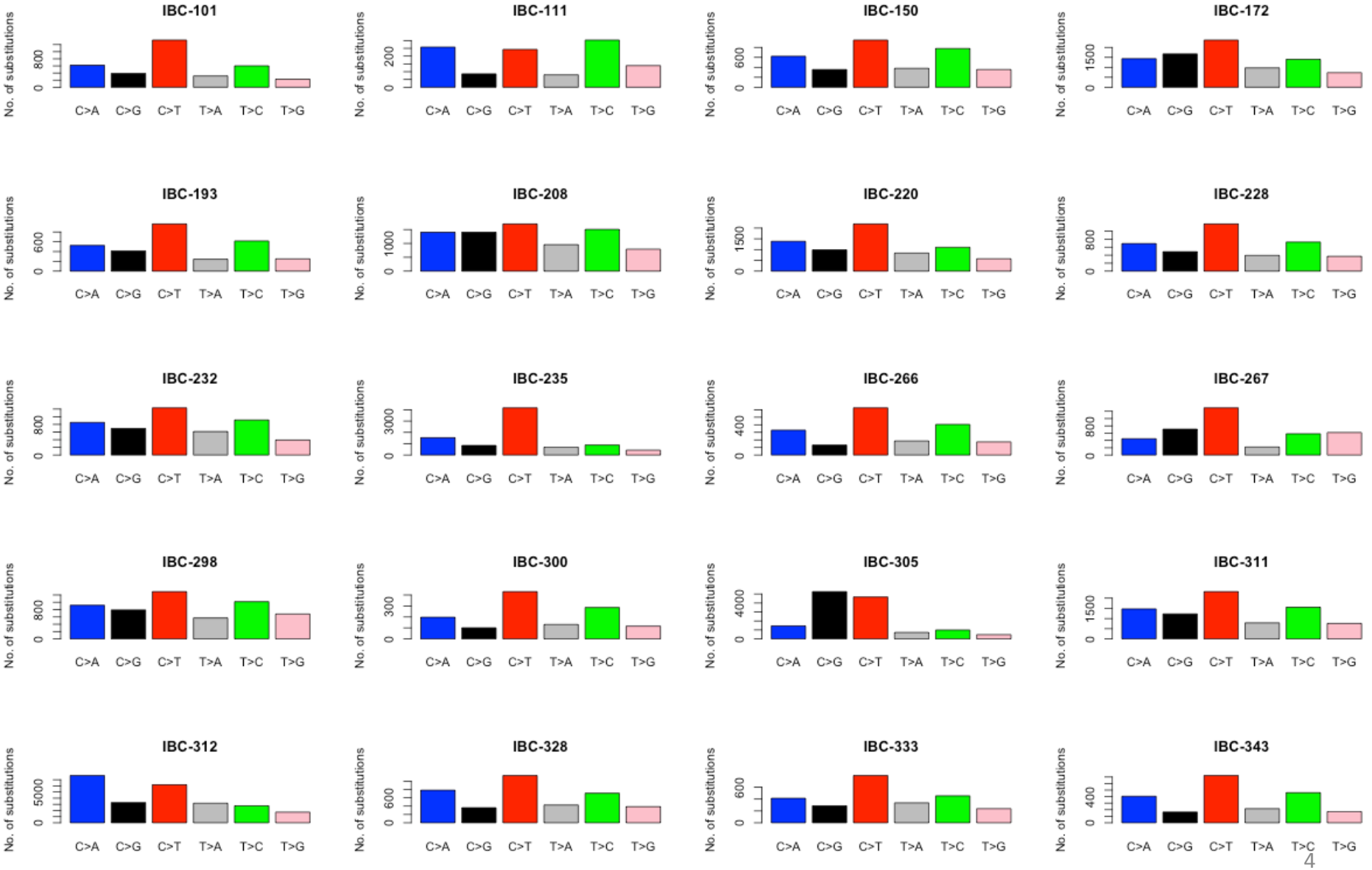
Overview of somatic variants analysis

- **Functional annotation & prioritization**
  - coding vs. non-coding
    - coding: synonymous; non-synonymous; LoF
    - noncoding: promoter; enhancer; TF binding site; ncRNA; etc.
  - target gene ("hub"); motif change; negative selection; recurrence
  - identification of candidate drivers
- **Deciphering somatic mutation profiles**
  - base substitution mutation spectra
  - tri-nucleotide context mutation spectra
  - mutational signatures
- **Frequency of mutated genes**
  - most frequently mutated genes
  - frequency of known cancer genes
- **Biological pathway & network analysis**

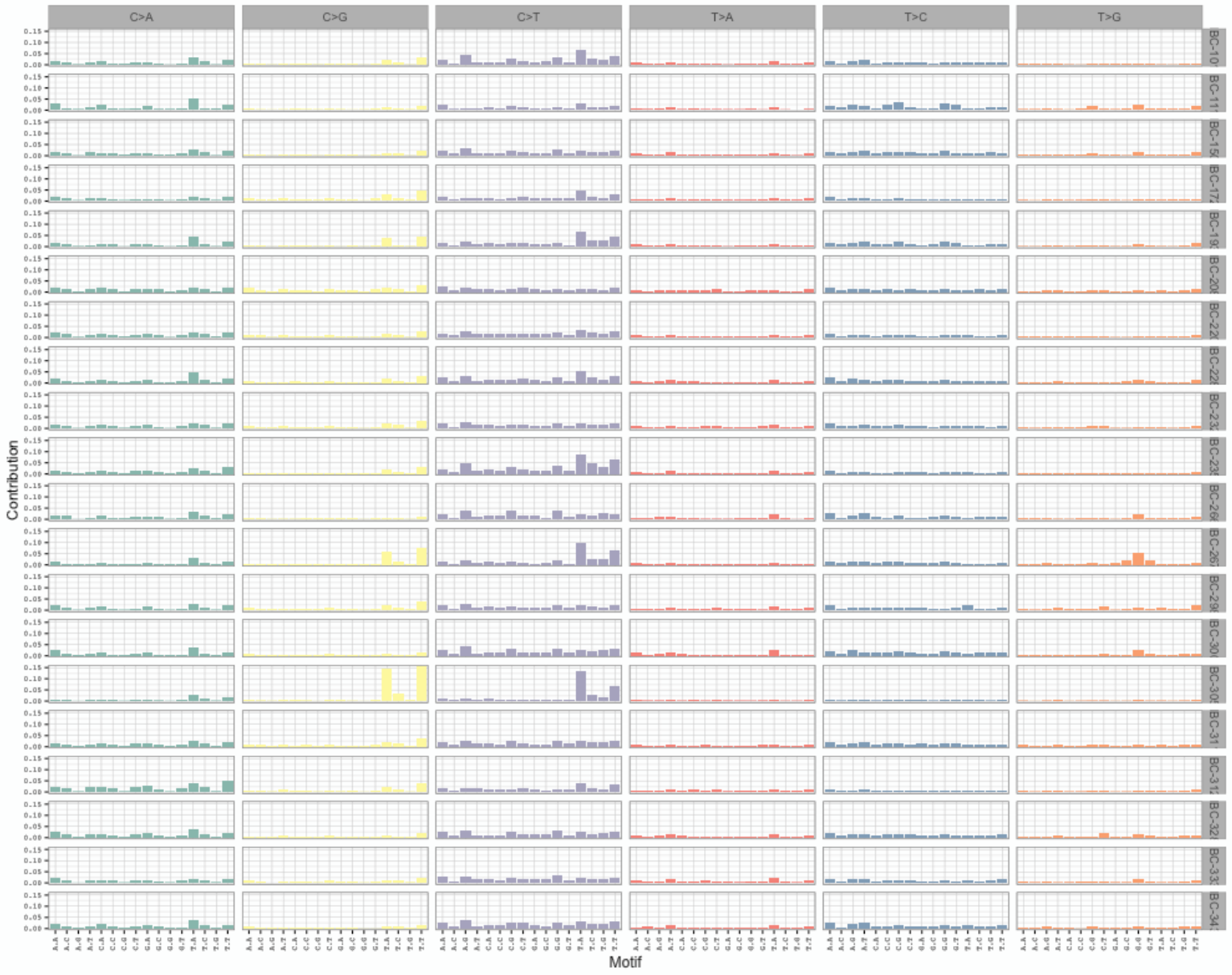
Overview

- Somatic variants calling pipeline
- Functional annotation & prioritization
  - coding vs. non-coding
    - coding: synonymous; non-synonymous; LoF
    - noncoding: promoter; enhancer; TF binding site; ncRNA; etc.
  - target gene ("hub"); motif change; negative selection; recurrence
  - identification of candidate drivers
- **Deciphering somatic mutation profiles**
  - base substitution mutation spectra
  - tri-nucleotide context mutation spectra
  - mutational signatures
- **Frequency of mutated genes**
  - most frequently mutated genes
  - frequency of known cancer genes
- Biological pathway & network analysis

Base substitution mutation spectra



Mutational Spectrum of 20 IBC samples



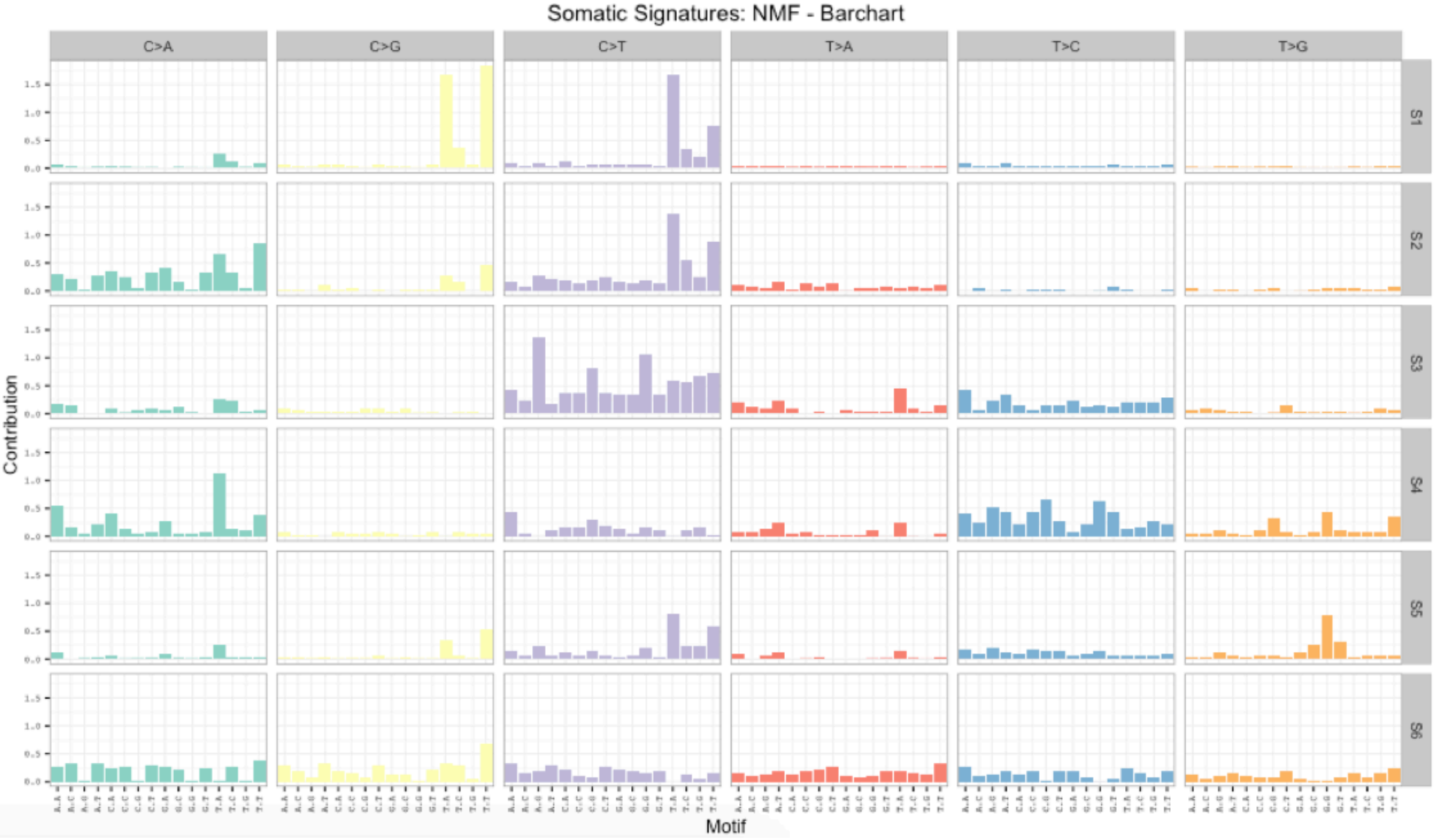
Inferring mutational signature

$$M = S \times E$$

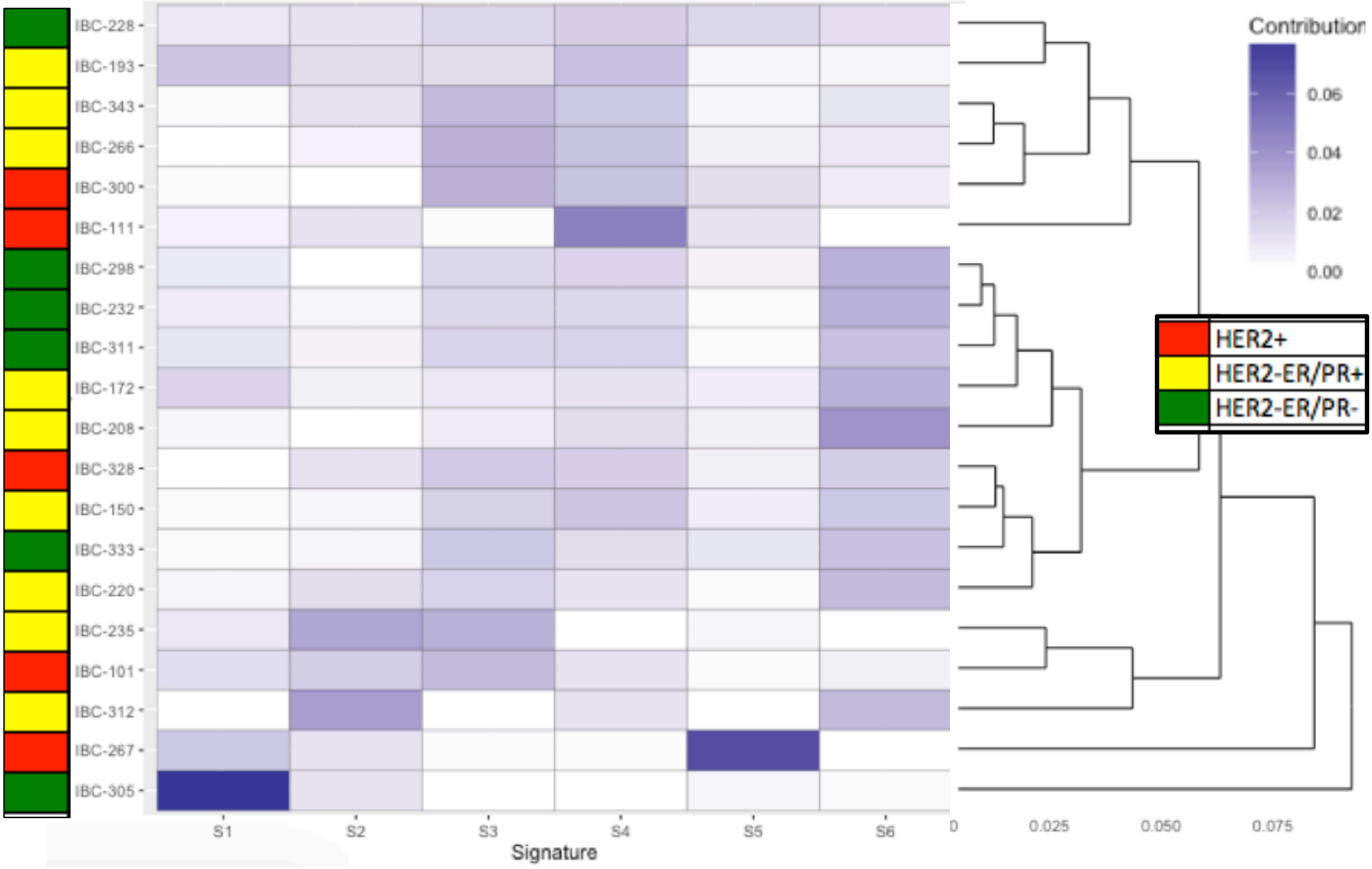
- M: mutational spectrum of all samples
  - 96 tri-nucleotide mutation contexts
- S: mutational signatures
  - 96 tri-nucleotide mutation contexts
- E: number of mutations contributed by each signature

- Goal: Finding **S** and **E** while only M is known
- Method:
  - Non-negative matrix factorization (NMF)
  - Principal component analysis (PCA)

Decomposition: Inferring somatic signatures

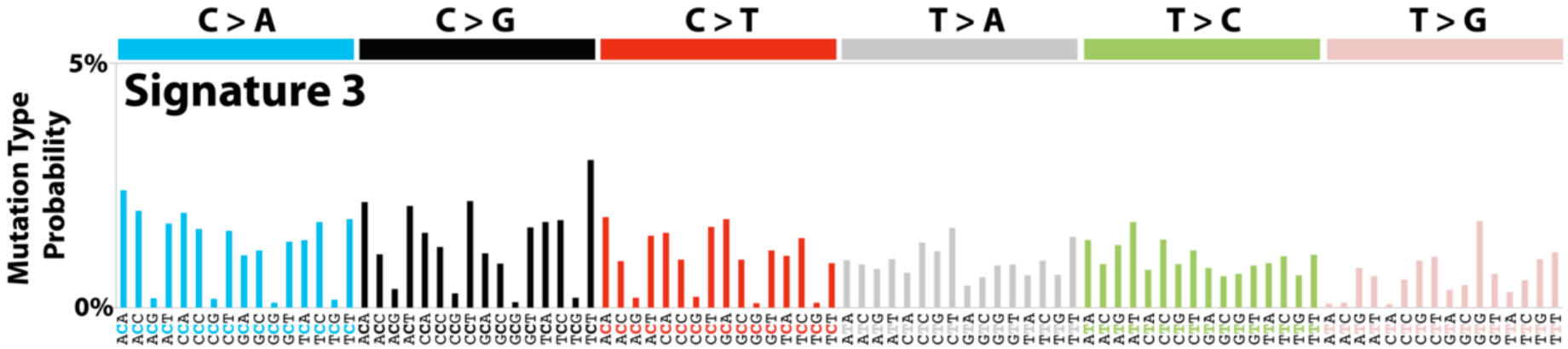


Clustering: by the contributions of signatures in each sample



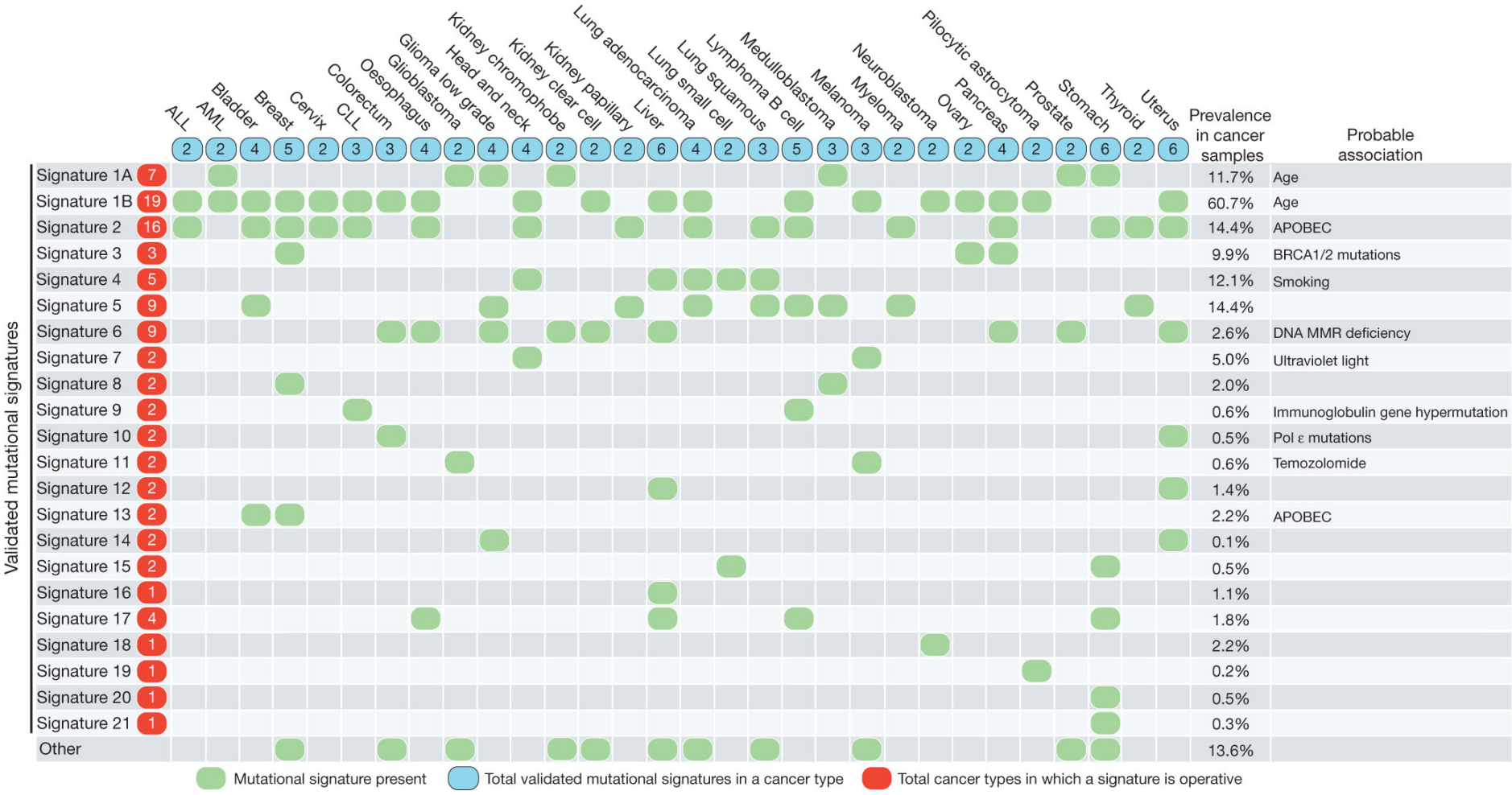
Signatures of mutational processes in human cancer

- COSMIC (Catalogue of somatic mutations in cancer) delivered **30** validated signatures
  - based on an analysis of 10,952 exomes and 1,048 whole-genomes across 40 distinct types of human cancer
  - profile & additional information (cancer type, proposed aetiology, etc.)



- **Cancer types:** found in breast, ovarian, and pancreatic cancers.
- **Proposed aetiology:** associated with failure of DNA double-strand break-repair by homologous recombination.
- **Comments:** strongly associated with germline and somatic BRCA1 and BRCA2 mutations in breast, pancreatic, and ovarian cancers. In pancreatic cancer, responders to platinum therapy usually exhibit Signature 3 mutations.

The presence of mutational signatures across human cancer types.



LB Alexandrov *et al. Nature* **000**, 1-7 (2013) doi:10.1038/nature12477

nature

Decomposition of mutational spectrum by validated mutational signatures

$$M = S \times E \qquad \text{or} \qquad \min \left\| \vec{\mathbf{M}} - \sum_{i=1}^{q} \left(\vec{\mathbf{S}}_i \times E_i \right) \right\|_2^F$$

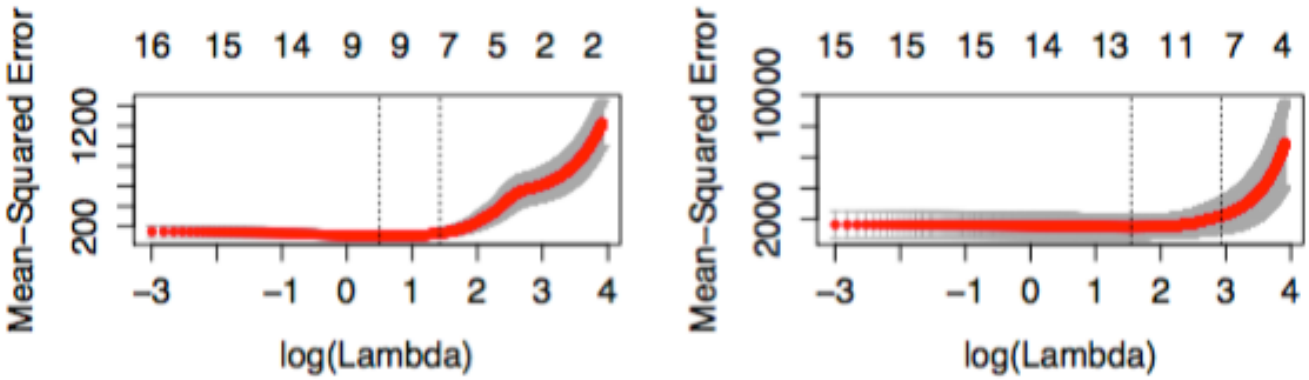
- M: mutational spectrum of all samples
  - 96 tri-nucleotide mutation contexts
- S: **30 validated** mutational signatures
  - 96 tri-nucleotide mutation contexts
- E: number of mutations contributed by each signature

- Goal: Finding **optimal E** while M and S are known
- Method:
  - Generalized linear model (GLM)
  - Linear programming/optimization

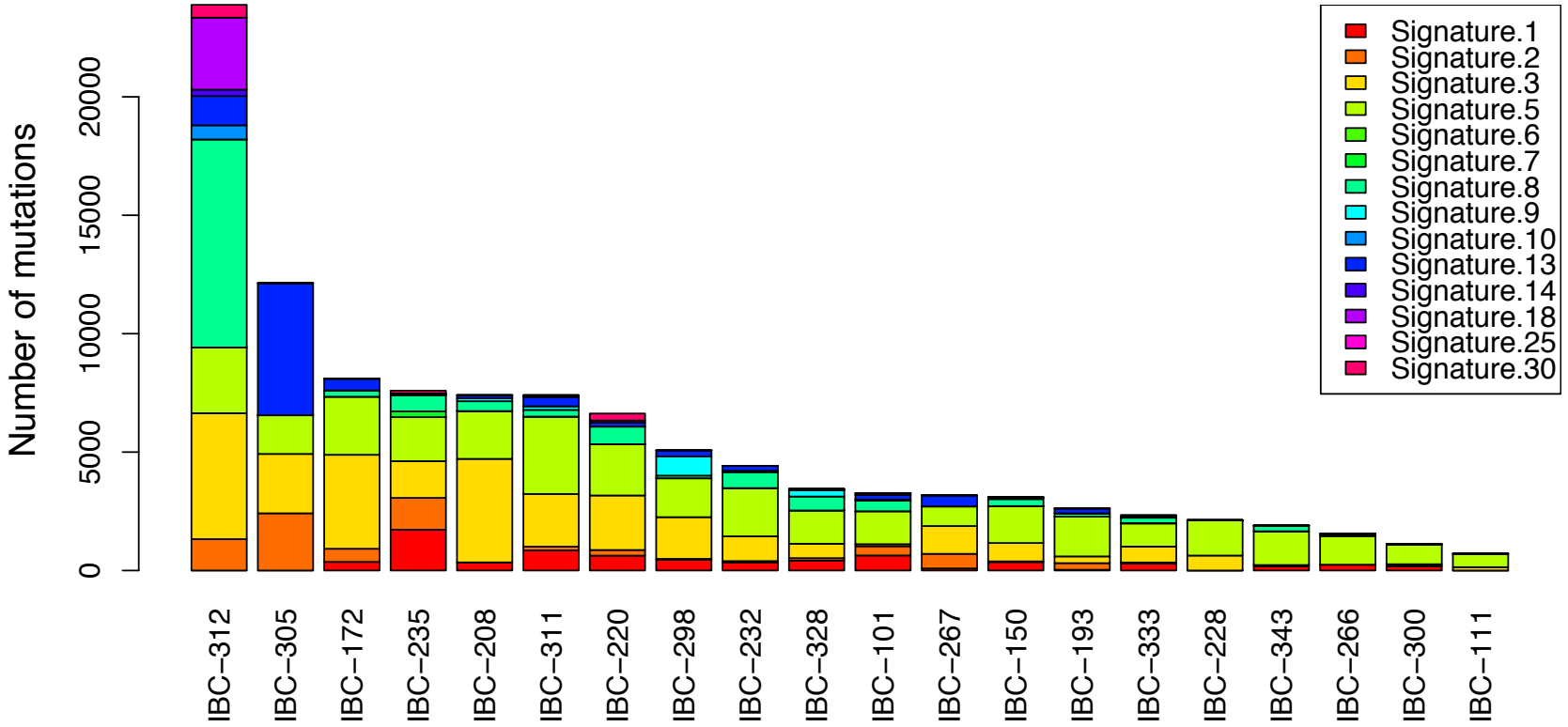
Generalized Linear Model

$$M = S \times E$$

- Fit a generalized linear model via **penalized** maximum likelihood.
  - *E* should be non-negative
  - no intercept for this linear model
  - promote sparsity in *E*
- The regularization path is computed for lasso penalty at a grid of values for the regularization parameter lambda.
  - Cross validation (CV) for finding optimal lambda value



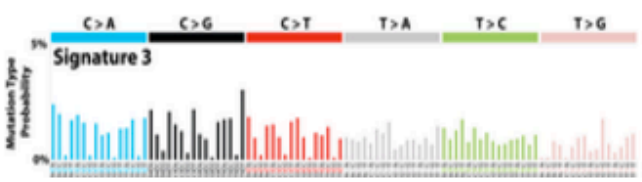
Contributions of mutational signatures to individual samples



IBC-312	IBC-305	IBC-172	IBC-235	IBC-208	IBC-311	IBC-220	IBC-298	IBC-232	IBC-328	IBC-101	IBC-267	IBC-150	IBC-193	IBC-333	IBC-228	IBC-343	IBC-266	IBC-300	IBC-111
23885	12134	8096	7595	7405	7405	6629	5072	4418	3451	3269	3173	3094	2619	2346	2126	1898	1546	1114	709
24819	13571	8596	8561	7770	8127	7081	5475	4693	3805	3477	3872	3442	3011	2532	4041	2139	1851	1262	1109
96.24%	89.41%	94.18%	88.72%	95.30%	91.12%	93.62%	92.64%	94.14%	90.70%	94.02%	81.95%	89.89%	86.98%	92.65%	52.61%	88.73%	83.52%	88.27%	63.93%

Dominant mutational signatures

Signature 3



Signature 3

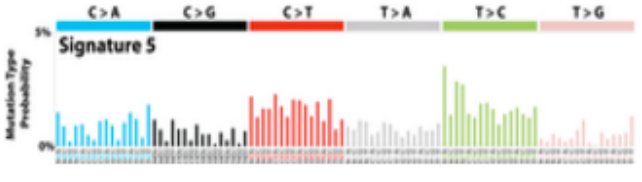
Cancer types: Signature 3 has been found in breast, ovarian, and pancreatic cancers.

Proposed aetiology: Signature 3 is associated with failure of DNA double-strand break-repair by homologous recombination.

Additional mutational features: Signature 3 associates strongly with elevated numbers of large (longer than 3bp) insertions and deletions with overlapping microhomology at breakpoint junctions.

Comments: Signature 3 is strongly associated with germline and somatic BRCA1 and BRCA2 mutations in breast, pancreatic, and ovarian cancers. In pancreatic cancer, responders to platinum therapy usually exhibit Signature 3 mutations.

Signature 5



Signature 5

Cancer types: Signature 5 has been found in all cancer types and most cancer samples.

Proposed aetiology: The aetiology of Signature 5 is unknown.

Additional mutational features: Signature 5 exhibits transcriptional strand bias for T>C substitutions at ApTpN context.

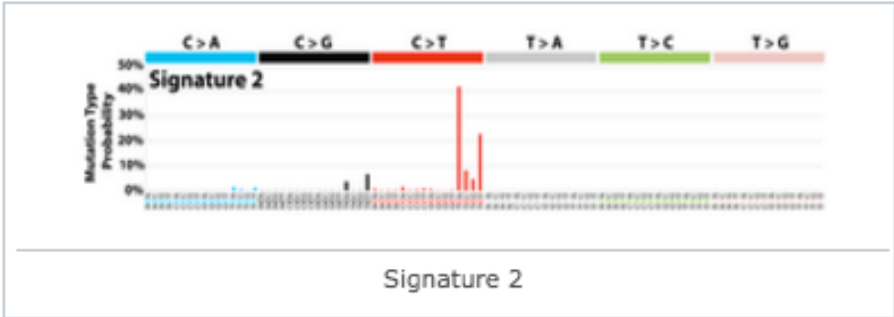
Comments: N/A

- Signature 5 is observed in 14.4% cancer samples:
  - bladder cancer, glioma, kidney, lung lymphoma B cell, medulloblastoma, myeloma, thyroid

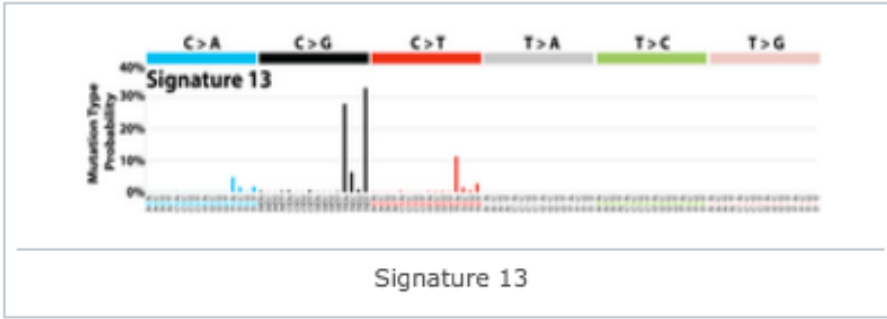
Not prevalent in breast cancer samples!

Coupled mutational signatures

Signature 2



Signature 13



- Signature 2 is usually found in the same samples as Signature 13.
- found in 22 cancer types and seems to be commonest in cervical and bladder cancers
- attributed to activity of the **AID/APOBEC** family of cytidine deaminases
- It has been proposed that activation of AID/APOBEC cytidine deaminases is due to
  - viral infection
  - retrotransposon jumping
  - tissue inflammation

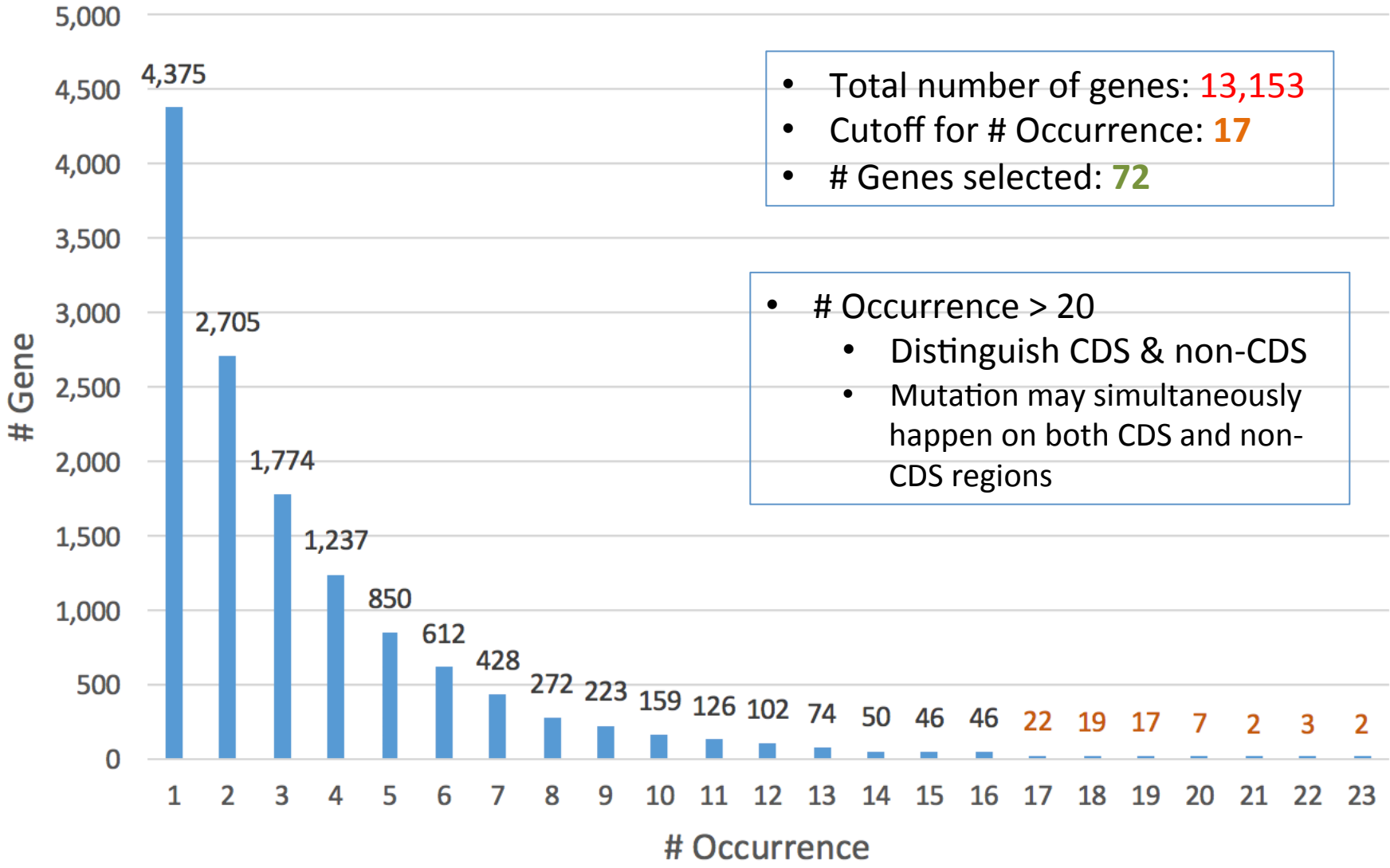
Overview

- Somatic variants calling pipeline
- Functional annotation & prioritization
  - coding vs. non-coding
    - coding: synonymous; non-synonymous; LoF
    - noncoding: promoter; enhancer; TF binding site; ncRNA; etc.
  - target gene ("hub"); motif change; negative selection; recurrence
  - identification of candidate drivers
- **Deciphering somatic mutation profiles**
  - base substitution mutation spectra
  - tri-nucleotide context mutation spectra
  - mutational signatures
- **Frequency of mutated genes**
  - most frequently mutated genes
  - frequency of known cancer genes
- Biological pathway & network analysis

Define: mutated genes

- Coding region
- Noncoding region
  - Intron
  - Promoter
  - UTR
  - Transcription factor binding site
  - DNase1 hypersensitive sites
  - ncRNA
  - Pseudogene
  - Enhancer
- Coding & Noncoding region

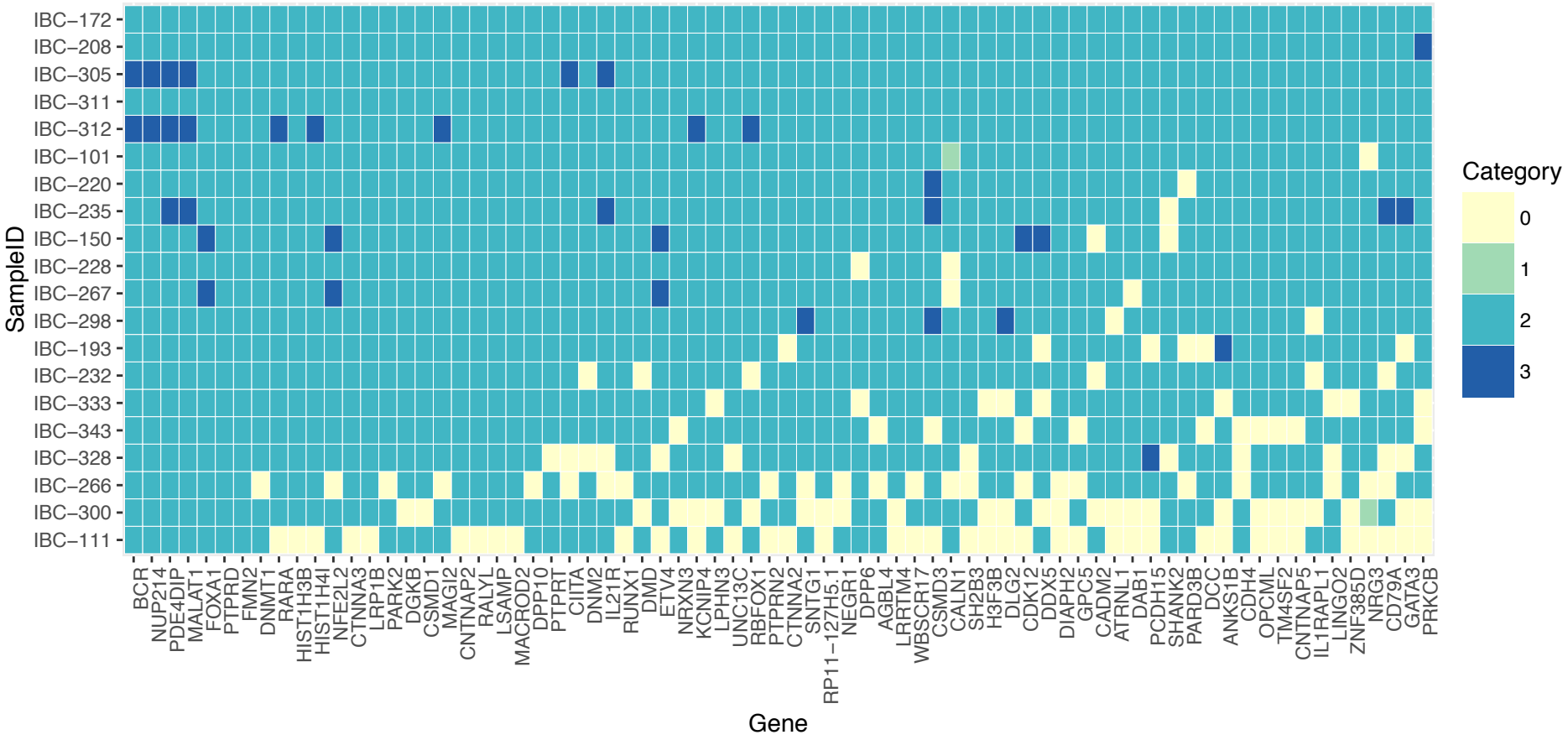
Frequently mutated genes



- Total number of genes: 13,153
- Cutoff for # Occurrence: 17
- # Genes selected: 72

- # Occurrence > 20
  - Distinguish CDS & non-CDS
  - Mutation may simultaneously happen on both CDS and non-CDS regions

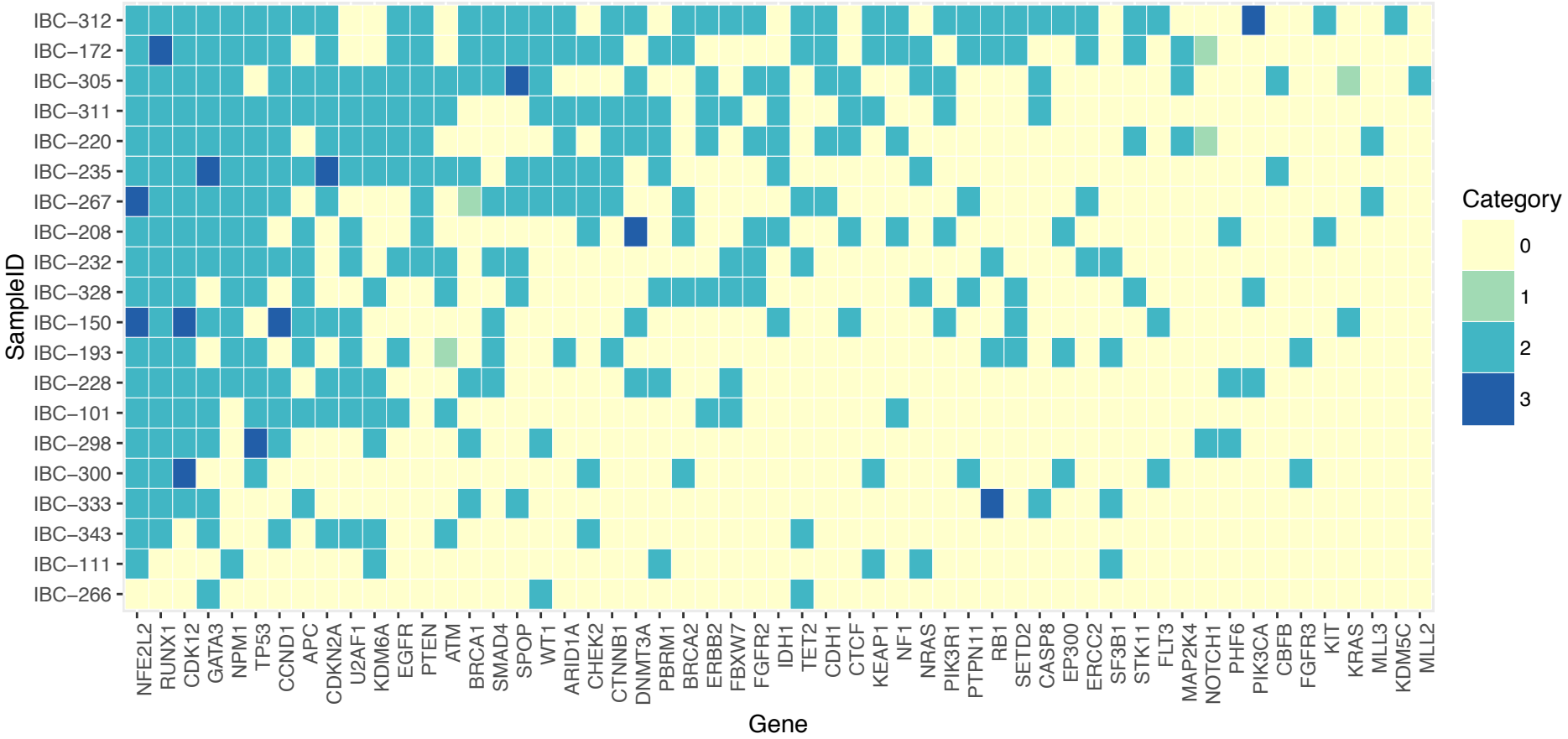
Most frequently mutated genes



High confidence pan-cancer genes

- Pan-cancer gene list
  - Cancer Gene Census
  - MuSIC
  - OncodriveFM
  - OncodriveCLUST
  - ActiveDriver
  - MutSig
  - High confidence driver/candidate driver
- High confidence driver & Selected in >=4 sources above
  - TP53: high confidence driver & selected in all 6 sources
  - ARID1A: high confidence driver & selected in 4 sources
    - CGC, MuSIC, OncodriveFM, MutSig
- 64 pan-cancer driver genes selected
  - 9 genes do **NOT** have any mutation reported in all 20 samples
  - AKT1, ATRX, BAP1, BRAF, CEBPA, HRAS, IDH2, PPP2R1A, VHL
  - 55 genes eventually selected

Mutational frequency of pan-cancer genes



NFE2L2 encodes a transcription factor which is a member of a small family of basic leucine zipper (bZIP) proteins. The encoded transcription factor regulates genes which contain antioxidant response elements (ARE) in their promoters; many of these genes encode proteins involved in **response to injury and inflammation** which includes the production of free radicals.

Overview

IBC vs. Non-IBC

- **Non-IBC samples selection**
  - ER, PR, HER2 status
  - age
  - race
- **Somatic variants calling pipeline**
  - exactly same as IBC pipeline
- **Functional annotation & prioritization**
- **Deciphering somatic mutation profiles**
- **Frequency of mutated genes**
- **Biological pathway & network analysis**