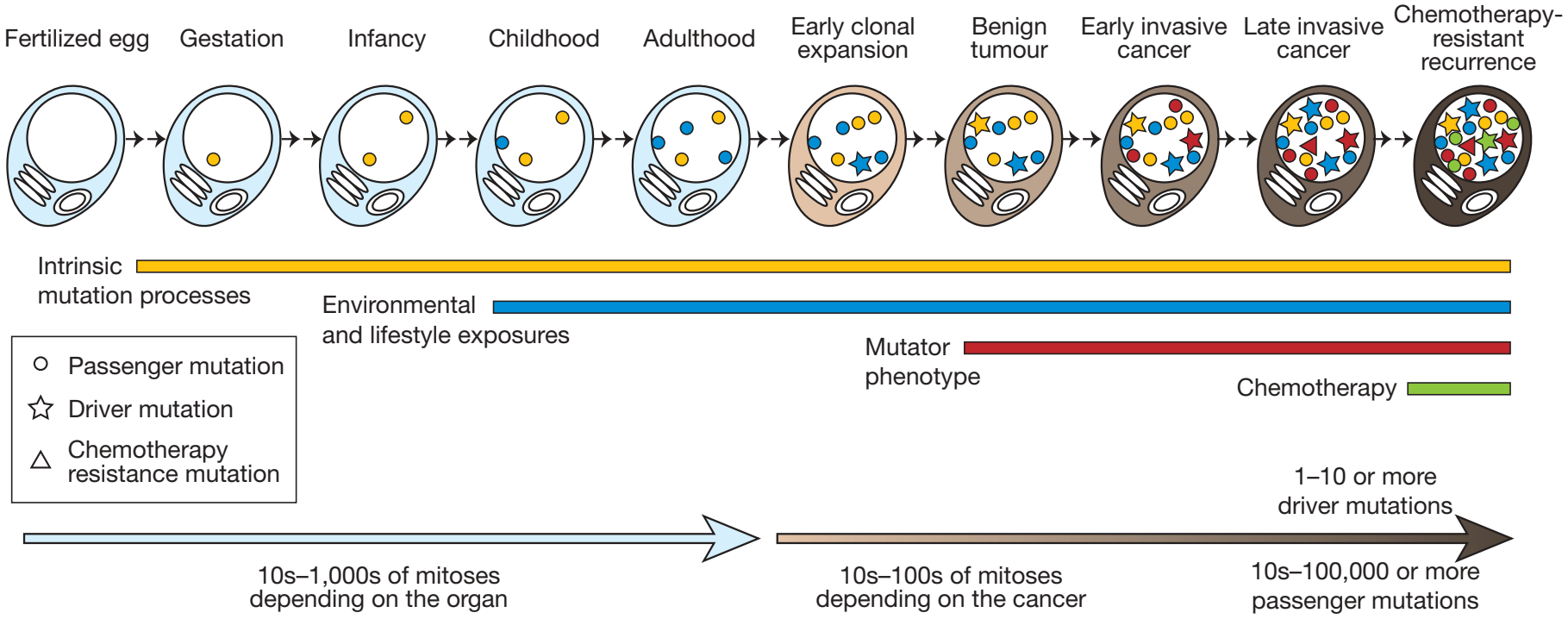
Deciphering mutational signatures in cancer with the hierarchical Dirichlet process

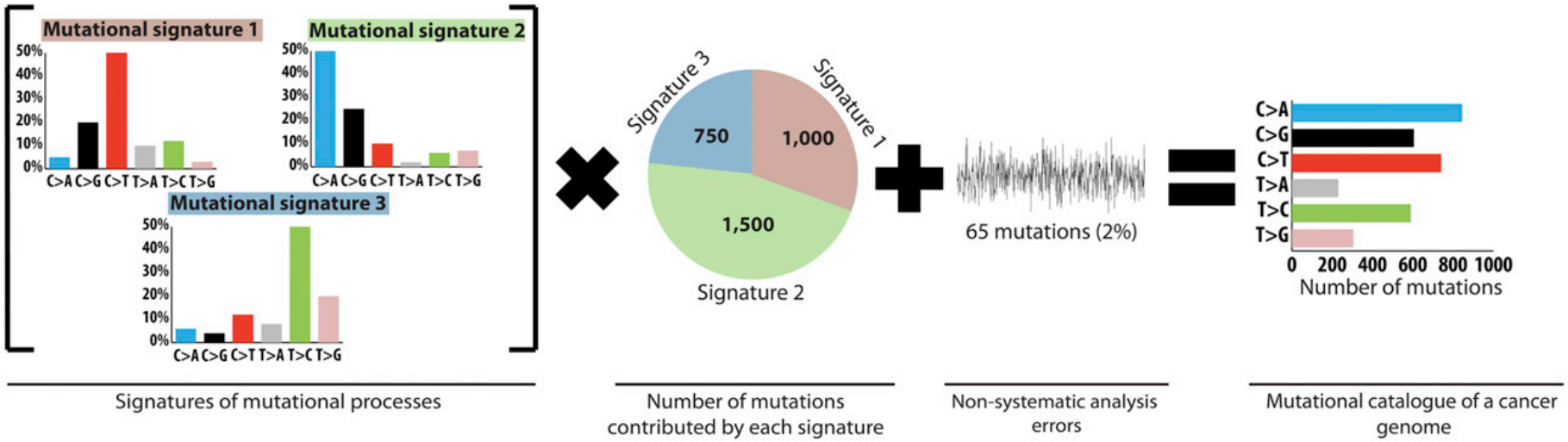
Nicola Roberts
Campbell group, Cancer Genome Project
Wellcome Trust Sanger Institute
Twitter: @NicolaRbrts, #quantgen

Cancer genomes are burdened with somatic mutations from a variety of mutational processes active during the lifetime of the cell



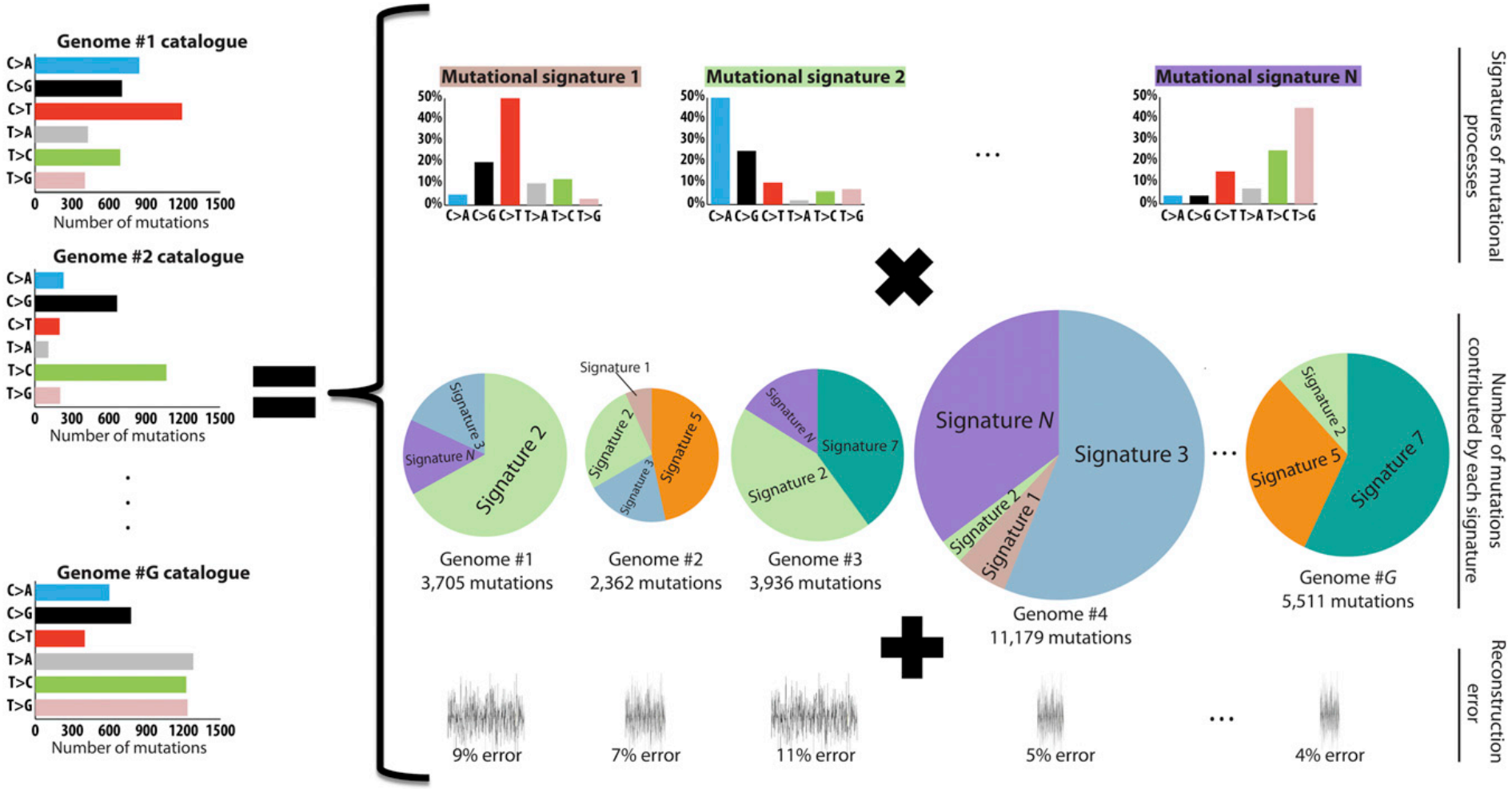
Stratton et al. (2009) Nature. doi:10.1038/nature07943

In a given cancer genome, the observed catalogue of somatic mutations is the sum of each mutational signature weighted by its activity during the lifetime of the cell lineage.



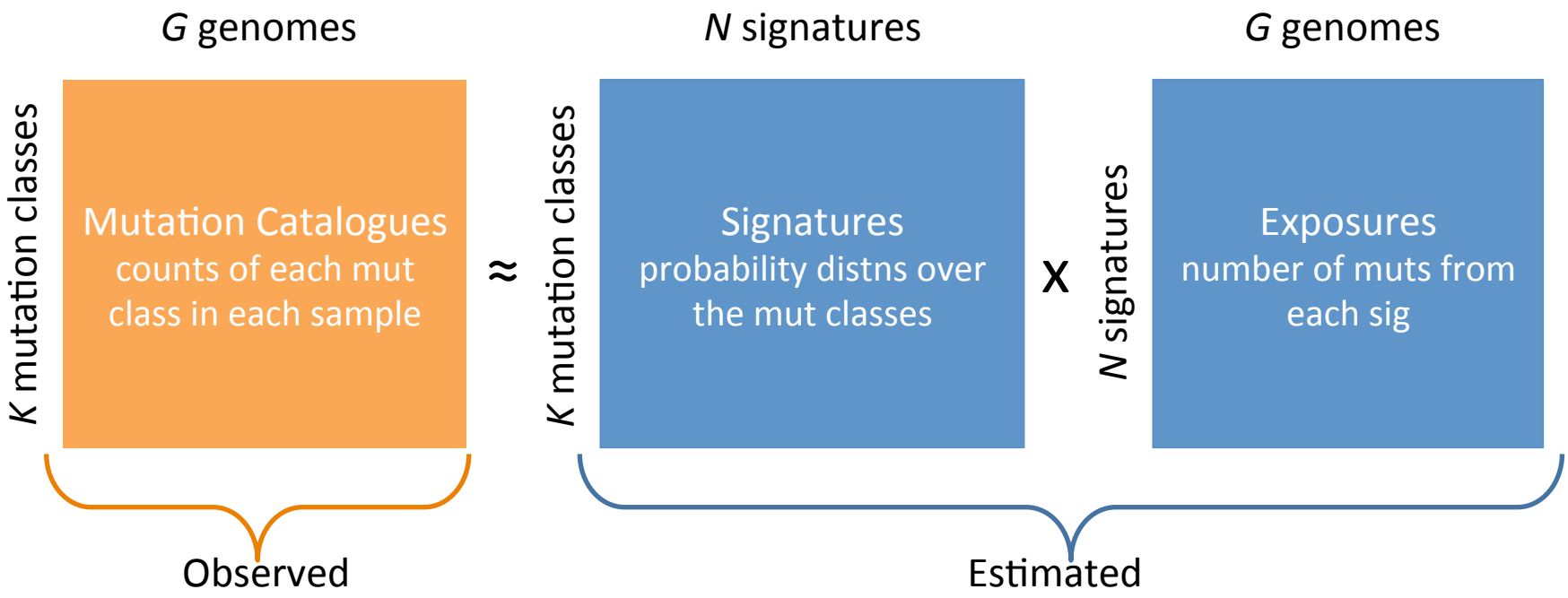
Alexandrov et al. (2013) Cell Reports. doi:10.1016/j.celrep.2012.12.008

Given a set of mutation catalogues from many cancer samples, can decipher signatures of the underlying mutational processes



Alexandrov et al. (2013) Cell Reports.
doi:10.1016/j.celrep.2012.12.008

Original approach: non-negative matrix factorization



Original papers:

- Alexandrov et al. Deciphering Signatures of Mutational Processes Operative in Human Cancer, Cell Reports (2013)
- Alexandrov et al. Signatures of mutational processes in human cancer, Nature (2013)

New approach: hierarchical Dirichlet process

Nonparametric Bayesian approach for mutational signatures analysis.

Advantages over NMF:

- Model **relationships between samples**
  - different cancer types, different samples from the same patient, different subclones from the same sample, etc.
- Statistical analysis of **significant differences** in signature prevalence across samples and groups
  - formal probabilistic model with credibility intervals for every parameter
- Simultaneous signature discovery and **matching to known signatures**

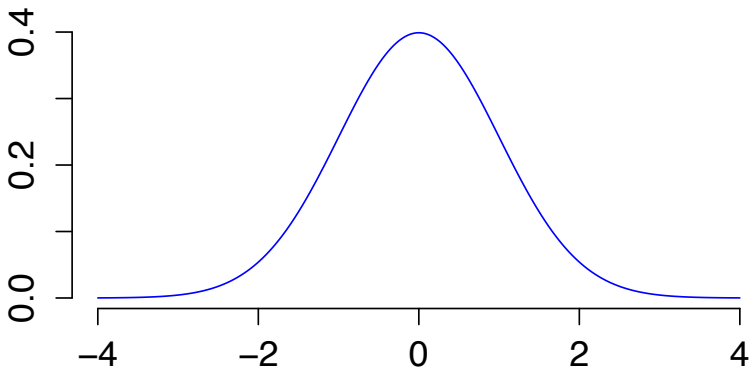
Original paper:
Teh et al. Hierarchical Dirichlet Processes, JASA (2006)

What is a Dirichlet process?

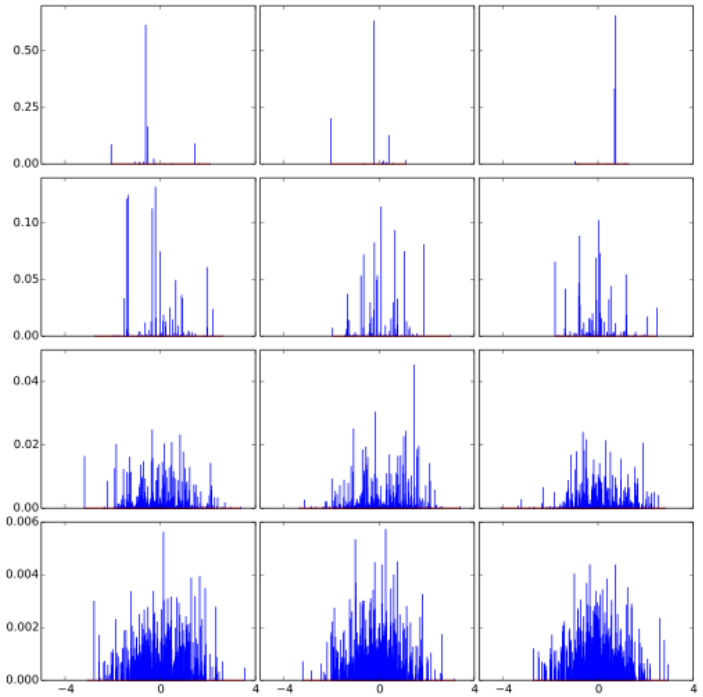
Defined by 'base' probability distribution *H* (and concentration parameter α)

A DP draw is a probability distribution built with random samples from *H* with increasingly small weights. It is a much clumpier and discretized distribution over the domain of *H*.

Example of a base distn *H*



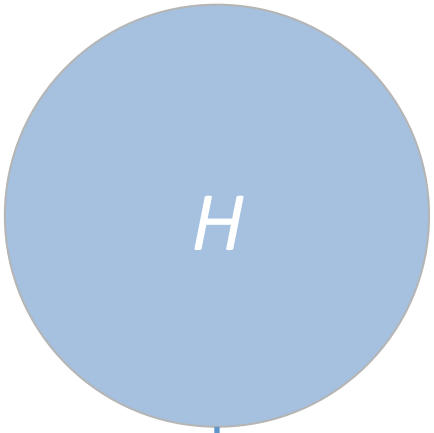
Examples of DP draws



http://commons.wikimedia.org/wiki/File:Dirichlet_process_draws.svg

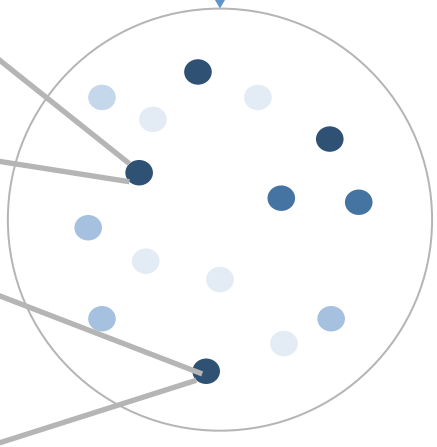
DP design for mutational signatures

Each signature is a discrete probability distribution over k mutation classes (e.g. k=6)



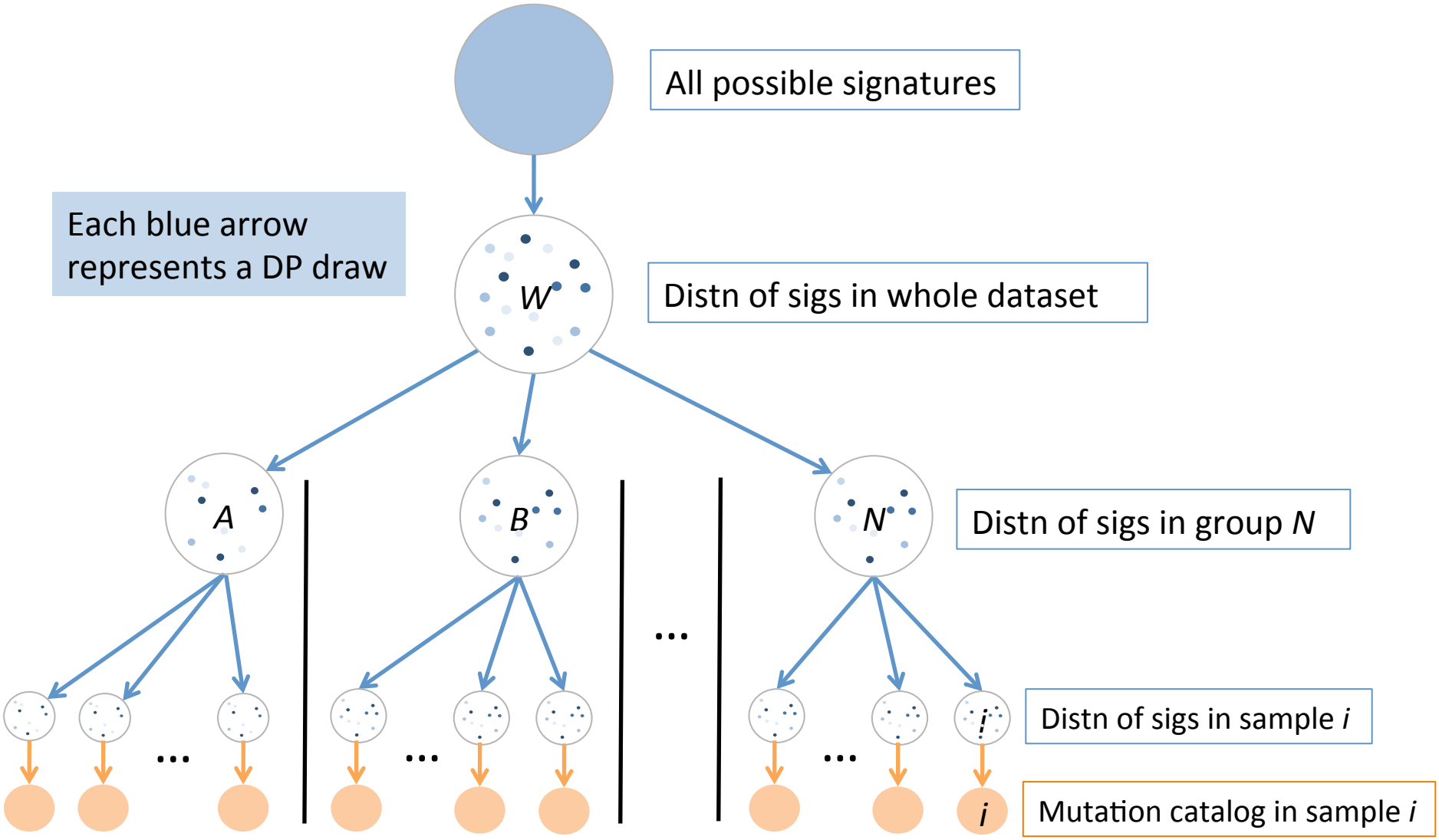
As a prior, base distribution H is uniform over all possible signatures (uniform Dirichlet)

Draw from a DP



Distribution of signatures in a collection of cancer genomes

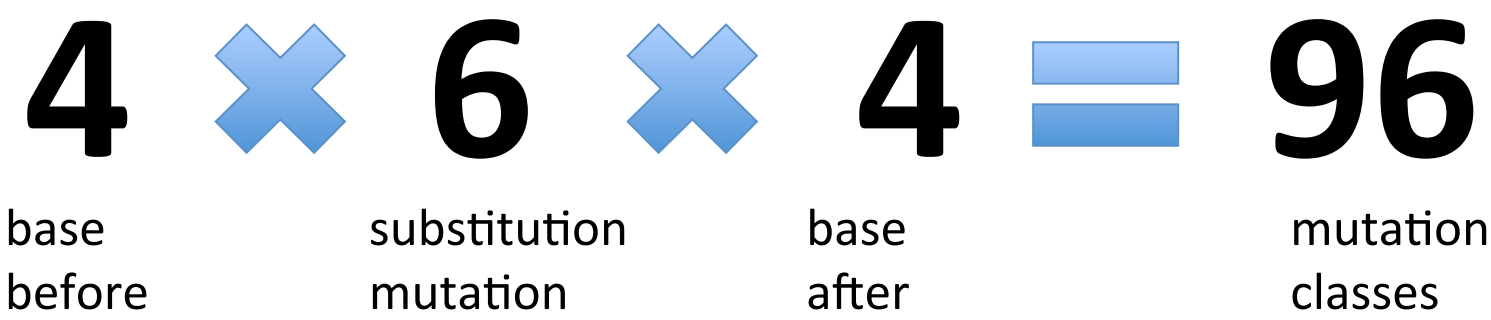
HDP design for multiple sample groups



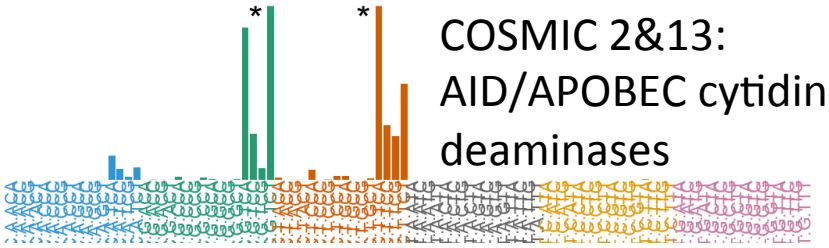
Example – TCGA exome seq data

Cancer	Breast	Colorect.	Glioma	Lung	Melanoma	Ovary	Prostate	Stomach	**Total**
Samples	844	559	217	636	396	471	330	212	**3665**
Total muts	39822	204630	20601	211762	280918	22307	15176	77345	**872561**

Tallied single base mutations in 96 categories defined by trinucleotide context

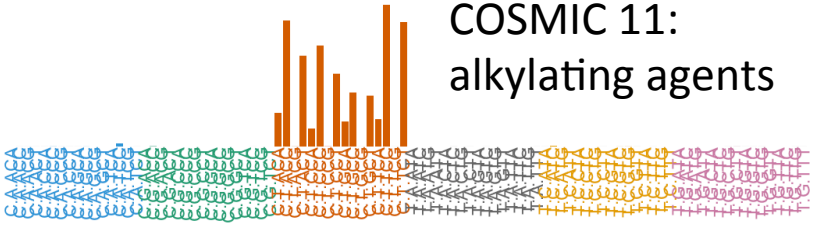


Signature 4



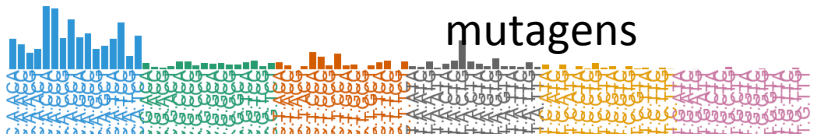
COSMIC 2&13:
AID/APOBEC cytidine deaminases

Signature 5



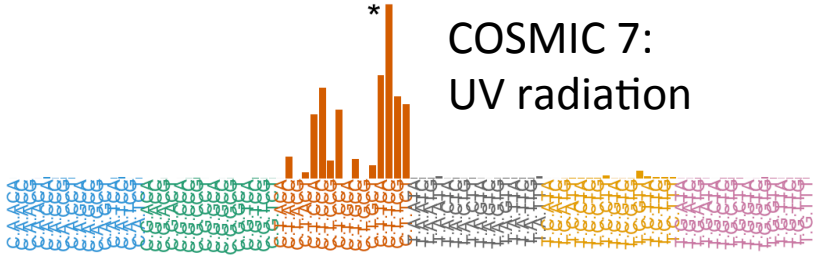
COSMIC 11:
alkylating agents

Signature 6



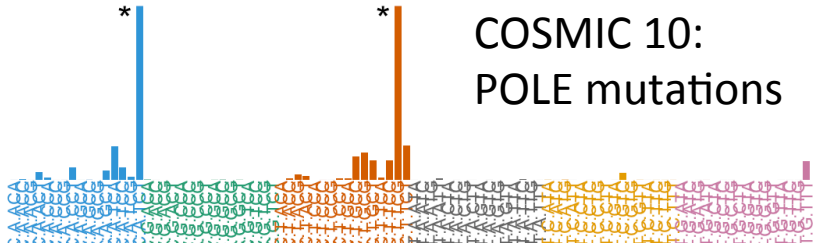
COSMIC 4:
tobacco mutagens

Signature 7



COSMIC 7:
UV radiation

Signature 8

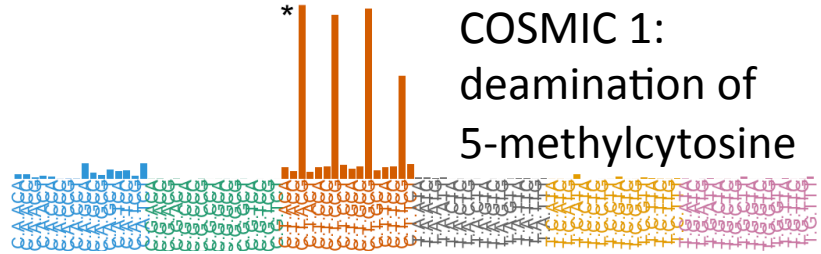


COSMIC 10:
POLE mutations

Signature 9

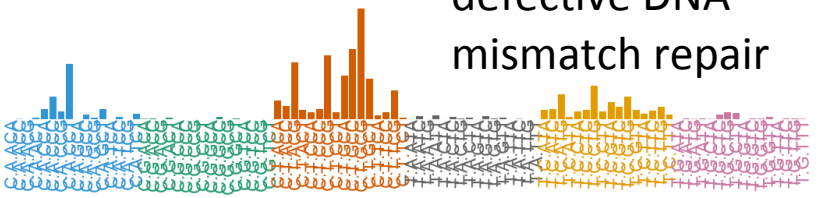
COSMIC 5:
intrinsic error?

Signature 10

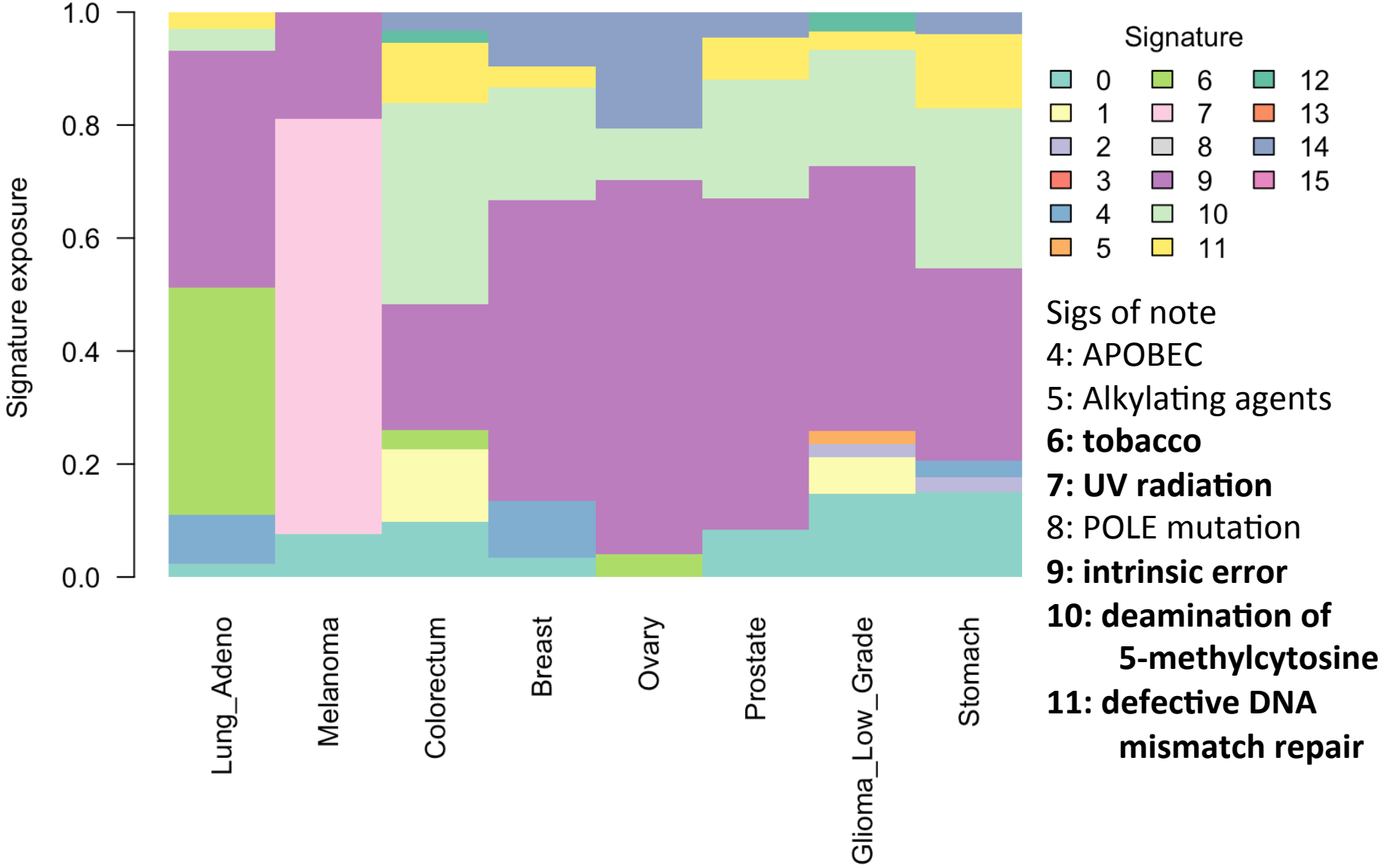


COSMIC 1:
deamination of 5-methylcytosine

Signature 11



COSMIC 6:
defective DNA mismatch repair



Sigs of note

4: APOBEC

5: Alkylating agents

6: tobacco

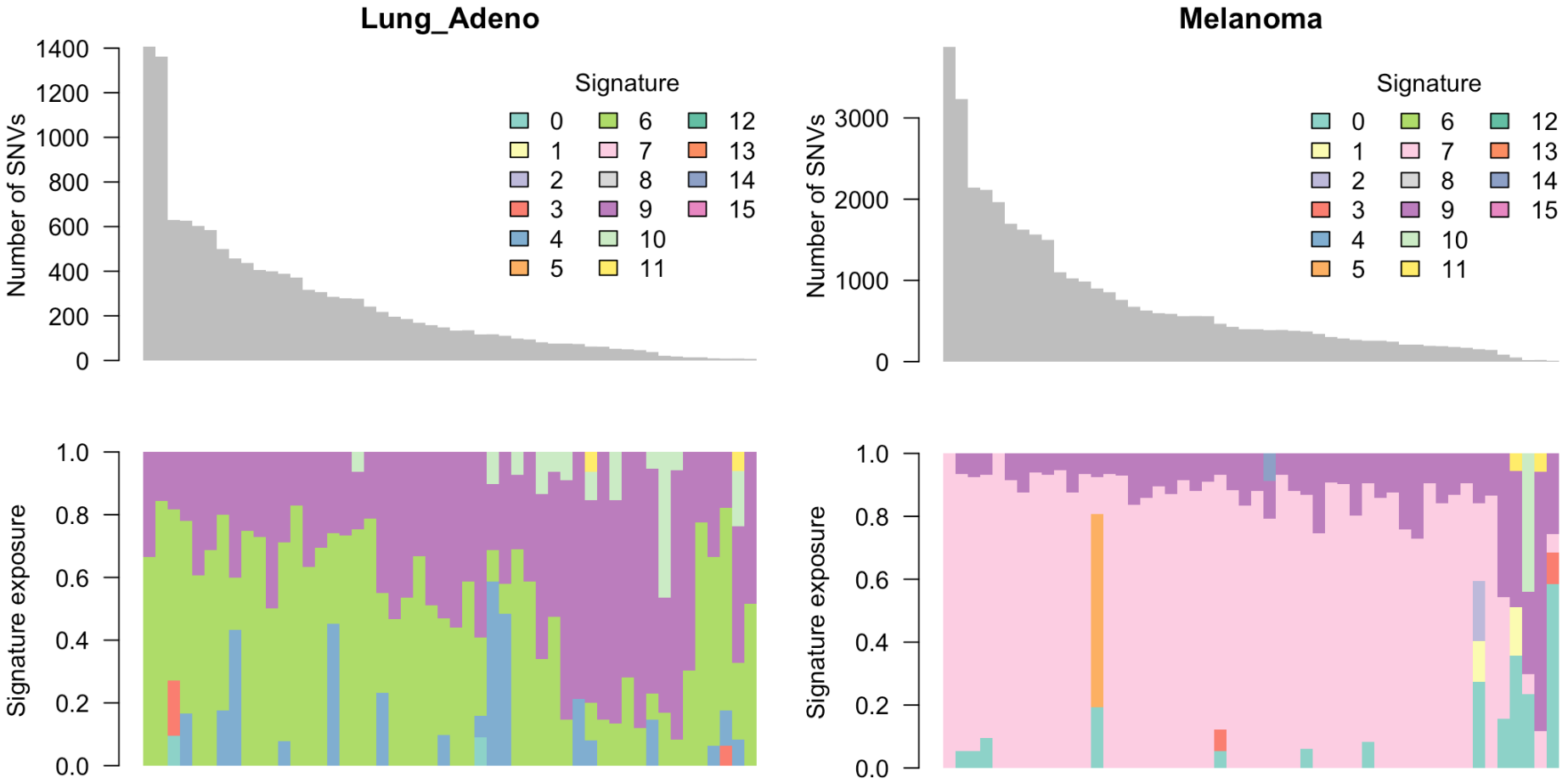
7: UV radiation

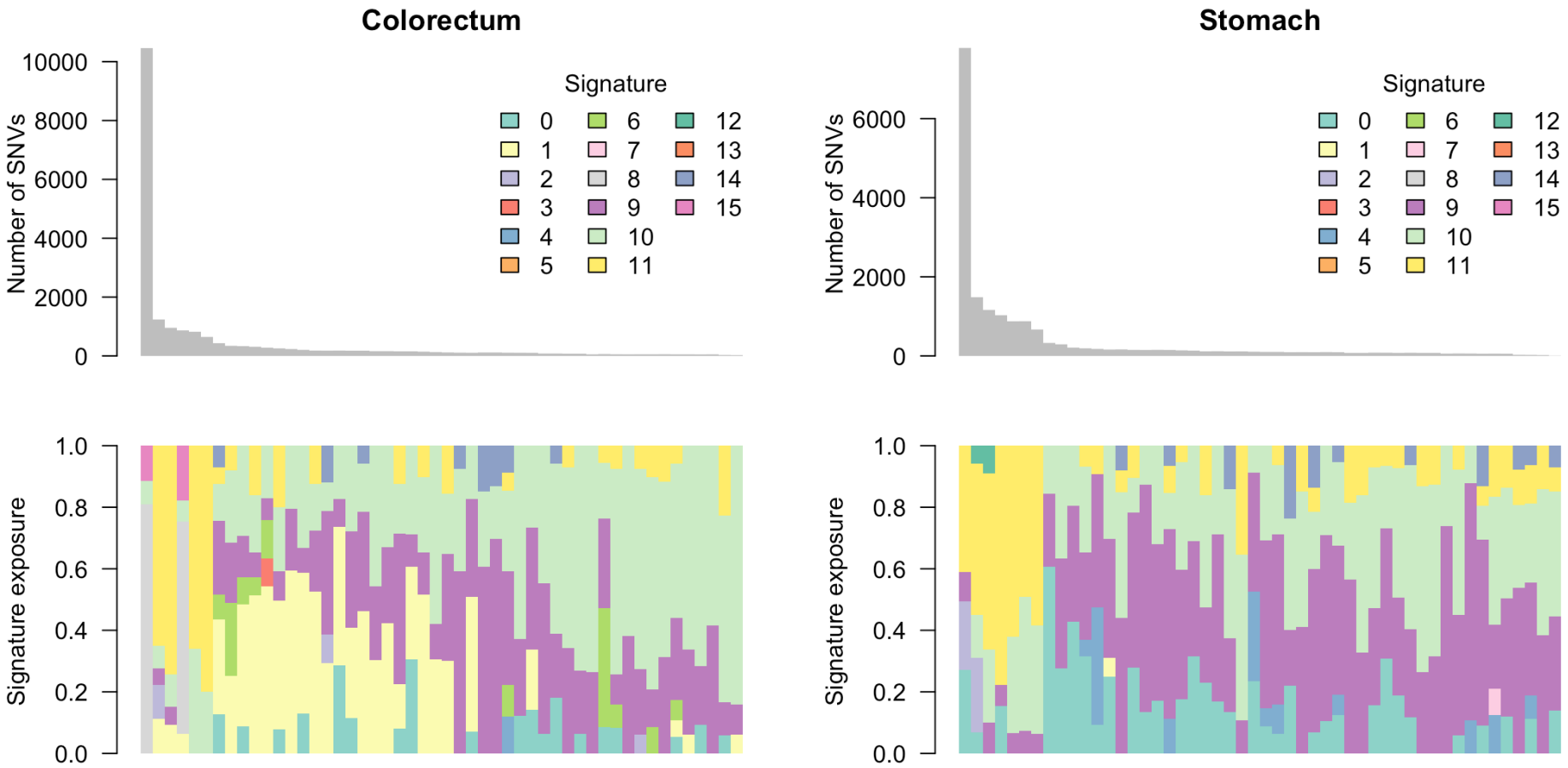
8: POLE mutation

9: intrinsic error

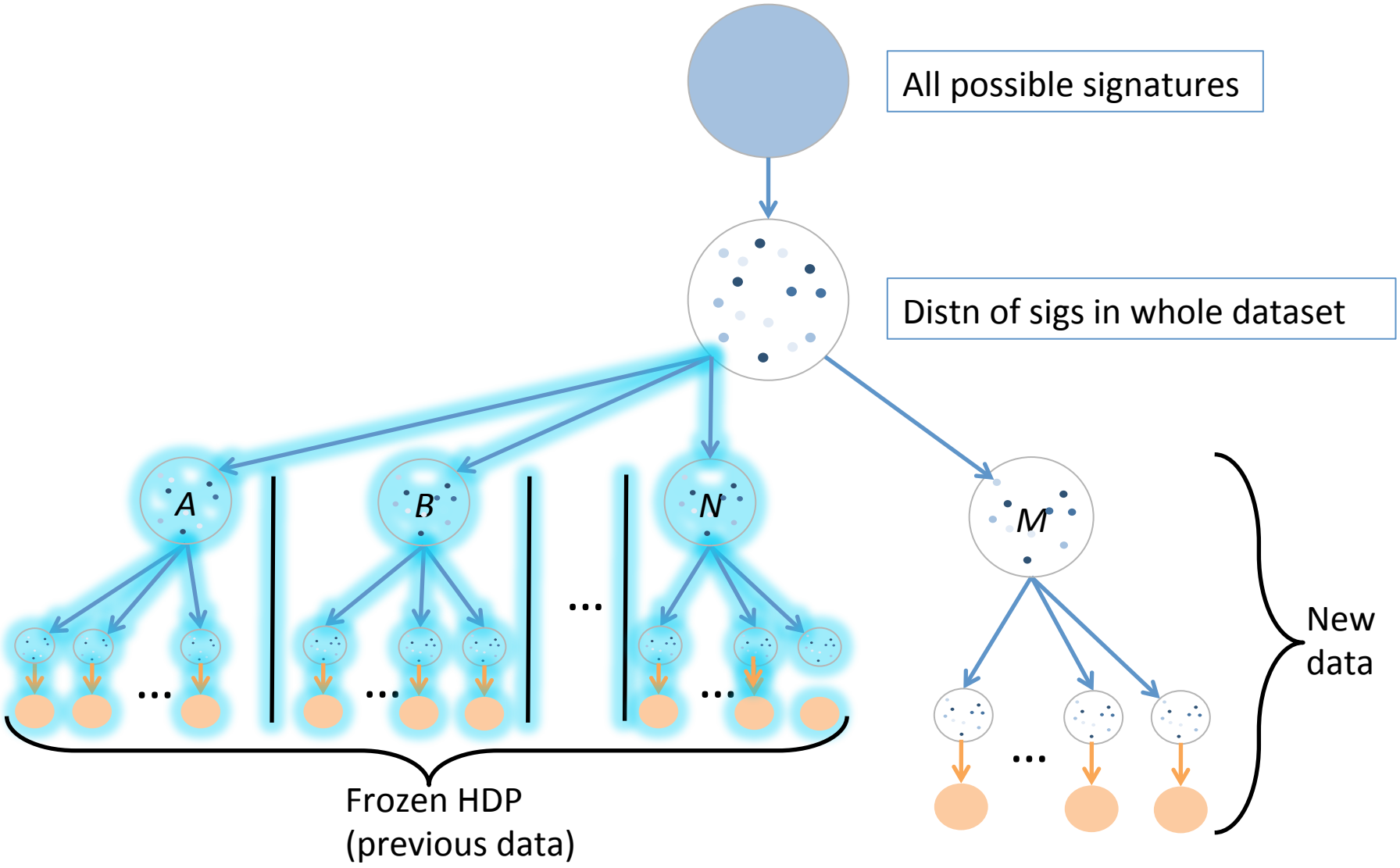
10: deamination of 5-methylcytosine

11: defective DNA mismatch repair


Lung_Adeno
Number of SNVs
1400
1200
1000
800
600
400
200
0
Signature
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
Signature exposure
1.0
0.8
0.6
0.4
0.2
0.0
Melanoma
Number of SNVs
3000
2000
1000
0
Signature
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
Signature exposure
1.0
0.8
0.6
0.4
0.2
0.0


Colorectum
Number of SNVs
10000
8000
6000
4000
2000
0
Signature
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
Signature exposure
1.0
0.8
0.6
0.4
0.2
0.0
Stomach
Number of SNVs
6000
4000
2000
0
Signature
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
Signature exposure
1.0
0.8
0.6
0.4
0.2
0.0

HDP design conditioned on previous data



Acknowledgements

Peter Campbell

Mike Stratton

Ludmil Alexandrov

Moritz Gerstung

David Wedge

My HDP R package available at https://github.com/nicolaroberts/hdp