ETH zürich

SIB Swiss Institute of Bioinformatics

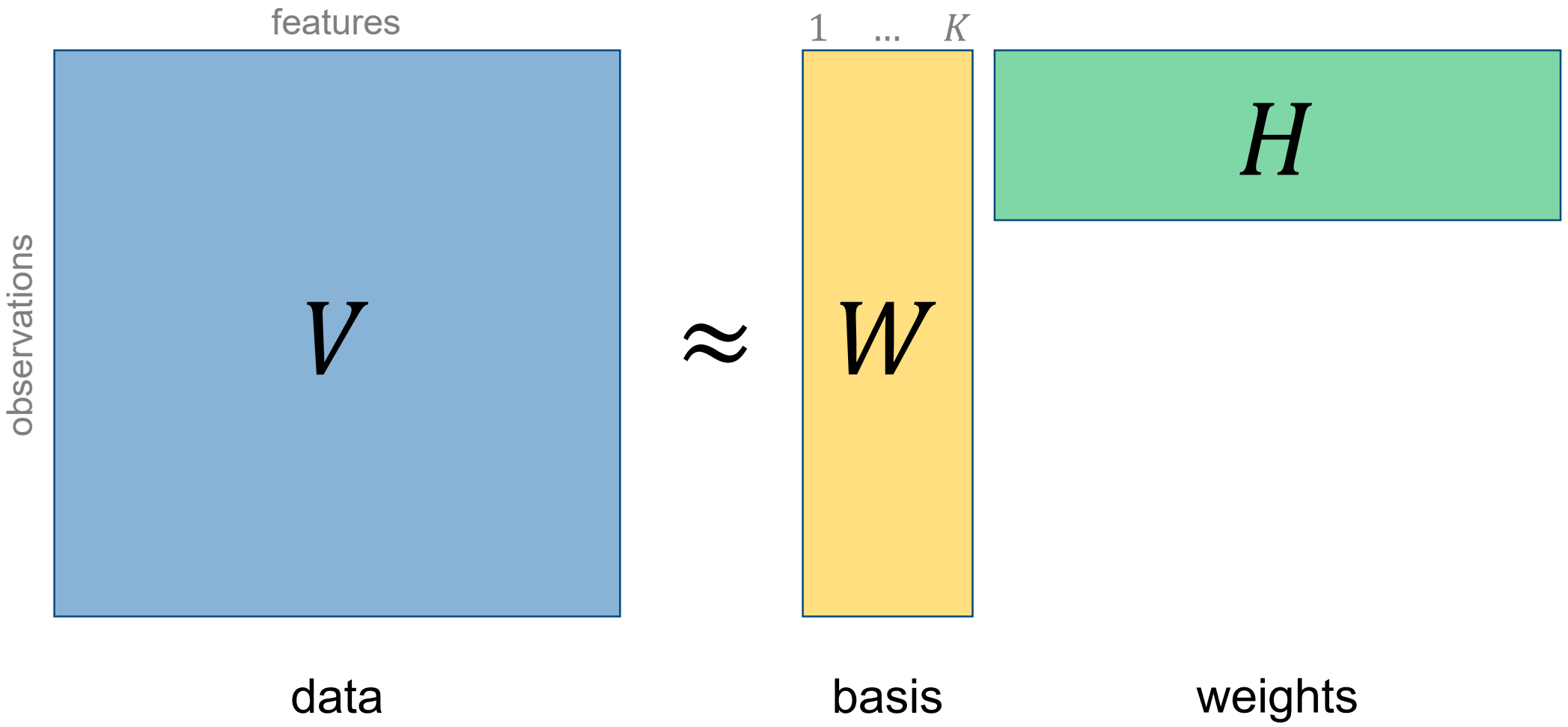
Computational Biology Group

Department of Biosystems Science and Engineering

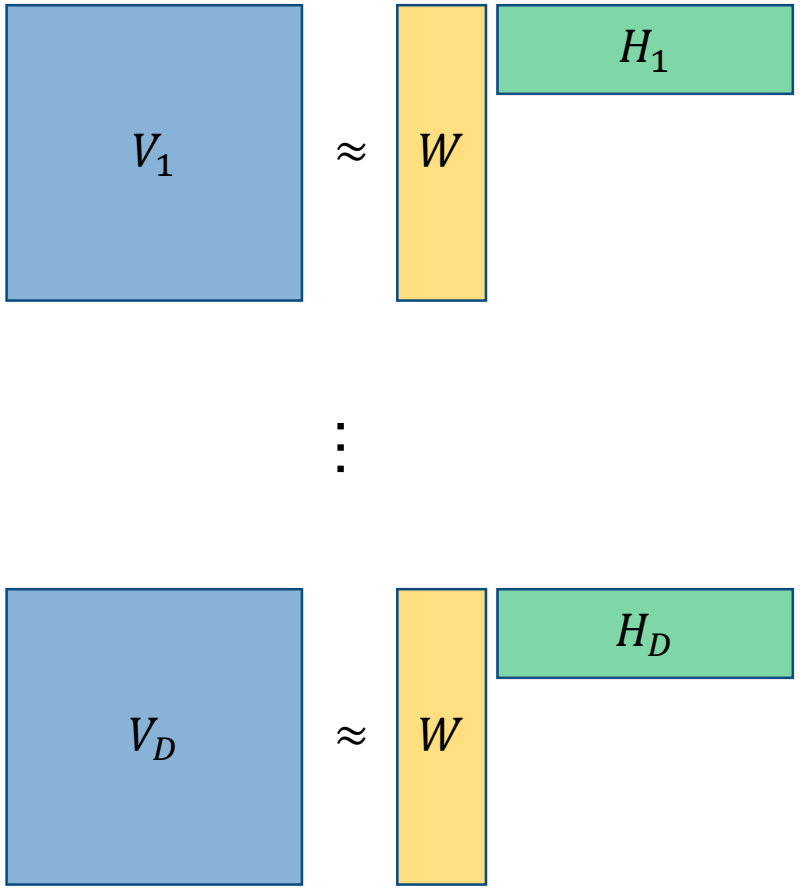
Non-negative matrix factorization

Niko Beerenwinkel

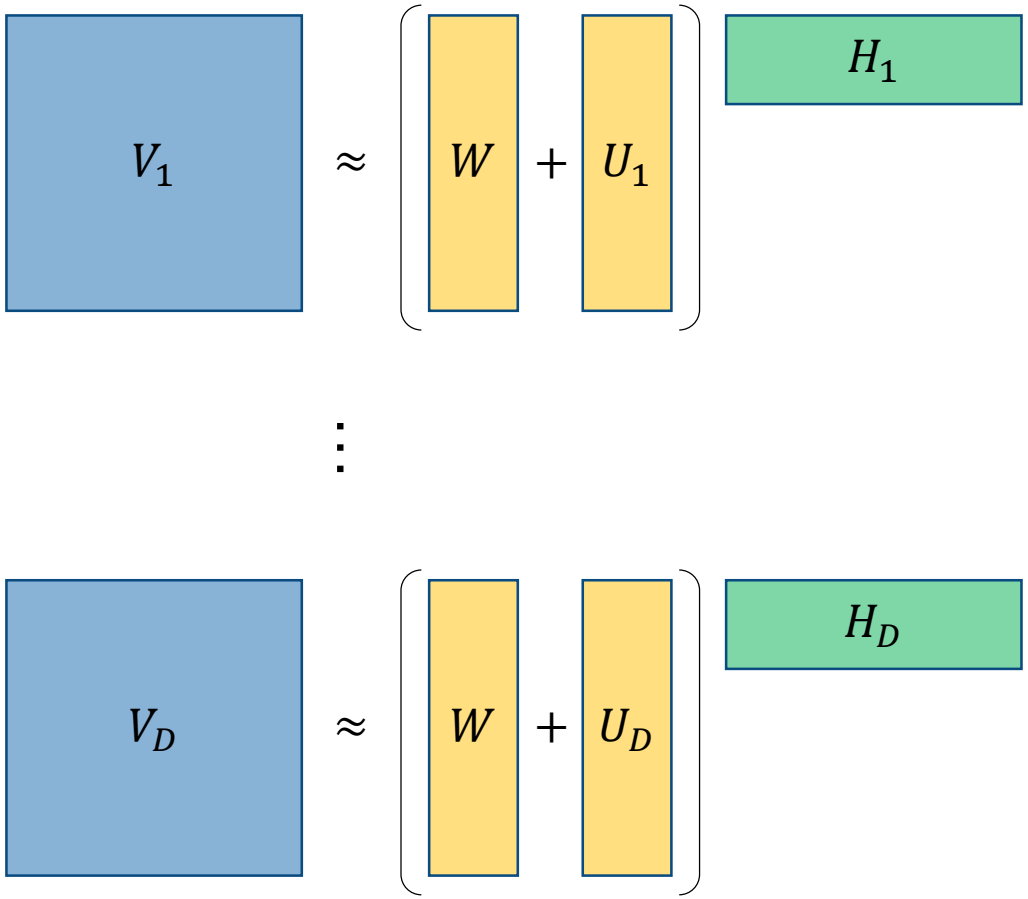
NMF: Given $V \geq 0$, find $W, H \geq 0$, such that



jNMF: Given $V_d \geq 0$, find $W, H_d \geq 0$, such that



iNMF: Given $V_d \geq 0$, find $W, U_d, H_d \geq 0$, such that



jNMF vs iNMF

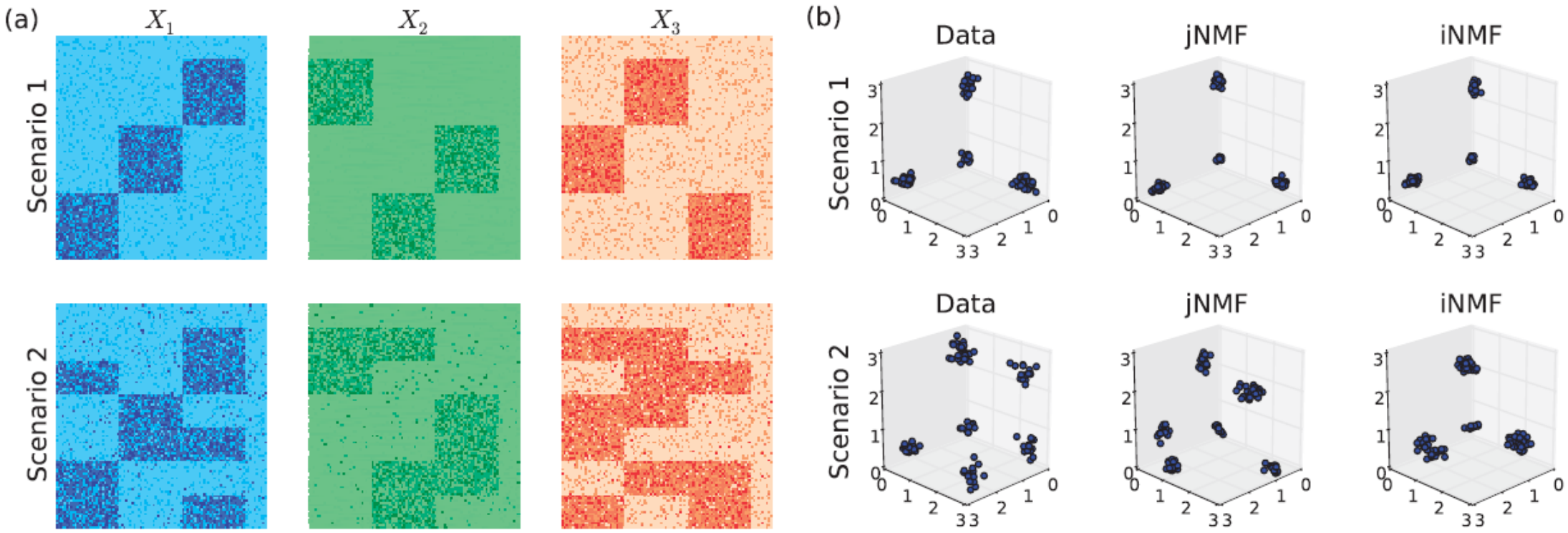
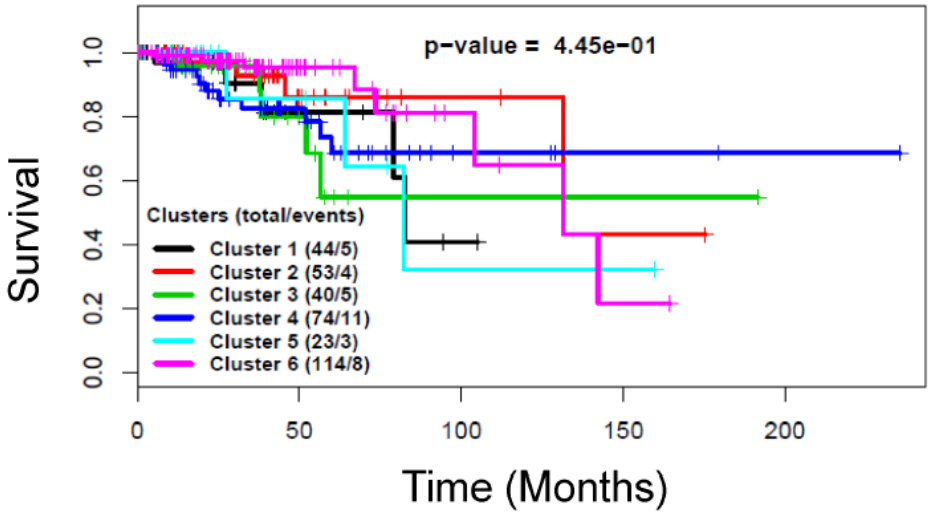
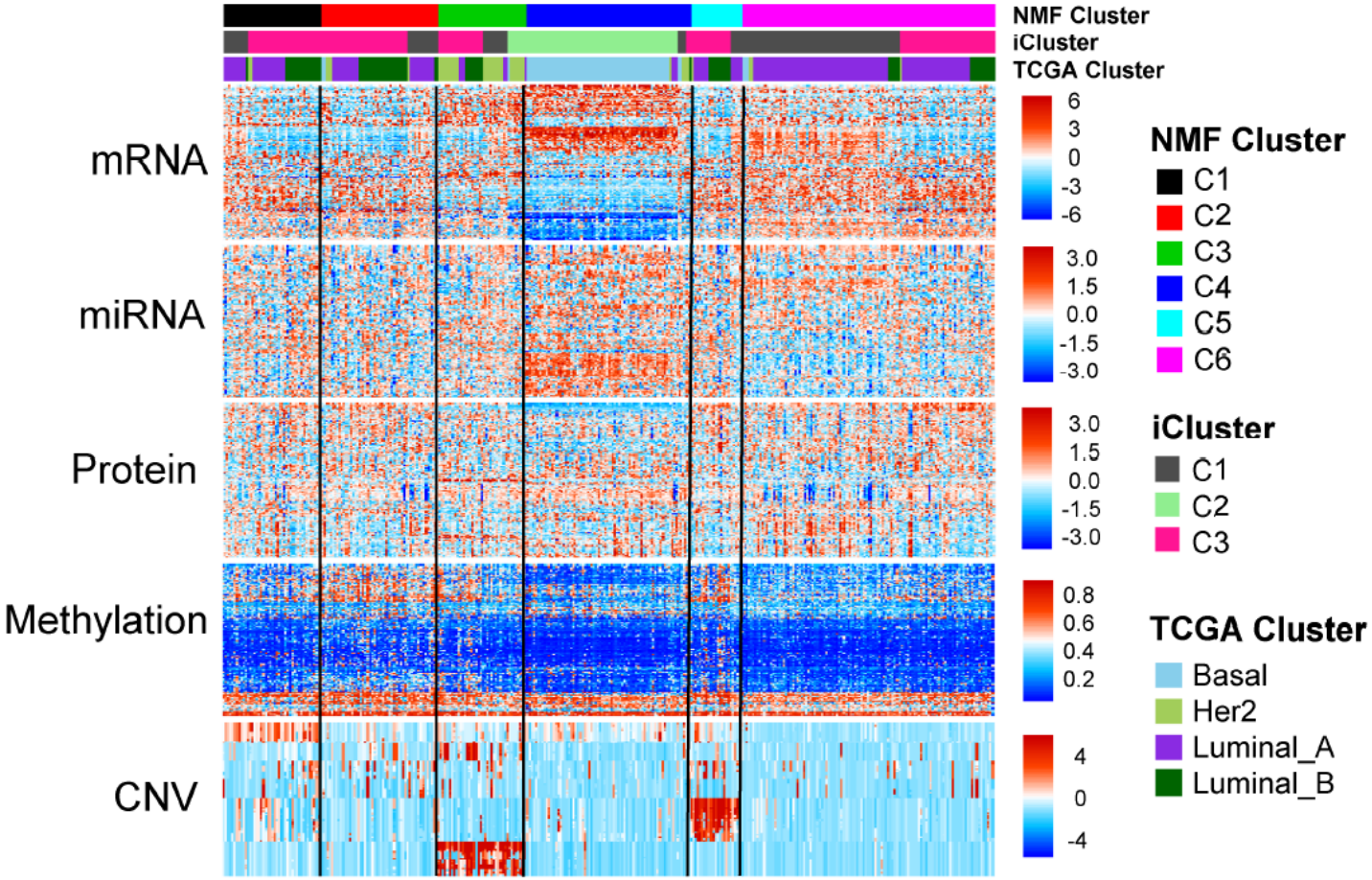


Fig. 1. (**a**) An example of multi-dimensional modules across three different data sources. Three modules are distinguishable in Scenario 1 as strong associations between subsets of variables across sources and a common subset of observations. Scenario 2 contains the same data with added random noise and confounding effects. (**b**) Low-dimensional representations of the data (X_2), jNMF approximations (W) and iNMF approximations (W). The modules are clearly detected by both methods in Scenario 1 but only by iNMF in Scenario 2 (Color version of this figure is available at *Bioinformatics* online.)

Multi-omics data integration using jNMF

- M = 348 breast cancer samples:
  - N_1 = 645 mRNAs
  - N_2 = 423 miRNAs
  - N_3 = 171 proteins
  - N_4 = 574 DNA methylation probes
  - N_5 = 409 CNVs
- jNMF revealed K = 6 clusters (modules) associated with survival:





Multi-omics data integration using iNMF

- M = 592 ovarian cancer samples:
  - DNA methylation (N_1 = 15,661 features)
  - Gene expression (N_2 = 14,821 features)
  - miRNA expression (N_3 = 799 features)
- iNMF revealed K = 4 clusters (modules)
- H_d determines assignment of genes to modules

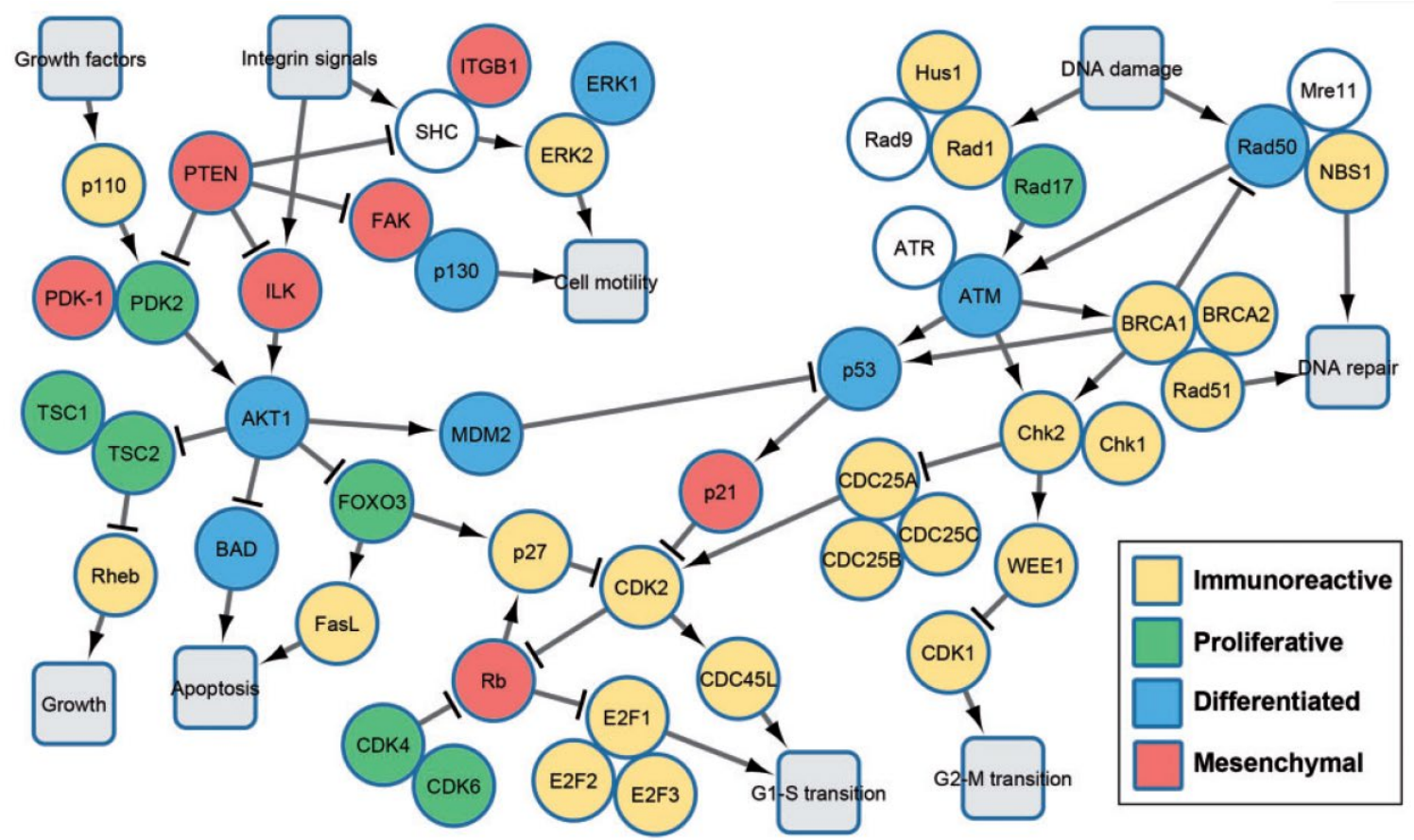
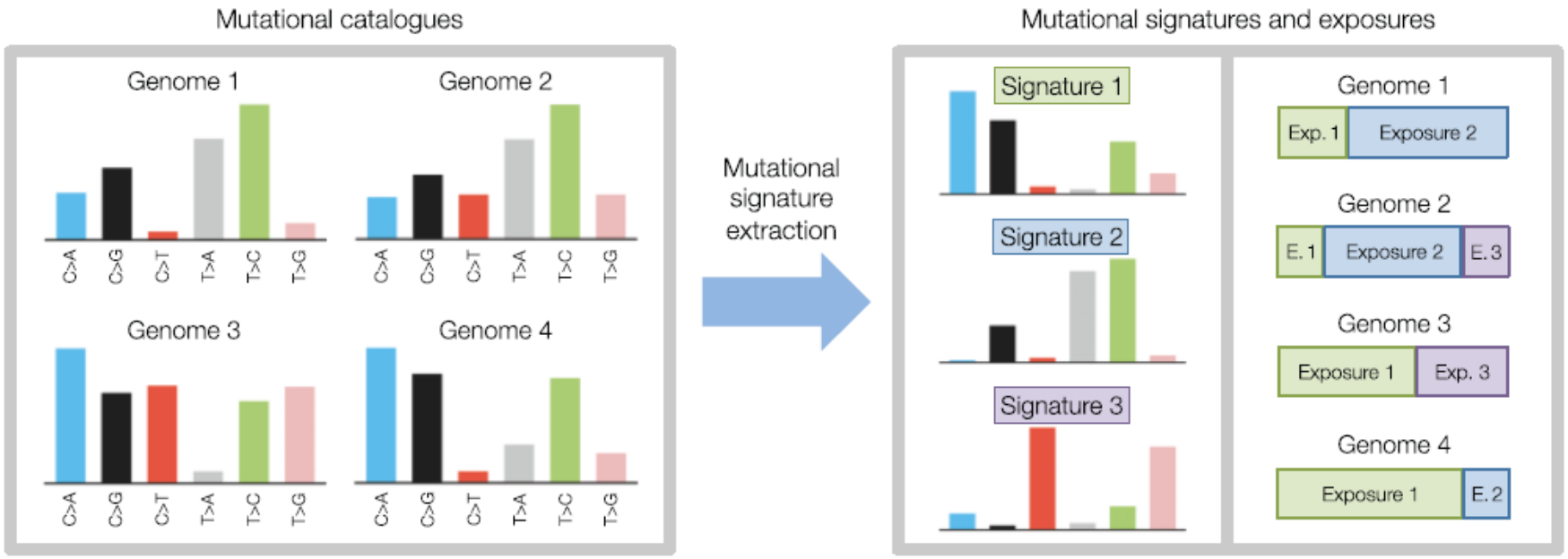
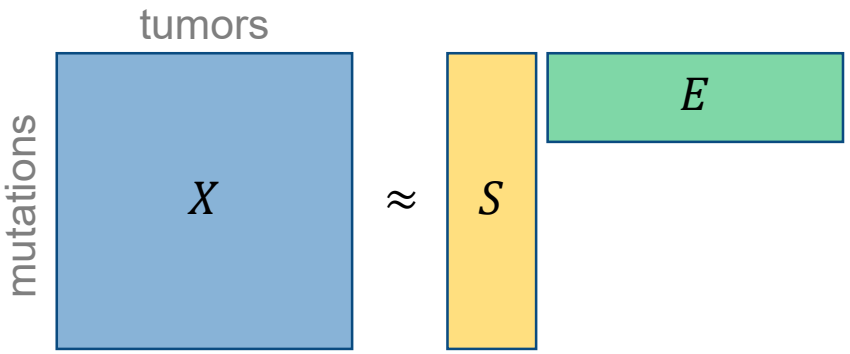


Fig. 3. Module memberships of genes (from iNMF) arranged according to pathways derived from BioCarta and relevant literature and include processes of DNA repair (top right), cell cycle regulation (bottom), cell survival and proliferation (left) and cell migration (top left) (Color version of this figure is available at *Bioinformatics* online.)

Mutational signatures





Baez-Ortega et al. 2019

Terminology and notation

A CpG site, *i.e.*, the " 5'—C—phosphate—G—3' " sequence of nucleotides, is indicated on one DNA strand (in yellow). On the reverse DNA strand (in blue), the complementary 5'—CpG—3' site is shown. A C-G base-paring between the two DNA strands is also indicated (right)

https://en.wikipedia.org/wiki/CpG_site

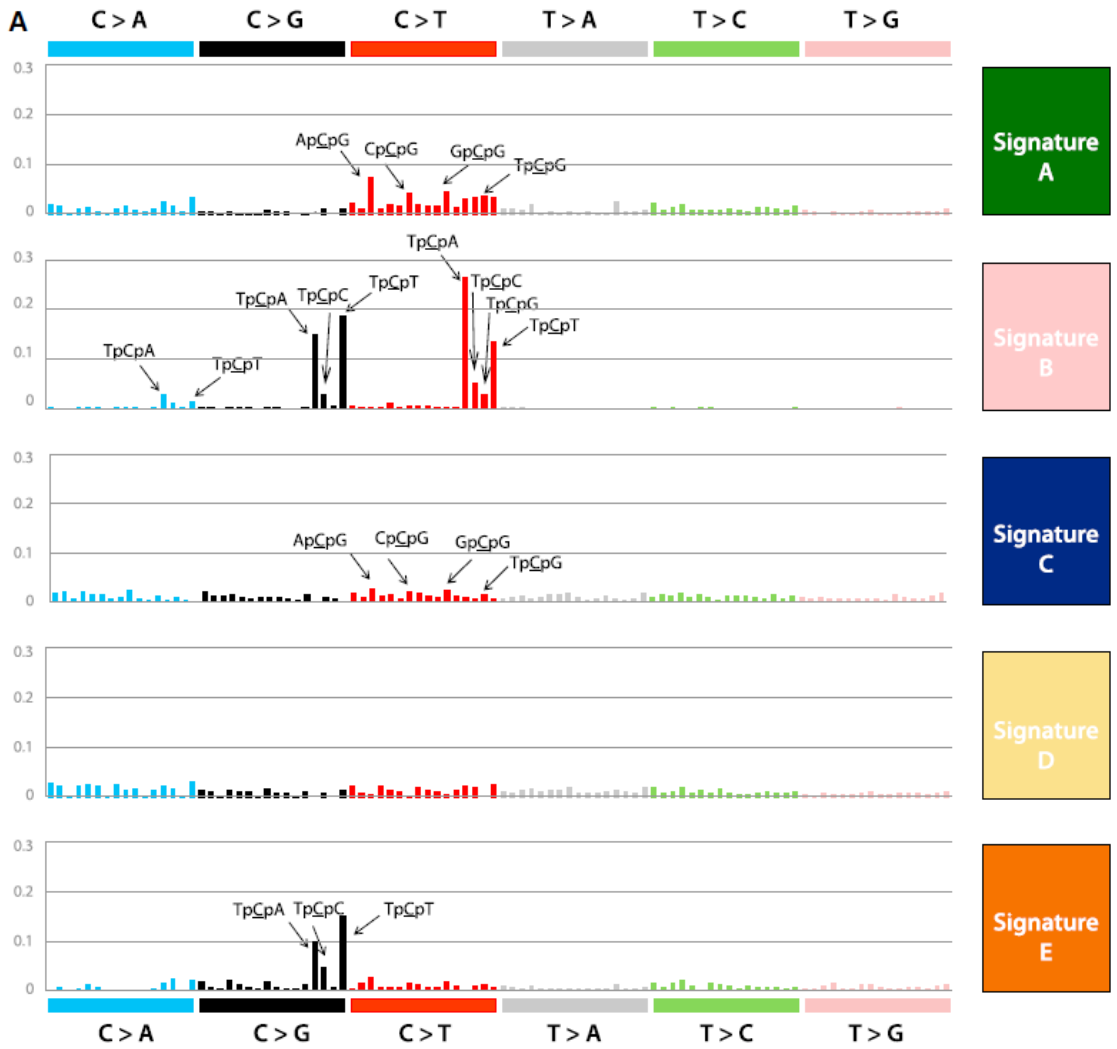
DNA base pairing: A:T, C:G
Pyrimidines = {C, T}

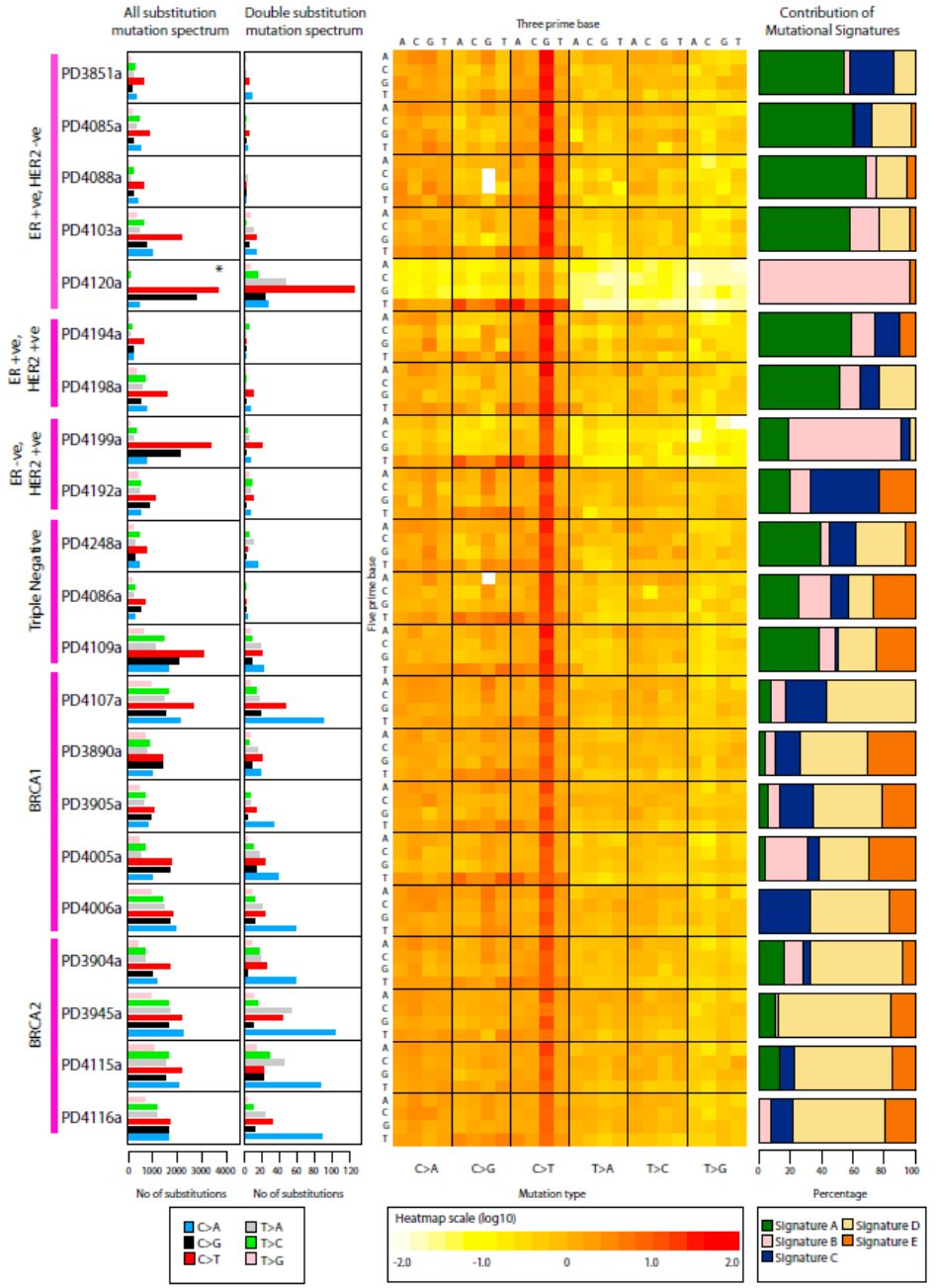
C:G > A:T mutation has pyrimidine change: C>A

ApCpG = 5'—A—phosphate—C—phosphate—G—3'
C is mutated

96 trinucleotide mutation channels

- 21 breast cancer samples
- Total of 183,916 somatic mutations
- NMF suggested 5 mutational signatures





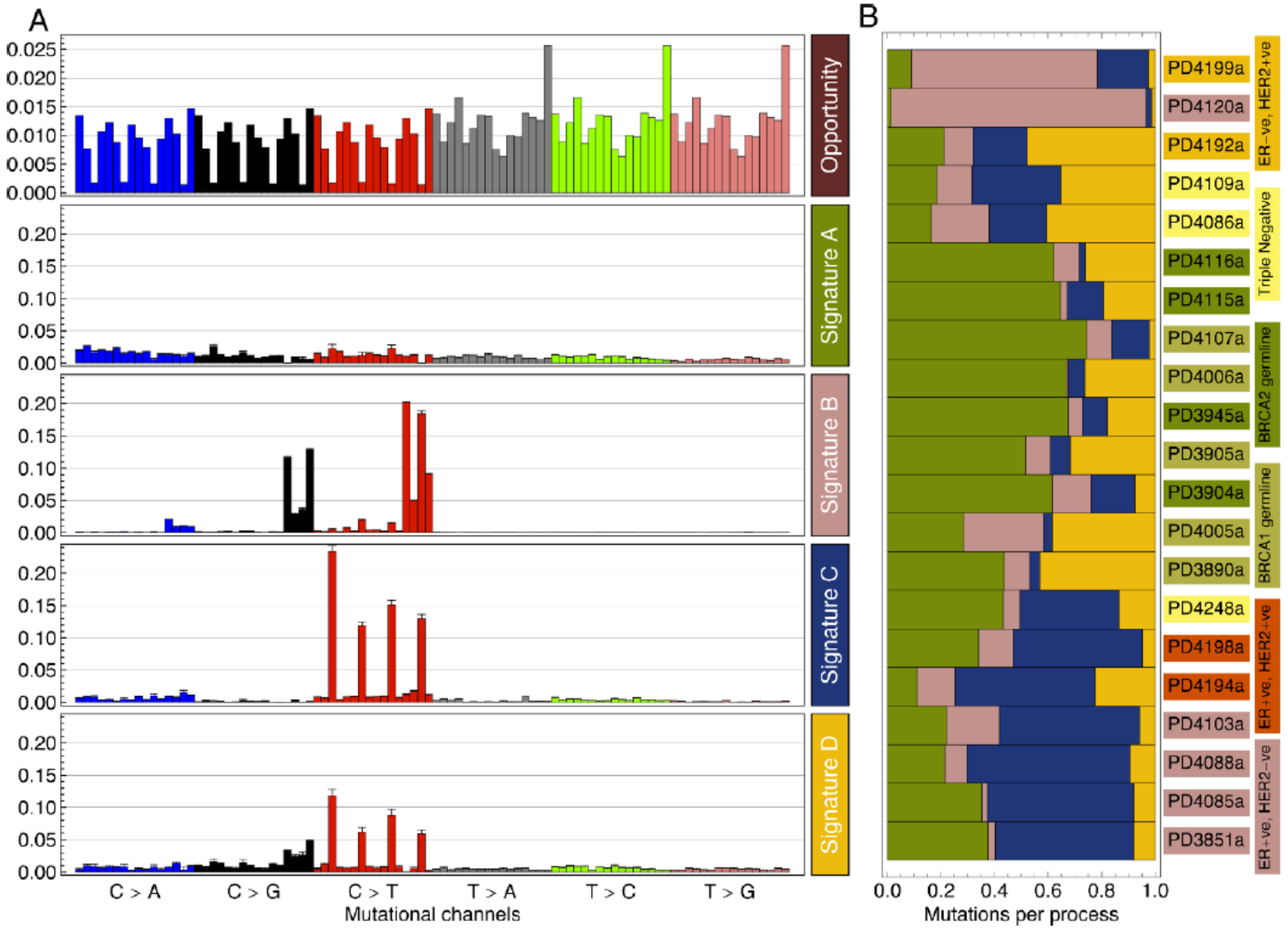


Figure 2 Mutational spectra and tumor composition. **(A)** The mutational opportunity spectrum of the human genome across 96 trinucleotide channels and the mutational signatures found in the breast cancer data. For each mutation, the 16 sequence contexts are ordered by the 5′ and then the 3′ base in the order A, C, G and T, that is, from ApCpA, ApCpC to TpCpG, TpCpT. The error bars were estimated analytically (see Additional file 1). **(B)** The contributions of each mutational process towards the mutations for each tumor. The correlation with cancer subtype matches that reported in [13].

Fischer et al. 2013

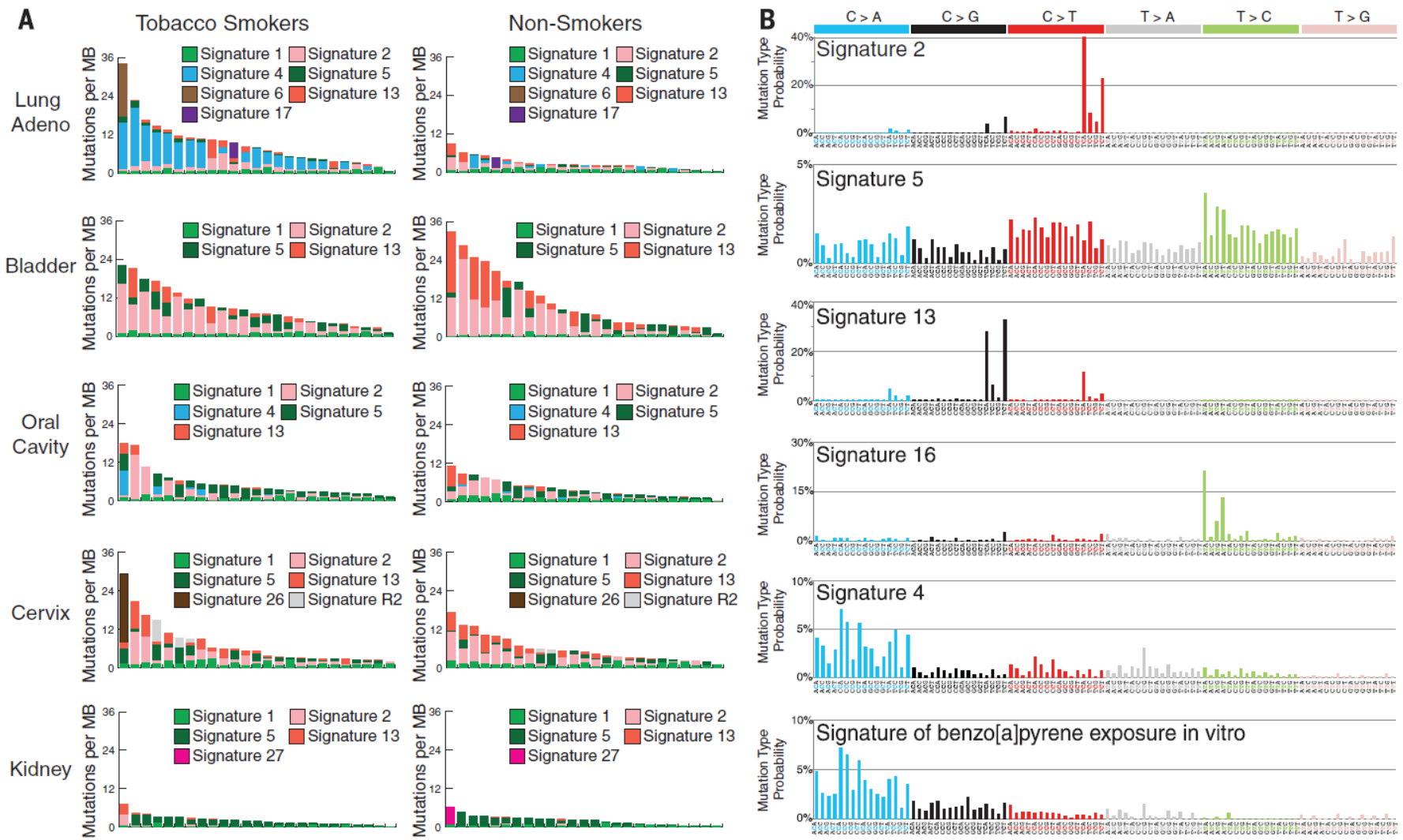


Fig. 2. Mutational signatures associated with tobacco smoking. (**A**) Each panel contains 25 randomly selected cancer genomes (represented by individual bars) from either smokers or nonsmokers in a given cancer type. The *y* axes reflect numbers of somatic mutations per megabase. Each bar is colored proportionately to the number of mutations per megabase attributed to the mutational signatures found in that sample. The naming of mutational signatures is consistent with previous reports (*16*–*18*). (**B**) Each panel contains the pattern of a mutational signature associated with tobacco smoking. Signatures are depicted using a 96-substitution classification defined by the substitution type and sequence context immediately 5′ and 3′ to the mutated base. Different colors are used to display the various types of substitutions. The percentages of mutations attributed to specific substitution types are on the *y* axes, whereas the *x* axes display different types of substitutions. Mutational signatures are depicted based on the trinucleotide frequency of the whole human genome. Signatures 2, 4, 5, 13, and 16 are extracted from cancers associated with tobacco smoking. The signature of benzo[*a*]pyrene is based on in vitro experimental data (*19*). Numerical values for these mutational signatures are provided in table S6.

References

- Yang 2015: https://doi.org/10.1093/bioinformatics/btv544
- Chalise 2017: https://doi.org/10.1371/journal.pone.0176278
- Baez-Ortega 2019: https://doi.org/10.1093/bib/bbx082
- Nik-Zainal 2012: https://doi.org/10.1016/j.cell.2012.04.024
- Fischer 2013: https://genomebiology.biomedcentral.com/articles/10.1186/gb-2013-14-4-r39
- Alexandrov 2016: https://www.science.org/doi/10.1126/science.aag0299