REFERENCEdb

A REpository oF simulatEd tumoR and normal gEnetic, geNomic, and transCriptomic datasEts

Yered Pita-Juarez

✉ yhpitajua@bidmc.harvard.edu

Ioannis S Vlachos

✉ ivlachos@bidmc.harvard.edu

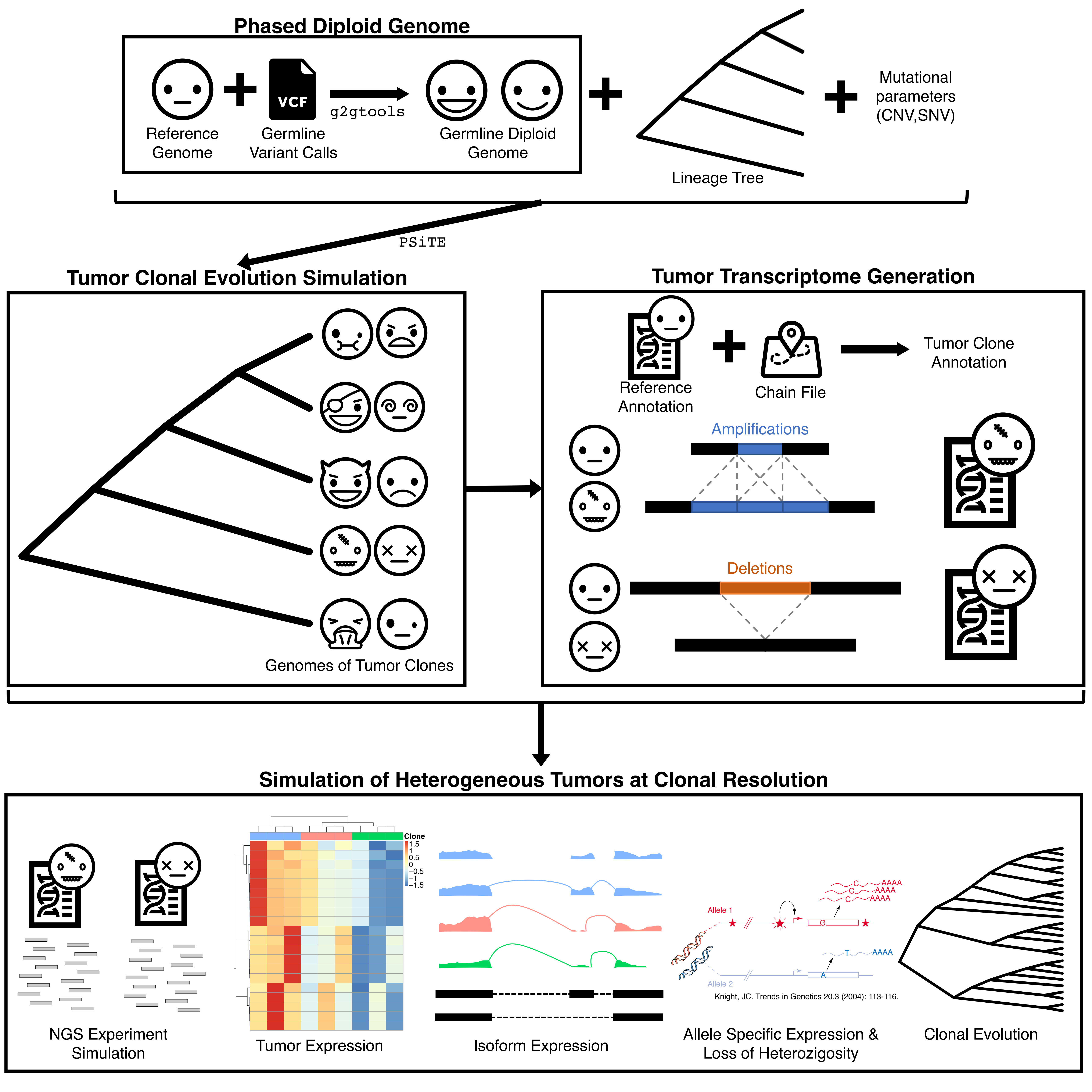
INTRO

- The advancement of high throughput technologies is revolutionizing cancer research and clinical practice.
- *In silico* methods extract information from these experimental datasets to study the complex genetic structure of tumors, clonal evolution, and differences in the genetic and transcriptional landscapes between cancer and normal tissues, as well as detect potential neoantigens.
- The advancement of these methods has been hampered by **the lack of diverse ground truth** datasets.
- **Realistic simulated cancer genomes and transcriptomes along with their normal counterparts** could guide the validation and improvement of such methods.

METHODS

- Phase reference genomes (1000G Phase 2) based on 1000G phase 3 variant calls with g2gtools (Choi, 2018) to generate a germline diploid genome
- Simulate tumor clonal evolution based on germline diploid genome, a lineage tree, and mutational parameters (SNV and CNV rates) with PSiTE (Yang et al., 2019)
- Generate a tumor transcriptome by assigning the transcript locations from the reference annotation (GENCODE/Ensembl) to the corresponding tumor genomes
- Assign transcript expression probabilities for each clone. Distributions permit homogeneous expression profiles across clones, while enabling expression diversity.
- Generate NGS experiments from the simulated cancer genomes and transcriptomes at clonal resolution (WGS, WES, RNA-Seq) or at bulk tumor level.
- **We provide a collection of ground truth:**
  - ❑ Clonal evolution trees
  - ❑ Tumor genomes
  - ❑ Tumor transcriptomes
  - ❑ NGS Data

A simulated **ground truth** collection of heterogeneous **tumor data** at **clonal resolution.**



Knight, JC. Trends in Genetics 20.3 (2004): 113-116.

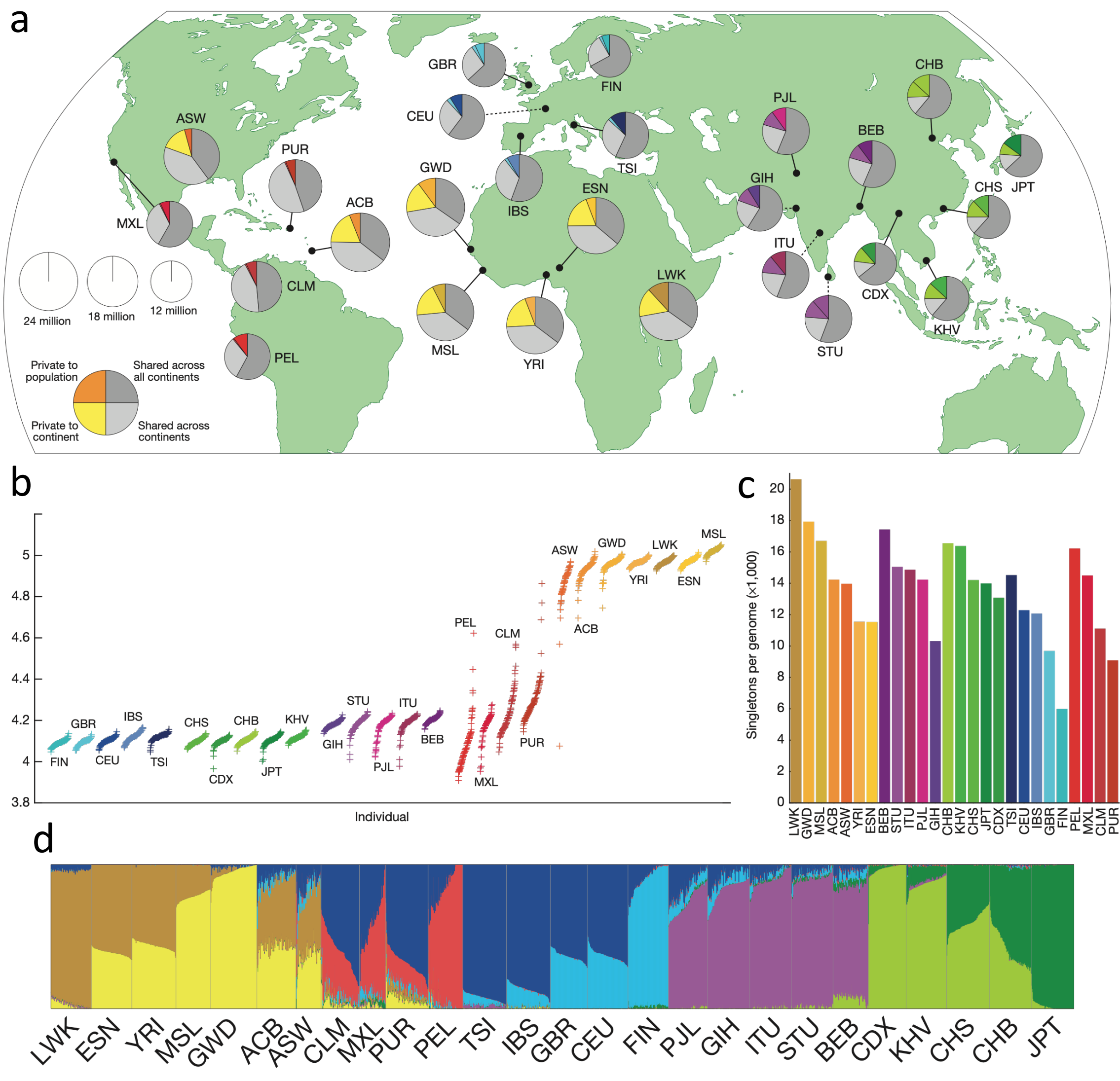
Non-coding RNA Laboratory

https://non-coding.org/

Cancer Research Institute & Harvard Initiative for RNA Medicine,
Broad Institute of MIT and Harvard,
Department of Pathology,
Beth Israel Deaconess Medical Center / Harvard Medical School,
Boston, MA 02115

DISCUSSION

- REFERENCEdb is the first collection of **simulated germline and concordant somatic genomes.**
- REFERENCEdb datasets have been created using diverse subjects and populations from the 1000 Genomes project, comprise germline, and tumor genetic information, gene annotation, and gene expression.
- They also incorporate **realistic clonal dynamics**, in the form of genetic and expression for each clone as well as their relationships.
- These datasets can become a gold standard to optimize and benchmark methods to detect mutations, clonal decomposition, gene expression, and neoantigens from high throughput data.



1000 Genomes Project Consortium. Nature 526.7571 (2015): 68.

Figure 1. a. Proportion of polymorphic variants in the 1000 Genomes Phase 3 samples. **b.** Number of variants per site. **c.** Average number of singletons per genome. **d.** Population structure inferred with 8 clusters (1000G Consortium, 2015)

Beth Israel Lahey Health

Beth Israel Deaconess Medical Center

HARVARD MEDICAL SCHOOL

BROAD INSTITUTE