Part 1.1

Comparative Analysis of Personal Genomes

CBB 752 Final Project

Yale University

9 May 2017

overview

1. introduction to **genomic databases**

2. retrieving **annotations** for variants

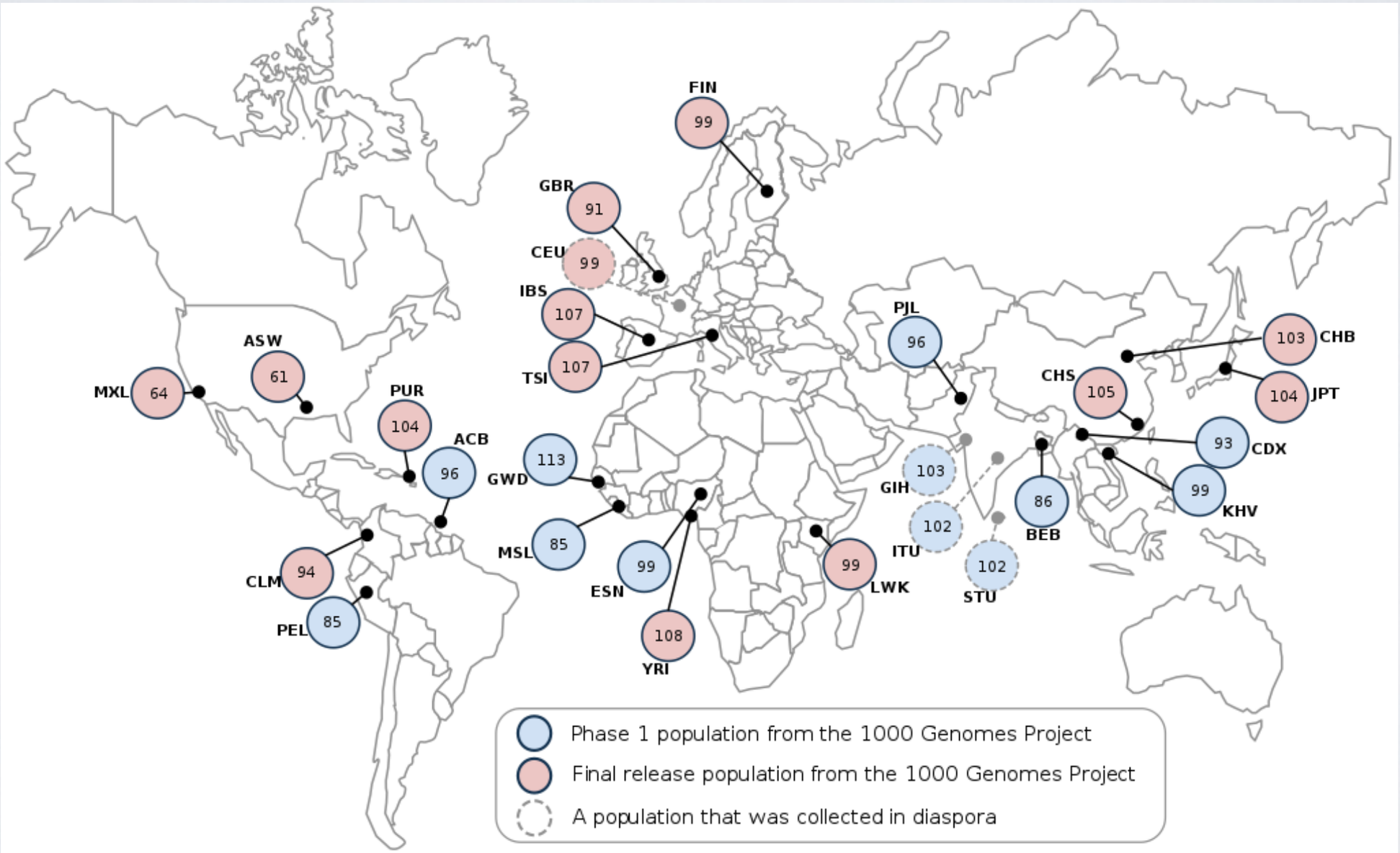
3. retrieving **population frequencies** for variants

1 | introduction to genomic databases

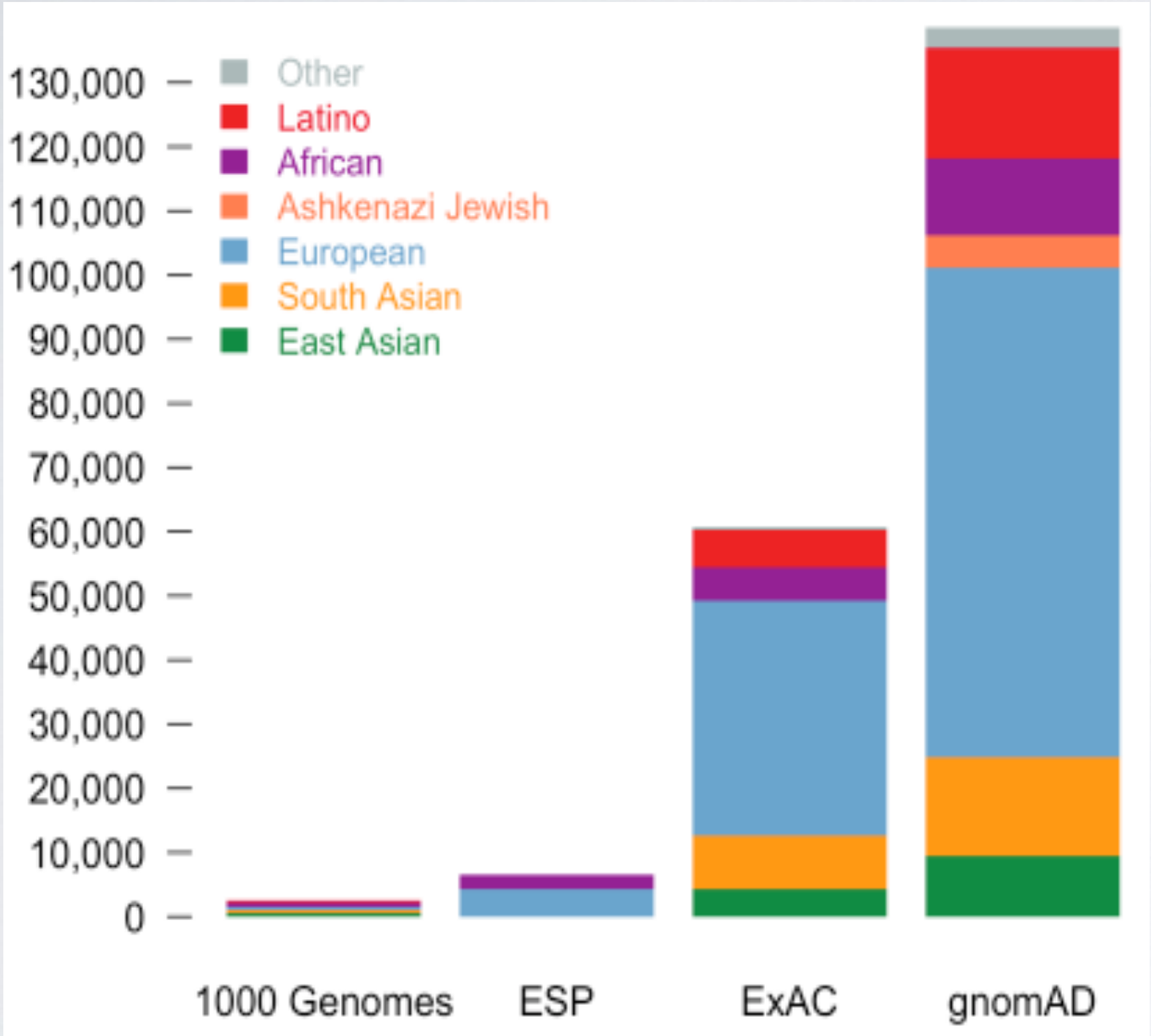
human genetic variation

Reference genetic variation:
catalogues the “normal” range
of variation in human genomes

Populations:
ancestral groups
cluster together



genomic databases



1000 Genomes Project (1KGP): individual-level genomes (2,504)

Exome Sequencing Project: aggregated population-level exomes (6,515)

Exome Aggregation Consortium (ExAC): more aggregated population-level exomes (60,706)

Genome Aggregation Database (gnomAD): update to ExAC; even more population-level exomes and genomes (138,632)

2 | retrieving annotations for variants

vcfR: query tool

Query gnomAD chr1

```
$ python vcfR.py -i sample_input.vcf -f
  https://storage.googleapis.com/gnomad-public/release-170228/vcf/
  genomes/gnomad.genomes.r2.0.1.sites.1.vcf.gz -o sample_out.vcf
```

Query 1000 Genomes chr1 phase 1

```
$ python vcfR.py -i sample_input.vcf -f
  ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/analysis_results/
  integrated_call_sets/ALL.chr1.integrated_phase1_v3.20101123.
  snps_indels_svs.genotypes.vcf.gz -o sample_out.vcf
```

results from vcfR

- Directly compared with reference VCF
- No need to download
- --anno option
- Calling myvariant database
- Matched terms in reference
- Annotation in JSON format
- Easier for further analysis

Sample_input:

- 190 variants in coding regions
- Different variants:

- chr1:g.12783G>A : intron_variant
- chr1:g.14464A>T : non_coding_transcript_exon_variant
- …

3 | retrieving population frequencies for variants

ANNOVAR

- Pinpointing a small subset of functionally important variants
- Mutation prediction approach for annotation
- Identify subsets of variants based on comparison to other variant databases:

- dbSNP
- 1000 Genome Project
- gnomAD - Genome Aggregation Database

- Pipeline: Download -> Convert -> Analyze

Download

- annotate_variation.pl -downdb 1000g2015aug humandb -buildver hg19
- annotate_variation.pl -buildver hg19 -downdb -webfrom annovar gnomad_genome humandb/

VCF conversion

- convert2annovar.pl -format vcf4 -withfreq data/indel.vcf > data/indel.avinput
- convert2annovar.pl -format vcf4 -withfreq data/snp.vcf > data/snp.avinput

Analysis

- annotate_variation.pl -filter -dbtype 1000g2015aug_all -buildver hg19 -out indel data/indel.avinput humandb/
- annotate_variation.pl -filter -dbtype hg19_gnomad_genome -buildver hg19 -out snp.gad data/snp.avinput humandb/

sample allele frequencies

Table 2: Three examples from chromosome 1 in 1000 Genomes:

Location	Reference	Alternate	Minor Allele Frequency
49554	A	G	0.063099
49298	T	C	0.782149
54490	G	A	0.096046

Table 3: Three different examples from chromosome 1 in gnomAD:

Location	Reference	Alternate	Minor Allele Frequency
10144	T	C	0.0007
92004	A	G	0.0387
108310	T	C	0.2959

sample allele frequencies

Source	1KGP known variants	1KGP unknown variants	gnomAD known variants	gnomAD unknown variants	Total
Indel	469,754	320,215	709,146	80,823	789,969
SNP	3,381,358	177,780	3,476,922	82,216	3,559,138

- Proportion of private variants is different by 2.5% and 30% for SNPs and indels, respectively, when comparing 1000genome to gnomAD.
- gnomAD contains much more variants, as we expect, since it comes from a much larger cohort.

summary

1. genomic databases are important for understanding the “normal” range of genetic variation

2. vcfR retrieves online references

3. ANNOVAR retrieves genome frequencies and annotations