Informatic Imperatives for Sequencing in Large-Scale Studies

Daniel MacArthur

Analytic and Translational Genetics Unit,
Massachusetts General Hospital
Broad Institute of Harvard and MIT
Harvard Medical School
1000 Genomes Project Consortium

Acknowledgments

- Shane McCarthy (Sanger)
- Mark DePristo, Ryan Poplin and Khalid Shakir (Broad)
- Mark Daly and David Altshuler (MGH/Broad)
- 1000 Genomes Analysis Group

A plausible near-future scenario

- we have exome sequence and complex phenotype data available for 100,000 individuals from multiple sources
- one **petabyte** of raw data
- goals:
  - create accurate, consistent variant calls across all samples
  - harmonized, cleaned phenotype data
  - data are not just accessible but *usable* by researchers from the wider community

Key challenges

- **Logistics**: moving, storing and crunching large data-sets
- **Harmonization**: integrating and cleaning data from multiple sources
- **Analysis**: extracting useful information from massive, structured data-sets
- **Access**: making data available and usable for many different audiences

Logistics: data management

- 100,000 exomes → 1 **petabyte** raw data
  - with 10 Gbit connection, 1-6 months
  - may be faster to use truck full of hard drives
- data storage is not free:
  - expect ~$1M/year for 100K exomes
- ~10x greater for WGS; less with compression
- **but** community now has extensive experience in handling large sequence data: logistical problems are soluble

Logistics: QC and meta-data

- sample-tracking:
  - genetic data: easy ID of duplicates, sample swaps, pedigree and population errors
  - phenotype data requires much more stringent tracking
- meta-data for each participant:
  - what consent have they provided?
  - where are their DNA/tissue/cell lines?
  - can they be recontacted for phenotyping?
- phenotype data likely to massively increase in near future

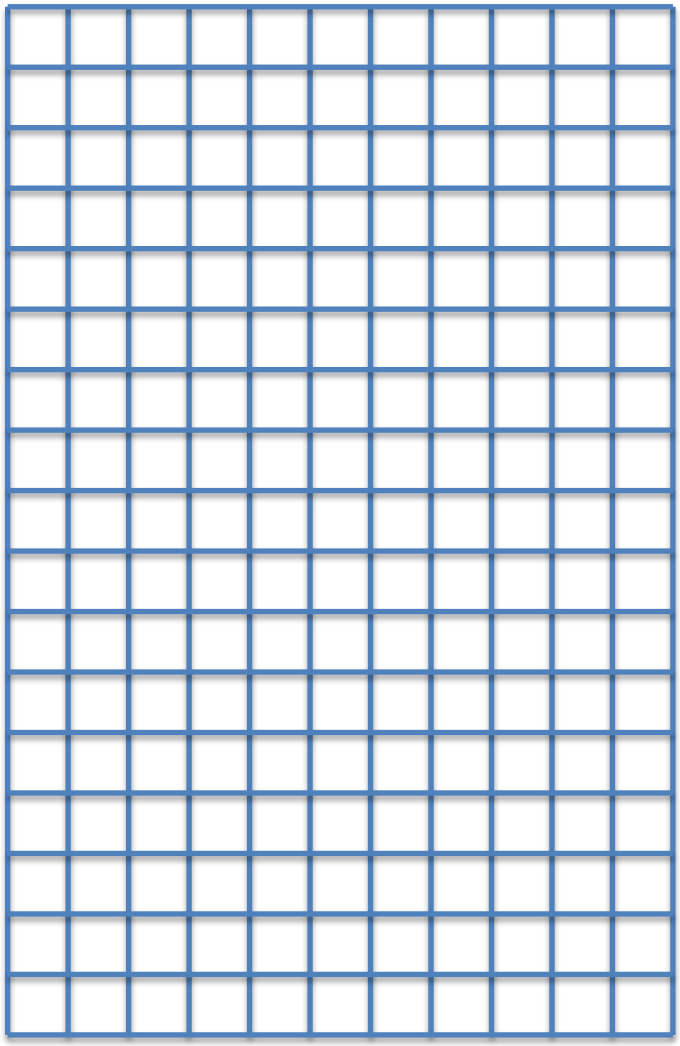
Harmonization

- both sequence and phenotype data are inconsistently generated between studies
- lack of consistency destroys ability to perform accurate cross-cohort analyses
- phenotype harmonization very difficult with current approaches
- sequence harmonization much more tractable, but **centralized processing and variant-calling are essential**

Squaring the matrix

Individuals

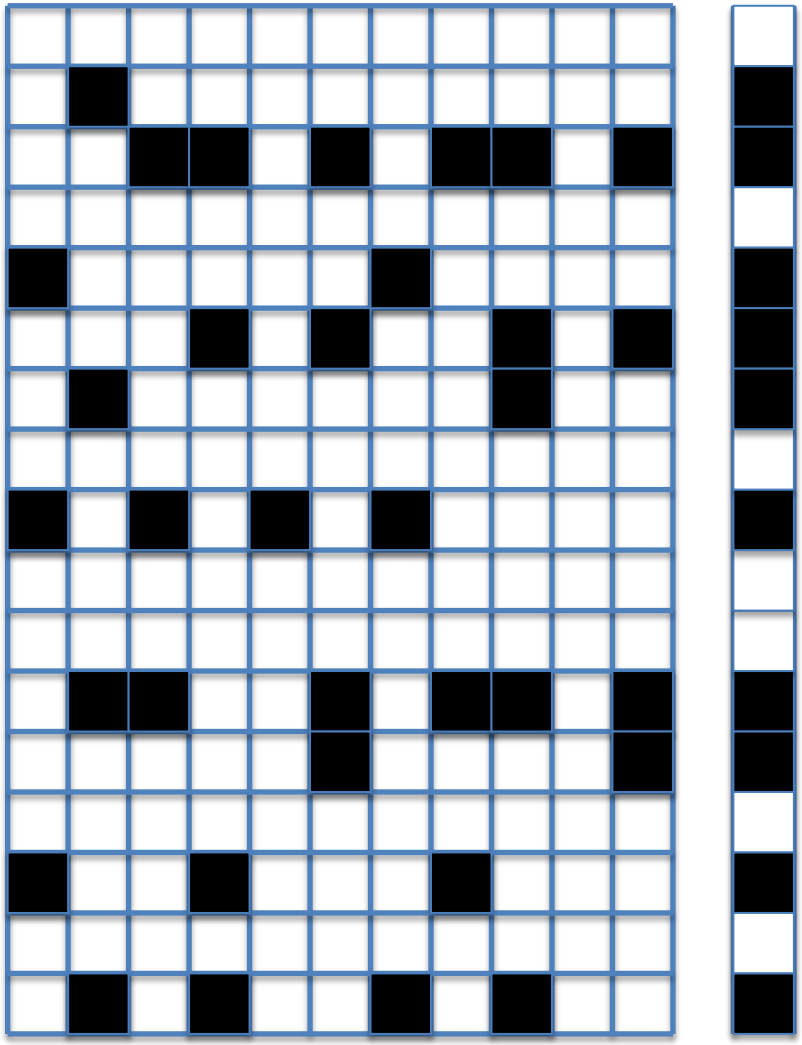
Variant positions



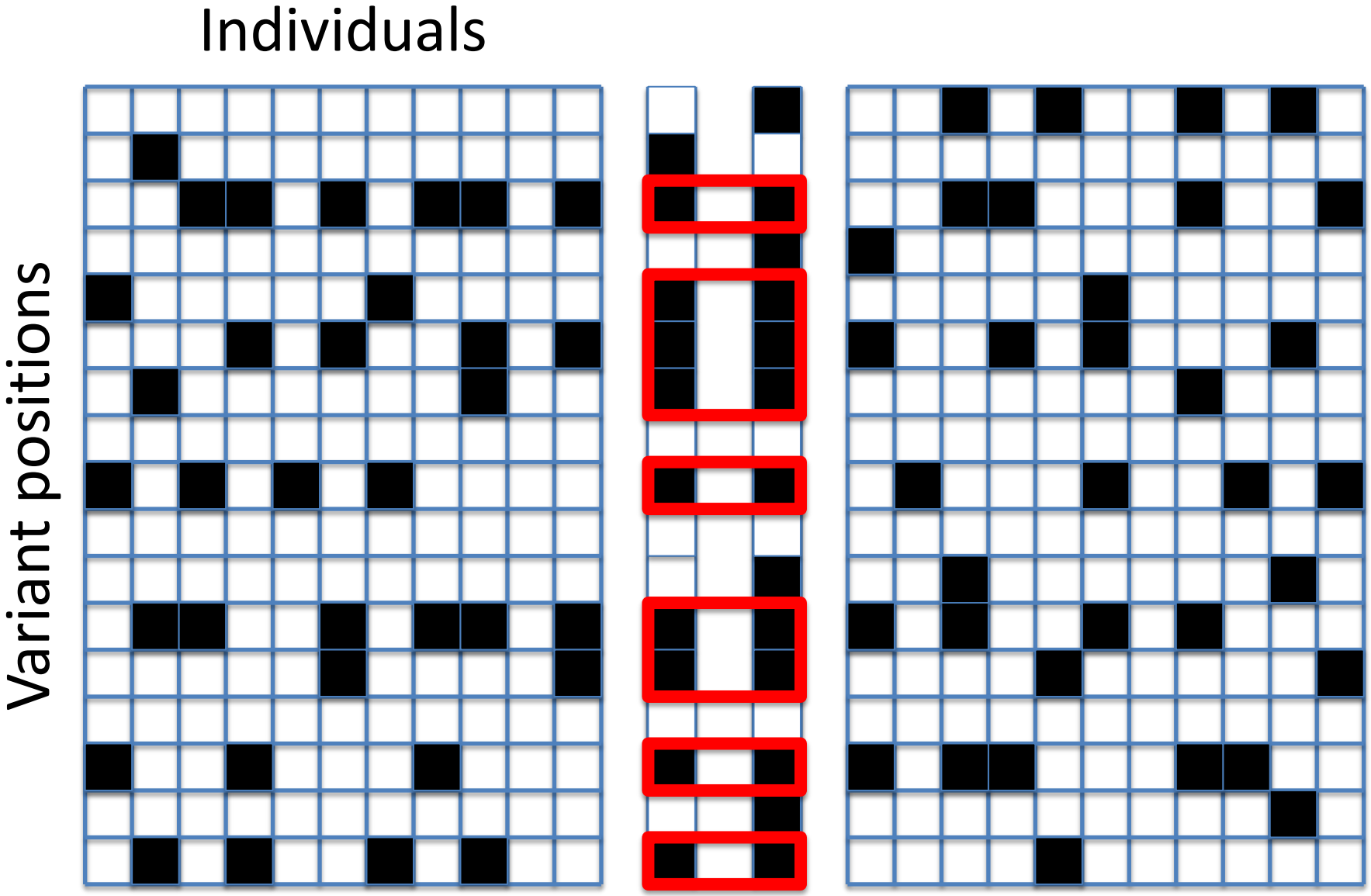
Squaring the matrix

Individuals

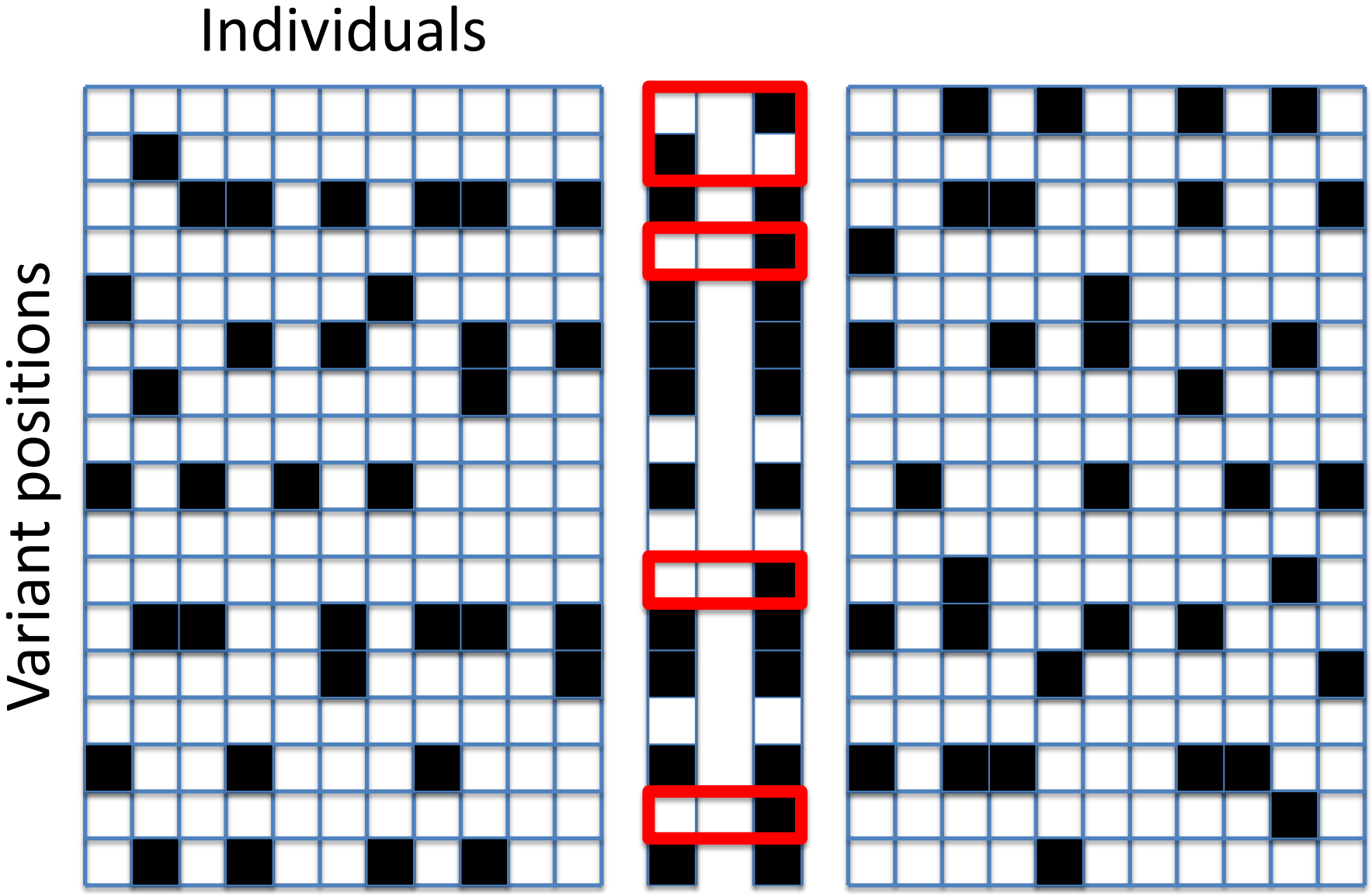
Variant positions



Squaring the matrix



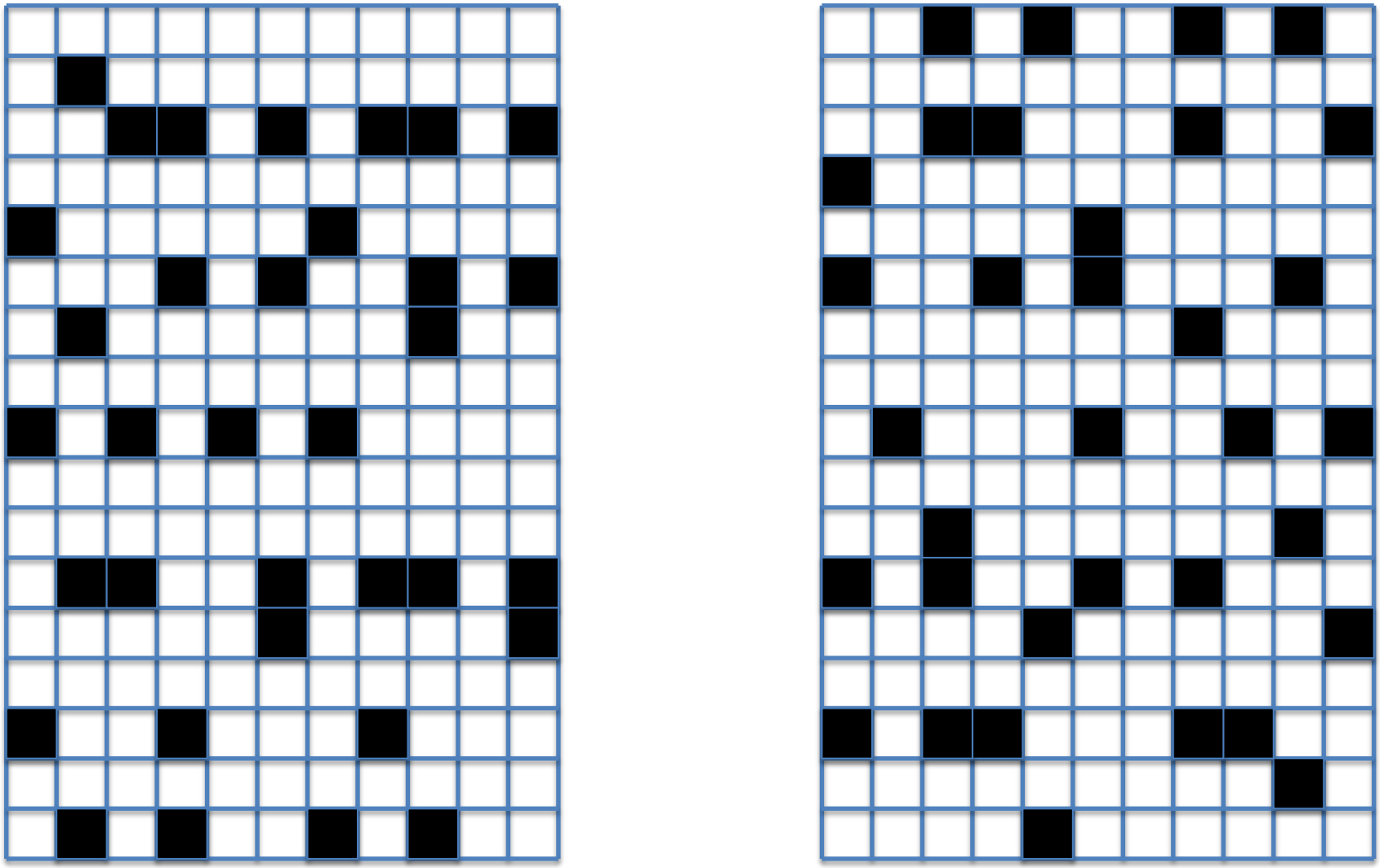
Squaring the matrix



Squaring the matrix

Individuals

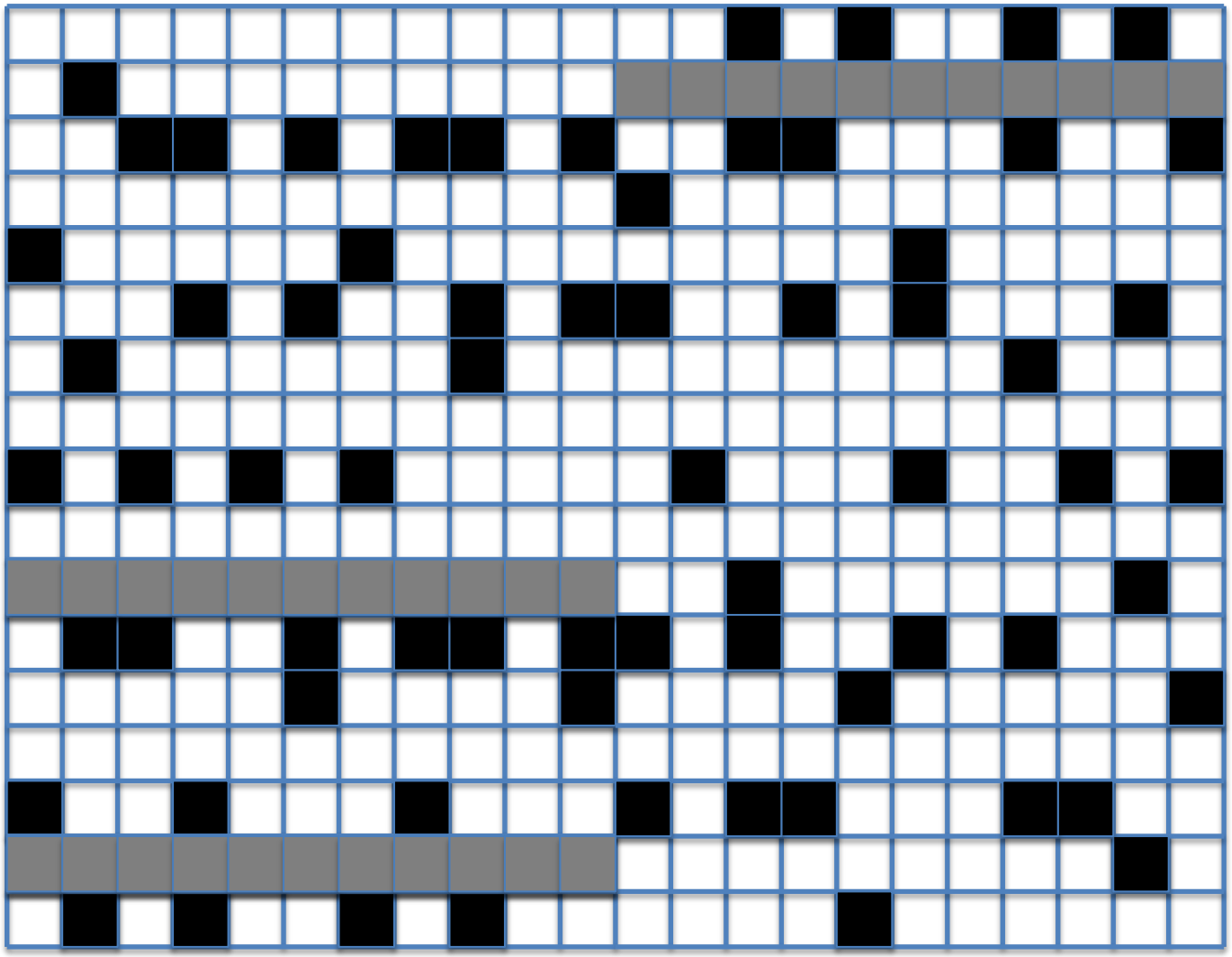
Variant positions



Squaring the matrix

Individuals

Variant positions

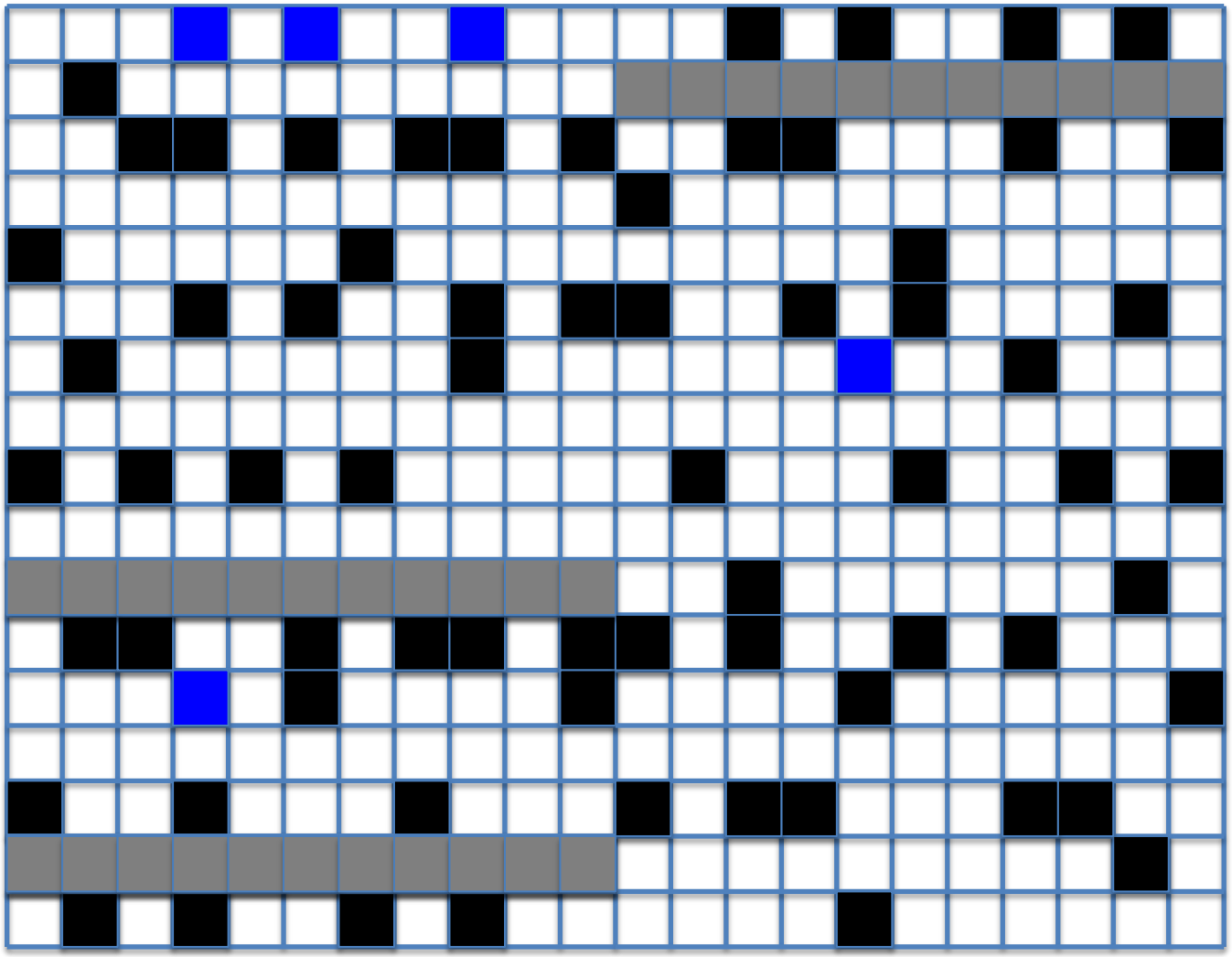


no data

Squaring the matrix

Individuals

Variant positions



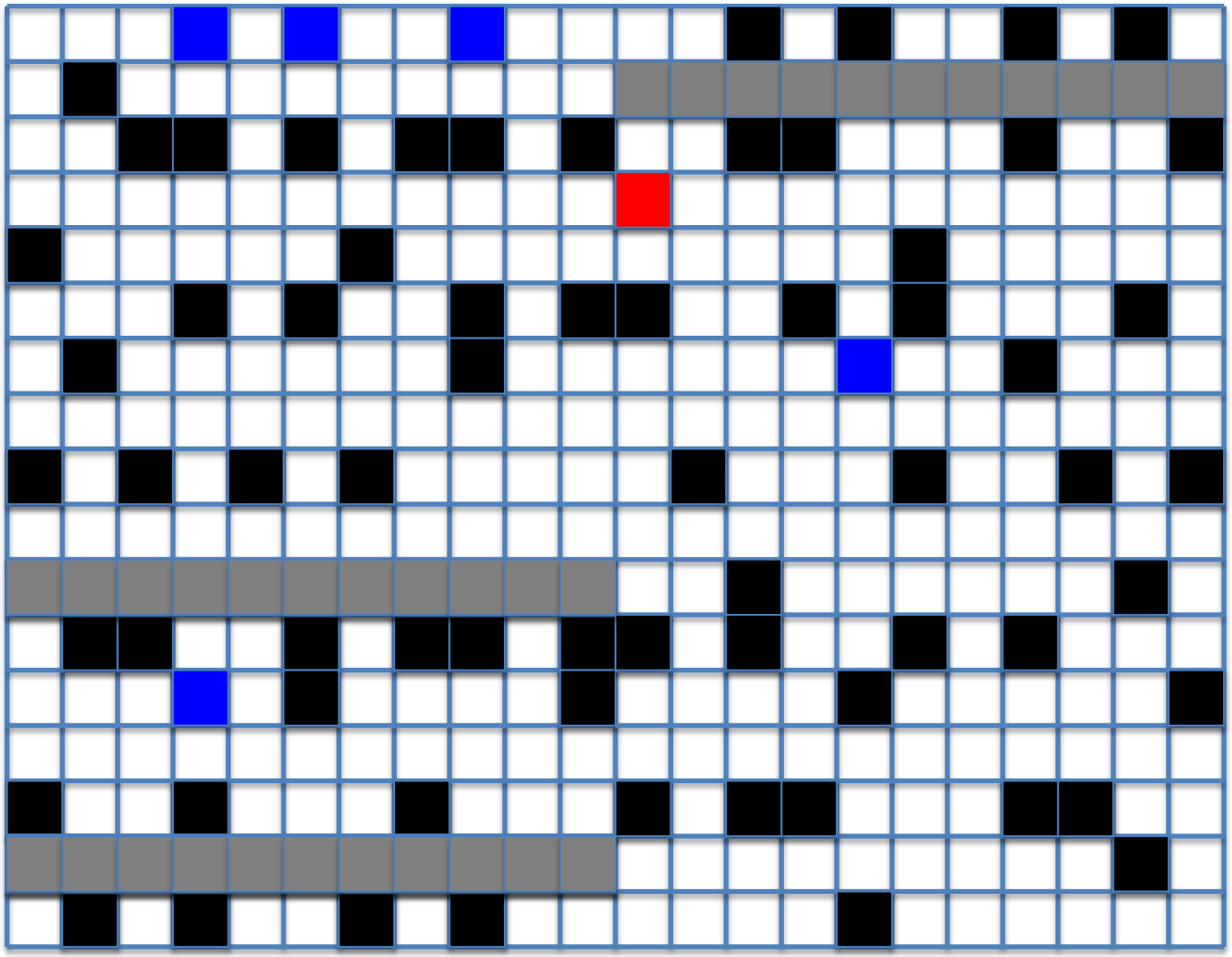
no data

new calls

Squaring the matrix

Individuals

Variant positions

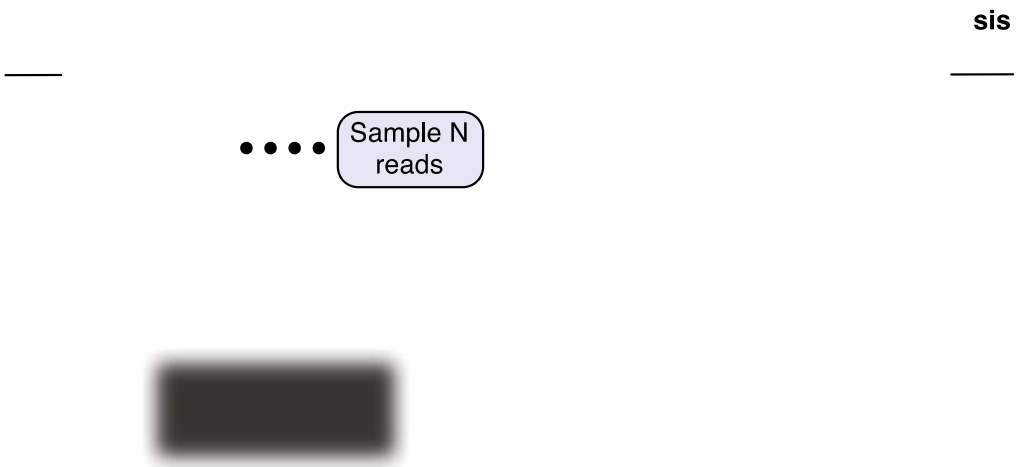


no data

new calls

false positives

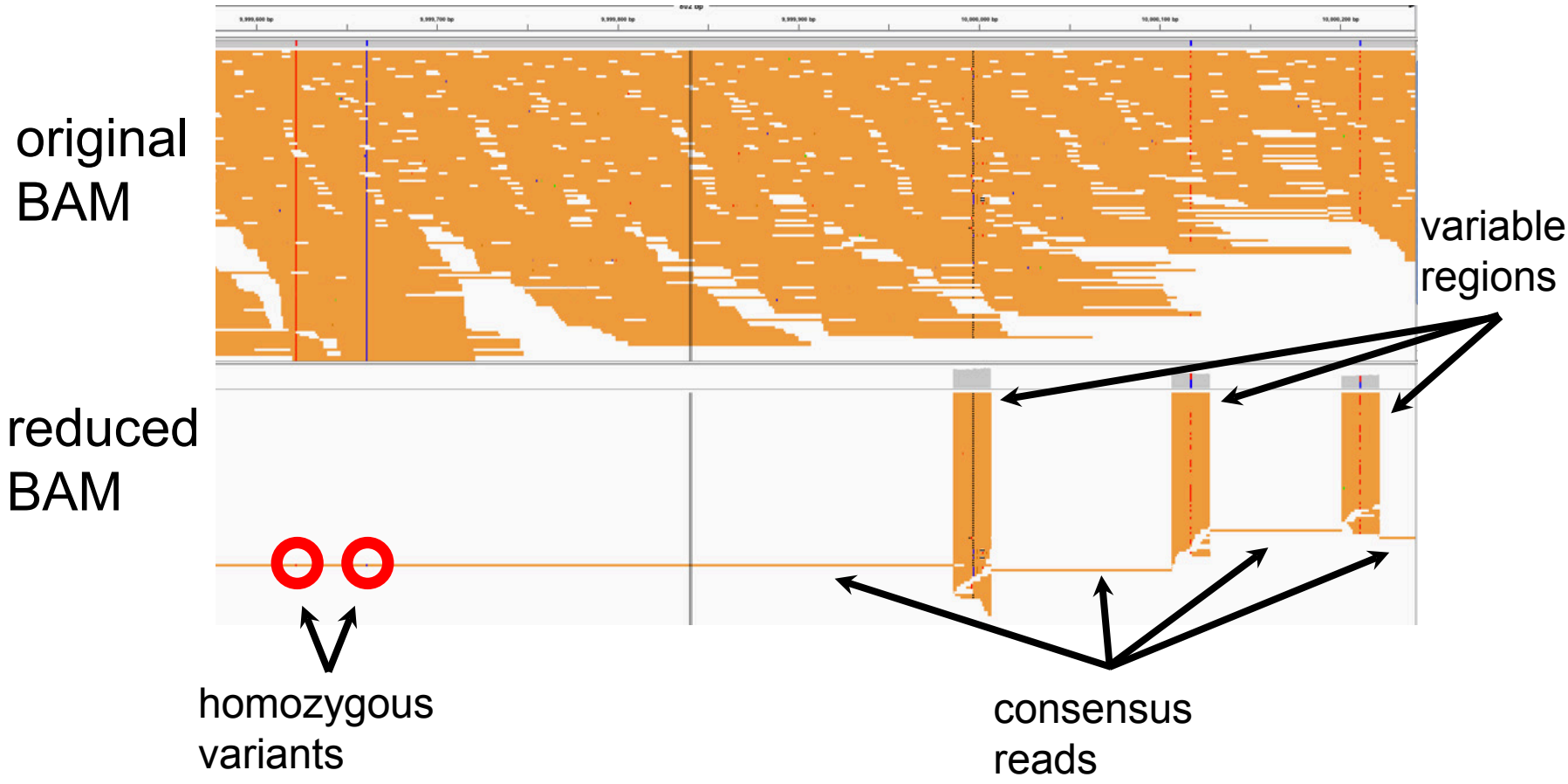
Broad variation discovery pipeline



Mark DePristo

DePristo, M., Banks, E., Poplin, R. et. al, (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet.

Reduced read BAMs



Enables variant-calling across thousands of samples simultaneously

Mauricio Carneiro

Pilot Analysis: 16,500 exomes

- 1000 Genomes
- type 2 diabetes
- autism
- schizophrenia
- controls

~16,500 exomes

aggregation
recalibration
joint calling

ALoFT

Mark DePristo
Khalid Shakir
Monkol Lek
David Altshuler
Mark Daly
Stacey Gabriel
Shaun Purcell
Steve McCarroll
Mike Boehnke
Mark McCarthy
others…

Preliminary results from chr1 (~10% of genome)

~2,800 CPU hours

~30 Gb max memory

Can this be done on a larger scale?

- currently preparing for exome-wide calling in ~20,000 individuals
  - large-scale validation
  - designing LoF arrays
- no fundamental technical barriers to scaling up to larger numbers
- caution: diversity of sequence platforms may soon increase

Analysis

- challenges (many discussed by Peter Donnelly yesterday):
  - variant-calling still immature
  - residual batch/technology effects
  - population structure for rare variants
  - rare variant aggregation
  - reference/annotation errors
  - large-scale validation
- broad phenotype data will impose major multiple testing burden

Access

- providing aggregated, harmonized variant calls will greatly empower **statistical geneticists**
- how do we empower the rest of the research community?
- consider typical use cases:
  - what missense/LoF variants are found in my favourite gene?
  - what phenotypes are they associated with?
  - which variants in my patient's genome (or my genome) are associated with disease?

Possible access models

- open access samples
- streamlined dbGaP process
- “licensed researcher” model:
  - researchers given full access to all data-sets consistent with their license
- central analysis server
  - analysis engine permits analyses that don’t de-identify samples
  - likely more powerful approach for non-statistical geneticists

Key messages

- very large-scale aggregation of sequence and phenotype data does not pose fundamental technical obstacles
- centralized processing and variant-calling is critical (and tractable)
- phenotype data will increase, and harmonization much more challenging
- the curse of multiple testing!
- substantial investment in new interfaces required to maximize research impact