THE UNIVERSITY OF CHICAGO
CENTER FOR CLINICAL CANCER GENETICS AND GLOBAL HEALTH

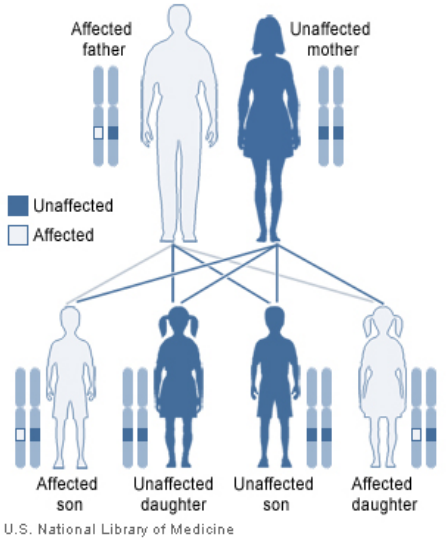
Improving Next-Generation Sequencing Variants Identification In Cancer Genes Using Globus Genomics

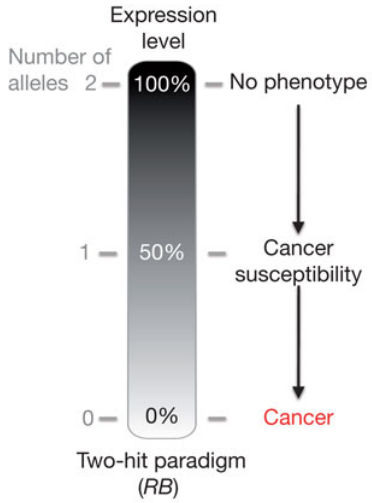
Toshio F Yoshimatsu[1], Yonglan Zheng[1], Alex Rodriguez[3], Vassily Trubetskoy[2], Ravi K Madduri[3], Paul J Dave[3], Nancy J Cox[2], Ian T Foster[3], Olufunmilayo I Olopade[1]

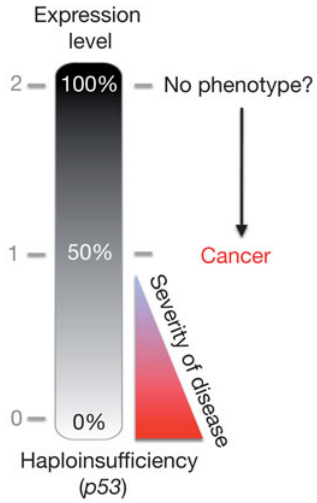
1. Department of Medicine, Section of Hematology/Oncology
2. Department of Medicine, Section of Genetic Medicine
3. Computation Institute

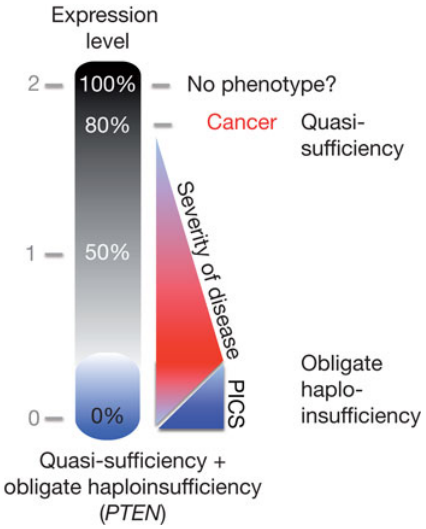
The University of Chicago

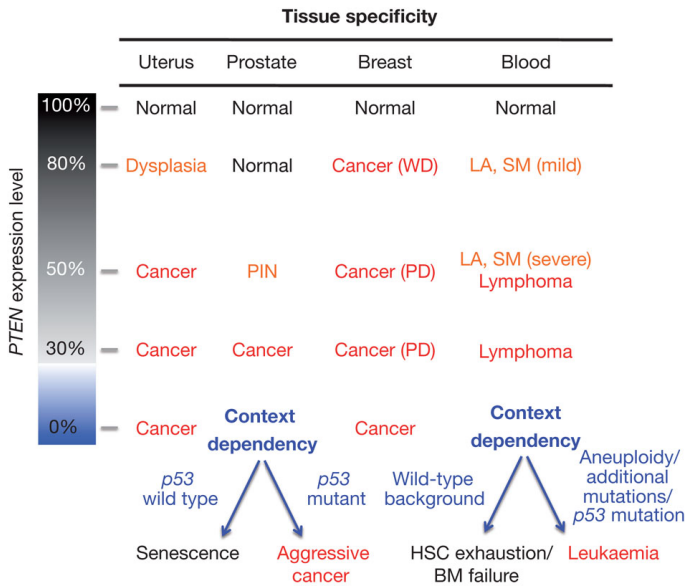
Cancer is a genetic disease





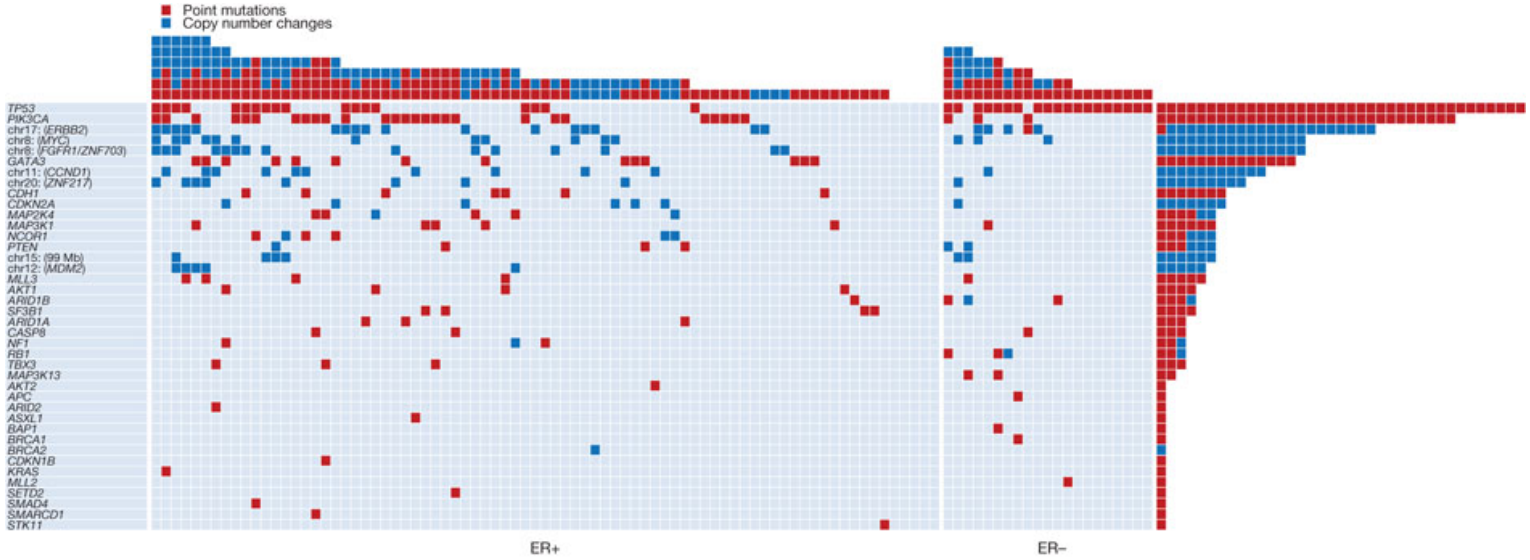


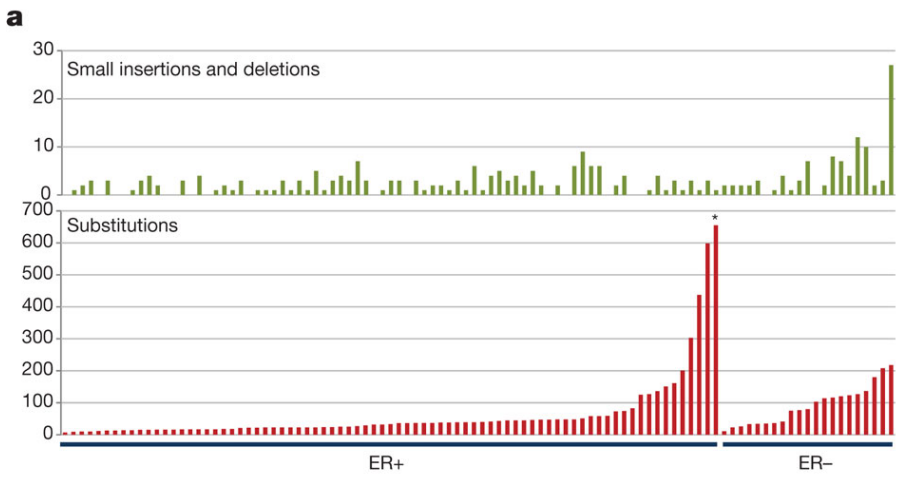


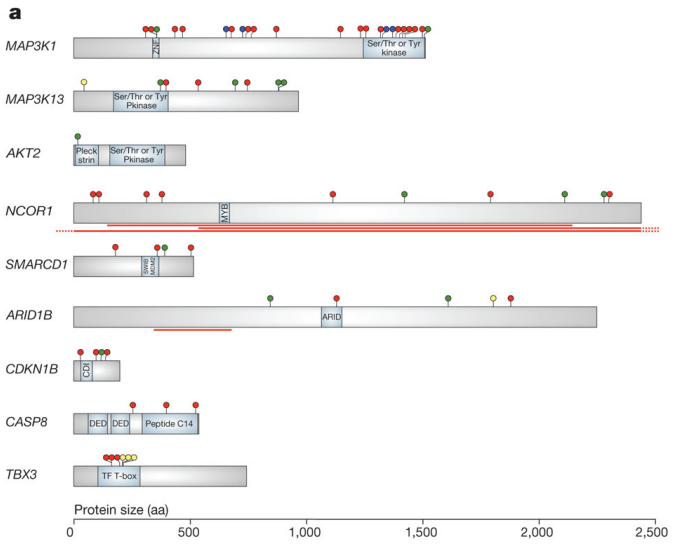


AH Berger *et al. Nature* **476**, 163-169 (2011) doi:10.1038/nature10275

Breast Cancer is extremely heterogeneous

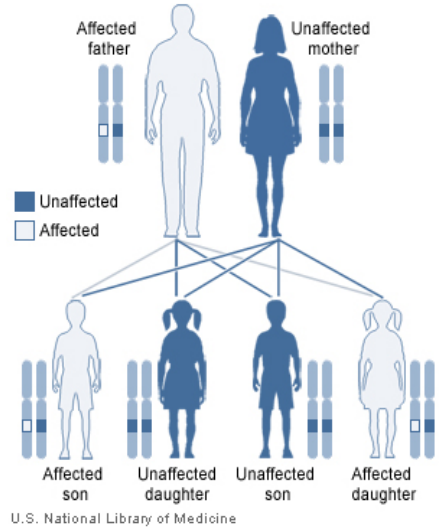


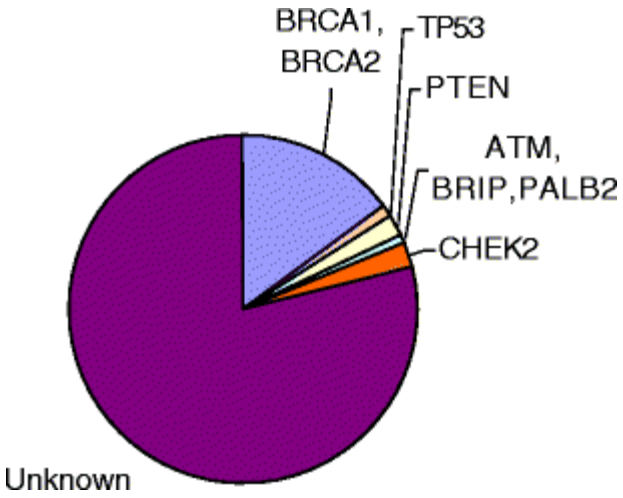


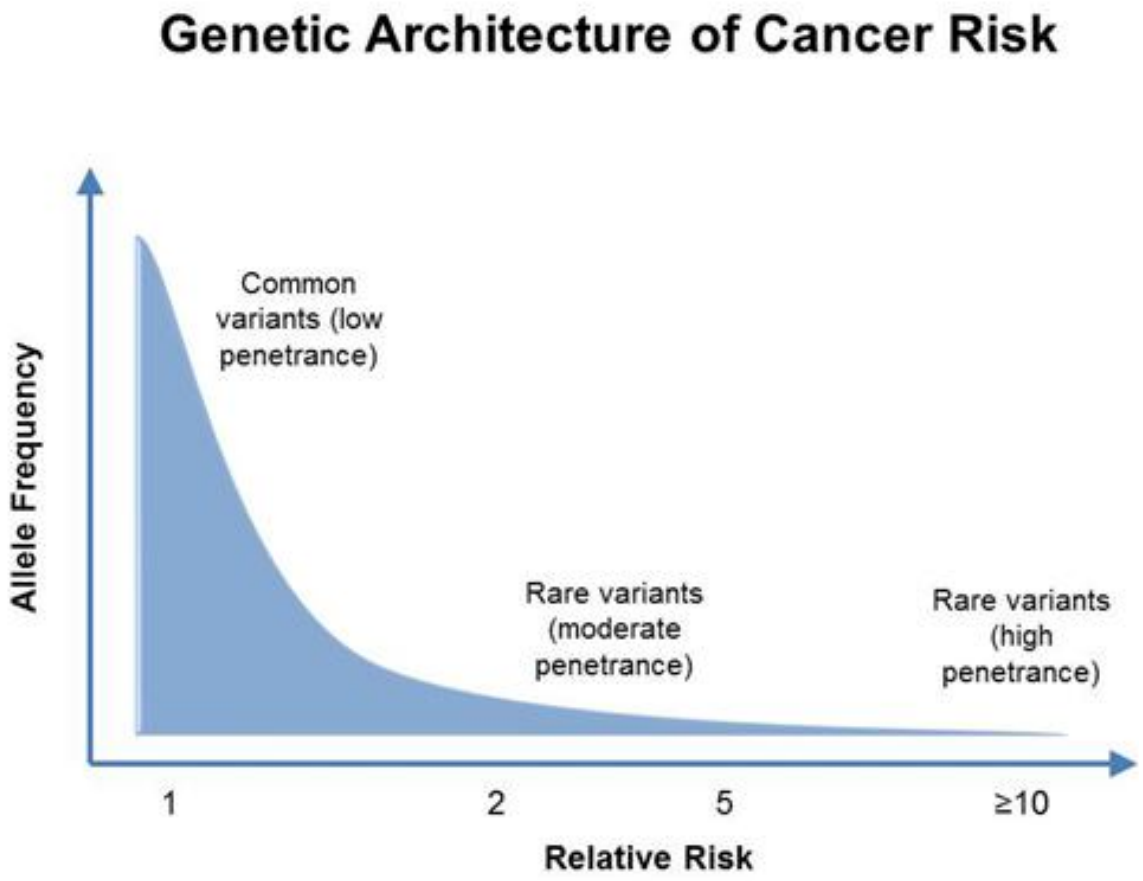


PJ Stephens *et al*. *Nature* **486**, 400-404 (2012)

Strategies for Breast Cancer risk assessment







Van der Groep *et al. Cell Oncol* **34,** 71-88 (2011).

BROCA Cancer Risk Panel

Developed by Dr. Mary-Claire King and colleagues at the University of Washington at Seattle

"This assay sequences all exons and flanking intronic sequences of 50 cancer genes. A total of 1.1 Mb (1.1 Million base pairs) are sequenced and the average coverage ranges from 320 to >1,000 sequencing reads per bp. Genomic regions are captured using biotinylated RNA oliognucleotides (SureSelect), prepared in paired-end libraries with ~200 bp insert size, and sequenced on an Illumina HiSeq2000 instrument with 100 bp read lengths."

(http://web.labmed.washington.edu/tests/genetics/BROCA)

BROCA in 200 Nigerian breast cancer cases

Phenotype	Number
early age onset (<50yrs)	158
familial	20
familial & early age onset	19
others	3

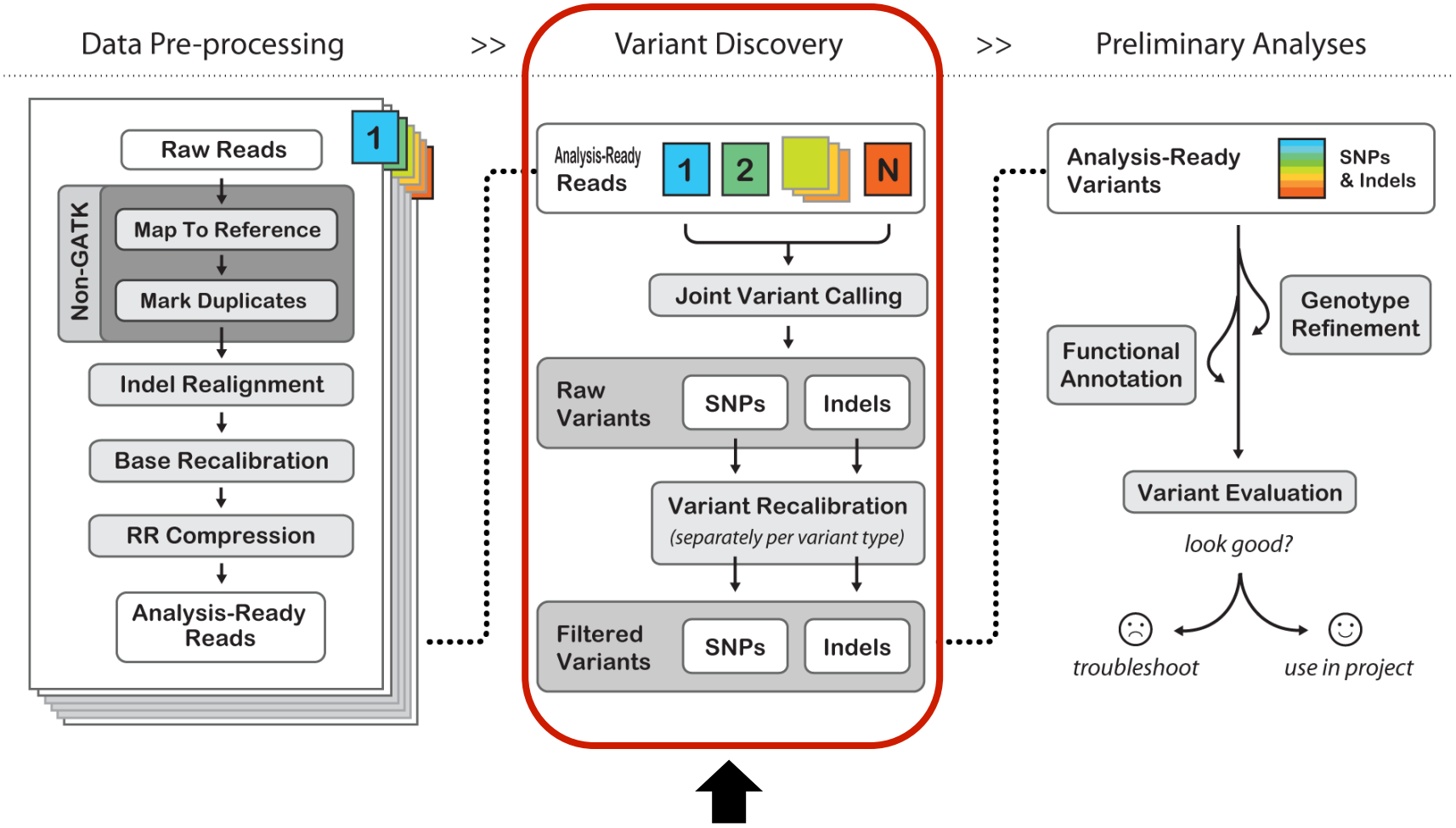
Real mutations are hidden in the noise

How to tell which mismatch is real mutation and which are just noise?
i.e. How can we reduce false positive and false negative detection rates?

Motivation

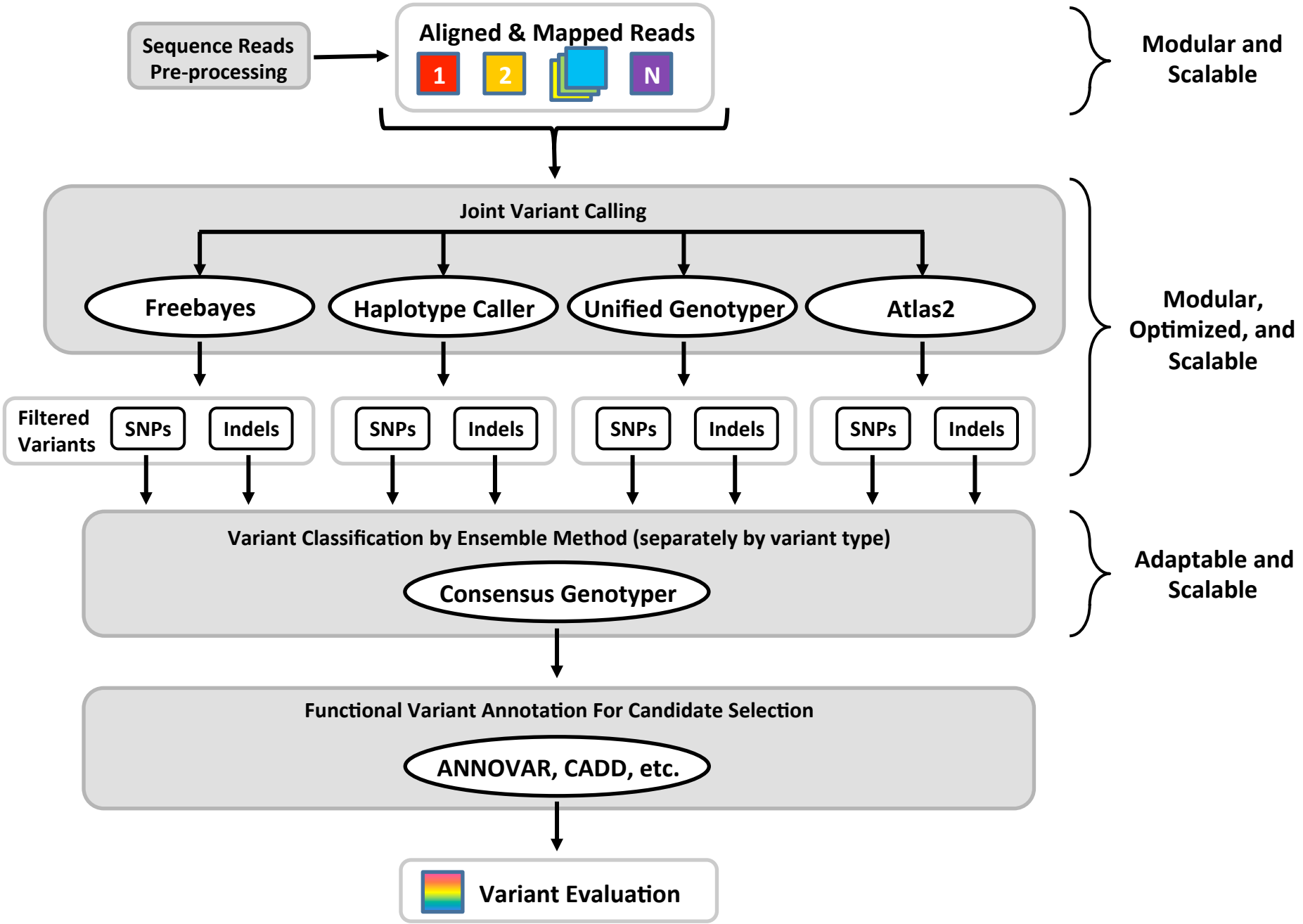
We want to improve sensitivity and specificity of variant identification using next-generation sequencing data

GATK Best Practice



Our work focuses on this step

*Image courtesy of Broad Institute GATK Team
(http://www.broadinstitute.org/gatk/guide/best-practices)


Sequence Reads Pre-processing
Aligned & Mapped Reads
1
2
N
Modular and Scalable
Joint Variant Calling
Freebayes
Haplotype Caller
Unified Genotyper
Atlas2
Modular, Optimized, and Scalable
Filtered Variants
SNPs
Indels
SNPs
Indels
SNPs
Indels
SNPs
Indels
Variant Classification by Ensemble Method (separately by variant type)
Consensus Genotyper
Adaptable and Scalable
Functional Variant Annotation For Candidate Selection
ANNOVAR, CADD, etc.
Variant Evaluation

Computation Performance

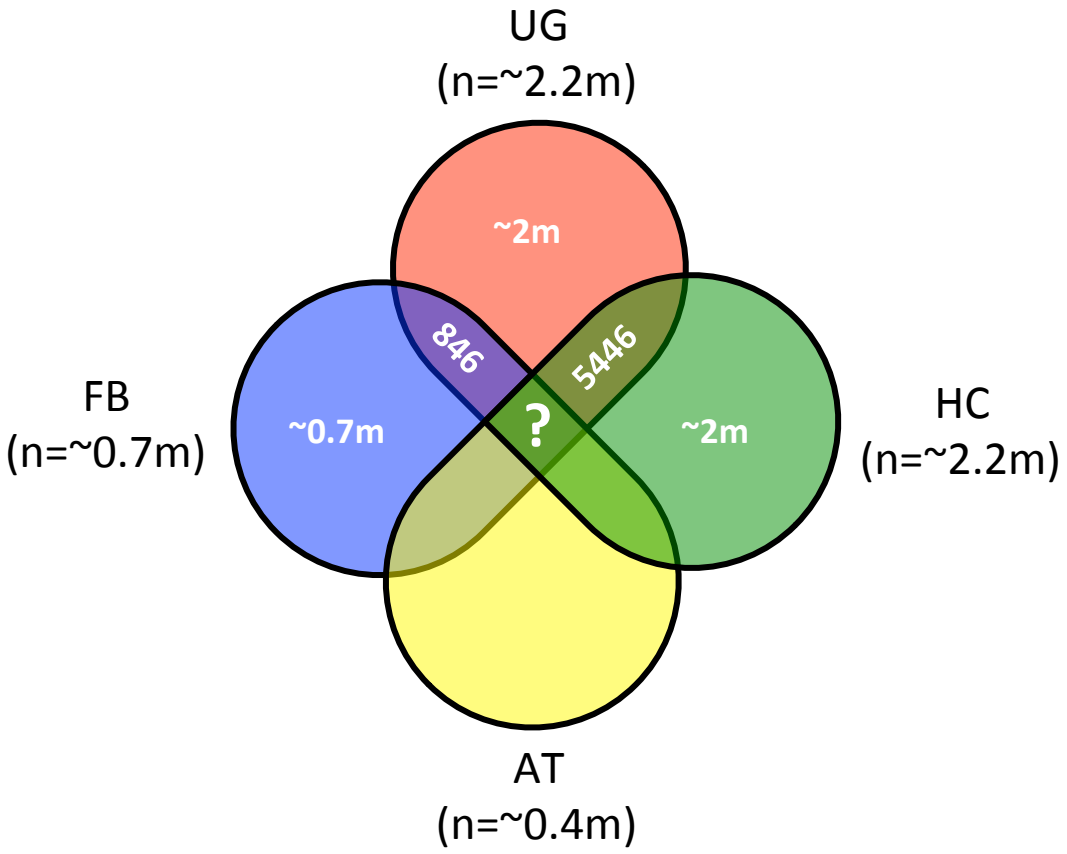
TABLE 2. Summary for the alignment of 200 BROCA target exome-seq Fastq files.

	GATK UG	GATK HC	Freebayes	ATLAS2-SNP
Input Type	BAM	BAM	BAM	BAM
Paralellization Level	Chromosome (24X)	Chromosome (24X)	Chromosome (24X)	Input File (200X)
Input Size	67.5 GB	67.5 GB	67.5 GB	67.5 GB
Data Generated	54 GB	19 GB	0.70 GB	25 GB
Output Size	615 MB	615 MB	315 MB	1.5 GB
Total CPU time	22 hours	27.3 hours	37.2 hours	708 hours
Walltime for analysis	6.75 hours	2.33 hours	4.33 hours	23 hours
Worker Nodes Used	13	23	24	200

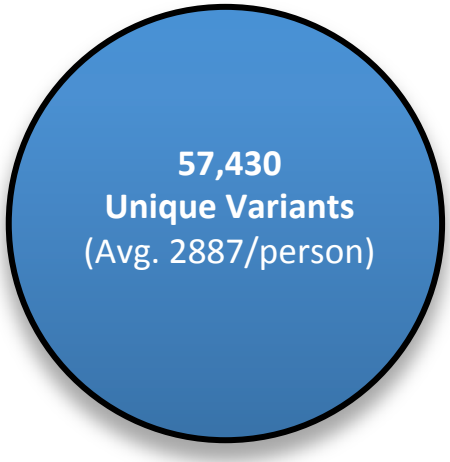
***Optimizations can still be achieved by running multiple chromosomes and samples on the same worker node**

Variant Calling Performance - SNPs

Globus Genomics



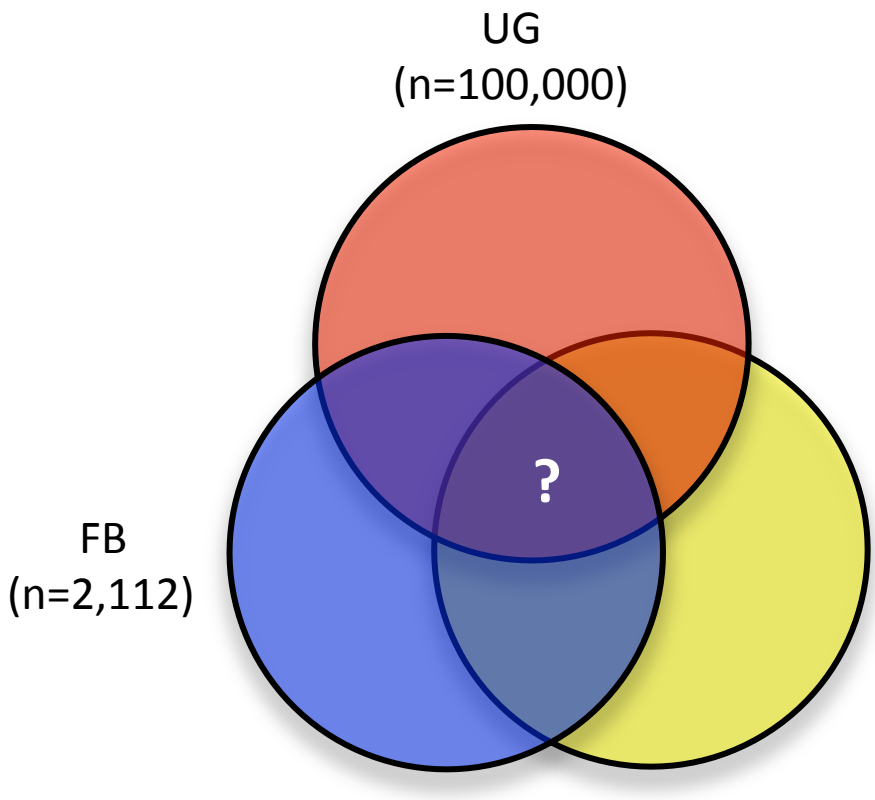
BROCA Default



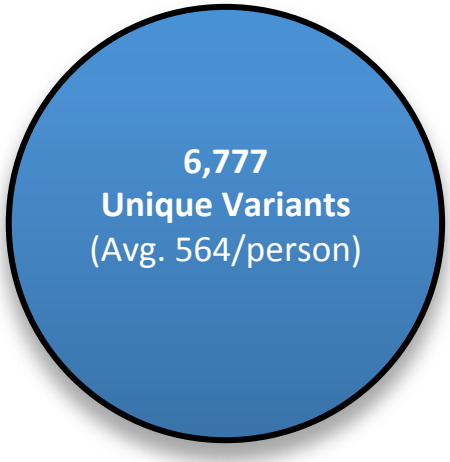
UG = Unified Genotyper; HC = Haplotype Caller; FB = Freebayes; AT = Atlas2

Variant Calling Performance - Indels

Globus Genomics



BROCA Default



UG = Unified Genotyper; FB = Freebayes; AT = Atlas2

Summary

- We believe our variant calling approach can be easily adapted and modified for generalization and implementation in other genome-wide and large-scale NGS variant analysis.
- Our method fully utilizes the capacity of Globus Genomics to make the workflow scalable, modular, and adaptable. Consequently, analysis time were dramatically shortened (often in the order of magnitude).
- Incorporation of Globus Online into the pipeline automates and facilitates transfer and sharing of large data.
- We are further developing the previous analysis method to include other variation types into consideration (e.g. indels, copy number change).
- Our ultimate goal is to help patients make informed decision, and not to scare them with false alarm.

Acknowledgements

University of Chicago

Dept. Medicine, Hem/Onc

Niu Qun

Yonglan Zheng

Funmi Olopade

Dept. Medicine, Genetic Medicine

Vassily Trubetskoy

Nancy Cox

Computation Institute

Alex Rodriguez

Ravi Madduri

Paul Dave

Ian Foster

University of Washington at Seattle

Mary-Claire King

Tom Walsh

Ming Lee

University of Ibadan

Oladosu Ojengbede

Temidayo Ogundiran

Abideen Oluwasola

Abayomi Odetunde

Support

Susan G Komen for the Cure

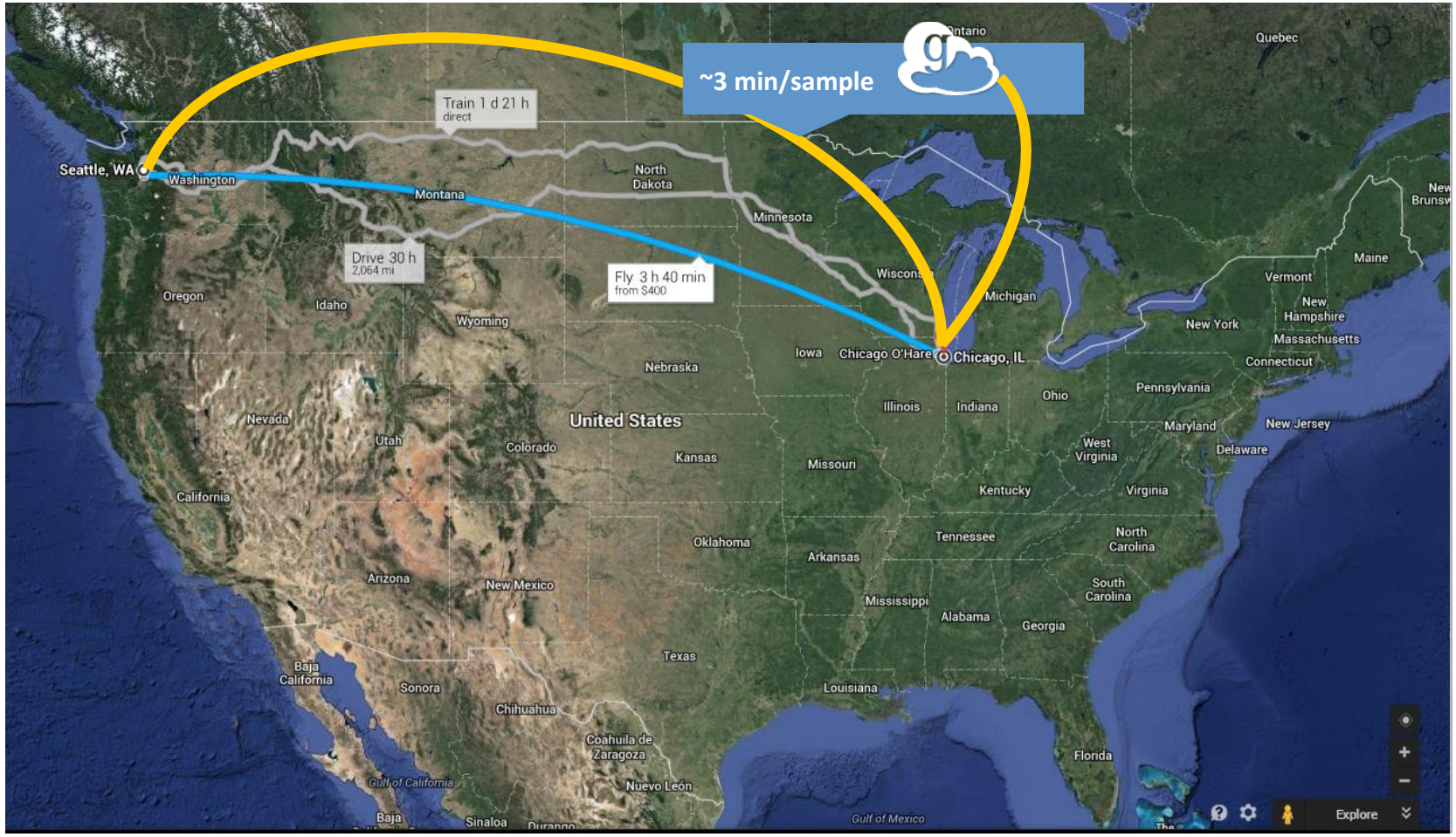
Breast Cancer Research Foundation

globus genomics

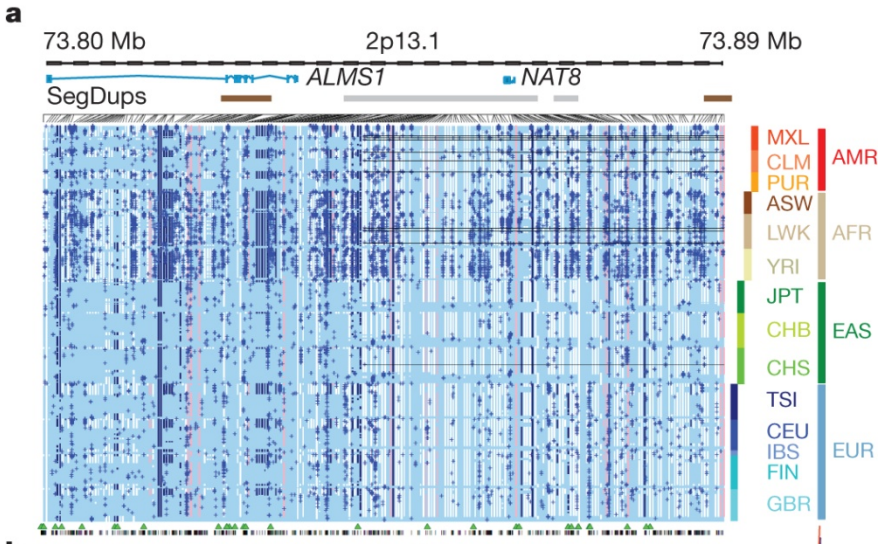
THE UNIVERSITY OF CHICAGO

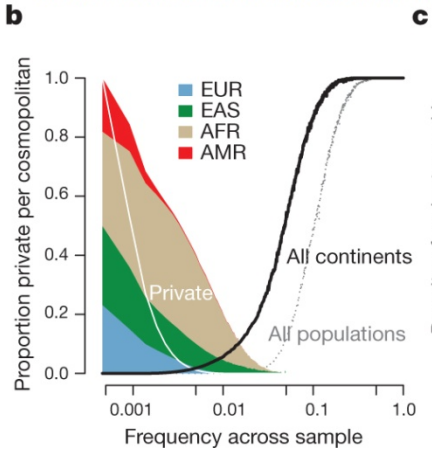
CENTER FOR CLINICAL CANCER GENETICS AND GLOBAL HEALTH

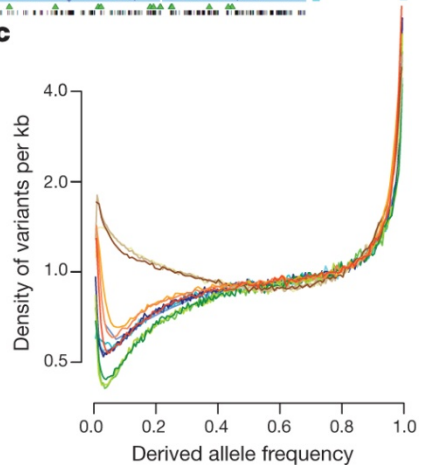
Data Transfer

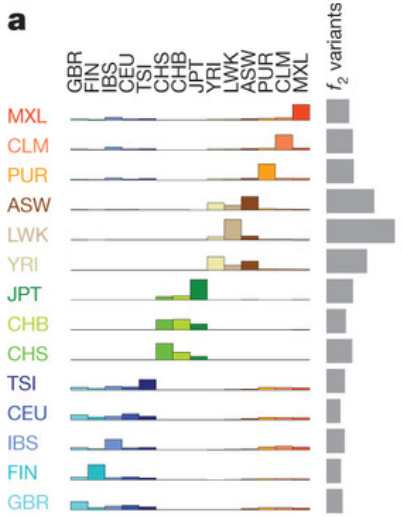


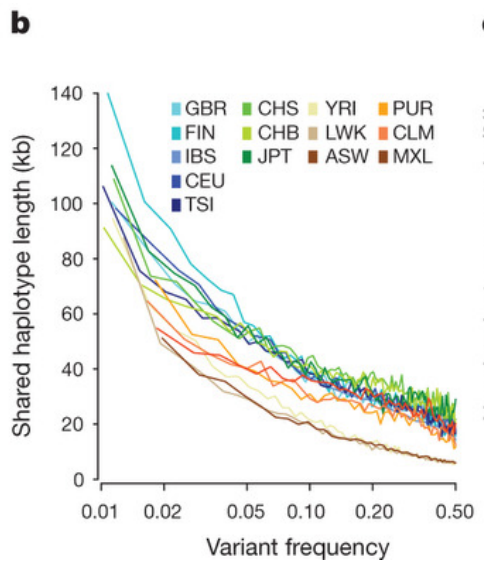
Africans are known to have complex genetic background

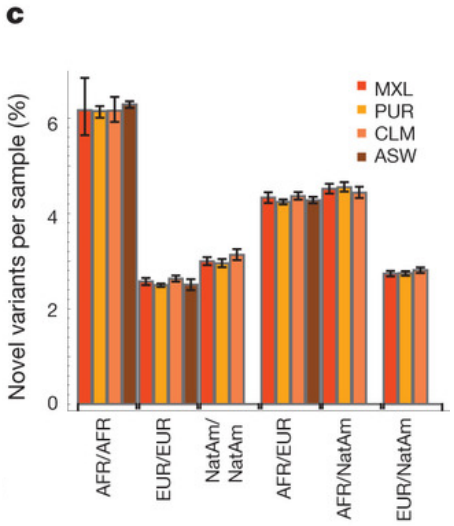








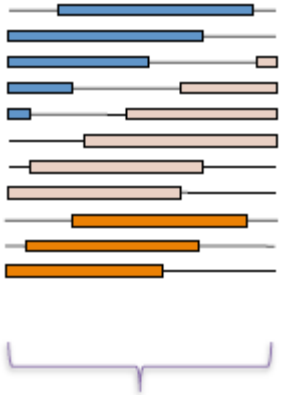




The 1000 Genomes Project Consortium Nature **491**, 56-65 (2012)

General workflow of NGS

Sequencing Output

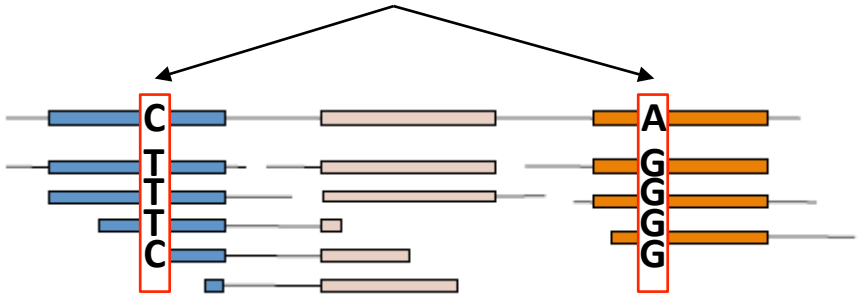


Enormous pile of short reads from NGS

Map & Align

Cleaning and Organizing Data

We are searching for variants



Reference genome

Reads mapped to reference