THE UNIVERSITY OF CHICAGO

Implementation of Consensus Variant Calling using Globus Genomics

Vassily Trubetskoy
April 16, 2014

Road Map

Context:

Disease/medical genetics, next generation sequencing

Consensus Variant Calling:

motivation, high level description

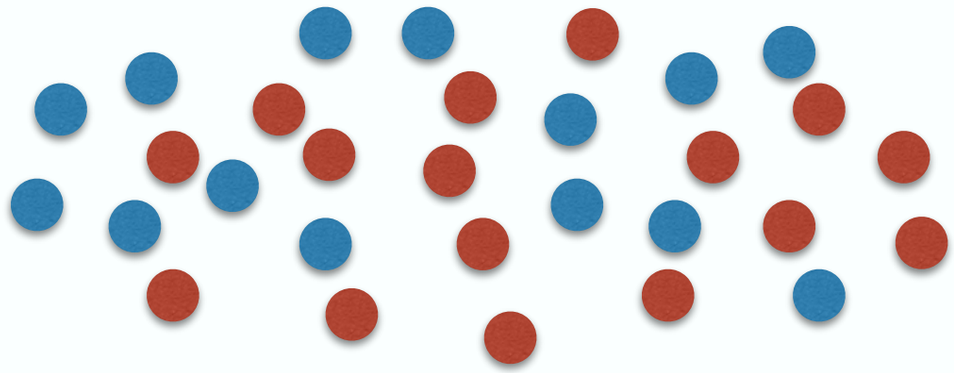
Globus genomics experiments:

Implementation, testing in ACE autism dataset

Context

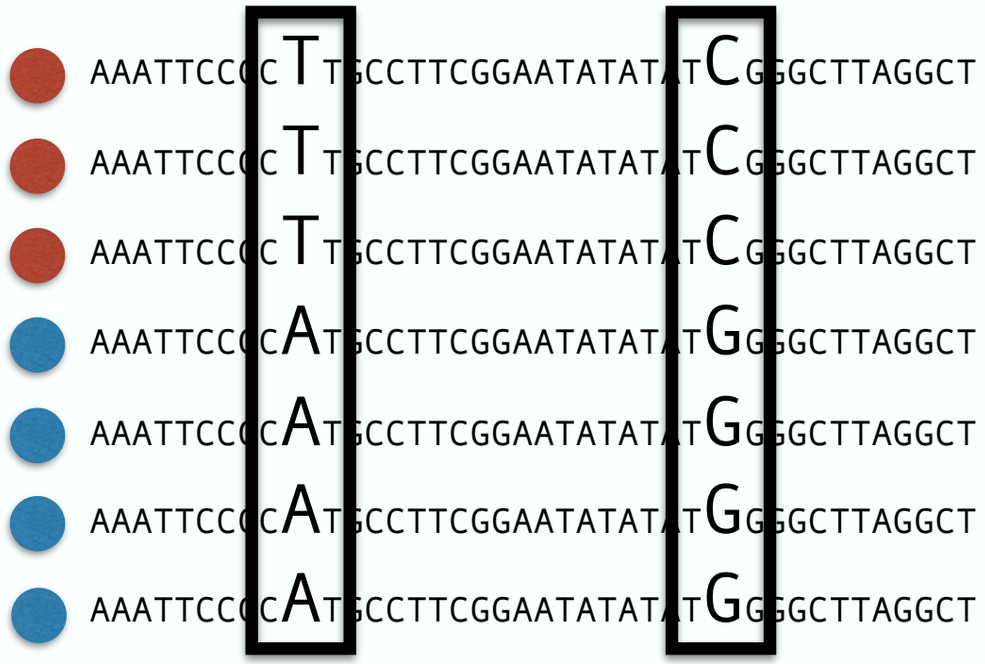
Disease and Medical Genetics

population

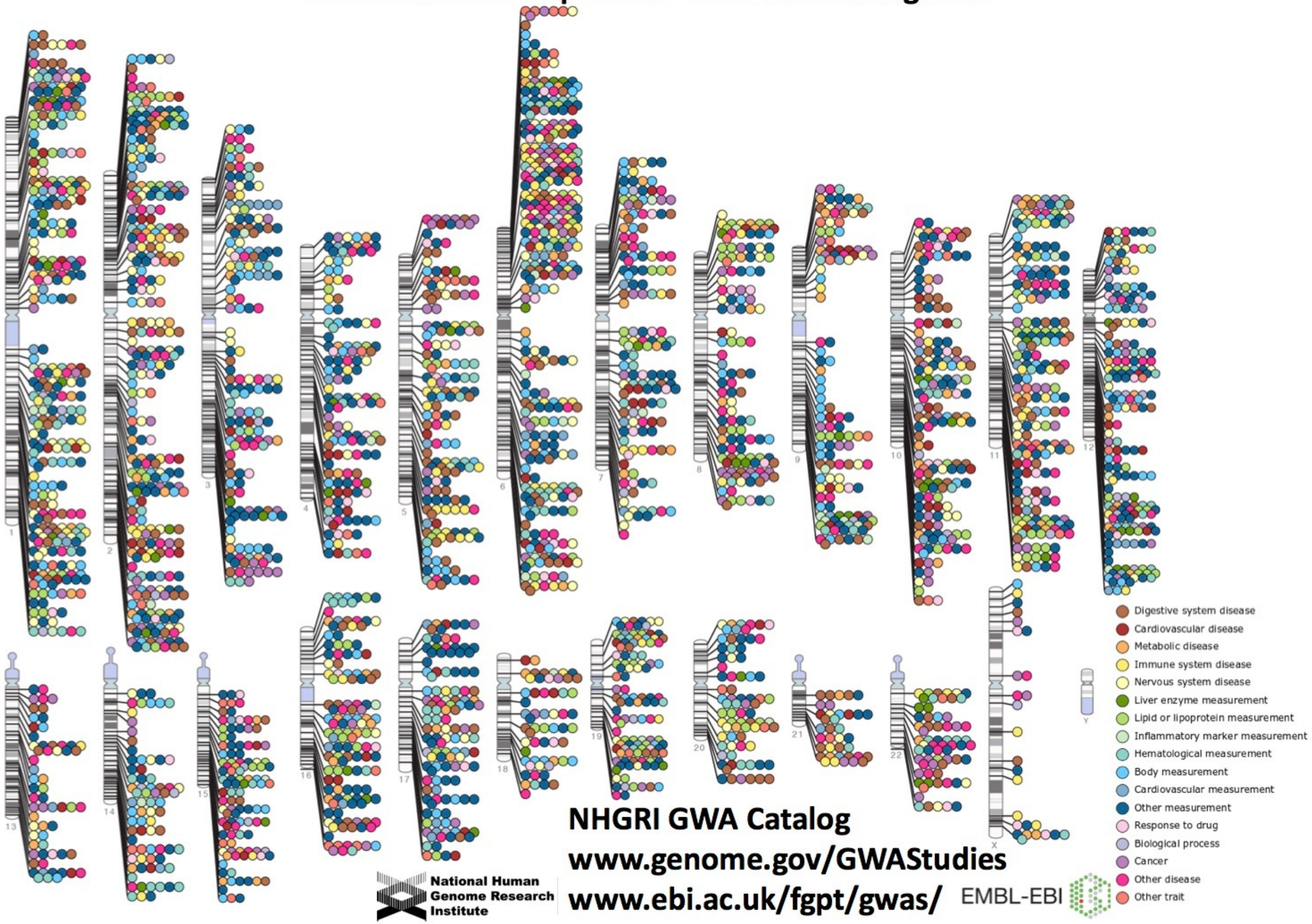


diseased

normal



Published Genome-Wide Associations through 12/2012
Published GWA at p≤5X10^{-8} for 17 trait categories

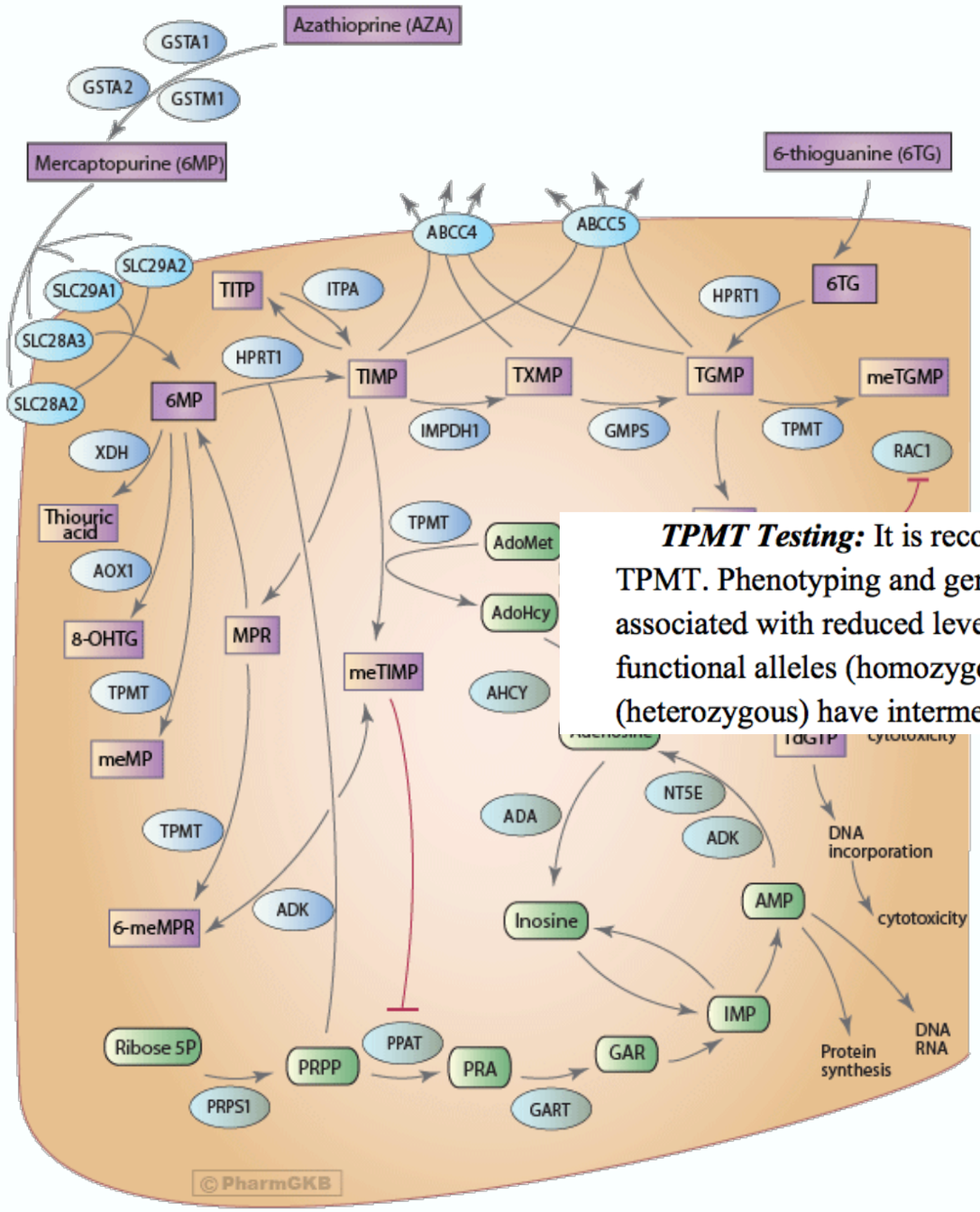


NHGRI GWA Catalog
www.genome.gov/GWAStudies
www.ebi.ac.uk/fgpt/gwas/

National Human Genome Research Institute

EMBL-EBI

Medical Genetics



IMURAN® (azathioprine)

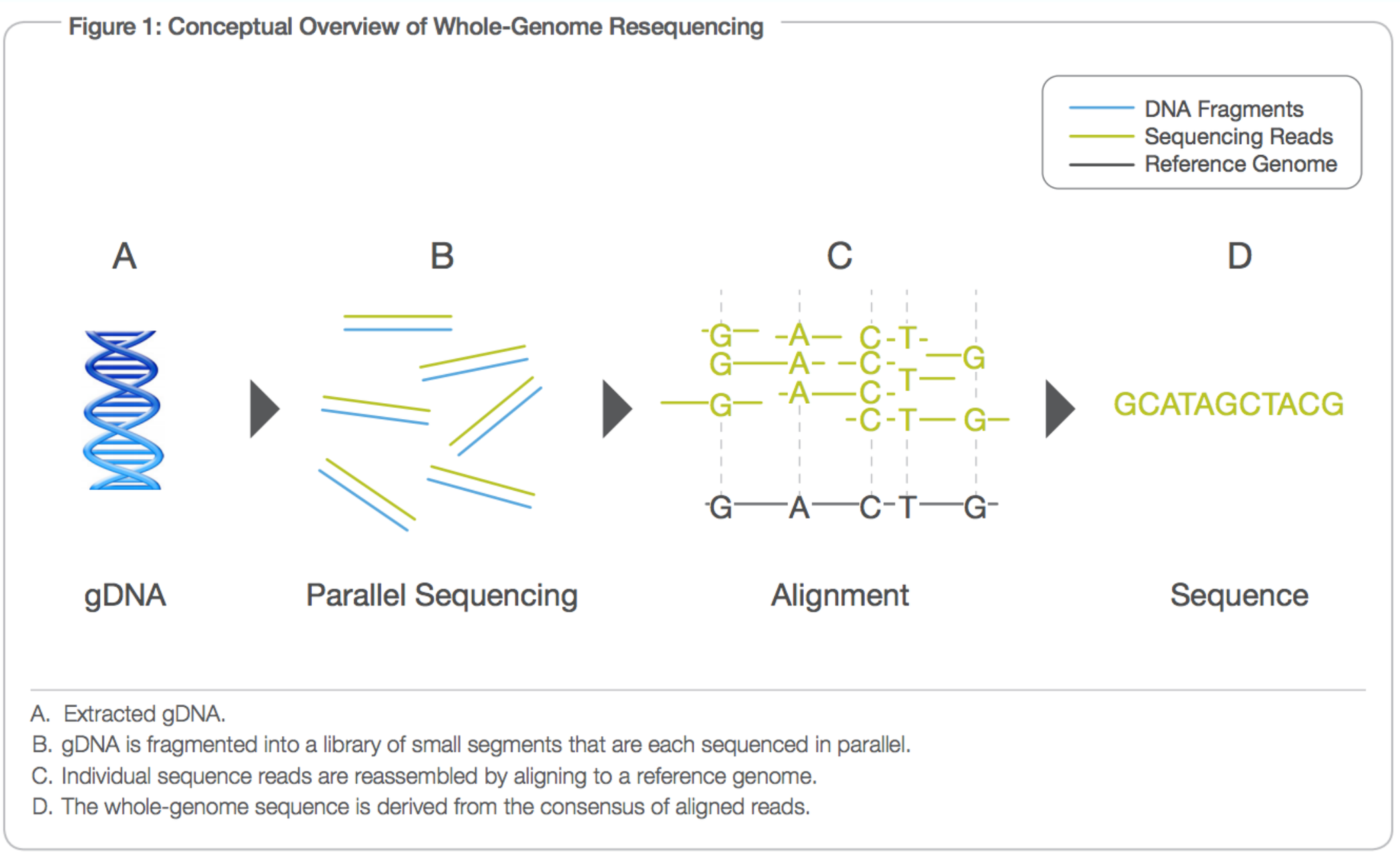
50-mg Scored Tablets

Rx only

TPMT Testing: It is recommended that consideration be given to either genotype or phenotype patients for TPMT. Phenotyping and genotyping methods are commercially available. The most common non-functional alleles associated with reduced levels of TPMT activity are *TPMT*2, TPMT*3A and TPMT*3C*. Patients with two non-functional alleles (homozygous) have low or absent TPMT activity and those with one non-functional allele (heterozygous) have intermediate activity. Accurate phenotyping (red blood cell TPMT activity) results are not

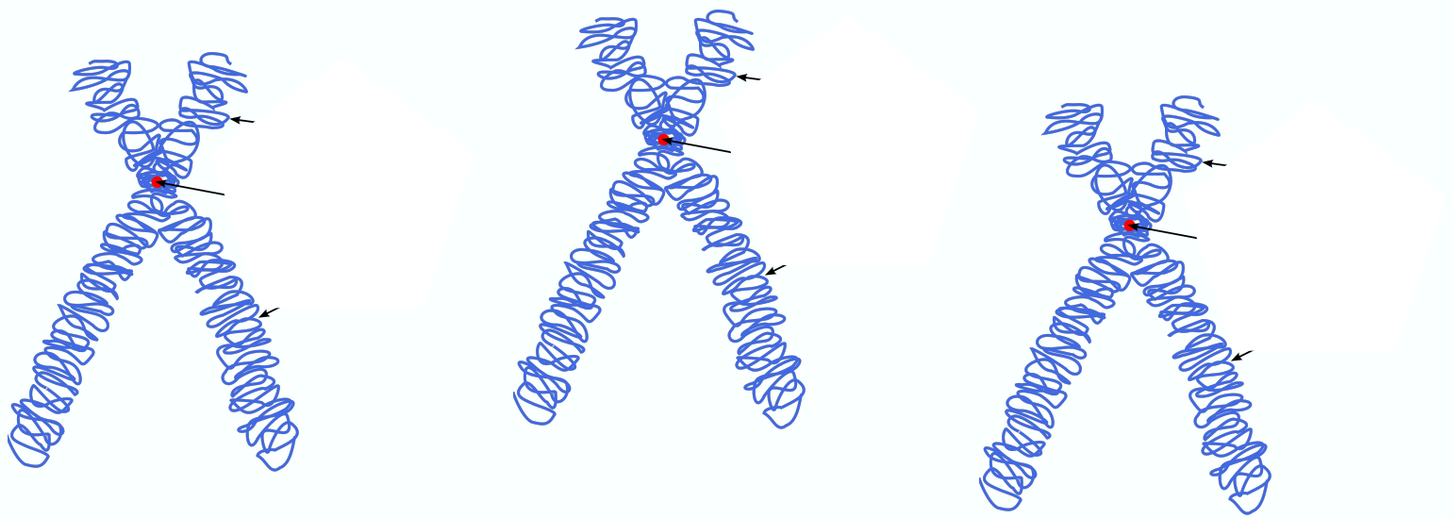
credit: PharmGKB

What is next generation sequencing?

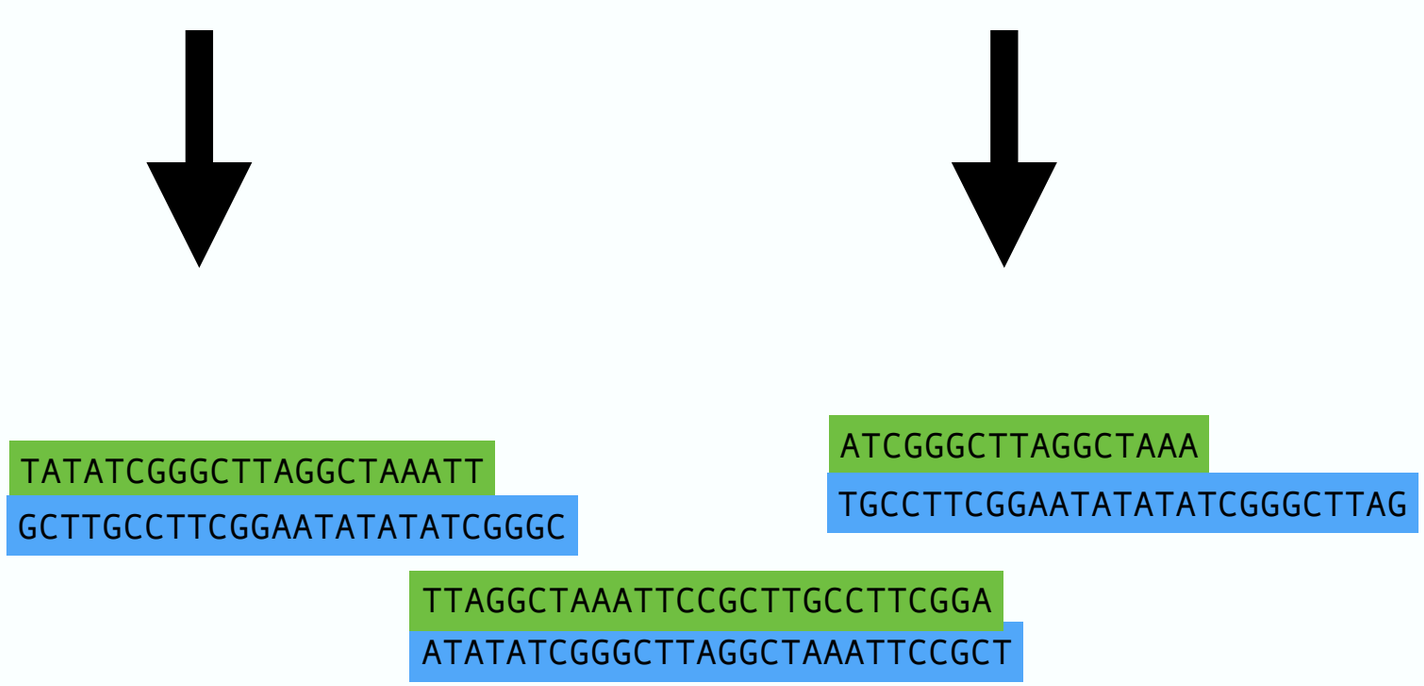


Credit: Illumina pub. 770-2012-008

Next gen sequencing

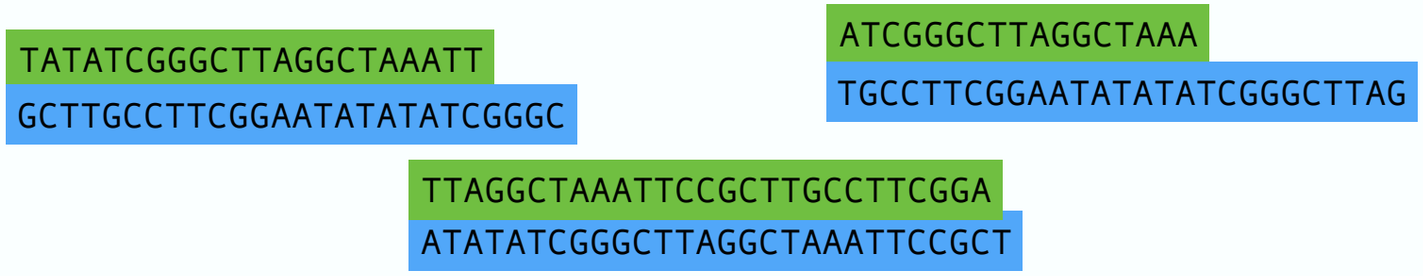


extract DNA

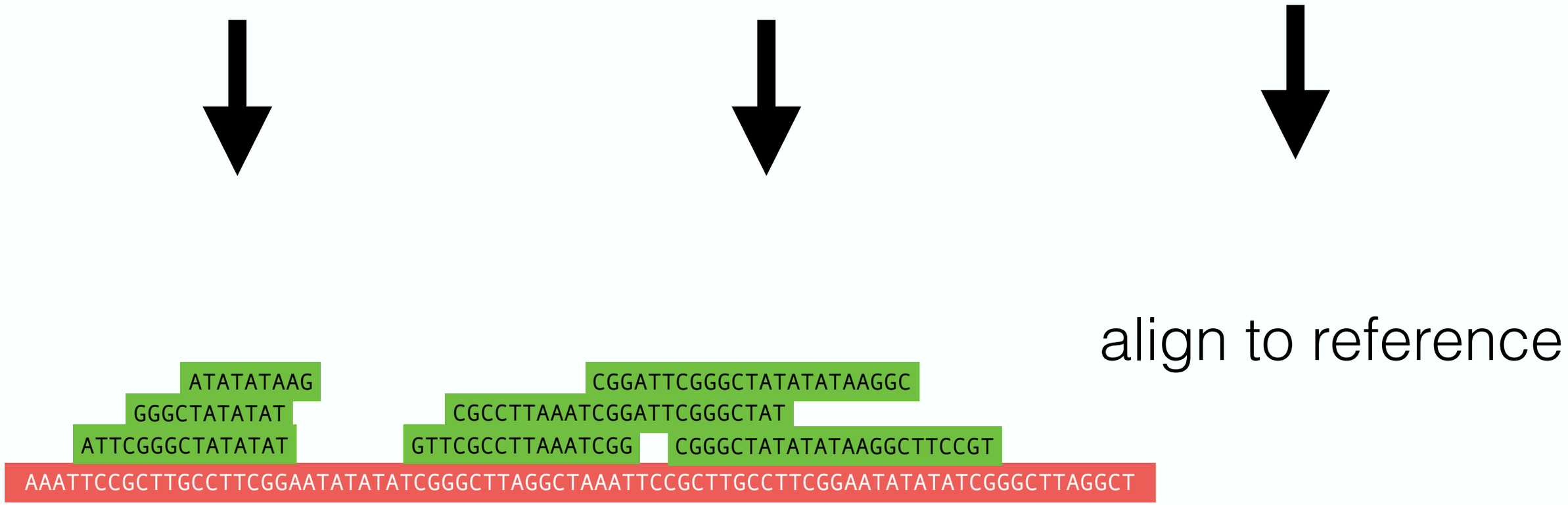


fragment and *sequence

Next gen sequencing



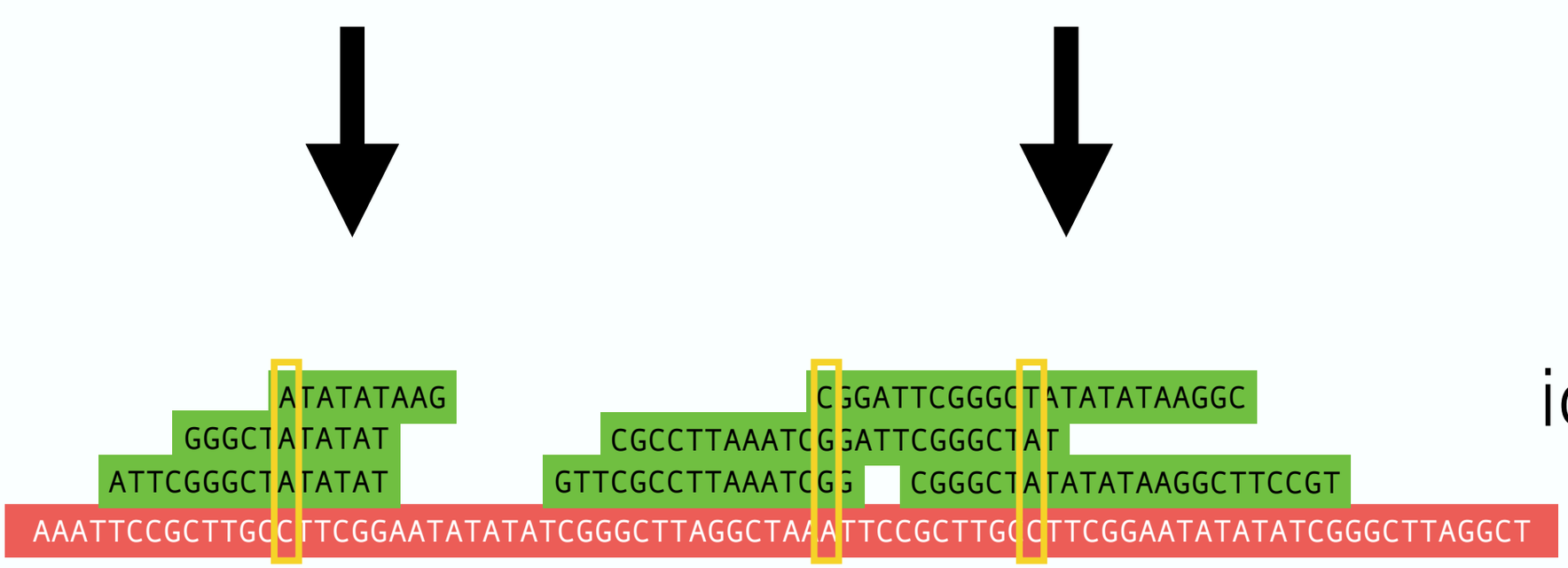
fragment and *sequence



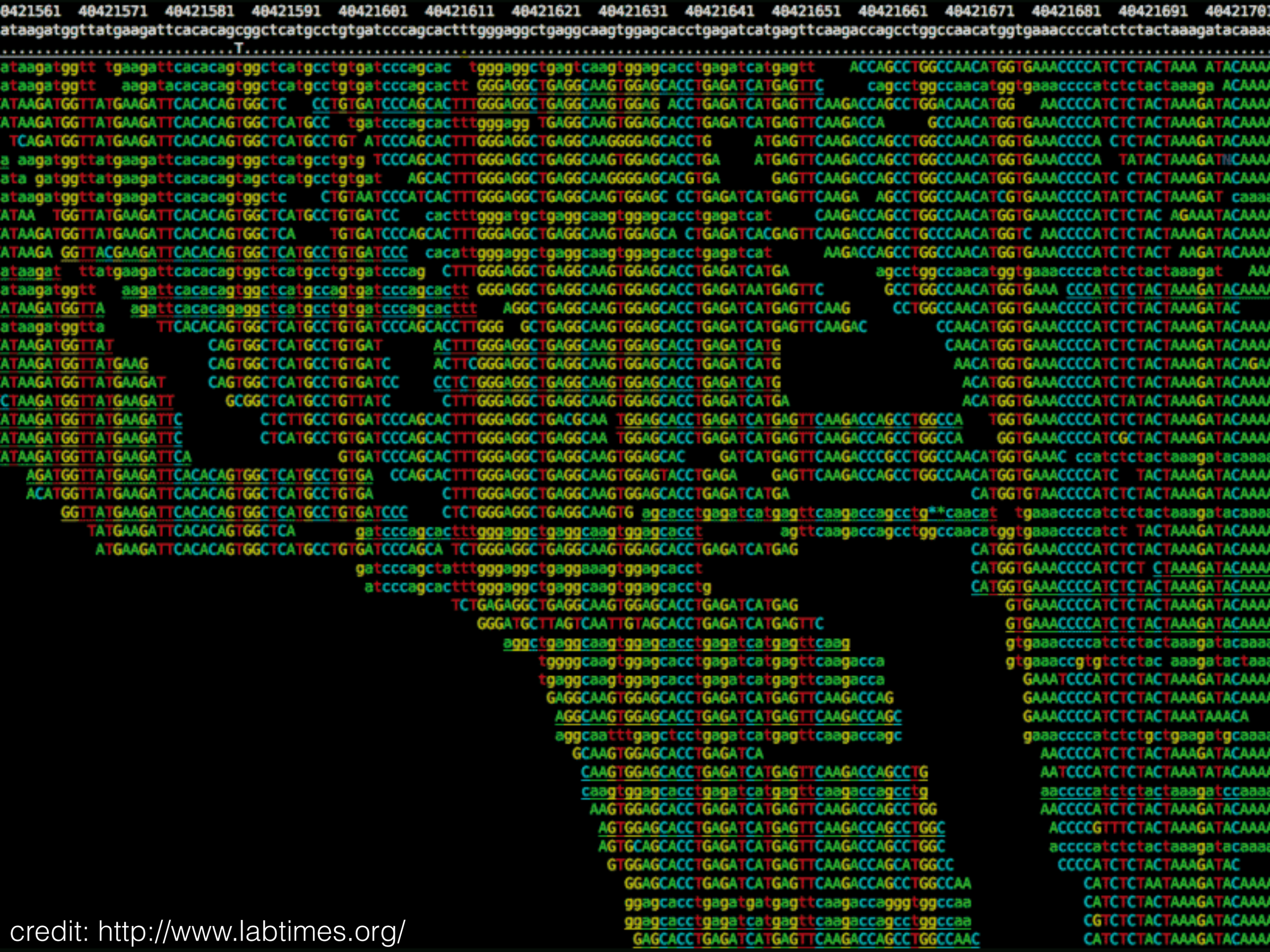
align to reference

Next gen sequencing

align to reference



identify variants


credit: http://www.labtimes.org/

Data scale

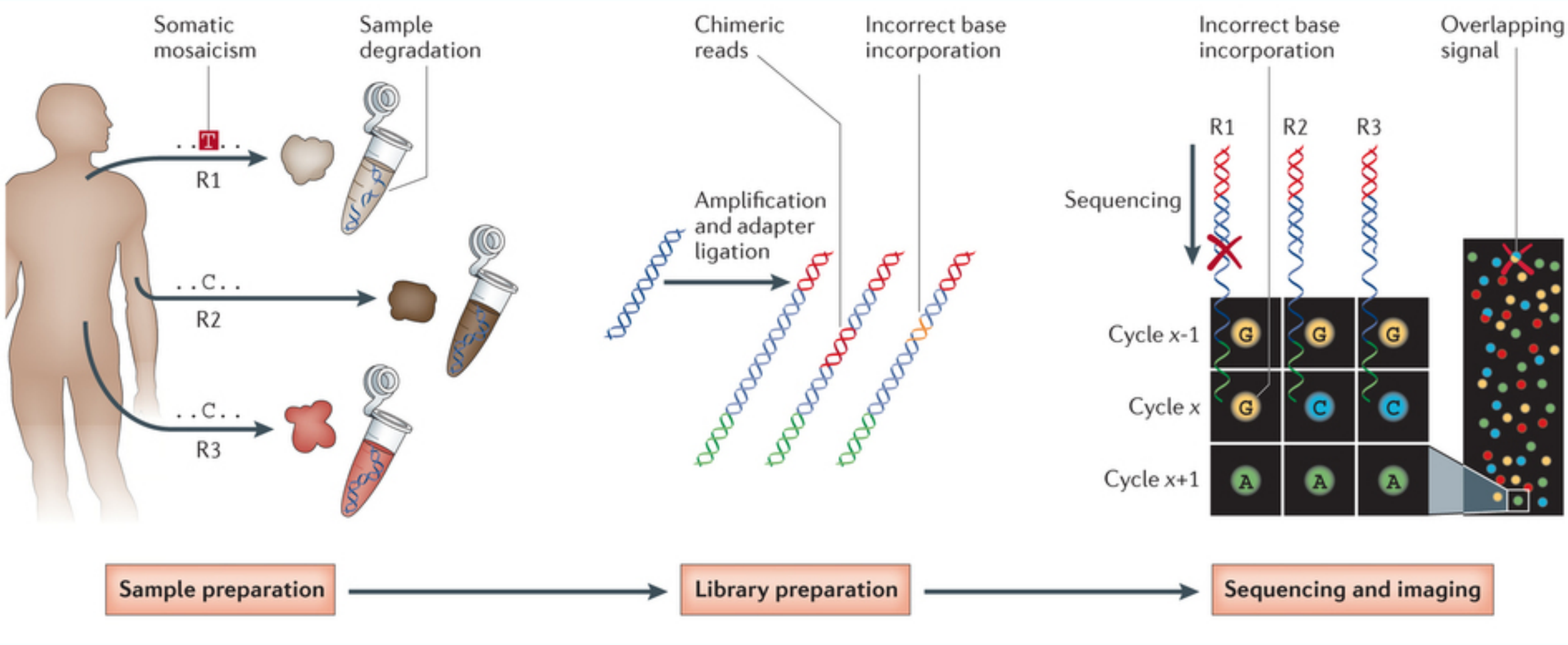
Type II Diabetes Genes Consortium

- 600 whole genomes with 250 TB in raw reads (~40 GB in genotype data)

The Cancer Genome Atlas Project

- large scale data collection in many patients, many cancers, and many tissues
- Sept 2013: 9k patients, 147k files, ~13Tb of genotype data

Cumulative Error



Robasky et. al Nature Reviews Genetics 2013

Consensus variant calling

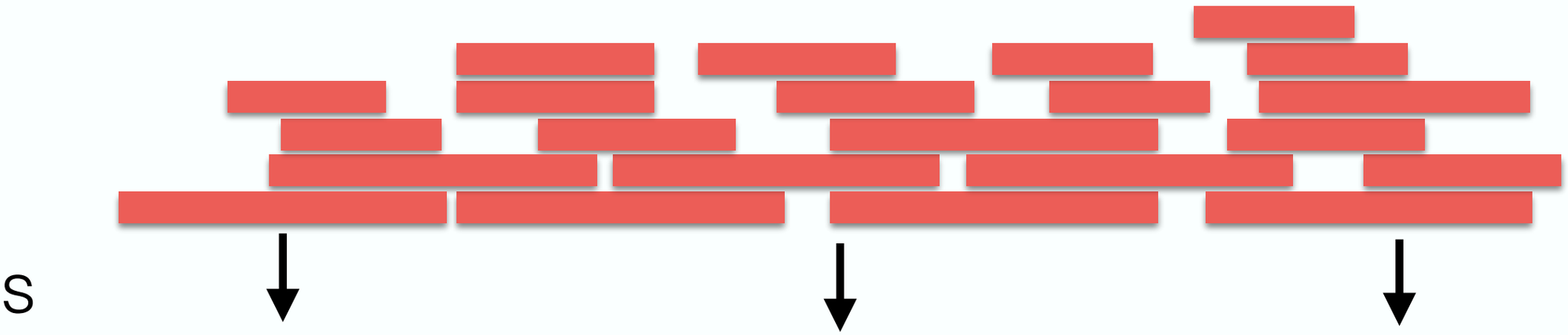
Motivation

- Cumulative sequencing error
- Performance of different models of variation
- Existing ensemble methods

Overview

aligned reads

fit models



consensus sites

```
SNP1:A|T
SNP2:C|T
SNP3:G|C
```

```
SNP1:A|T
SNP2:C|T
SNP3:G|C
```

```
SNP1:A|T
SNP2:C|T
SNP3:G|C
```

consensus genotypes

```
SNP1:A|T
SNP2:C|T
SNP3:G|C
```

Globus Genomics

Motivation

Accessibility

Usable by non-programmers

Single interface for similar tasks

Open source

Reproducibility

Share *exact* methods

Data provenance

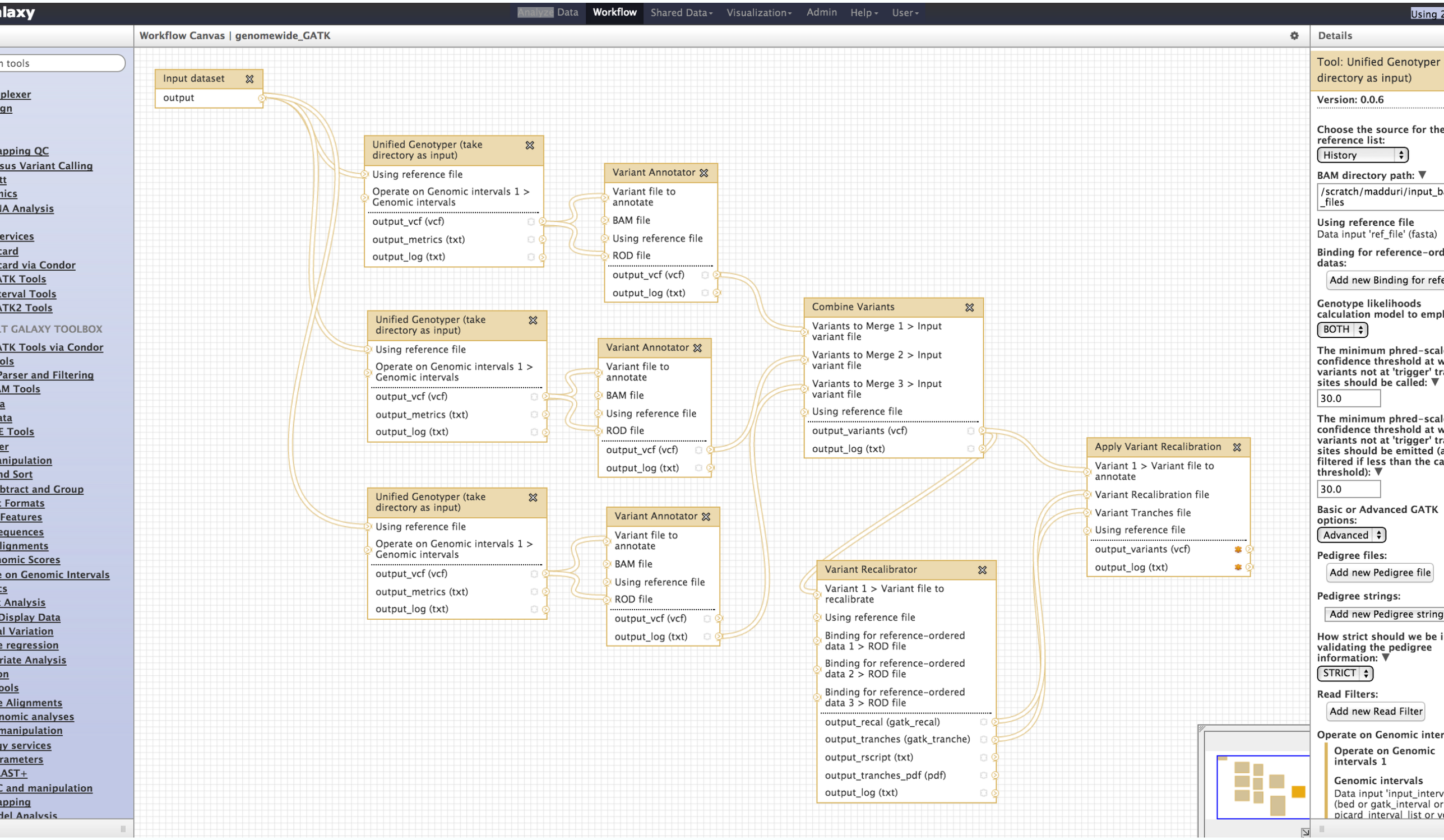
Efficiency

Scale storage

Parallel execution

Iterate workflow construction

Accessibility -> Galaxy + Globus Online



Reproducibility -> Galaxy + Globus Online

Galaxy | Analyze Data | Workflow | Shared Data | Visualization | Admin | Help | User | Using 421.2 GB

Tools

search tools

- Globus
- demultiplexer
- Novoalign
- Atlas2
- Consensus Genotyper for Exome Variants (CGES)
- Polymutt
- Proteomics
- NGS: RNA Analysis
- Miso
- NCBO services

Saved Histories

search history names and tags

Advanced Search

Name	Datasets		Tags	Sharing	Size on Disk	Created	L
freebayes_resubmission	61	30	0 Tags		1.5 GB	6 days ago	6
freebayes_resubmission_testing	7	12	0 Tags		575.5 MB	Mar 25, 2014	A
atlas_resubmission_testing	5		0 Tags	Accessible	4.5 GB	Mar 26, 2014	A
GATK_resubmission	134	17	0 Tags		4.6 GB	Apr 03, 2014	A

History

GATK_resubmission
4.6 GB

- 131: Apply Variant Recalibration on data 30, data 126, and others (log)
- 130: Apply Variant Recalibration on data 30, data 126, and others (Variants File)
- 129: Variant Recalibrator on data 30, data 26, and others (log)
- 128: Variant Recalibrator on data 30, data 26, and others (PDF

globus | Manage Data | Groups | Support | vasya

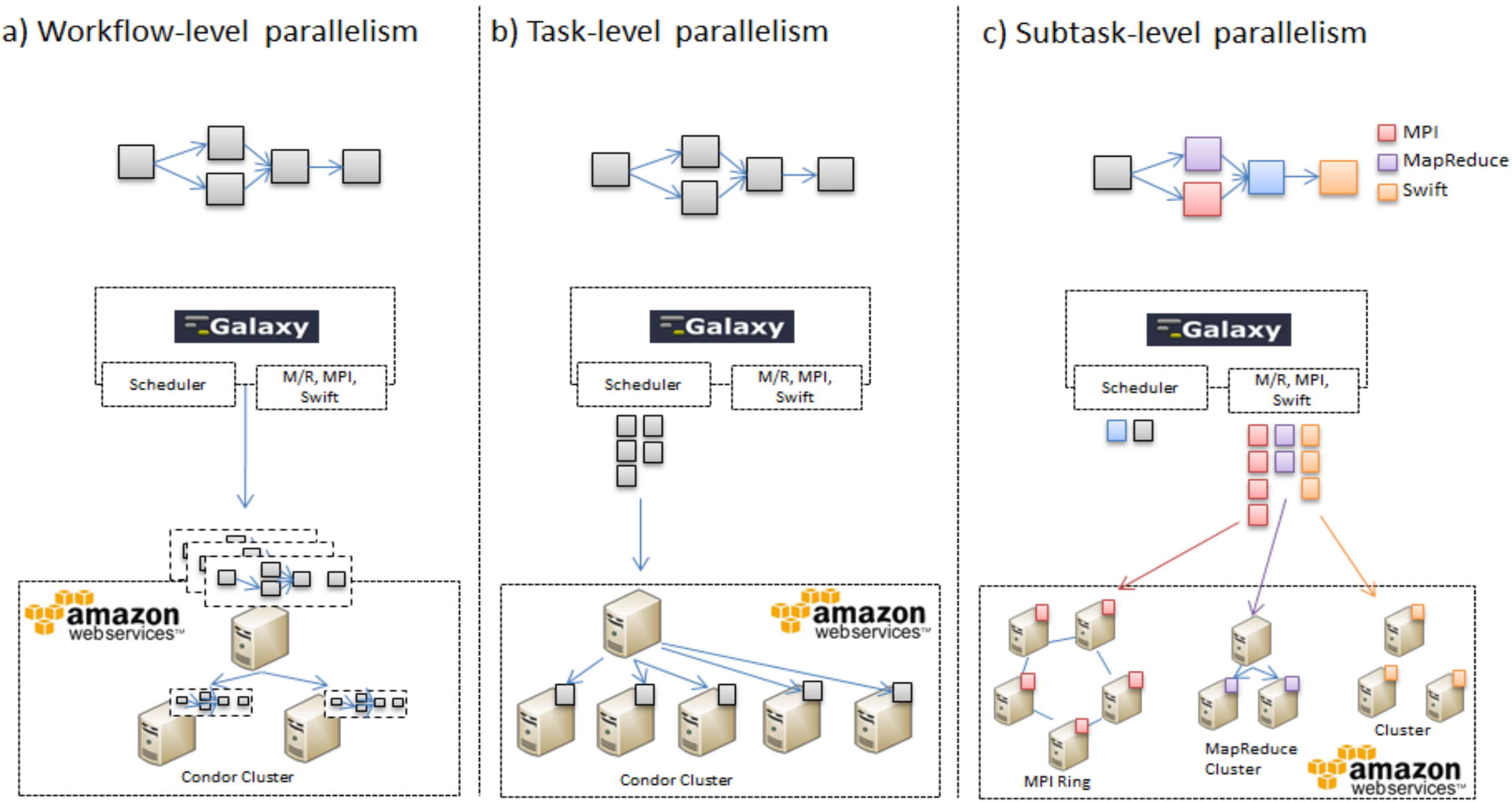
Transfer Files | Activity | Manage Endpoints | Dashboard

Activity

Sort By: start date & time | filter this list

- vasya#genemed7 to galaxy#cox0116
  transfer completed 3 months ago
- autism_atlas
  transfer cancelled 6 months ago

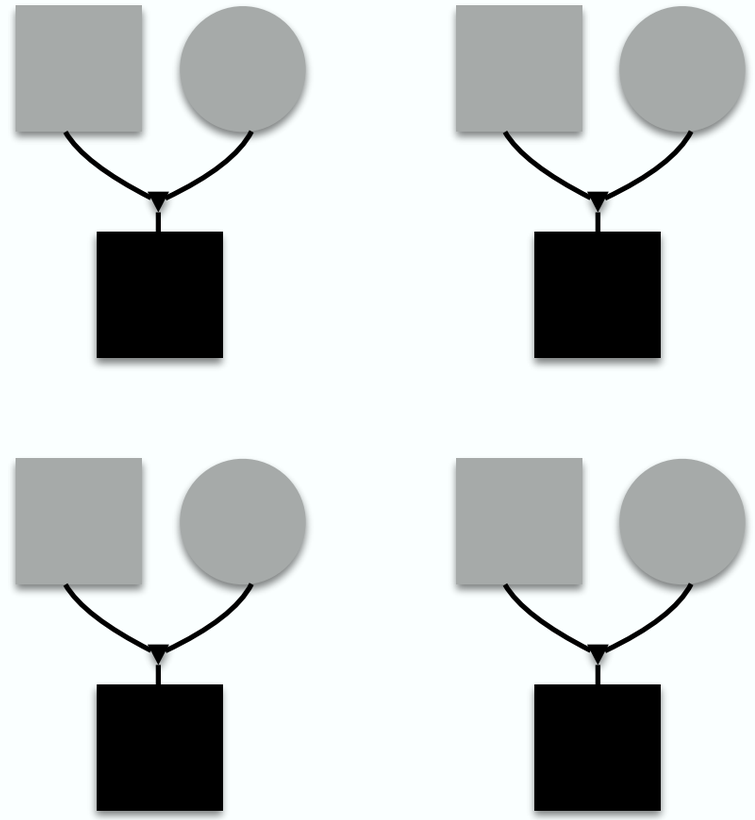
Efficiency -> Amazon Elastic Compute

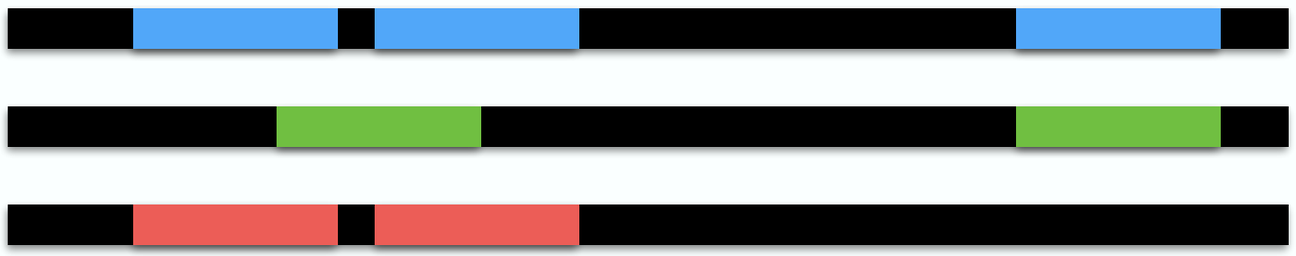


Consensus calling with Globus Genomics

Autism ACE

- Autism Centers of Excellence consortium
- 132 samples with 40 complete trios
- Illumina Whole Exome Capture product
- 1.8 TB of raw data

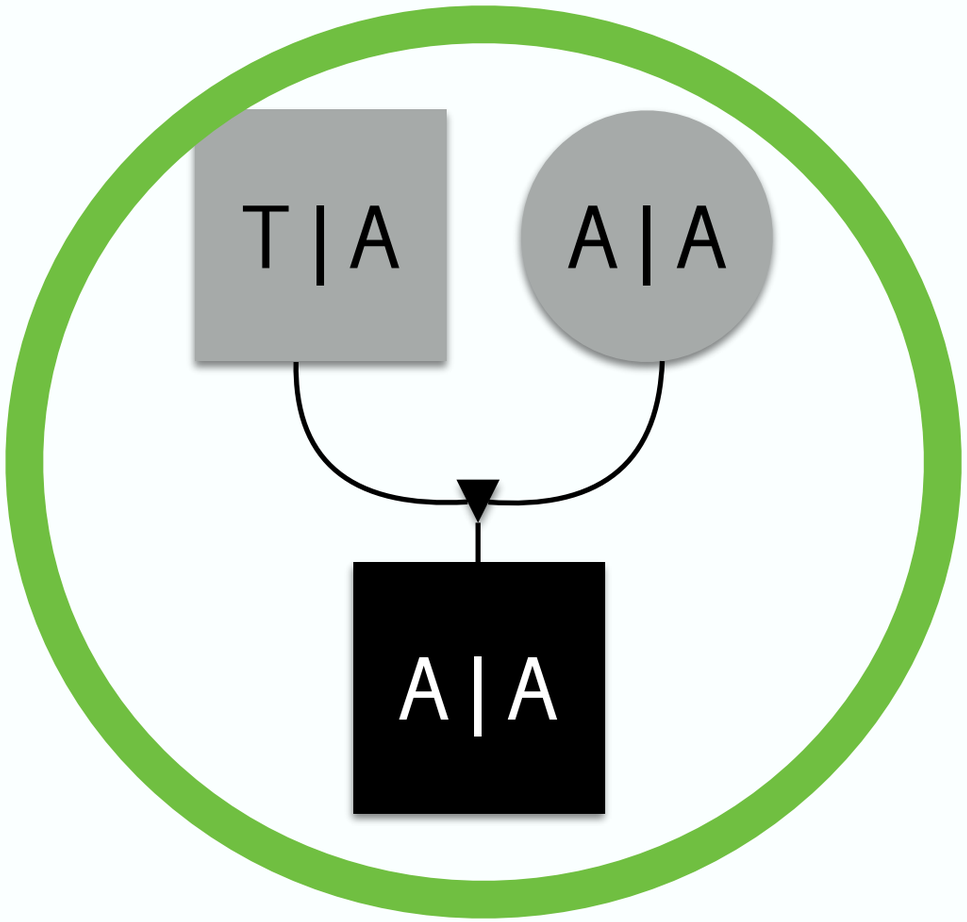


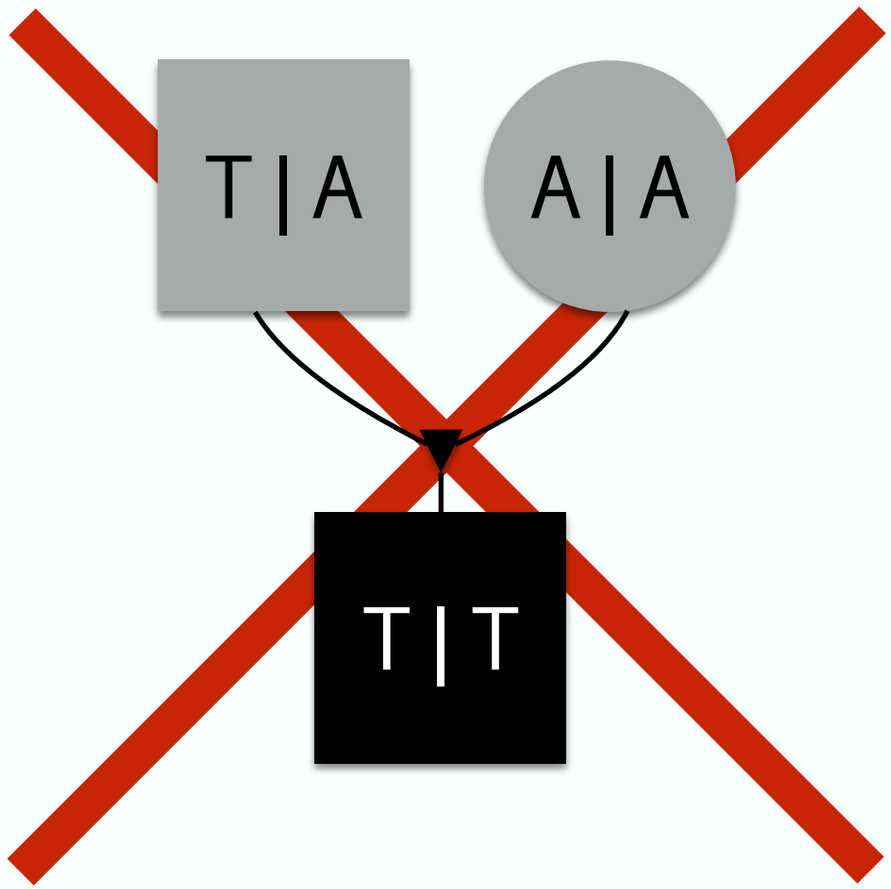


Estimating Genotype Error

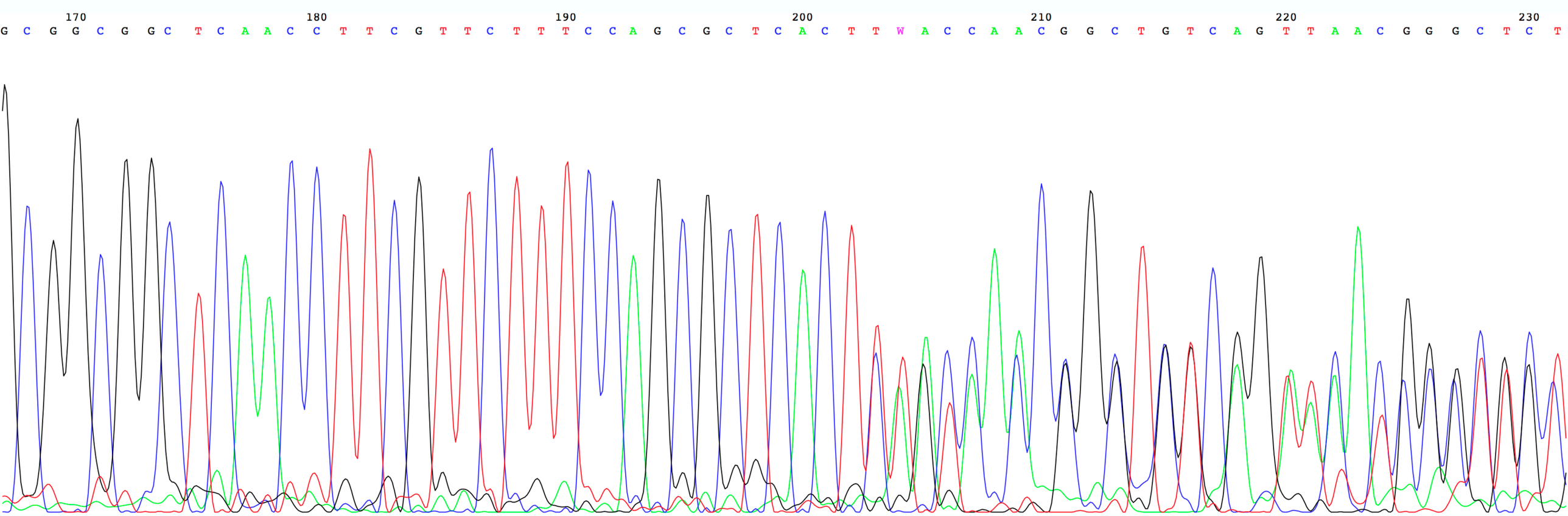
- Mendelian transmission
- Sanger validated genotypes
- Variant rediscovery

Estimating genotyping error: mendelian transmission





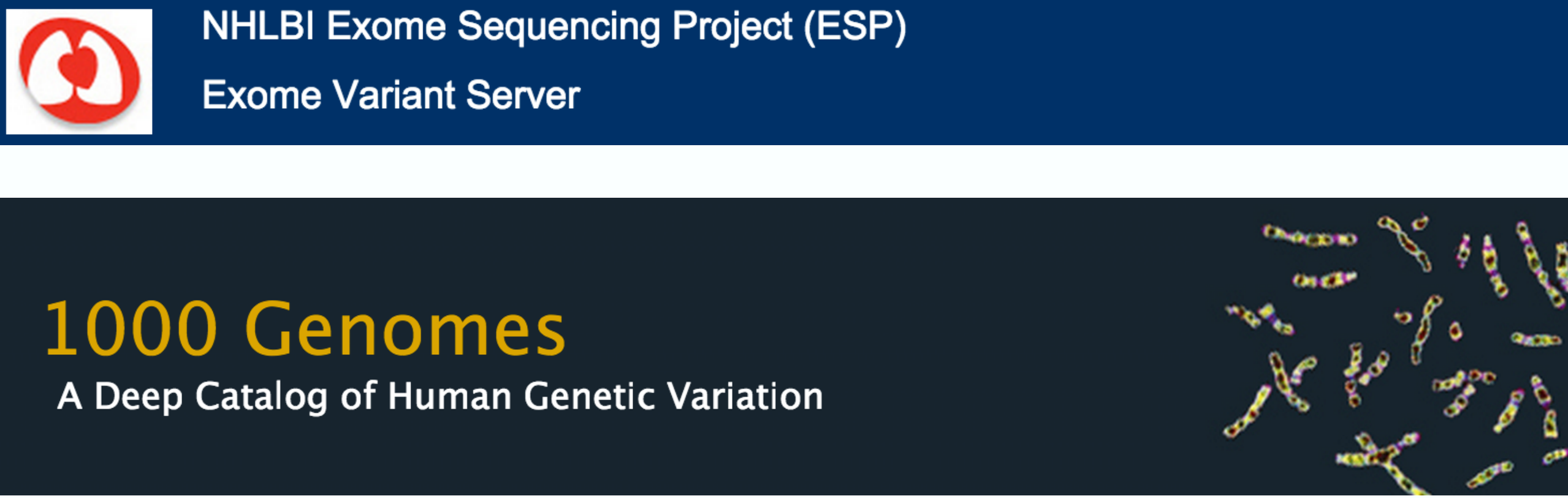
Estimating genotyping error: Sanger validated variants



credit: http://yost.genetics.utah.edu/

Estimating genotyping error: Prior information

Rediscovering variants from the 1000 Genomes and Exome Variant Server projects



Results

Call Set	# of Sites	**Mendel Rate**	EVS discover	1000G discover	**PPV**
AtlasSNP2	214,149	**0.245%**	72.7%	69.0%	**92.3%**
Freebayes	140,803	**0.449%**	78.9%	74.3%	**81.1%**
GATK	265,625	**1.03%**	61.3%	58.3%	**84.4%**
Consensus	129,706	**0.0459%**	82.9%	78.1%	**93.5%**

Runtimes

Model	Runtime (days)	CPU time (days)	Nodes used
GATK UG	0.875	10.1	23
Freebayes	1.31	30.1	23
ATLAS	4.6	x	135

Concurrent work

nature biotechnology

Home | Current issue | Research ▾ | Archive ▾ | For authors ▾ | About the journal ▾

NATURE BIOTECHNOLOGY | COMPUTATIONAL BIOLOGY | ANALYSIS

日本語要約

Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls

Justin M Zook, **Brad Chapman**, **Jason Wang**, **David Mittelman**, **Oliver Hofmann**, **Winston Hide** & **Marc Salit**

Affiliations | **Contributions** | **Corresponding author**

Nature Biotechnology **32**, 246–251 (2014) | doi:10.1038/nbt.2835

Received 14 December 2013 | Accepted 27 January 2014 | Published online 16 February 2014

nature REVIEWS GENETICS

Access provided t

nature.com ▸ journal home ▸ archive ▸ issue ▸ perspectives ▸ opinion ▸ full text

NATURE REVIEWS GENETICS | PERSPECTIVES | OPINION

ARTICLE SERIES: **Applications of next-generation sequencing**

The role of replicates for error mitigation in next-generation sequencing

Kimberly Robasky, **Nathan E. Lewis** & **George M. Church**

Affiliations | **Corresponding author**

Nature Reviews Genetics **15**, 56–62 (2014) | doi:10.1038/nrg3655

Published online 10 December 2013

Research

Highly accessed | Open Access

Genome Medicine
Volume 5
Issue 3

Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing

Jason O'Rawe[1,2], **Tao Jiang**[3], **Guangqing Sun**[3], **Yiyang Wu**[1,2], **Wei Wang**[4], **Jingchu Hu**[3], **Paul Bodily**[5], **Lifeng Tian**[6], **Hakon Hakonarson**[6], **W Evan Johnson**[7], **Zhi Wei**[4], **Kai Wang**[8,9]* and **Gholson J Lyon**[1,2,9]*

Viewing options
Abstract
Full text

Acknowledgements

ACE Consortium

- Ed Cook
- Jim Sutcliffe

Globus Group

- Ravi Madduri
- Ian Foster
- Alex Rodriguez
- Paul Dave

Cox Lab

- Nancy Cox
- Lea Davis

Funding

- Amazon EC2 Academic grants

PUBLIC vtrubets / galaxy.consensus

Watch 1 | Unstar 4 | Fork 0

consensus calling tool for cox Galaxy instance. — Edit

193 commits | **3** branches | **0** releases | **1** contributor

branch: **master** galaxy.consensus /

Update README.md

vtrubets authored 3 months ago — latest commit 115bf149b3

consensus_tool	Merge branch 'master' of https://github.com/vtrubets/galaxy.consensus	3 months ago
data	added some small test data	4 months ago
.gitignore	Update .gitignore	8 months ago
README.md	Update README.md	3 months ago
setup.py	Initial commit for the setup file. Starting to package this thing up …	10 months ago

README.md

Description:

This is an implementation of an ensemble variant calling method. Specifically, it takes VCF files generated by various calling algorithms and merges them according to specified thresholds on variant and genotype concordance. The resulting VCF can range from a strict consensus among inputs, to a union of all possible observations.

SSH clone URL

git@github.com:vtr

You can clone with HTTPS, SSH, or Subversion.

Clone in Desktop

Download ZIP