



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

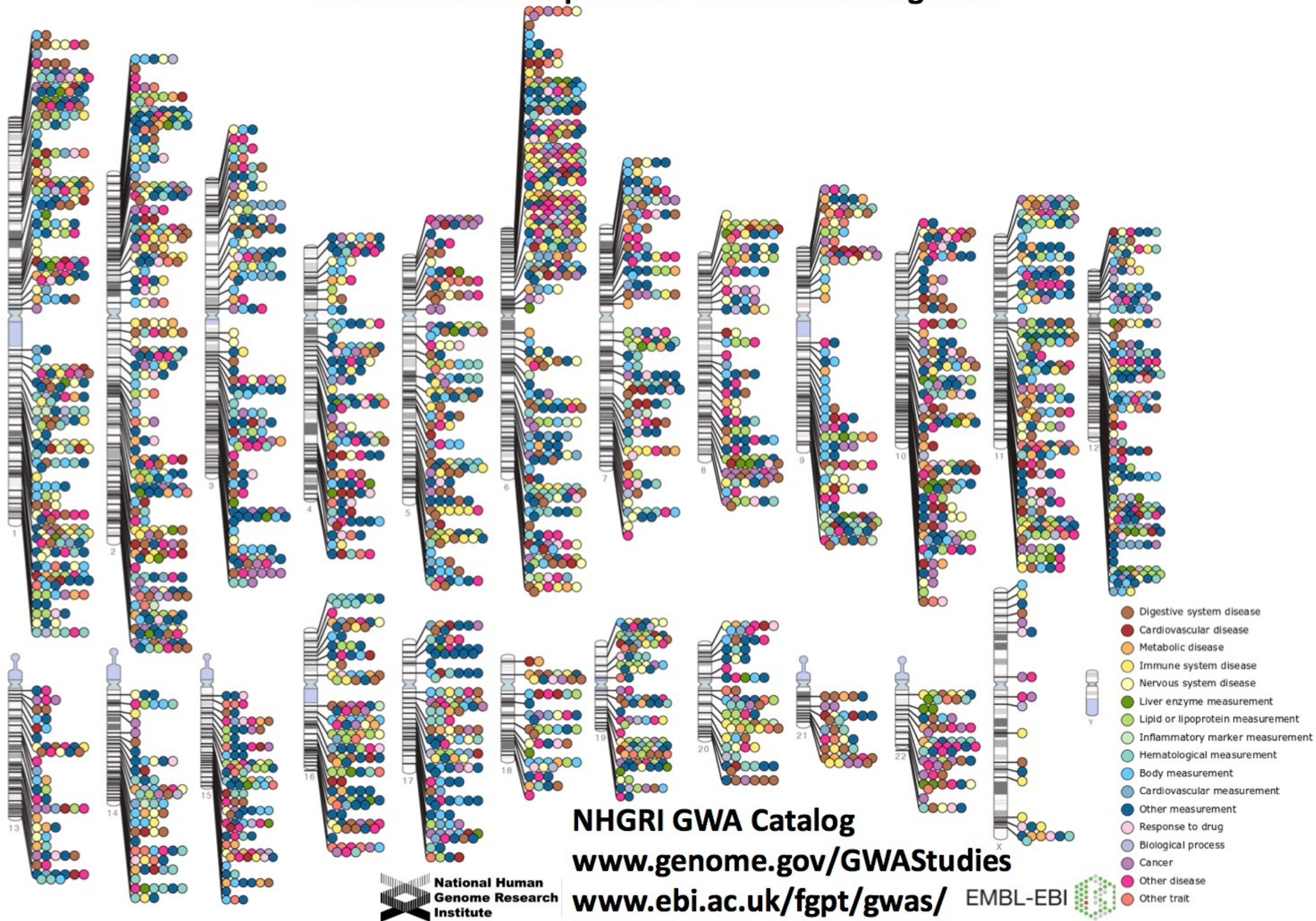
Globus genomics experiments:

Implementation, testing in ACE autism dataset

Context

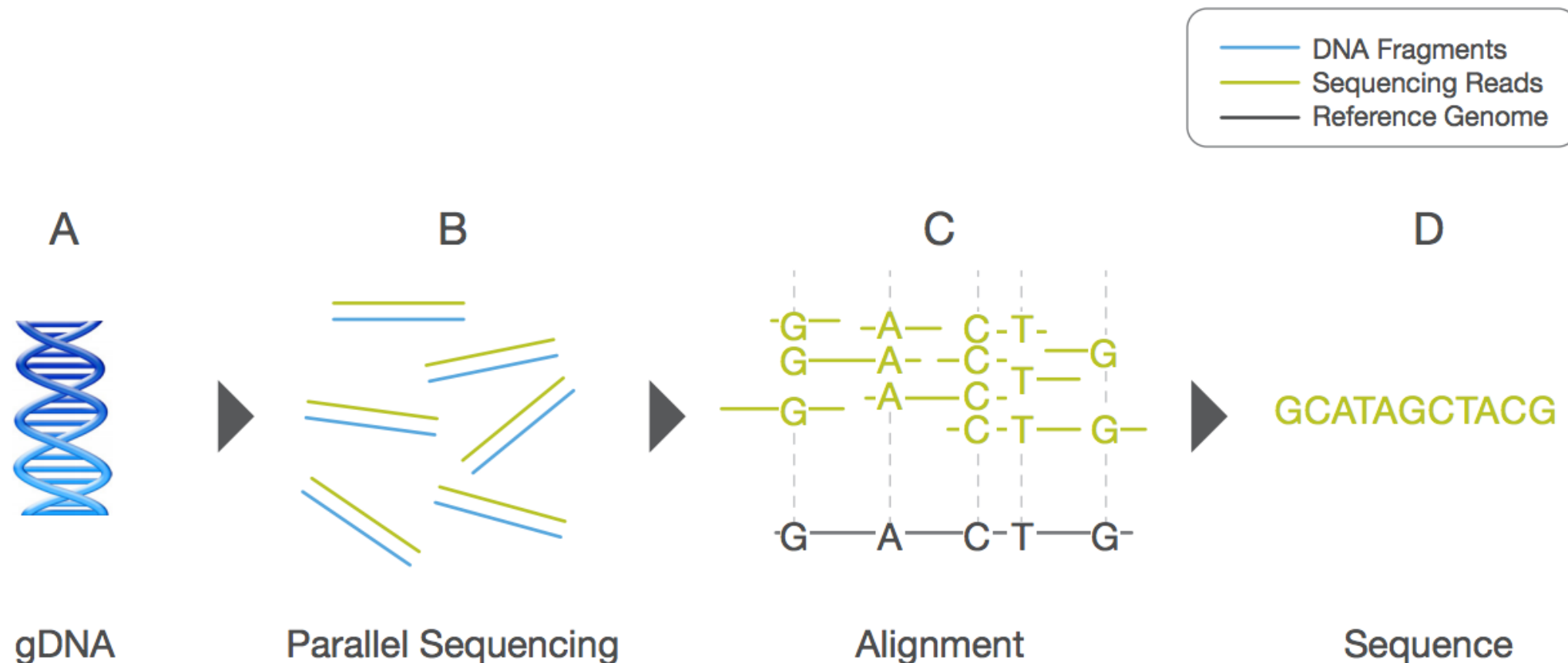
Published Genome-Wide Associations through 12/2012

Published GWA at $p \leq 5 \times 10^{-8}$ for 17 trait categories



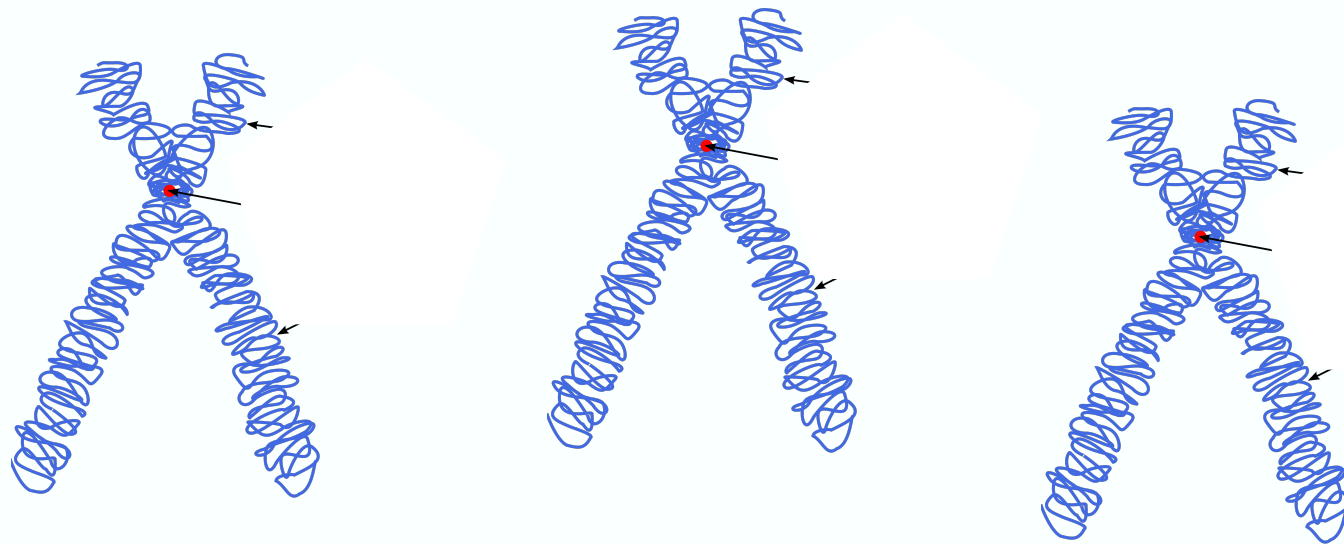
What is next generation sequencing?

Figure 1: Conceptual Overview of Whole-Genome Resequencing

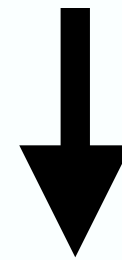
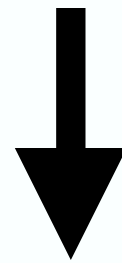
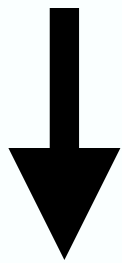


- Extracted gDNA.
- gDNA is fragmented into a library of small segments that are each sequenced in parallel.
- Individual sequence reads are reassembled by aligning to a reference genome.
- The whole-genome sequence is derived from the consensus of aligned reads.

Next gen sequencing



extract DNA



TATATCGGGCTTAGGCTAAATT
GCTTGCCTTCGGAATATATATCGGGC

ATCGGGCTTAGGCTAAA
TGCCTTCGGAATATATATCGGGCTTAG

TTAGGCTAAATTCCGCTTGCCTTCGGA
ATATATCGGGCTTAGGCTAAATTCCGCT

fragment and
*sequence

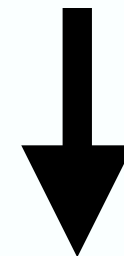
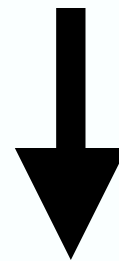
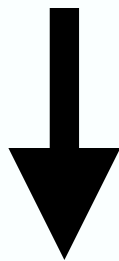
Next gen sequencing

TATATCGGGCTTAGGCTAAATT
GCTTGCCTTCGGAATATATATCGGGC

ATCGGGCTTAGGCTAAA
TGCCTTCGGAATATATATCGGGCTTAG

TTAGGCTAAATTCCGCTTGCCTTCGGA
ATATATCGGGCTTAGGCTAAATTCCGCT

fragment and
*sequence



ATATATAAG
GGGCTATATAT
ATTCGGGCTATATAT

CGGATTCGGGCTATATATAAGGC
CGCCTTAAATCGGATTCGGGCTAT
GTTCGCCTTAAATCGG CGGGCTATATATAAGGCTTCCGT

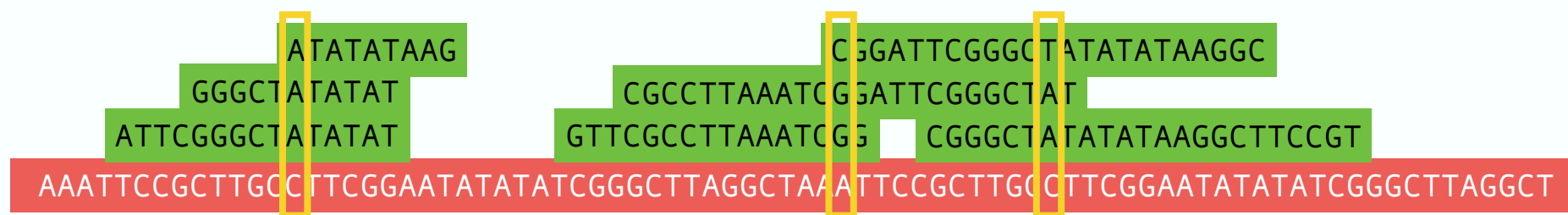
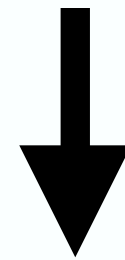
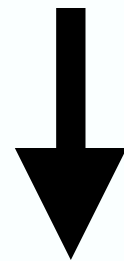
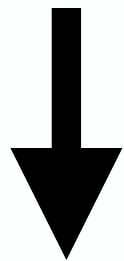
AAATTCGCTTGCCTTCGGAATATATATCGGGCTTAGGCTAAATTCCGCTTGCCTTCGGAATATATATCGGGCTTAGGCT

align to reference

Next gen sequencing



align to reference



identify variants

SNP

SNP SNP

40421561 40421571 40421581 40421591 40421601 40421611 40421621 40421631 40421641 40421651 40421661 40421671 40421681 40421691 40421701
ataagatggttatgaagattcacacagcggctcatgcctgtgatcccagcactttgggaggctgaggcaagtggagcacctgagatcatgagttcaagaccagcctggccaacatgggtgaaaccccatctctactaaagatacaaaa
.....T.....
ataagatggtt tgaagattcacacagtggtcctcatgcctgtgatcccagcac tgggaggctgagtcagtggagcacctgagatcatgagtt ACCAGCCTGGCCAACATGGTGAACCCCATCTCTACTAAA ATACAAA
ataagatggtt aagatacacacagtggtcctcatgcctgtgatcccagcactt GGGAGGCTGAGGCAAGTGGAGCACCTGAGATCATGAGTTC cagcctggccaacatgggtgaaaccccatctctactaaaga ACAA
ATAAGATGGTTATGAAGATTCACACAGTGGCTC CCTGTGATCCCAGCACTTTGGGAGGCTGAGGCAAGTGGAG ACCTGAGATCATGAGTTC AAGACCAGCCTGGACAACATGG AACCCCATCTCTACTAAAGATACAAAA
ATAAGATGGTTATGAAGATTCACACAGTGGCTCATGCC tgatcccagcactttgggagg TGAGGCAAGTGGAGCACCTGAGATCATGAGTTC AAGACCA GCCAACATGGTGAACCCCATCTCTACTAAAGATACAAAA
TCAGATGGTTATGAAGATTCACACAGTGGCTCATGCCGT ATCCCAGCACTTTGGGAGGCTGAGGCAAGGGGAGCACCTG ATGAGTTC AAGACCAGCCTGGCCAACATGGTGAACCCCA CTCTACTAAAGATACAAAA
a aagatggttatgaagattcacacagtggtcctcatgcctgtg TCCCAGCACTTTGGGAGCCTGAGGCAAGTGGAGCACCTGA ATGAGTTC AAGACCAGCCTGGCCAACATGGTGAACCCCA TATACTAAAGATNCAAAA
ata gatggttatgaagattcacacagtagctcatgcctgtgat AGCACTTTGGGAGGCTGAGGCAAGGGGAGCACGTGA GAGTTC AAGACCAGCCTGGCCAACATGGTGAACCCCATC CTACTAAAGATACAAAA
ataagatggttatgaagattcacacagtggtc CTGTAATCCCATCACTTTGGGAGGCTGAGGCAAGTGGAGC CCTGAGATCATGAGTTC AAGA AGCCTGGCCAACATCGTGAACCCCATATCTACTAAAGAT caaaa
ATAA TGGTTATGAAGATTCACACAGTGGCTCATGCCGTGTGATCC cactttgggatgctgaggcaagtggagcacctgagatcat CAAGACCAGCCTGGCCAACATGGTGAACCCCATCTCTAC AGAAATACAAA
ATAAGATGGTTATGAAGATTCACACAGTGGCTCA TGTGATCCCAGCACTTTGGGAGGCTGAGGCAAGTGGAGCA CTGAGATCACGAGTTC AAGACCAGCCTGCCAACATGGTC AACCCCATCTCTACTAAAGATACAAAA
ATAAGA GGTACGAAGATTCACACAGTGGCTCATGCCGTGTGATCCC cacattgggaggctgaggcaagtggagcacctgagatcat AAGACCAGCCTGGCCAACATGGTGAACCCCATCTCTACT AAGATACAAAA
ataagat ttatgaagattcacacagtggtcctcatgcctgtgatcccag CTTTGGGAGGCTGAGGCAAGTGGAGCACCTGAGATCATGA agcctggccaacatgggtgaaaccccatctctactaaagat AAA
ataagatggtt aagattcacacagtggtcctcatgcctgtgatcccagcactt GGGAGGCTGAGGCAAGTGGAGCACCTGAGATAATGAGTTC GCCTGGCCAACATGGTGAAC CCCATCTCTACTAAAGATACAAAA
ATAAGATGGTTA agattcacacagaggctcatgcctgtgatcccagcacttt AGGCTGAGGCAAGTGGAGCACCTGAGATCATGAGTTC AAG CCTGGCCAACATGGTGAACCCCATCTCTACTAAAGATAC
ataagatggtta TTCACACAGTGGCTCATGCCGTGTGATCCCAGCACCTTGGG GCTGAGGCAAGTGGAGCACCTGAGATCATGAGTTC AAGAC CCAACATGGTGAACCCCATCTCTACTAAAGATACAAAA
ATAAGATGGTTAT CAGTGGCTCATGCCGTGTGAT ACTTTGGGAGGCTGAGGCAAGTGGAGCACCTGAGATCATG CAACATGGTGAACCCCATCTCTACTAAAGATACAAAA
ATAAGATGGTTATGAAG CAGTGGCTCATGCCGTGTGATC ACTTCGGGAGGCTGAGGCAAGTGGAGCACCTGAGATCATG AACATGGTGAACCCCATCTCTACTAAAGATAACAGAA
ATAAGATGGTTATGAAGAT CAGTGGCTCATGCCGTGTGATCC CCTCTGGGAGGCTGAGGCAAGTGGAGCACCTGAGATCATG ACATGGTGAACCCCATCTCTACTAAAGATACAAAA
CTAAGATGGTTATGAAGATT GCGGCTCATGCCGTGTTATC CTTTGGGAGGCTGAGGCAAGTGGAGCACCTGAGATCATGA ACATGGTGAACCCCATCTATACTAAAGATACAAAA
ATAAGATGGTTATGAAGATTC CTCTTGCCTGTGATCCCAGCACTTTGGGAGGCTGACGCAA TGGAGCACCTGAGATCATGAGTTC AAGACCAGCCTGGCCA TGGTGAACCCCATCTCTACTAAAGATACAAAA
ATAAGATGGTTATGAAGATTC CTGATGCCGTGTGATCCCAGCACTTTGGGAGGCTGAGGCAA TGGAGCACCTGAGATCATGAGTTC AAGACCAGCCTGGCCA GGTGAACCCCATCTCTACTAAAGATACAAAA
ATAAGATGGTTATGAAGATTC GTGATCCCAGCACTTTGGGAGGCTGAGGCAAGTGGAGCAC GATCATGAGTTC AAGACCAGCCTGGCCAACATGGTGAAC ccatctctactaaagatacaaaa
AGATGGTTATGAAGATTCACACAGTGGCTCATGCCGTGTGA CCAGCACTTTGGGAGGCTGAGGCAAGTGGAGTACCTGAGA GAGTTC AAGACCAGCCTGGCCAACATGGTGAACCCCATC TACTAAAGATACAAAA
ACATGGTTATGAAGATTCACACAGTGGCTCATGCCGTGTGA CTTTGGGAGGCTGAGGCAAGTGGAGCACCTGAGATCATGA CATGGTGAACCCCATCTCTACTAAAGATACAAAA
GGTTATGAAGATTCACACAGTGGCTCATGCCGTGTGATCCC CTCTGGGAGGCTGAGGCAAGTG agcacctgagatcatgagttcaagaccagcctg*+caacat tgaaccccatctctactaaagatacaaaa
TATGAAGATTCACACAGTGGCTCA gatcccagcactttgggaggctgaggcaagtggagcacct agttcaagaccagcctggccaacatgggtgaaaccccatct TACTAAAGATACAAAA
ATGAAGATTCACACAGTGGCTCATGCCGTGTGATCCCAGCA TCTGGGAGGCTGAGGCAAGTGGAGCACCTGAGATCATGAG CATGGTGAACCCCATCTCTACTAAAGATACAAAA
gatcccagcactttgggaggctgaggcaagtggagcacct CATGGTGAACCCCATCTCT CTAAGATACAAAA
atcccagcactttgggaggctgaggcaagtggagcacctg CATGGTGAACCCCATCTCTACTAAAGATACAAAA
TCTGAGAGGCTGAGGCAAGTGGAGCACCTGAGATCATGAG GTGAACCCCATCTCTACTAAAGATACAAAA
GGGATGCTTAGTCAATTGTAGCACCTGAGATCATGAGTTC GTGAACCCCATCTCTACTAAAGATACAAAA
aggctgaggcaagtggagcacctgagatcatgagttcaag tgaaccccatctctactaaagatacaaaa
tggggcaagtggagcacctgagatcatgagttcaagacca gtgaaaccgtgtctctac aaagatactaaa
tgaggcaagtggagcacctgagatcatgagttcaagacca GAAATCCCATCTCTACTAAAGATACAAAA
GAGGCAAGTGGAGCACCTGAGATCATGAGTTC AAGACCAG GAAACCCCATCTCTACTAAAGATACAAAA
AGGCAAGTGGAGCACCTGAGATCATGAGTTC AAGACCAG GAAACCCCATCTCTACTAAATAAACA
aggcaatttgagctcctgagatcatgagttcaagaccagc gaaaccccatctctgctgagatgcaaaa
GCAAGTGGAGCACCTGAGATCA AACCCCATCTCTACTAAAGATACAAAA
CAAGTGGAGCACCTGAGATCATGAGTTC AAGACCAGCCTG AATCCCATCTCTACTAAATATACAAA
caagtggagcacctgagatcatgagttcaagaccagcctg aaccccatctctactaaagatcccaaa
AAGTGGAGCACCTGAGATCATGAGTTC AAGACCAGCCTGG AACCCCATCTCTACTAAAGATACAAAA
AGTGGAGCACCTGAGATCATGAGTTC AAGACCAGCCTGGC ACCCCGTTTCTACTAAAGATACAAAA
AGTGCAGCACCTGAGATCATGAGTTC AAGACCAGCCTGGC accccatctctactaaagatacaaaa
GTGGAGCACCTGAGATCATGAGTTC AAGACCAGCATGGCC CCCCATCTCTACTAAAGATAC
GGAGCACCTGAGATCATGAGTTC AAGACCAGCCTGGCCAAC CATCTCTAATAAAGATACAAAA
ggagcacctgagatgatgagttcaagaccaggggtggccaa CATCTCTACTAAAGATACAAAA
ggagcacctgagatcatgagttcaagaccagcctggccaa CGTCTCTACTAAAGATACAAAA
GAGCACCTGAGATCATGAGTTC AAGACCAGCCTGGCCAAC CATCTCTACTAAAGATACAAAA

Data scale

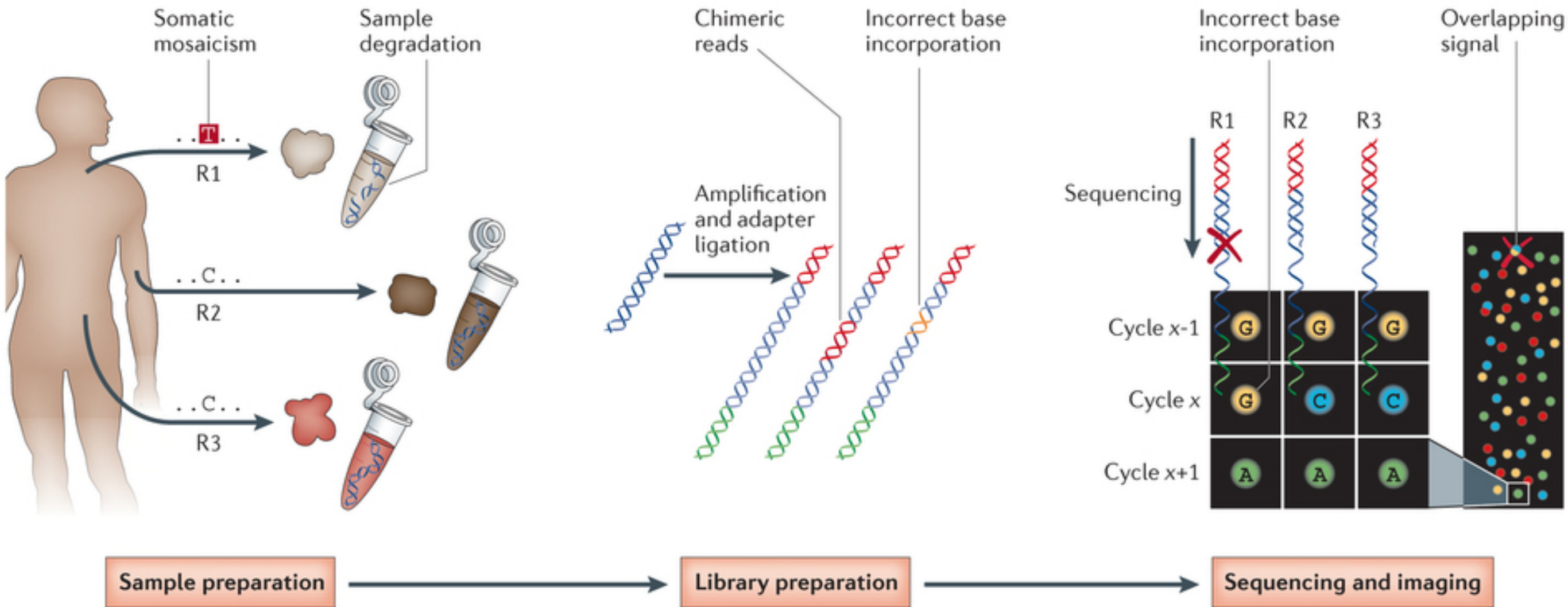
Type II Diabetes Genes Consortium

- 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



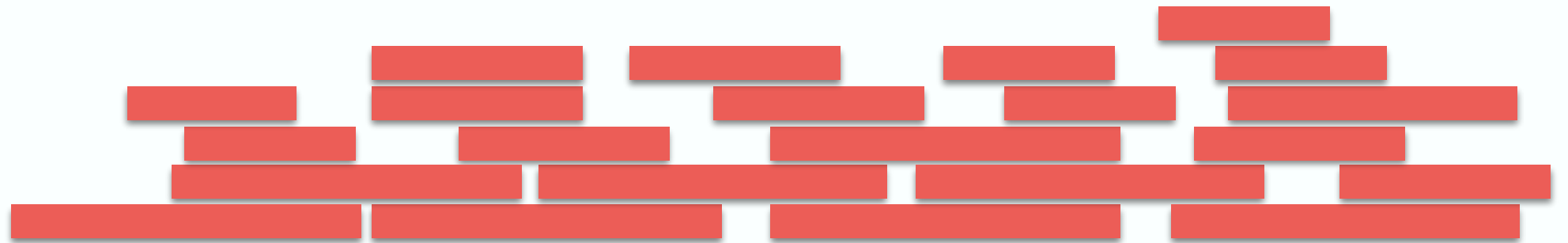
Consensus variant calling

Motivation

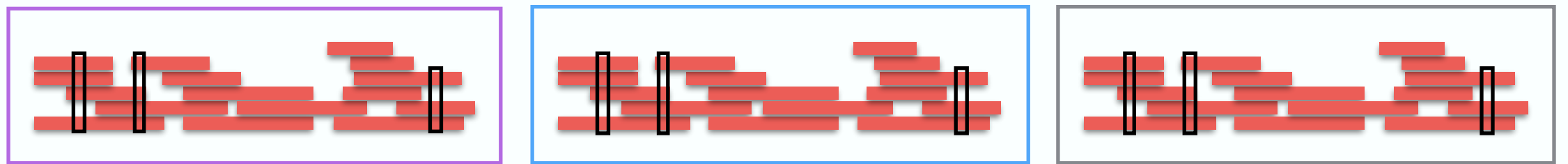
- Cumulative sequencing error
- Performance of different models of variation
- Existing ensemble methods

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

The screenshot displays a Galaxy workflow canvas titled "Workflow Canvas | genomewide_GATK". The workflow is composed of several interconnected tools:

- Input dataset** (output) feeds into three parallel paths.
- Each path starts with a **Unified Genotyper (take directory as input)** tool, which uses a reference file and operates on genomic intervals to produce VCF, metrics, and log files.
- The VCF outputs from these tools are fed into **Variant Annotator** tools, which also use a reference file and ROD files to produce annotated VCF and log files.
- The annotated VCF files are then combined in a **Combine Variants** tool, which merges the input variant files and produces a combined VCF and log file.
- The combined VCF is processed by a **Variant Recalibrator** tool, which recalibrates variants based on reference-ordered data and produces recalibrated VCF, tranches, and log files.
- Finally, the recalibrated VCF is used in an **Apply Variant Recalibration** tool, which applies the recalibration to the variant file and produces the final VCF and log files.

The right sidebar shows the configuration for the **Unified Genotyper (take directory as input)** tool:

- Tool:** Unified Genotyper (take directory as input)
- Version:** 0.0.6
- Choose the source for the reference list:** History
- BAM directory path:** /scratch/madduri/input_bam_files
- Using reference file:** Data input 'ref_file' (fasta)
- Binding for reference-ordered data:** Add new Binding for reference-ordered data
- Genotype likelihoods calculation model to employ:** BOTH
- The minimum phred-scaled confidence threshold at which variants not at 'trigger' tranches should be called:** 30.0
- The minimum phred-scaled confidence threshold at which variants not at 'trigger' tranches should be emitted (and filtered if less than the confidence threshold):** 30.0
- Basic or Advanced GATK options:** Advanced
- Pedigree files:** Add new Pedigree file
- Pedigree strings:** Add new Pedigree string
- How strict should we be in validating the pedigree information:** STRICT
- Read Filters:** Add new Read Filter
- Operate on Genomic intervals 1**
- Genomic intervals:** Data input 'input_intervals' (bed or gatk_interval or picard_interval_list or vcf)

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

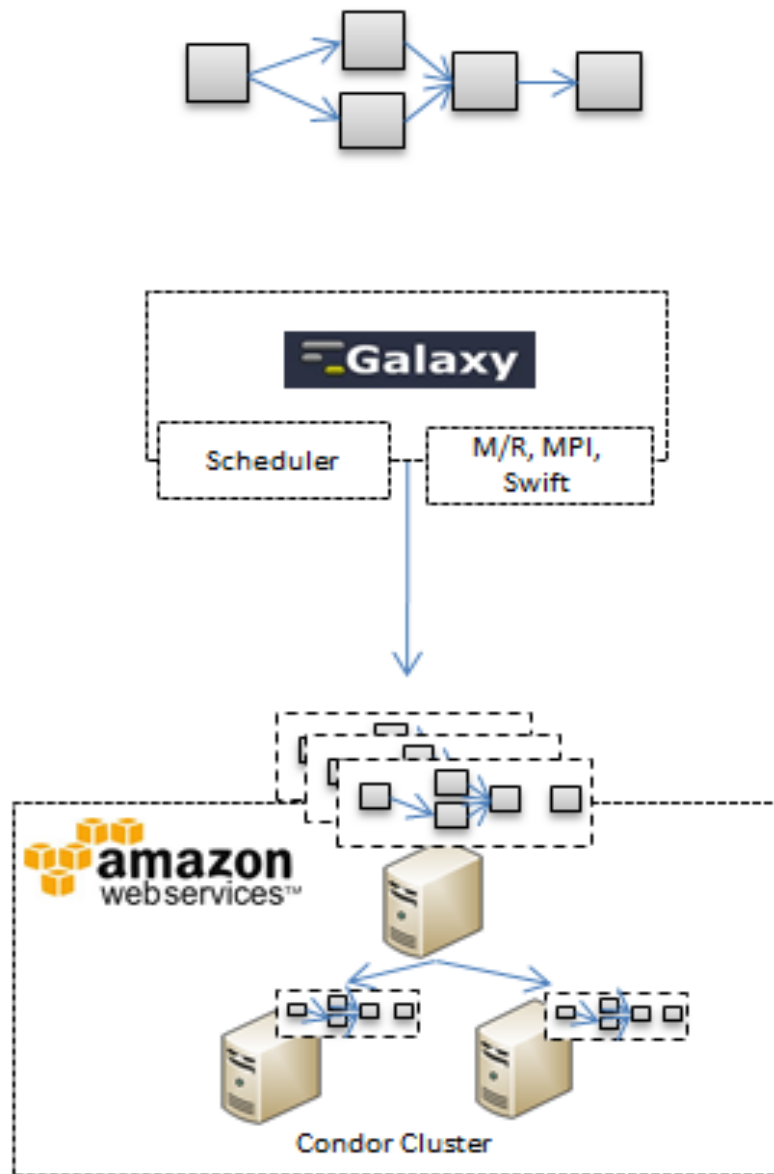
Activity

Sort By start date & time

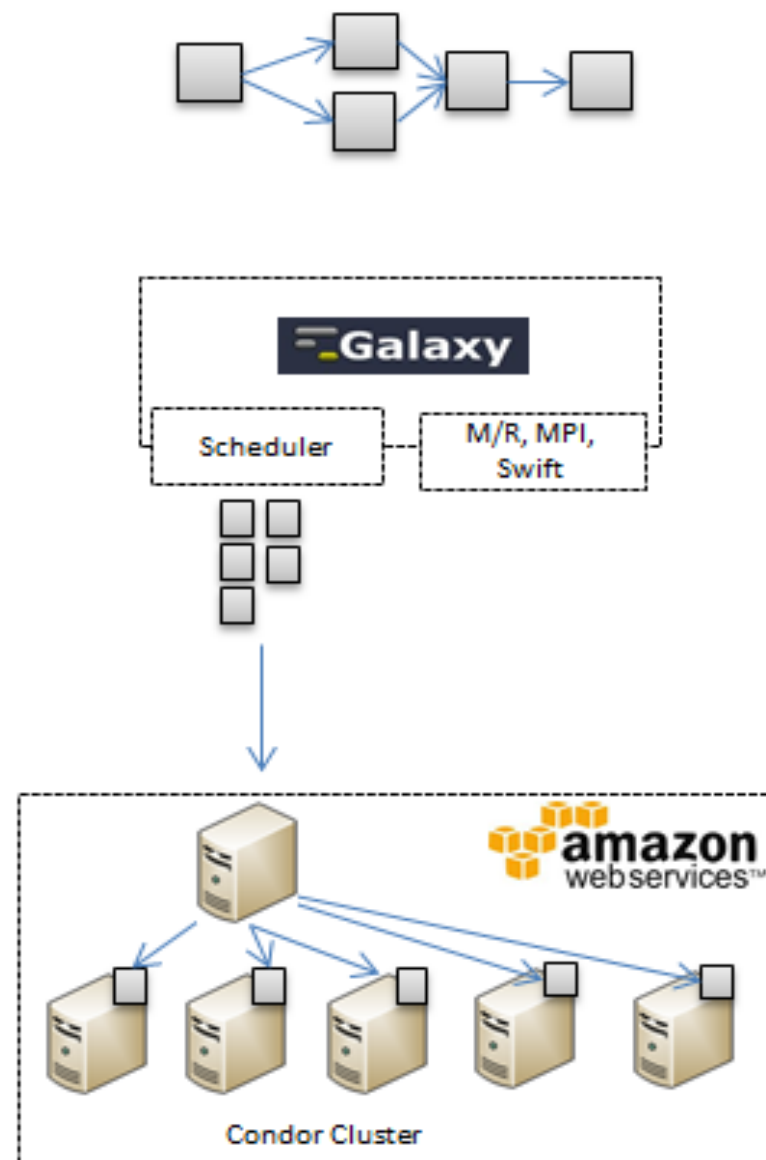
- vasya#genemed7 to galaxy#cox0116
transfer completed 3 months ago
- autism_atlas
transfer cancelled 6 months ago
- galaxy#cox0116 to coxlab#genegate

Efficiency -> Amazon Elastic Compute

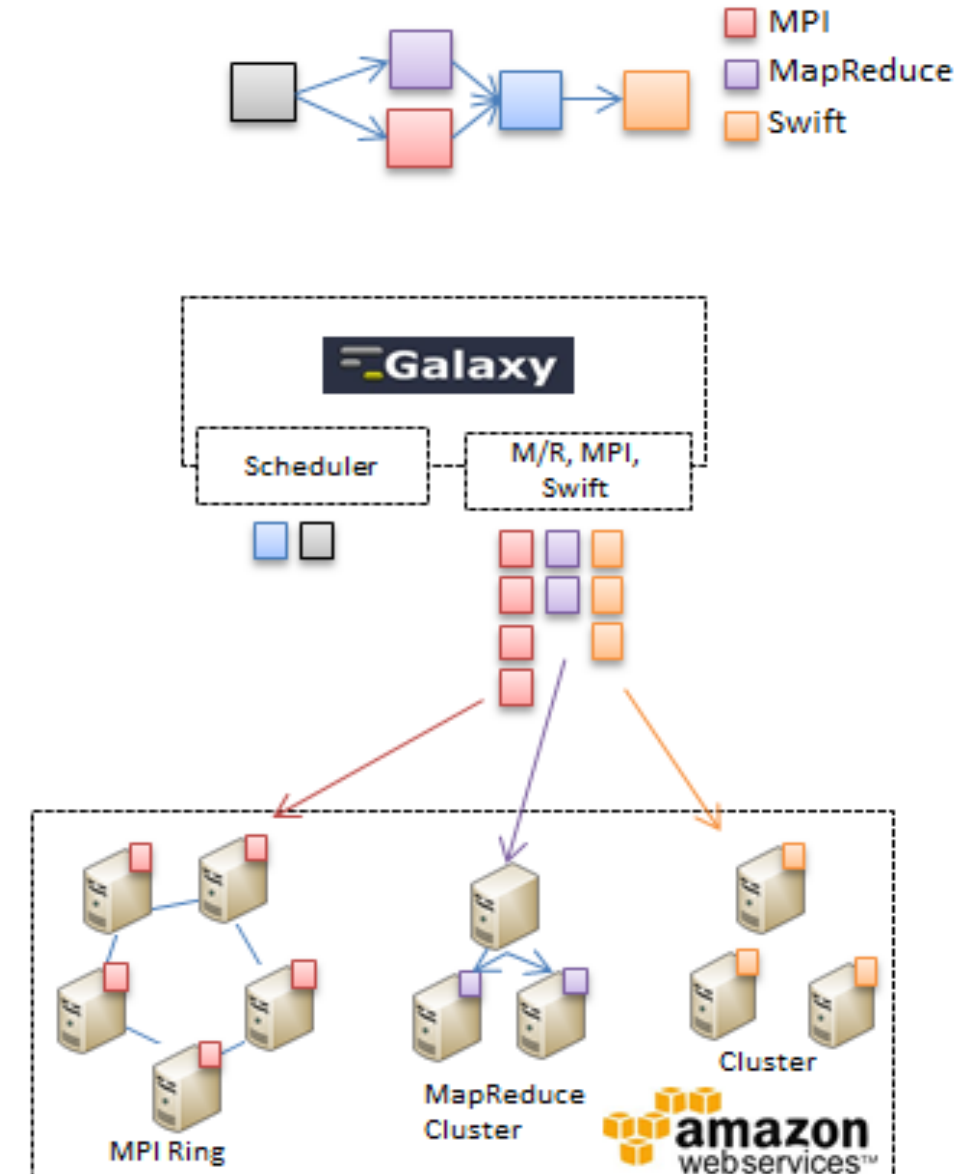
a) Workflow-level parallelism



b) Task-level parallelism



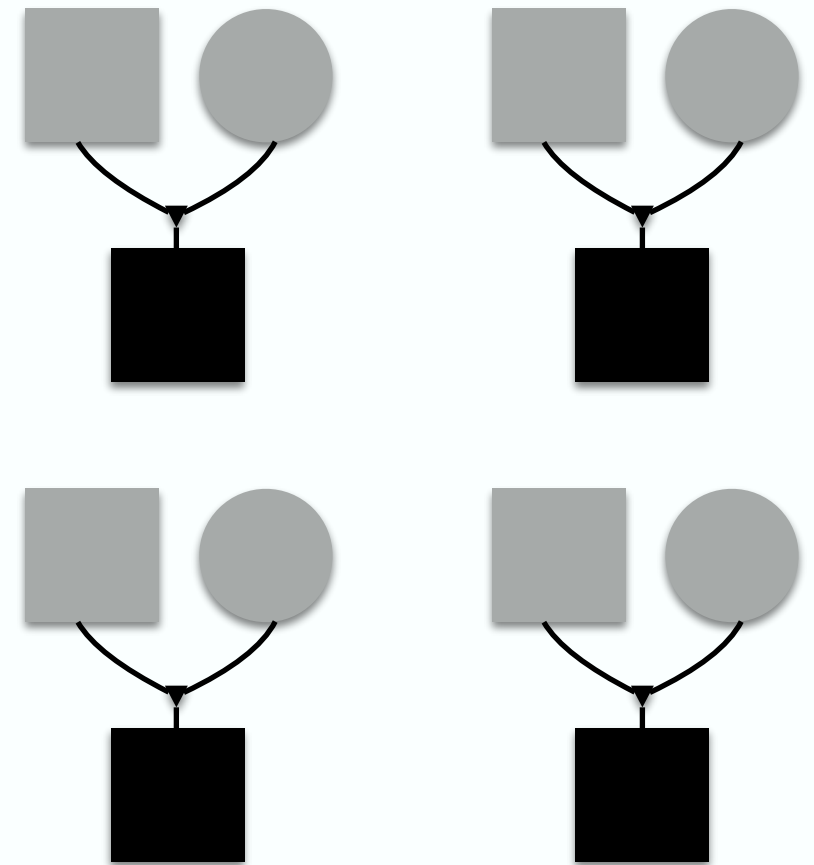
c) Subtask-level parallelism



Consensus calling with Globus Genomics

Autism ACE

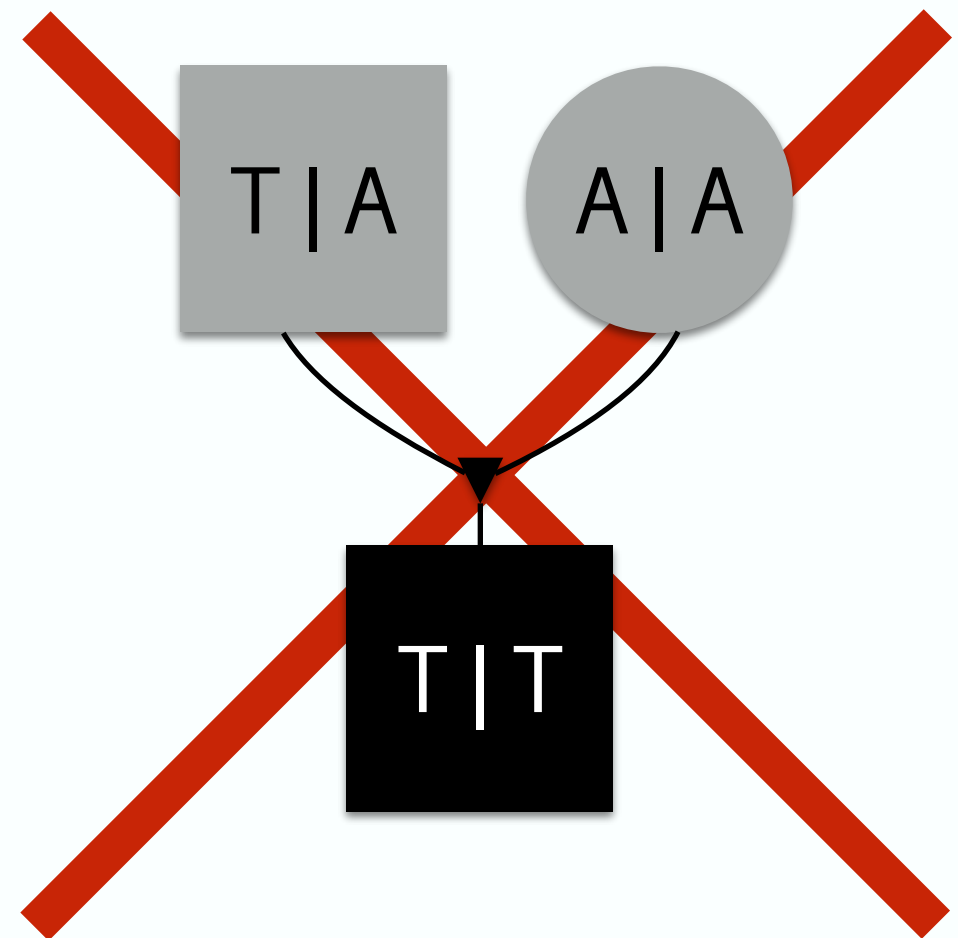
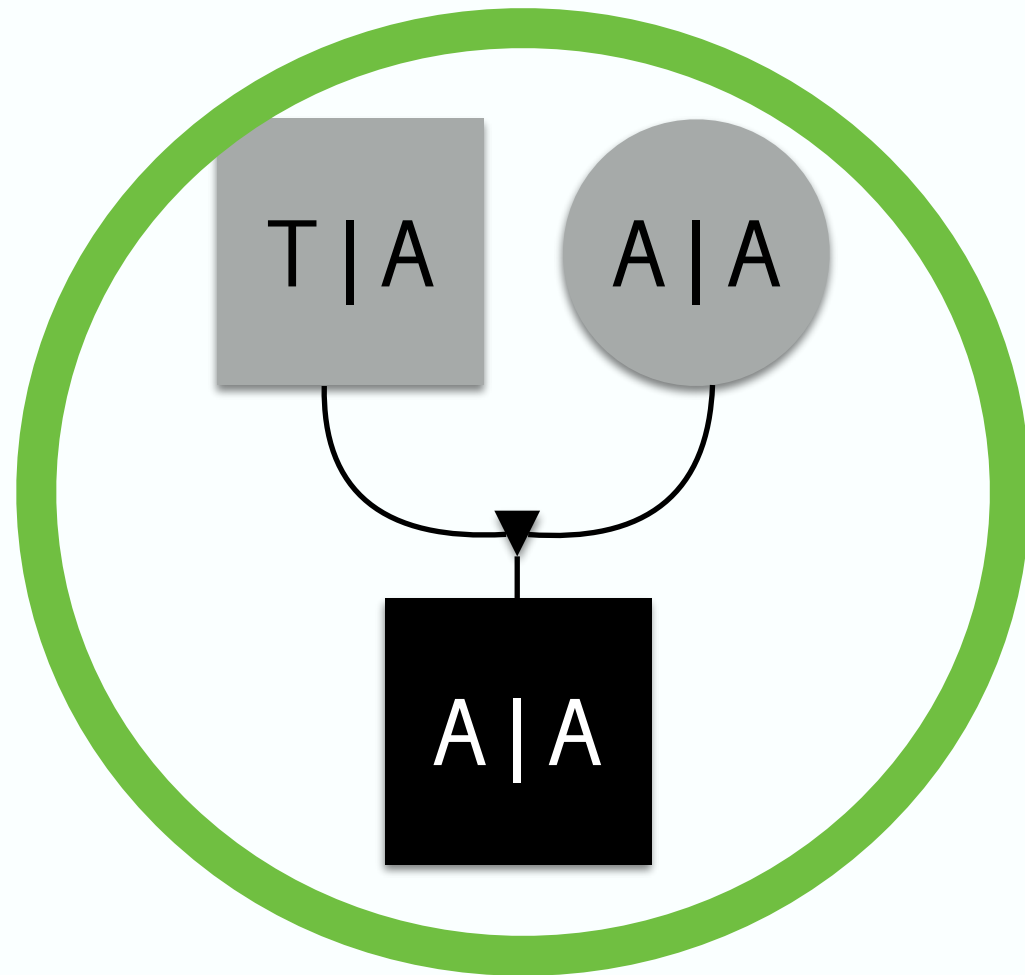
- Autism Centers of Excellence consortium
- 132 samples with 40 complete trios
- Illumina Whole Exome Capture product
- 1.8 TB of raw data



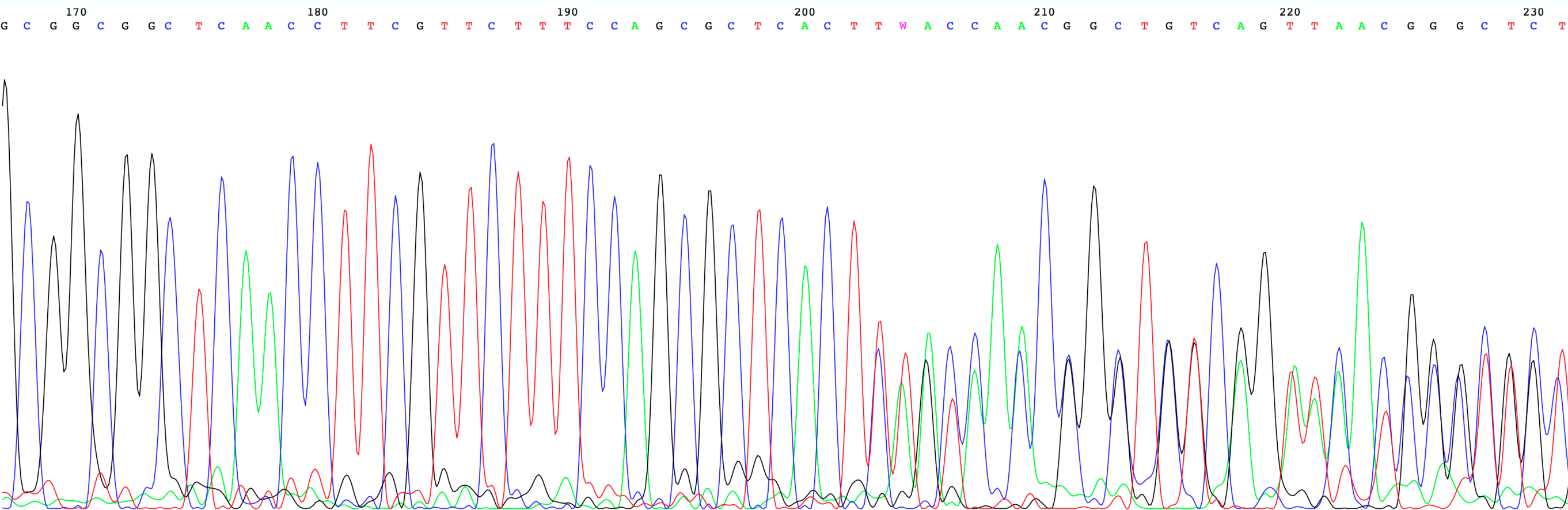
Estimating Genotype Error

- Mendelian transmission
- Sanger validated genotypes
- Variant rediscovery

Estimating genotyping error: mendelian transmission



Estimating genotyping error: Sanger validated variants



Estimating genotyping error: Prior information

Rediscovering variants from the 1000 Genomes and Exome Variant Server projects



NHLBI Exome Sequencing Project (ESP)
Exome Variant Server

1000 Genomes

A Deep Catalog of Human Genetic Variation



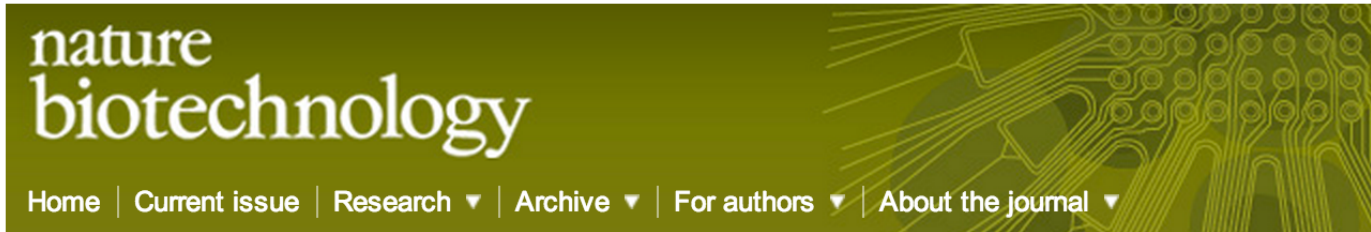
Results

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 | 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.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

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*}

Genome Medicine
Volume 5
Issue 3

Viewing options
[Abstract](#)
[Full text](#)

Acknowledgements

ACE Consortium

- Ed Cook
- Jim Sutcliffe

Cox Lab

- Nancy Cox
- Lea Davis

Globus Group

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

Funding

- Amazon EC2
Academic
grants

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.

<> Code

Issues 4

Pull Requests 0

Wiki

Pulse

Graphs

Network

Settings

SSH clone URL

git@github.com:vtr

You can clone with [HTTPS](#), [SSH](#), or [Subversion](#).

Clone in Desktop

Download ZIP