

Introducing Glow

William Brandler, Karen Feng
Data + AI Meetup
August 11, 2020

Outline

Current state of genomics

- The transition from academia to industry
- Genomics for drug discovery
- Genomics as a big data problem

How Glow solves problems with big genomics data

- From ad hoc command line tools to production data science
- Partnership between Databricks and Regeneron
- Innovations in Glow

Genomics is taking off in industry

 **Yaniv (((Erlich)))** @erlichya · Mar 8, 2019

Here is a partial list of academic PIs in genomics/bioinfo who left (or LoA) their position to join the industry in the last ~2 years:

- @cdbustamante
- @jgschraiber
- @KMS_Meltzy
- @gabecasis
- @markmccarthyoxf
- @joe_pickrell
- @srikosuri
- @adamauton
- @Piwdb
- @erlichya

23 94 270 

Replies

 **Tomaz** @tomaz_berisa · Mar 8, 2019

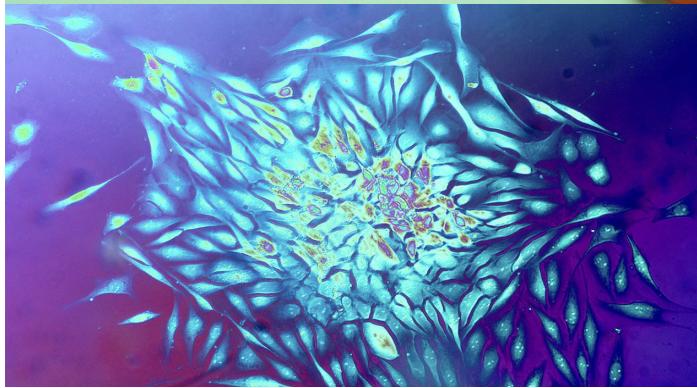
Replying to @erlichya @cdbustamante and 7 others
Also @marchini

0 10 

 **Adam Auton** @adamauton · Mar 8, 2019

Replying to @erlichya @cdbustamante and 6 others
No regrets.

3 17 



Ultimately, genomics will be fully integrated into clinical practice



GENOME-SEQUENCING ANNIVERSARY

A Celebration of the Genome, Part I

Ten years ago, the first peer-reviewed reports of the sequencing of the human genome were published. At that time we announced that “humanity has been given a great gift,” which has proven to be the case in many ways but has also presented a great challenge. To commemorate the event, we have asked a cross section of insightful individuals—representing many viewpoints—to look at what it has meant to them or their communities to have access to human genome sequences. We will be publishing their comments throughout the month of February.

—Barbara R. Jasny and Laura M. Zahn



Faces of the Genome

Francis S. Collins
Director, National Institutes of Health, Bethesda, MD, USA.

formed in July 2010, using stem cells from the cord blood of a matched, healthy donor. Seven months later, Nic appears to be on the road to recovery. While he is still on immunosuppressants, doctors report the new stem cells are stably engrafted, blood counts are good, and there's been no



<https://science.sciencemag.org/content/331/6017/546.2>

Genomics is revolutionizing drug development

90% of experimental medicines fail during clinical trials



>\$100bn spent on pharma R&D **per year**
Only 10-20 new drugs are approved each year



2X greater chance of drugs being approved if they have supporting genetic evidence



The power of big genomic data

MIT
Technology
Review

Topics+ The Download Magazine Events More+ [Subscribe](#)

Rewriting Life

UK Biobank Supercharges Medicine with Gene Data on 500,000 Brits



PharmaTimes
online

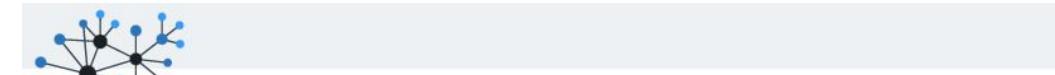
Search news, articles, experts, insights or people GO

Magazine Web Exclusives News Competitions Appointments Jobs Business Insights Webinars Thought Leadership

Free news subscriptions Free RSS feeds

UK's genome sequencing project will hit 100,000 by year-end

4th July 2018



genomeweb

Business & Policy Technology Research Diagnostics Disease Areas Applied Markets

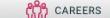
Home » Tools & Technology » Sequencing » Pilot GenomeAsia 100K Project Highlights Insights Gleaned From Diverse Dataset

Pilot GenomeAsia 100K Project Highlights Insights Gleaned From Diverse Dataset

REGENERON

> SCIENTISTS > PATIENTS > MEDICAL PROFESSIONALS > INVESTORS & MEDIA CAREERS

ABOUT SCIENCE MEDICINES RESPONSIBILITY



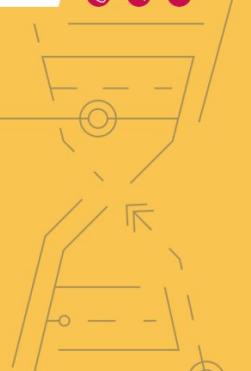
Regeneron Perspectives

A MILLION IS JUST THE BEGINNING



AUTHOR: JOHN OVERTON, PHD, VICE PRESIDENT, SEQUENCING AND LAB OPERATIONS

PUBLISHED ON: FEBRUARY 28, 2020



Pharma are adopting population-scale genomics

Population genetics is used in industry to

- build drug discovery pipelines
- prioritize R&D investments

Through understanding

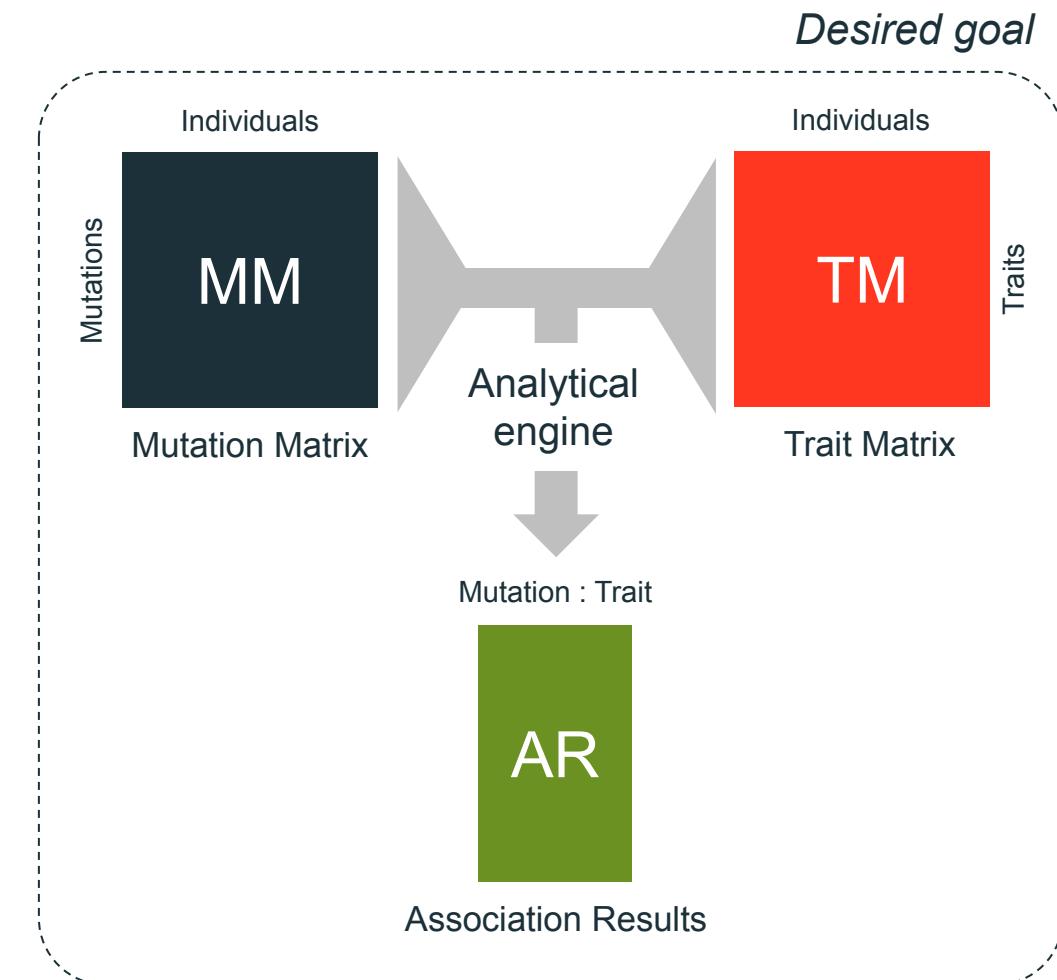
- the exact population frequency of specific genetic disorders
- the relationship between clinical phenotypes and genetics
- how patients respond in clinical trials based on their genetics
- what genes are amenable to drug discovery

Bigger pipeline - Increased likelihood of drug approval - Billions of dollars of savings

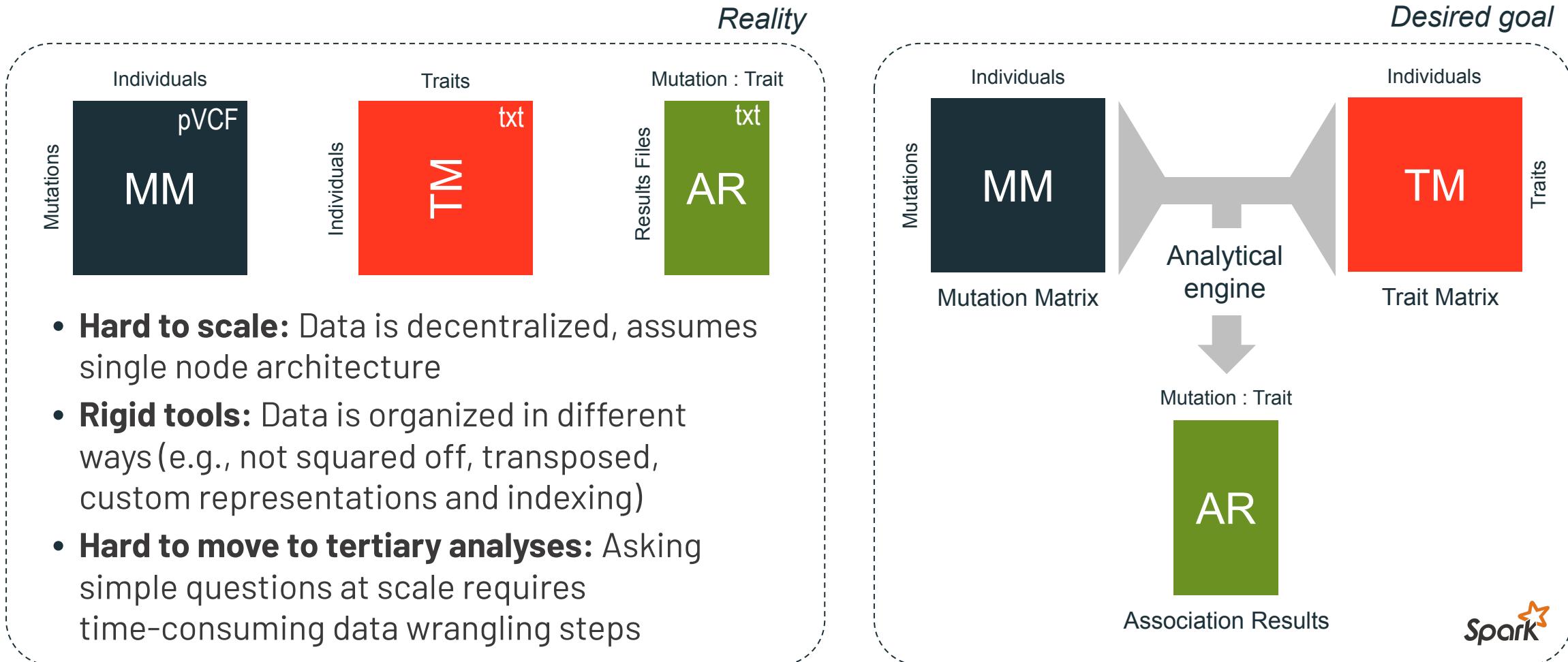
The Regeneron Genetics Center mines the world's genomics data to extract insights

Approach:

- Sequence a large number of individuals from many cohorts (>70 to date)
- Obtain paired phenotypic data (e.g. de-identified electronic medical records)
- Run all-vs-all association tests between all mutations and traits
- Mine association results to extract actionable insights
- Design for scalability & automation

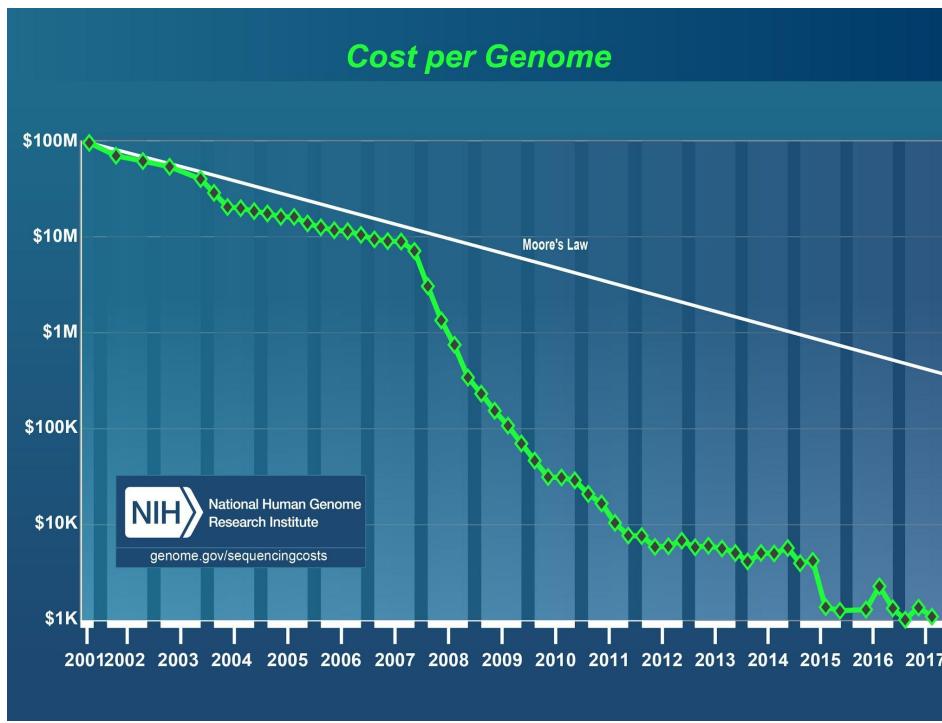


But it's complicated to gain novel insights from genomics data

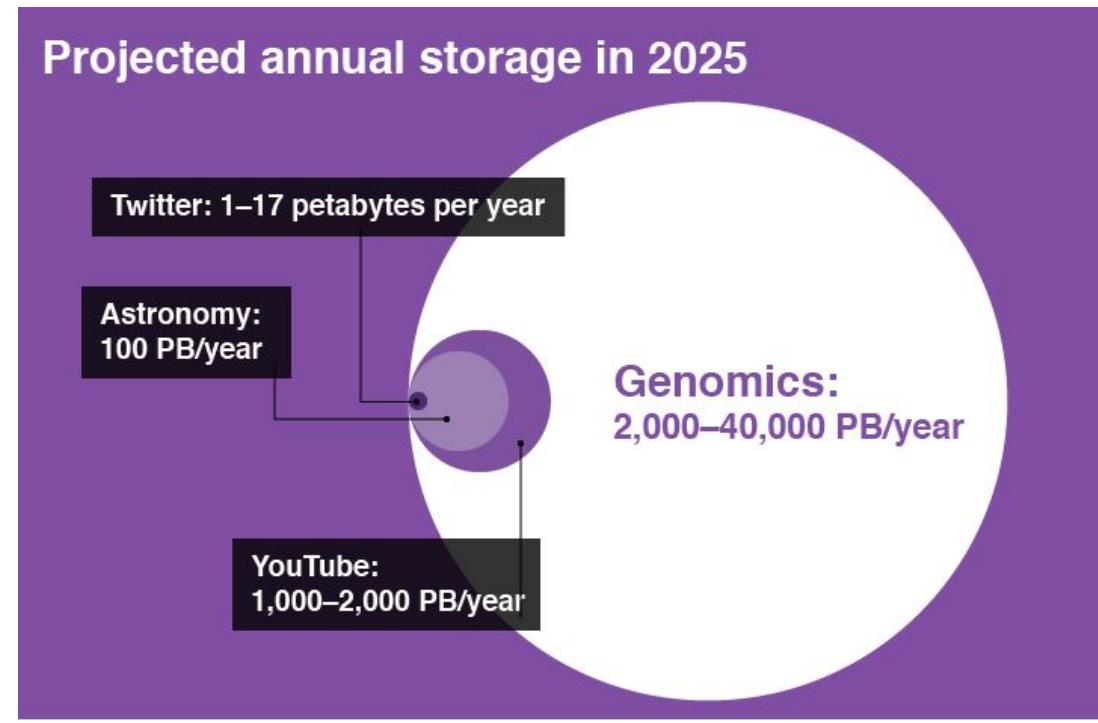


Genomics is a big data problem

From \$2.7B to <\$1,000



40,000 Petabytes / year by 2025



Genomic analysis on big data is hard!

- Existing tools are often difficult to
 - Scale
 - Integrate

Genomic analysis on big data is hard!

- Existing tools are often difficult to
 - Scale
 - Integrate

Welcome to Biostar!



Community Log In Sign Up

Question: How to split vcf file by chromosome?

 I have tried several options on the web including a few post her on Biostar to split my VCF file by chromosome, but could not do it properly. Say I have a vcf file called myvcf.vcf.gz and want to split that per chromosome, what would be the best way to split it by chromosome?
3

 vcf • 14k views

 ADD COMMENT • link • Not following ▾ modified 14 months ago by cartoonist • 60 • written 3.7 years ago by MAPK • 1.4k

 Related post: Splitting VCF file to decrease file size to run it on VEP and wANNOVAR
 ADD REPLY • link written 10 months ago by zx8754 ♦ 8.2k

 **16** 
`bgzip -c myvcf.vcf > myvcf.vcf.gz
tabix -p vcf myvcf.vcf.gz
tabix myvcf.vcf.gz chr1 > chr1.vcf`

 It will give `chr1.vcf` file containing variants for `chr1`. You can loop the last command over all the chromosomes. If you need `vcf header` also, use `-h` flag with last command.

 ADD COMMENT • link modified 13 months ago by RamRS ♦ 24k • written 3.7 years ago by venu ♦ 6.3k

 How do you extract chrX, chrY and chrM? Doesn't seem to work for those.
2 ADD REPLY • link written 3.1 years ago by MAPK • 1.4k

Genomic analysis on big data is hard!

- Existing tools are often difficult to
 - Scale
 - Integrate

Welcome to Biostar!

Community Log In Sign Up

Question: How to split vcf file by chromosome?

```
for i in chr2L chr2R chr3L chr3R chr4 chrX;do
    GenomeAnalysisTK -R ${ref_seq} \
        -T SelectVariants \
        -V my_flies.vcf \
        -L $i \
        -o my_flies.${i}.vcf
    done;
```

chromosomes. If you need `vcf header` also, use `-h` flag with last command.

venu ♦ 6.3k
Germany

ADD COMMENT • link modified 13 months ago by RamRS ♦ 24k • written 3.7 years ago by venu ♦ 6.3k

How do you extract chrX, chrY and chrM? Doesn't seem to work for those.

2 ADD REPLY • link written 3.1 years ago by MAPK • 1.4k

Genomic analysis on big data is hard!

- Existing tools are often difficult to
 - Scale
 - Integrate

Welcome to Biostar!

Biostars
BIOINFORMATICS EXPLAINED

Community Log In Sign Up

Question: How to split vcf file by chromosome?

```
for i in chr2L chr2R chr3L chr3R chr4 chrX;do  
for i in {1..16};  
do vcftools --vcf VCF_FILE --chr $i  
--recode --recode-INFO-all --out VCF_$i;  
done  
-o my_tries.$i.vcf  
done;
```

chromosomes. If you need `vcf header` also, use `-h` flag with last command.

venu ♦ 6.3k
Germany

ADD COMMENT • link modified 13 months ago by RamRS ♦ 24k • written 3.7 years ago by venu ♦ 6.3k

How do you extract chrX, chrY and chrM? Doesn't seem to work for those.
2 ADD REPLY • link

written 3.1 years ago by MAPK • 1.4k

Genomic analysis on big data is hard!

- Existing tools are often difficult to
 - Scale
 - Integrate

Welcome to Biostar!

Biostars
BIOINFORMATICS EXPLAINED

Community Log In Sign Up

Question: How to split vcf file by chromosome?

```
for i in chr2L chr2R chr3L chr3R chr4 chrX;do  
    bgzip -c myvcf.vcf > myvcf.vcf.gz  
    tabix -p vcf myvcf.vcf.gz  
    tabix myvcf.vcf.gz chr1 > chr1.vcf  
done  
-o my_tries.$i.vcf  
done;
```

chromosomes. If you need `vcf header` also, use `-h` flag with last command.

venu ♦ 6.3k
Germany

ADD COMMENT • link modified 13 months ago by RamRS ♦ 24k • written 3.7 years ago by venu ♦ 6.3k

How do you extract chrX, chrY and chrM? Doesn't seem to work for those.
2 ADD REPLY • link

written 3.1 years ago by MAPK • 1.4k

Genomic analysis on big data is hard!

- Existing tools are often difficult to
 - Scale
 - Integrate

Welcome to Biostar!

Community Log In Sign Up

Question: How to split vcf file by chromosome?

```
for i in chr2L chr2R chr3L chr3R chr4 chrX;do  
for begin in myvcf.vcf myvcf.vcf --  
do begin > myvcf.vcf.gz  
--r java -jar SnpSift.jar split file.vcf  
--r begin > myvcf.vcf.gz  
done  
done  
done  
done;  
done;
```

chromosomes. If you need `vcf header` also, use `-h` flag with last command.

venu ♦ 6.3k
Germany

ADD COMMENT · link modified 13 months ago by RamRS ♦ 24k · written 3.7 years ago by venu ♦ 6.3k

How do you extract chrX, chrY and chrM? Doesn't seem to work for those.
2 ADD REPLY · link

written 3.1 years ago by MAPK · 1.4k

Genomic analysis on big data is hard!

- Existing tools are often difficult to
 - Scale
 - Integrate

PLINK 1.9 home plink2-users GitHub **File formats** PLINK 1.9 index PLINK 2.0

Introduction, downloads
S: 17 Jun 2019 (b6.10)
D: 17 Jun 2019
[Recent version history](#)
[What's new?](#)
[Future development](#)
[Limitations](#)
[Note to testers](#)
[\[Jump to search box\]](#)

General usage
[Getting started](#)
[Citation instructions](#)

Standard data input
PLINK 1 binary (.bed)
Autoconversion behavior
PLINK text (.ped, .tped...)
VCF (.vcf(.gz), .bcf)
Oxford (.gen(.gz), .bgen)
23andMe text
Generate random
Unusual chromosome IDs
Recombination map
Allele frequencies
Phenotypes
Covariates
Clusters of samples
Variant sets
Binary distance matrix
IBD report (.genome)

Input filtering
Sample ID file
Variant ID file
Positional ranges file
Cluster membership
Set membership
Attribute-based
Chromosomes
SNPs only
Simple variant window
Multiple variant ranges
Sample/variant thinning
Covariates (-filter)
Missing genotypes

Jump to: [.adjusted](#) | [.allele.no.snp](#) | [.assoc](#) | [.assoc.dosage](#) | [.assoc.fisher](#) | [.assoc.linear](#) | [.assoc.logistic](#) | [.auto.R](#) | [.bcf](#) | [.beagle.dat](#) | [.bed](#) | [.bim](#) | [.blocks*](#) | [.chr-*.](#)dat | [.chr-*.](#)map | [.clst](#) | [.clumped*](#) | [.cluster*](#) | [.cmh](#) | [.cmh2](#) | [.cnv](#) | [.cnv.indiv](#) | [.cnv.overlap](#) | [.cnv.summary](#) | [.cov](#) | [.dfam](#) | [.diff](#) | [.dist](#) | [.dupvar](#) | [.eigenvec*](#) | [.epi.*](#) | [.fam](#) | [.flipscan](#) | [.frq](#) | [.frq.cc](#) | [.frq.count](#) | [.frq.strat](#) | [.frqx](#) | [.fst](#) | [.gen](#) | [.genome](#) | [.grm](#) | [.grm.N.bin](#) | [.grm.bin](#) | [.gvar](#) | [.het](#) | [.hh](#) | [.hom](#) | [.hom.indiv](#) | [.hom.overlap*](#) | [.hom.summary](#) | [.homog](#) | [.hwe](#) | [.ibc](#) | [.imiss](#) | [.info](#) | [.lasso](#) | [.ld](#) | [.ldset](#) | [.lgen](#) | [.list](#) | [.lmiss](#) | [.map](#) | [.mdist](#) | [.mdist.missing](#) | [.mds](#) | [.*mendel](#) | [.meta](#) | [.mibs](#) | [.missing](#) | [.missing.hap](#) | [.model](#) | [.mperm](#) | [.nearest](#) | [.occur.dosage](#) | [.out.dosage](#) | [.ped](#) | [.perm](#) | [.pphe](#) | [.prob](#) | [.profile](#) | [.qassoc](#) | [.qassoc.gxe](#) | [.qassoc.means](#) | [.qfam.*](#) | [.range.report](#) | [.raw](#) | [.recode.*.txt](#) | [.recode.phase.inp](#) | [.recode.strct_in](#) | [.ref](#) | [.rel](#) | [.rlist](#) | [.sample](#) | [.set](#) | [.set.\(permlperm\)](#) | [.set.table](#) | [.sexcheck](#) | [.simfreq](#) | [.tags.list](#) | [.tdt](#) | [.tdt.poo](#) | [.tfam](#) | [.tped](#) | [.traw](#) | [.twolocus](#) | [.var.ranges](#) | [.vcf](#)

.adjusted (basic multiple-testing corrections)
Produced by [--adjust](#).

A text file with a header line, and then one line per set or polymorphic variant with the following 8-11 fields:

CHR	Chromosome code. <i>Not present with set tests.</i>
'SNP'/'SET'	Variant/set identifier
UNADJ	Unadjusted p-value
GC	Devlin & Roeder (1999) genomic control corrected p-value. <i>Requires an additive model.</i>
QQ	P-value quantile. <i>Only present with 'qq-plot' modifier.</i>
BONF	Bonferroni correction
HOLM	Holm-Bonferroni (1979) adjusted p-value
SIDAK_SS	Šidák single-step adjusted p-value
SIDAK_SD	Šidák step-down adjusted p-value
FDR_BH	Benjamini & Hochberg (1995) step-up false discovery control
FDR_BY	Benjamini & Yekutieli (2001) step-up false discovery control

Genomic analysis on big data is hard!

- Existing tools are often difficult to
 - Scale
 - Integrate

The screenshot shows the UCSC PLINK 1.9 homepage with a sidebar on the left containing links like "Introduction, download", "Recent version history", "What's new?", "Future development", "Limitations", "Note to testers", and "Jump to search box". The main content area has a header "Frequently Asked Questions: Data File Formats" and a "Topics" section. Under "General formats", there is a large list of file formats including Axt, BAM, BED, BED detail, bedGraph, barChart, bigBed, bigGenePred, bigPsl, bigMaf, bigChain, bigNarrowPeak, bigWig, and Chain formats. Under "ENCODE-specific formats", there is a list including ENCODE broadPeak, gappedPeak, narrowPeak, pairedTagAlign, peptideMapping, RNA elements, and tagAlign formats. Under "Download-only formats", there is a list including .2bit, .fasta, .fastQ, and .nib formats.

PLINK 1.9 homepage
Frequently Asked Questions: Data File Formats

Topics

General formats

- Axt format
- BAM format
- BED format
- BED detail format
- bedGraph format
- barChart and bigBarChart format
- bigBed format
- bigGenePred table format
- bigPsl table format
- bigMaf table format
- bigChain table format
- bigNarrowPeak table format
- bigWig format
- Chain format

ENCODE-specific formats

- ENCODE broadPeak format
- ENCODE gappedPeak format
- ENCODE narrowPeak format
- ENCODE pairedTagAlign format
- ENCODE peptideMapping format
- ENCODE RNA elements format
- ENCODE tagAlign format

Download-only formats

- .2bit format
- .fasta format
- .fastQ format
- .nib format

Genomic analysis on big data is hard!

- Existing tools are often difficult to
 - Scale
 - Integrate

The screenshot shows the GDC homepage with the PLINK 1.9 homepage visible. The main content area displays a table titled "Imported Data Types and File Formats". The table has three columns: Data Type, Data Subtype, and Format.

Data Type	Data Subtype	Format
Raw Sequencing data	Aligned Reads	BAM
	Unaligned Reads	FASTQ
	Coverage WIG	WIGGLE
Simple Nucleotide Variation	Genotypes	TSV
	Simple Germline Variation	MAF, VCF
	Simple Somatic Mutation	
Raw Microarray Data	Raw Intensities	
	CGH Array QC	
	Intensities Log2Ratio	
	Expression Control	
	Intensities	
	Normalized Intensities	
Download-	Probeset Summary	Metrics
	Methylation Array QC	
	Metrics	

Additional text on the page includes:

- Identifiable by file extension.
(.ats.)
- ssoc.linear | .assoc.logistic |
| .clumped* | .cluster* | .cmh
st | .dupvar | .eigenvect* |
enome | .grm | .grm.N.bin |
| .homog | .hwe | .ibc | .imiss
mds | .*mendel | .meta |
ut.dosage | .ped | .perm |
e.report | .raw | .recode.*.txt |
| lmpperm} | .set.table |
.var.ranges | .vcf
- with the following 8-11 fields:
- quires an additive model.

Genomic analysis on big data is hard!

- Existing tools are often difficult to
 - Scale
 - Integrate

The screenshot shows a section of the GDC website titled "Imported Data Types and File Formats". It features a table with two columns: "Data Type" and "Data Subtype". The table lists various file formats categorized by their type and subtype. A note at the top of the table states: "The following table lists data type and data subtype categories available to users through GDC. Not all programs support all types." The table includes entries for BAM, BED, VCF, and other sequencing and variant formats.

Data Type	Data Subtype
Raw Sequencing data	Aligned Reads
Raw Sequencing data	Unaligned Reads
Coverage WIG	
Simple Nucleotide Variation	Genotypes
Simple Nucleotide Variation	Simple Germline Variation
Simple Nucleotide Variation	Simple Somatic Mutation
Raw Intensities	
CGH Array QC	
Intensities Log2Ratio	
Expression Control	
Raw Microarray Data	Intensities
	Normalized Intensities
	Probeset Summary
	Methylation Array QC Metrics

IGSR: The International Genome Sample Resource
Providing ongoing support for the 1000 Genomes Project data

Data file formats

Alignment files: BAM and CRAM
BAM files are binary representations of the Sequence Alignment/Map format. These files and the associated described in this [Bioinformatics publication](#). Additional information about SAM/BAM is available at the SAMtools development site: <http://samtools.sourceforge.net>

CRAM files are similar to BAM files but give a compressed representation of the alignment. This compression reference the sequence data is aligned to. The file format was designed to reduce the disk footprint of align who provide further information on the [format](#).

Information on working with IGSR CRAM files are available on the [FTP site](#).

Variant Call Format (VCF)
The VCF format is a tab delimited format for storing variant calls and individual genotypes. It is able to store single nucleotide variants to large scale insertions and deletions. [Further details about VCF are available](#).

Data file specifications
The specifications for these file formats continue to develop. Current specifications for SAM/BAM, CRAM are [hts-specs](#).

Summary file formats

BAS
.bas files contain statistics relating to .bam or .cram files, with one line per readgroup and columns separate is a header that describes each column. The first six columns provide meta information about each readgroup columns providing various statistics about the readgroup. Where data isn't available to calculate the result fc value will be 0. Further information is available on the [FTP site](#).

Summary index files
Various types of index file exist on the site, primarily listing available sequence data and alignments. The ind delimited files where the data is arranged in columns. Immediately before the body of the file there is a head with #, that gives the column names. In addition, index files may have further information at the head of the file with ## and can provide descriptions of the columns, the date the index was generated and other pieces of appropriate to the file and data set. Further information is available on the [FTP site](#).

Genomic analysis on big data is hard!

- Existing tools are often difficult to
 - Scale
 - Integrate

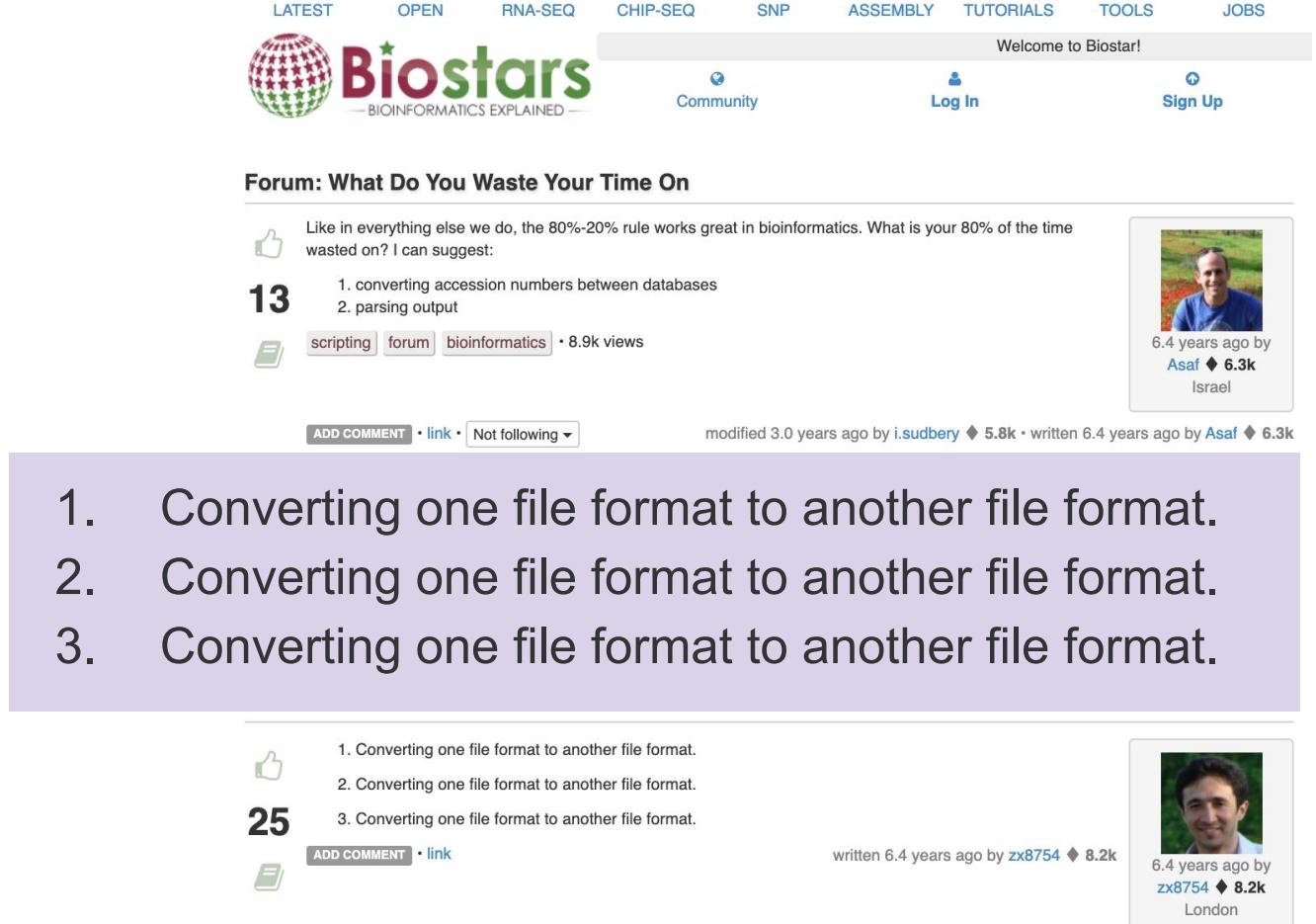
The screenshot shows a forum post titled "Forum: What Do You Waste Your Time On". The post asks, "Like in everything else we do, the 80%-20% rule works great in bioinformatics. What is your 80% of the time wasted on? I can suggest:" followed by a numbered list: 1. converting accession numbers between databases, 2. parsing output. The post has 13 upvotes and 8.9k views. It was written 6.4 years ago by Asaf (6.3k) from Israel.

Below it, another user, Michael Dondrup, has posted a link to www.biostars.org. This post has 44 upvotes and 46k views. It was written 6.4 years ago by Michael Dondrup from Bergen, Norway.

At the bottom, a third user, zx8754, has listed three ways to waste time in bioinformatics: 1. Converting one file format to another file format, 2. Converting one file format to another file format, 3. Converting one file format to another file format. This post has 25 upvotes and 8.2k views. It was written 6.4 years ago by zx8754 from London.

Genomic analysis on big data is hard!

- Existing tools are often difficult to
 - Scale
 - Integrate



The screenshot shows a forum post titled "Forum: What Do You Waste Your Time On". The post contains a list of three items, all of which are "Converting one file format to another file format". The post has 13 upvotes and 25 comments. The author is Asaf, from Israel, who posted 6.4 years ago. The post has 8.9k views and was modified 3.0 years ago by i.sudbery.

Forum: What Do You Waste Your Time On

Like in everything else we do, the 80%-20% rule works great in bioinformatics. What is your 80% of the time wasted on? I can suggest:

13 1. converting accession numbers between databases
2. parsing output

scripting **forum** **bioinformatics** • 8.9k views

6.4 years ago by **Asaf** ♦ 6.3k Israel

modified 3.0 years ago by **i.sudbery** ♦ 5.8k • written 6.4 years ago by **Asaf** ♦ 6.3k

ADD COMMENT • [link](#) • Not following ▾

1. Converting one file format to another file format.
2. Converting one file format to another file format.
3. Converting one file format to another file format.

25 1. Converting one file format to another file format.
2. Converting one file format to another file format.
3. Converting one file format to another file format.

ADD COMMENT • [link](#)

written 6.4 years ago by **zx8754** ♦ 8.2k

6.4 years ago by **zx8754** ♦ 8.2k London

Genomic analysis on big data is hard!

- Existing tools are often difficult to
 - Scale
 - Integrate

Welcome to Biostar!



Biostars
BIOINFORMATICS EXPLAINED

Community Log In Sign Up

Question: VCF files: Change Chromosome Notation

There are two VCF files that I like to merge them, using GATK or VCFtools. The problem is, they have different chromosomal notation, one has Chr, the other does not. This question could be similar to [this one](#)

10 Is there any quick awk/sed commands that you suggest ?! Also I appreciate if you make comment, which of these two (GATK/VCFtools) is more reliable for this task.

next-gen sequence • 19k views

ADD COMMENT

I am trying to solve some issue with vcf file to map

Give a statistical geneticist an awk line, feed him for a day, teach a statistical geneticist how to awk, feed him for a lifetime...

awk

```
awk '{gsub(/CP003827/, "8"); gsub(/CP003822/, "3"); gsub(/CP003824/, "5");
gsub(/CP003834/, "Mt"); gsub(/CP003833/, "14"); gsub(/CP003829/, "10");
gsub(/CP003826/, "7"); gsub(/CP003820/, "1"); gsub(/CP003828/, "9");
gsub(/CP003825/, "6"); gsub(/CP003821/, "2"); gsub(/CP003830/, "11");
gsub(/CP003832/, "13"); gsub(/CP003831/, "12"); gsub(/CP003823/, "4");
print;}' original.vcf > original_newchr.vcf
```

I had no knowledge of awk before stumbling onto this post so there might be a more elegant way to do this, but this seems to work, which is good enough for me!

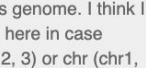
ADD REPLY • link modified 3 months ago by **RamRS** ♦ 24k • written 3.6 years ago by **acgerstein** • 0



5.5 years ago by **Quak** • 300 United States



years ago by **Quak** • 300



ccus genome. I think I st it here in case (1, 2, 3) or chr (chr1, some names in a vcf

Glow is designed to solve these problems





Mission: Industrialize genomics by integrating bioinformatics into data science

Core principles:

- Build on Apache Spark
- Flexibly and natively support genomics tools and file formats
- Provide single-line functions for common genomics workloads
- Build an open-source community

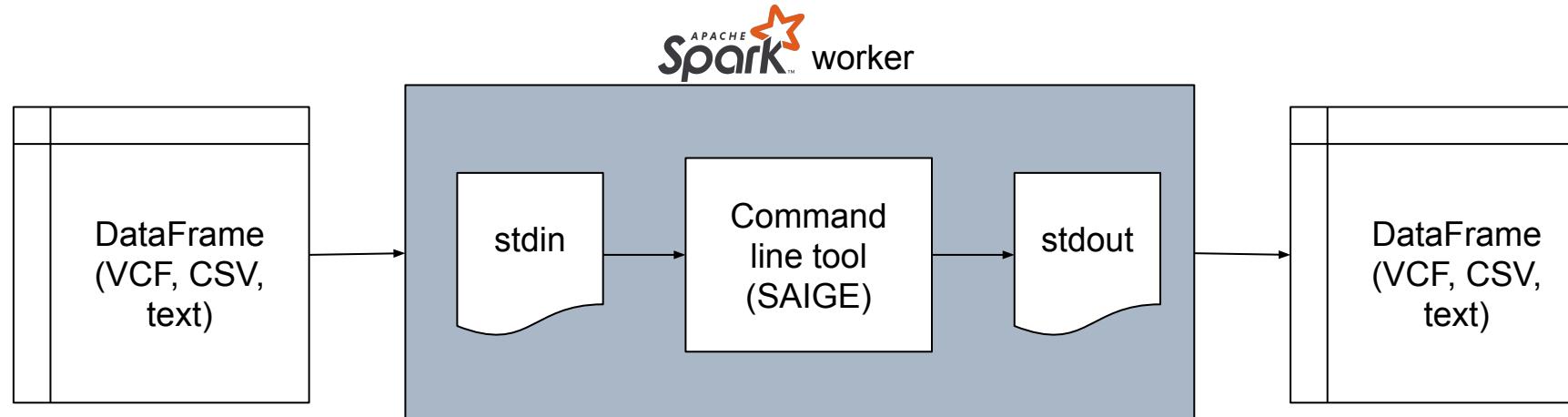
Glow is built on Apache Spark

- Join with phenotype data
- Transition data to ML/data science ecosystem
- High-level APIs in multiple languages
 - Python, R, SQL, Scala, Java



Glow is backwards-compatible with legacy tools and file formats

- File formats: VCF, BGEN, PLINK, GFF
- Parallelize command line bioinformatics tools with the pipe transformer



Glow provides APIs for common genomics workloads

Variant data manipulation

- Quality control
- Liftover
- Variant normalization
- Split multiallelic variants

Tertiary analysis

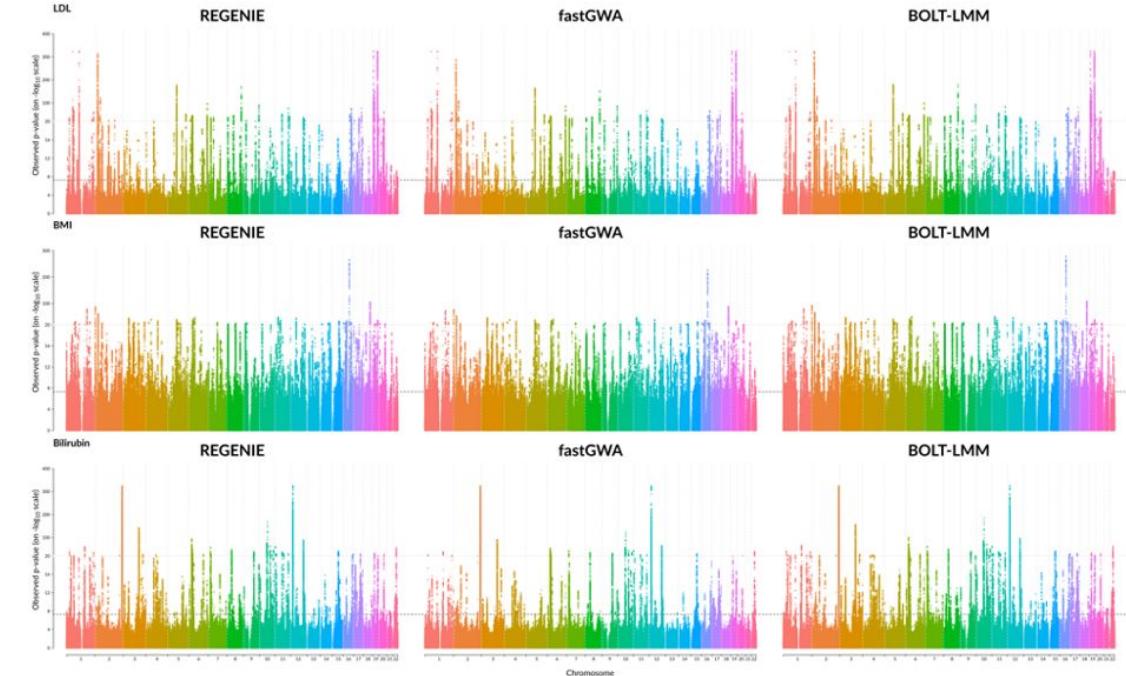
- Integration with annotation, clinical & phenotype data
- Genome-wide association studies

Introducing GloWGR

An industrial-scale, ultra-fast & sensitive method for genetic association studies

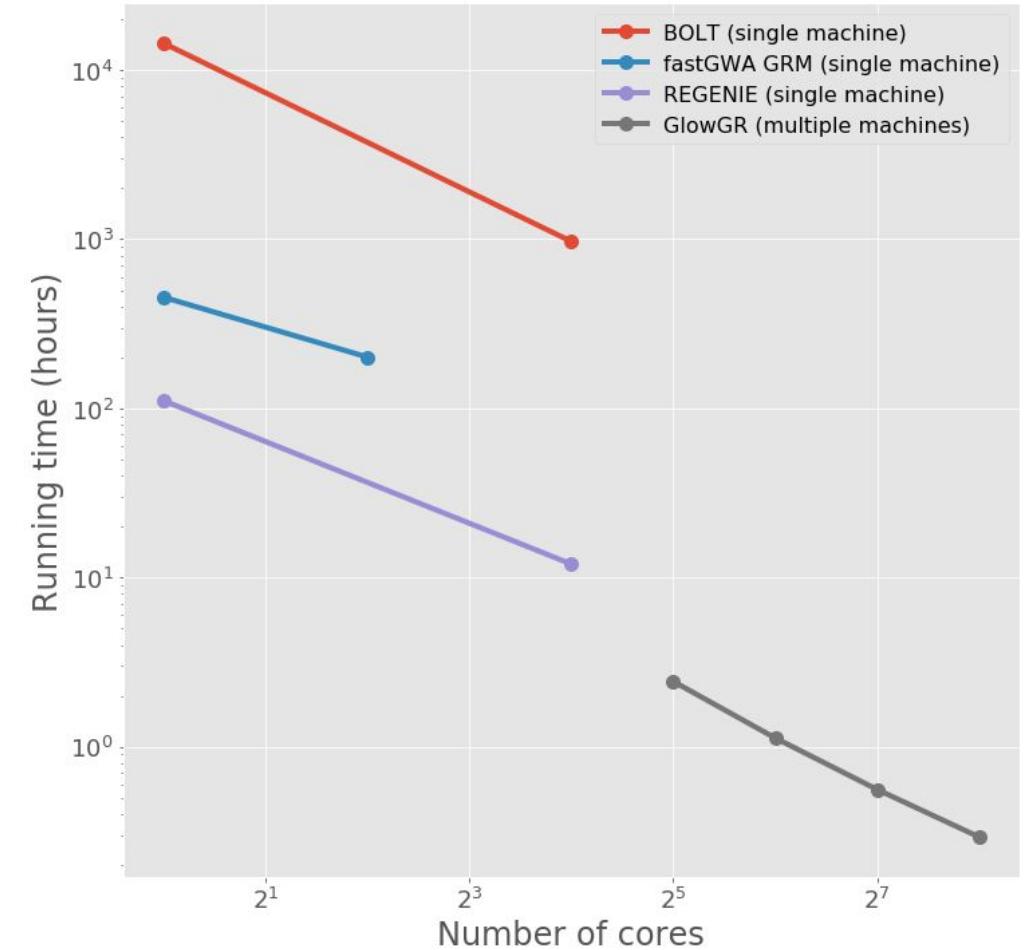
GloWGR is a major innovation in bioinformatics driven by industry collaboration

- Accurately accounts for population structure
- Computationally tractable on biobank-scale datasets with thousands of phenotypes
- Parallelized version of the regenie method using Apache Spark



GloWGR is a major innovation in bioinformatics driven by industry collaboration

- Accurately accounts for population structure
- Computationally tractable on biobank-scale datasets with thousands of phenotypes
- Parallelized version of the regenie method using Apache Spark

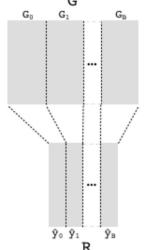
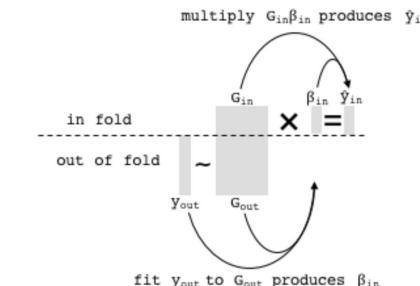
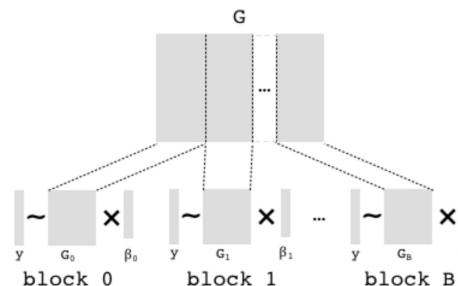


GloWGR is a major innovation in bioinformatics driven by industry collaboration

- Accurately accounts for population structure
- Computationally tractable on biobank-scale datasets with thousands of phenotypes
- Parallelized version of the regenie method using Apache Spark

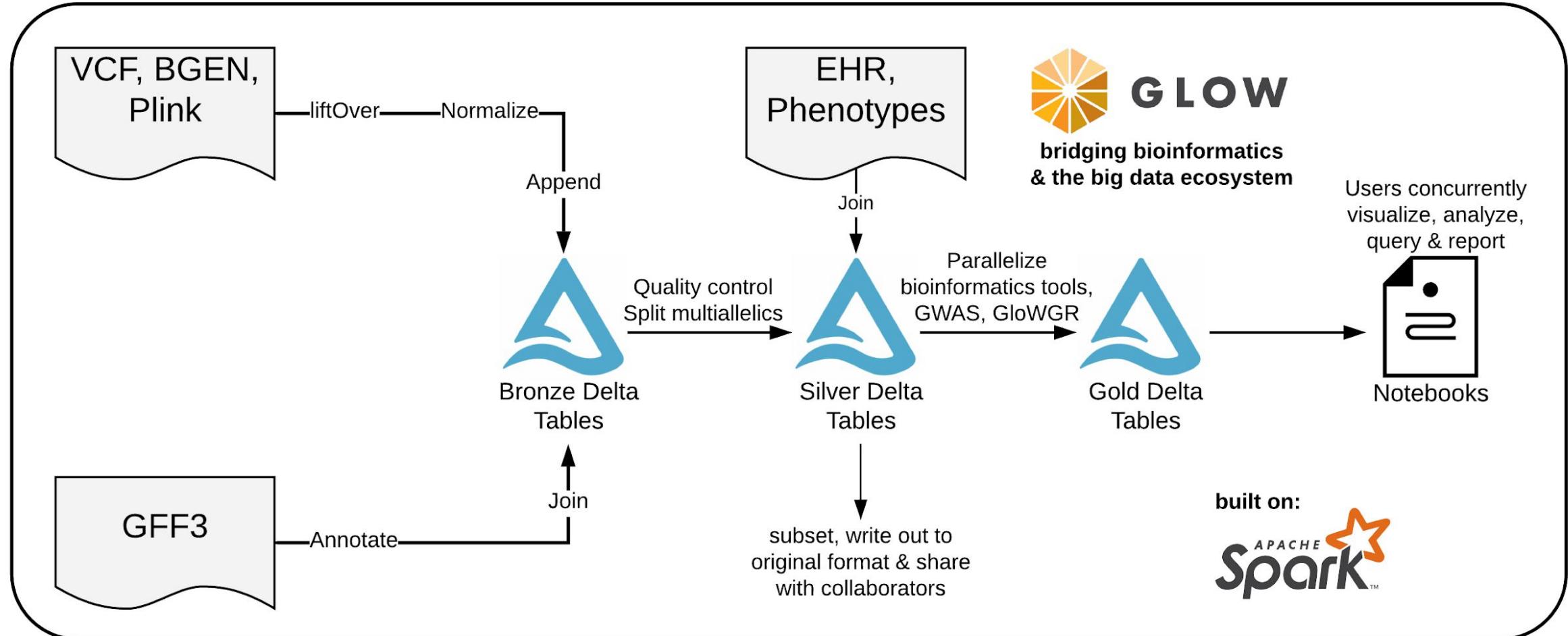
$$\begin{matrix} \mathbf{y} \\ \vdots \\ \mathbf{y}_N \end{matrix} \sim \begin{matrix} \mathbf{G} \\ \vdots \\ \mathbf{g}_{NM} \end{matrix} \times \begin{matrix} \boldsymbol{\beta} \\ \vdots \\ \beta_N \end{matrix}$$

Traditional WGR Method



Decomposition of WGR in
GloWGR/REGENIE

End-to-end genomics workflows in Glow





+



GLOW

projectglow.io