Bystro Tutorial: 2018 ACM-BCB Conference

Alex V. Kotlar
Department Human Genetics, Emory University

Thomas S. Wingo
Division of Neurology, Atlanta VA Medical Center
Departments of Neurology and Human Genetics, Emory University

Financial Disclosures

External Industry Relationships *	Company Name(s)	Role
Equity, stock, or options in biomedical industry companies or publishers	None	N/A
Board of Directors or officer	None	N/A
Royalties from Emory or from external entity	None	N/A
Industry funds to Emory for my research	None	N/A
Other	None	N/A

What is our goal?

Understand how genetic variation influences traits and diseases.

What will you learn?

- Overview of how Bystro helps achieve our goal.
- This will be done with some slide and some live demo.

- Follow along here:
- https://github.com/akotlar/bystro-tutorial
- https://github.com/akotlar/bystro/blob/akotlar-tutorial-patch-1/TUTORIAL.md

- Citation:

Kotlar et al., Genome Biol. Bystro: rapid online variant annotation and natural-language filtering at whole-genome scale. 2018 Feb 6;19(1):14

Our approach: overview

1) Select variants of interest from the genome

2) Apply analytical models on those variants

Our approach: detail

1) Annotate genomic variants with relevant features

2) Select variants of interest on those features

Annotation + Selection

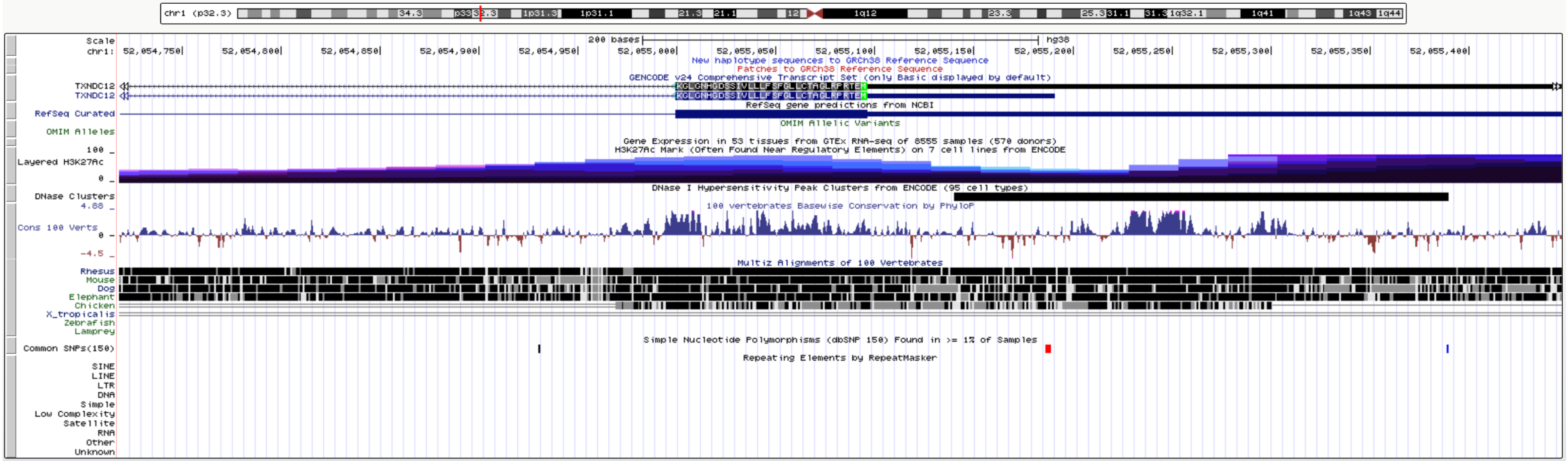
What does the data look like?

#CHROM	POS	REF	ALT	Sample1
Chr1	52055333	G	C	1|1
Chr3	196238448	T	+T	0|1
…				

UCSC Genome Browser on Human Dec. 2013 (GRCh38/hg38) Assembly

move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x 100x

chr1:52,054,719-52,055,447 729 bp. enter position, gene symbol, HGVS or search terms go



move start < 2.0 >

Click on a feature for details. Click or drag in the base position track to zoom in. Click side bars for track options. Drag side bars or labels up or down to reorder tracks. Drag tracks left or right to new position. Press "?" for keyboard shortcuts.

move end < 2.0 >

track search | default tracks | default order | hide all | add custom tracks | track hubs | configure | multi-region | reverse | resize | refresh

collapse all

Use drop-down controls below and press refresh to alter tracks displayed.
Tracks with lots of items will automatically be displayed in more compact modes.

expand all

Mapping and Sequencing

refresh

Genes and Gene Predictions

refresh

GENCODE v24 pack	NCBI RefSeq dense	Other RefSeq hide	All GENCODE... hide	AUGUSTUS hide	CCDS hide
CRISPR... hide	Geneid Genes hide	Genscan Genes hide	IKMC Genes Mapped hide	LRG Transcripts hide	MGC Genes hide
Non-coding RNA... hide	Old UCSC Genes hide	ORFeome Clones hide	Pfam in UCSC Gene hide	RetroGenes V9 hide	SGP Genes hide
SIB Genes hide	TransMap... hide	UCSC Alt Events hide	UniProt hide		

dbSNP SNP rs527993343 Search

Create alert Advanced Help

Display Settings: Summary Send to:

rs527993343 *[Homo sapiens]*

1.

GGAATTGTCTGGGATGTGCCGGGCC[A/G]AGCACTGTAGCTCTCGCCGGTGTCT

Chromosome:	1:920030
Gene:	LOC100130417 (GeneView) LOC284600 (GeneView)
Functional Consequence:	nc transcript variant,upstream variant 2KB
Validated:	by 1000G,by frequency
Global MAF:	A=0.0004/2
HGVS:	NC_000001.10:g.855410G>A, NC_000001.11:g.920030G>A, NR_026874.2:n.-338C>T, NR_122045.1:n.-338C>T, XR_001737589.1:n.6896G>A

Search details

rs527993343[All Fields]

Search See more...

Recent activity

Turn Off Clear

rs527993343 (1) SNP

265926 OR 32657 265923 265926[alleleid] (4) ClinVar

Variant annotation: common tools

1) Time: months runtime
2) Space: terabytes
3) Complex installation
4) Complex output
5) Complex selection/filtering

```
Last login: Tue May  2 10:55:55 on ttys000
Alexs-MacBook-Pro:~ alexkotlar$ ( time pigz -d -c /mn
t/annotator/comparisons/phase1.vcf.gz | perl variant_
effect_predictor.pl  --fork 8 --cache --dir ./.vep --
plugin CADD,.vep/whole_genome_SNV
> s.tsv.gz --no_stats --offline --refSeq --fasta .vep
/homo_sapiens/86_GRCh37/Homo_sapiens.GRCh37.75.dna.pr
imary_assembly.fa --check_ref -o ../../../../comparis
ons/res
> ults/vep/1000G_phase1_all/run1/vep-annotation.t>>>>
> ults/vep/1000G_phase1_all/run1/vep-annotation.txt ;
 ) &> ../../../../comparisons/results/vep/1000G_phase
1_all/run1/annotation-time.log
```

Variant selection: same challenges

```
gemini query -q "select chrom, start, end, ref, alt,
(gts).(*) from variants" \
--gt-filter "gt_types.mom == HET and \ gt_types.dad == HET
and \gt_types.kid == HOM_ALT" \
$db
```

Outputs are complex

Chr	Start	End	Ref	Alt	Func.refGen	Gene.refGen	GeneDetail.r	ExonicFunc.r	AAChange.re	phastCons10	phyloP100w	ExAC_ALL	ExAC_AFR
1	5935162	5935162	A	T	splicing;intrc	NM_001291	NM_001291	.	.	Name=chr1.!	Name=chr1.!	0.8362	0.8949
1	12065948	12065948	C	T	exonic	NM_001127	.	nonsynonym	MFN2:NM_0	Name=chr1.:	Name=chr1.:	8.24E-06	0
1	46655126	46655129	TCAC	-	splicing;intrc	NM_001290	.	.	.	Name=chr1.4	Name=chr1.4	.	.
1	68912524	68912532	GAGCCAGAG	-	exonic	NM_000329	.	nonframeshi	RPE65:NM_0	Name=chr1.(	Name=chr1.(	.	.
1	68912527	68912532	CCAGAG	-	exonic	NM_000329	.	nonframeshi	RPE65:NM_0	Name=chr1.(	Name=chr1.(	.	.
1	109817590	109817590	G	T	UTR3	NM_001408	NM_001408	.	.	Name=chr1.:	Name=chr1.:	.	.
1	145597476	145597479	AAGT	-	intronic	NM_001039	.	.	.	Name=chr1.:	Name=chr1.:	0.0006	0.0066
1	153791301	153791302	TG	-	exonic	NM_020699	.	frameshift d	GATAD2B:NI	Name=chr1.:	Name=chr1.:	.	.
1	156104667	156104690	TGAGAGCCG	CCCC	exonic	NM_001257	.	frameshift su	LMNA:NM_1	Name=chr1.:	Name=chr1.:	.	.
1	156108541	156108541	-	G	exonic;intror	NM_001257	dist=884	frameshift in	LMNA:NM_1	Name=chr1.:	Name=chr1.:	.	.
1	161279695	161279695	T	A	exonic	NM_000530	.	nonsynonym	MPZ:NM_00	Name=chr1.:	Name=chr1.:	.	.
1	169519049	169519049	T	T	exonic	NM_000130	.	synonymous	F5:NM_0001	Name=chr1.:	Name=chr1.:	.	.
1	226125468	226125468	G	A	exonic	NM_001172	.	synonymous	LEFTY2:NM_	Name=chr1.:	Name=chr1.:	0.0094	0.0926
10	89623036	89623038	GCA	-	UTR5;upstre	NM_001126	NM_001126	.	.	Name=chr10	Name=chr10	.	.
11	62457852	62457852	C	A	exonic;ncRN	NM_001122	NM_001130	nonsynonym	BSCL2:NM_0	Name=chr11	Name=chr11	.	.
11	108178710	108178710	-	T	exonic;down	NM_000051	dist=536	frameshift in	ATM:NM_00	Name=chr11	Name=chr11	.	.
11	111735981	111735981	G	A	intronic	NM_001077	.	.	.	Name=chr11	Name=chr11	0.0228	0.0048
12	11023080	11023080	C	A	ncRNA_intro	NR_037918	.	.	.	Name=chr12	Name=chr12	.	.
12	22018713	22018713	C	-	intronic	NM_005691	.	.	.	Name=chr12	Name=chr12	.	.
12	52912946	52912946	T	C	splicing;intrc	NM_000424	NM_000424	.	.	Name=chr12	Name=chr12	.	.
12	103234293	103234293	C	-	exonic	NM_000277	.	frameshift d	PAH:NM_00	Name=chr12	Name=chr12	.	.
12	103234293	103234293	C	-	exonic	NM_000277	.	frameshift d	PAH:NM_00	Name=chr12	Name=chr12	.	.
12	103311124	103311124	T	C	UTR5	NM_000277	NM_000277	.	.	Name=chr12	Name=chr12	.	.

Bystro

Online analysis

(annotation + selection)

Simply fast

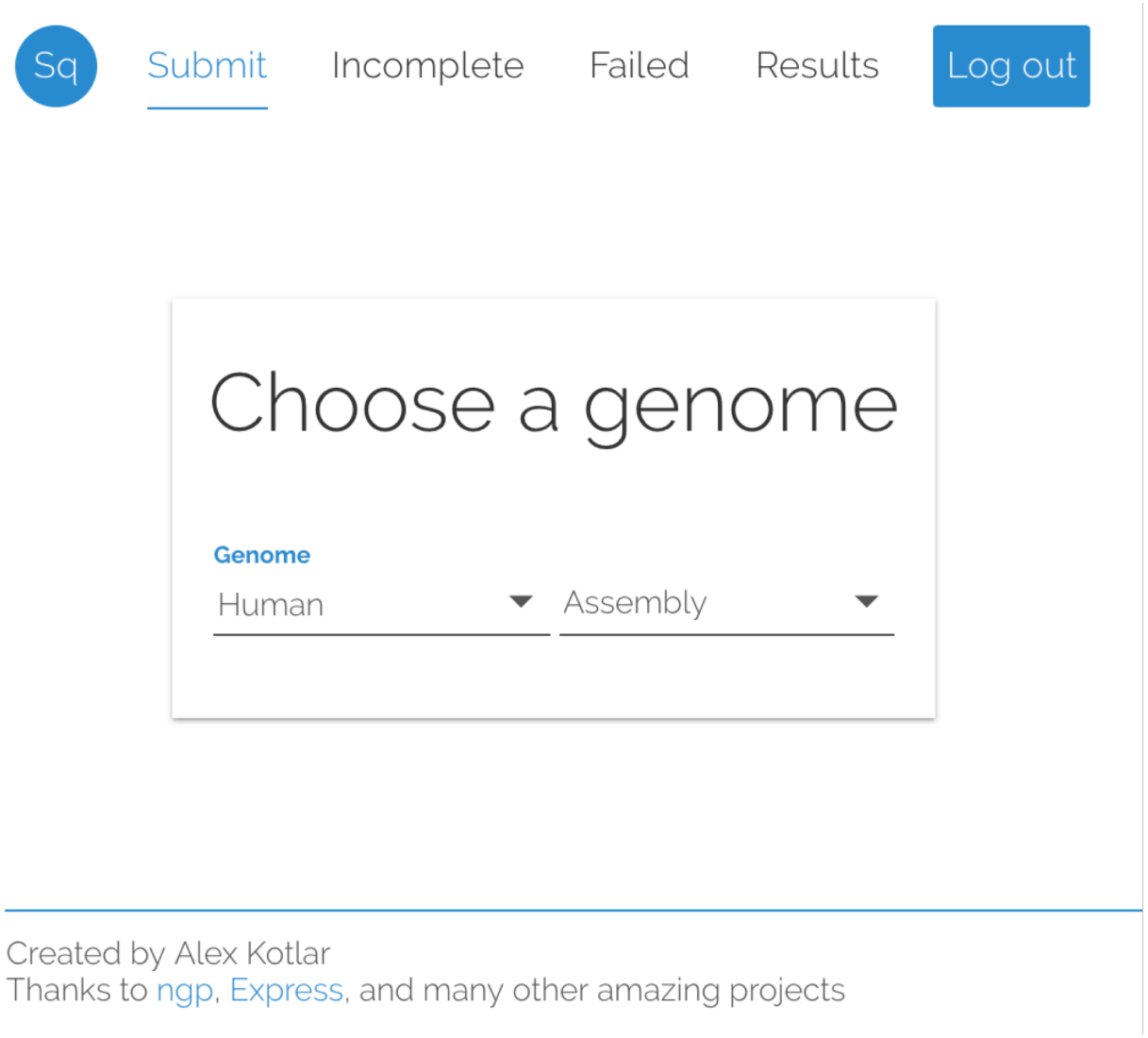
~2M variants/min (4 cores, .snp)

- 900k/min : vcf, 1 WGS sample
- 650k/min : vcf, 2504 WGS samples
- O(nlog(n))

Google-like variant selection

21TB+ submissions

~1GB memory (min: ~200MB)



Bystro

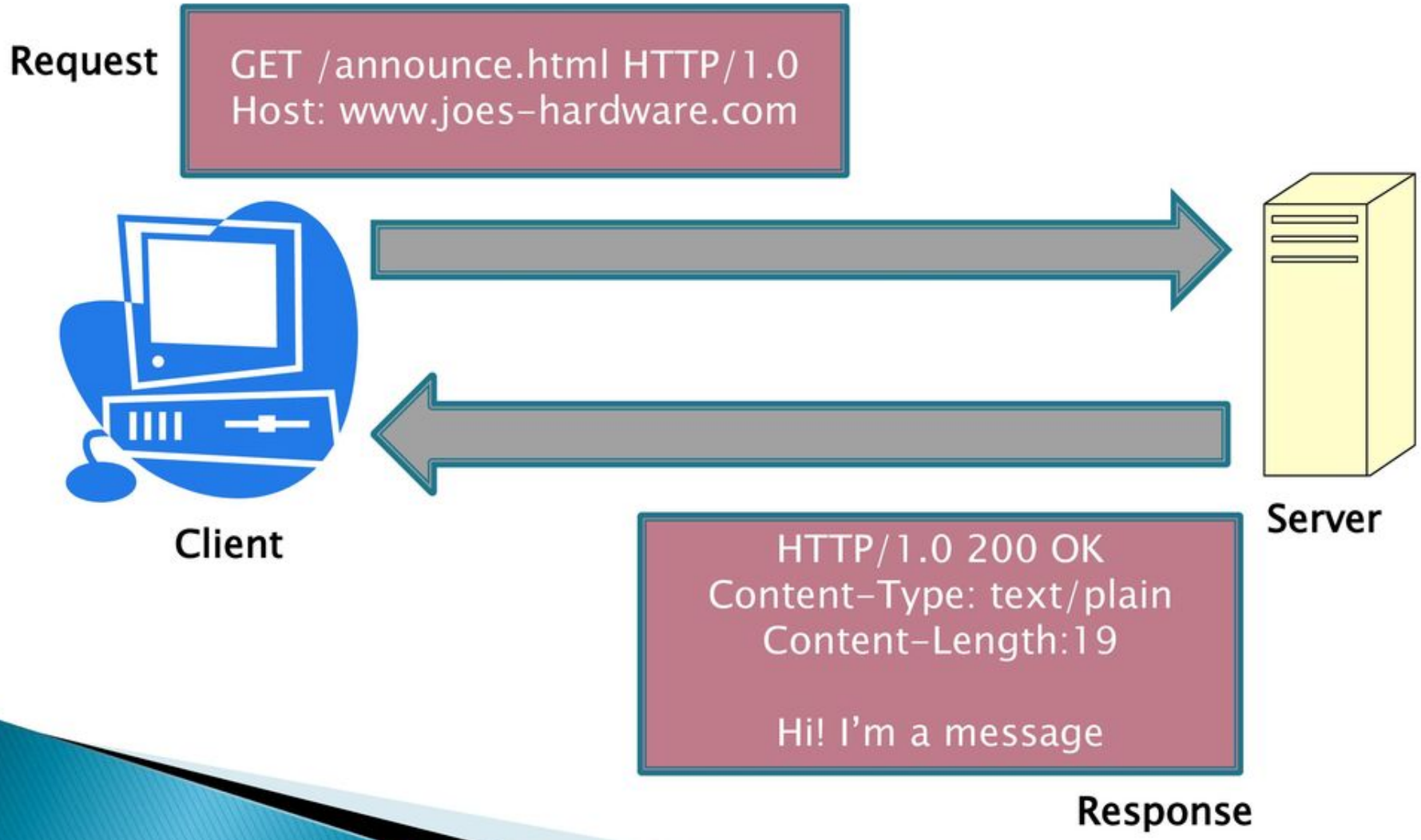
Variant annotation & filtering for any size data.

Start Guide Try

Architectural Overview

The “cloud”: fault-tolerant microservices

Request/ Response Messages



Architecture Summary

Async single-page web application

- AngularJS
- 60fps render

Beanstalkd atomic queue

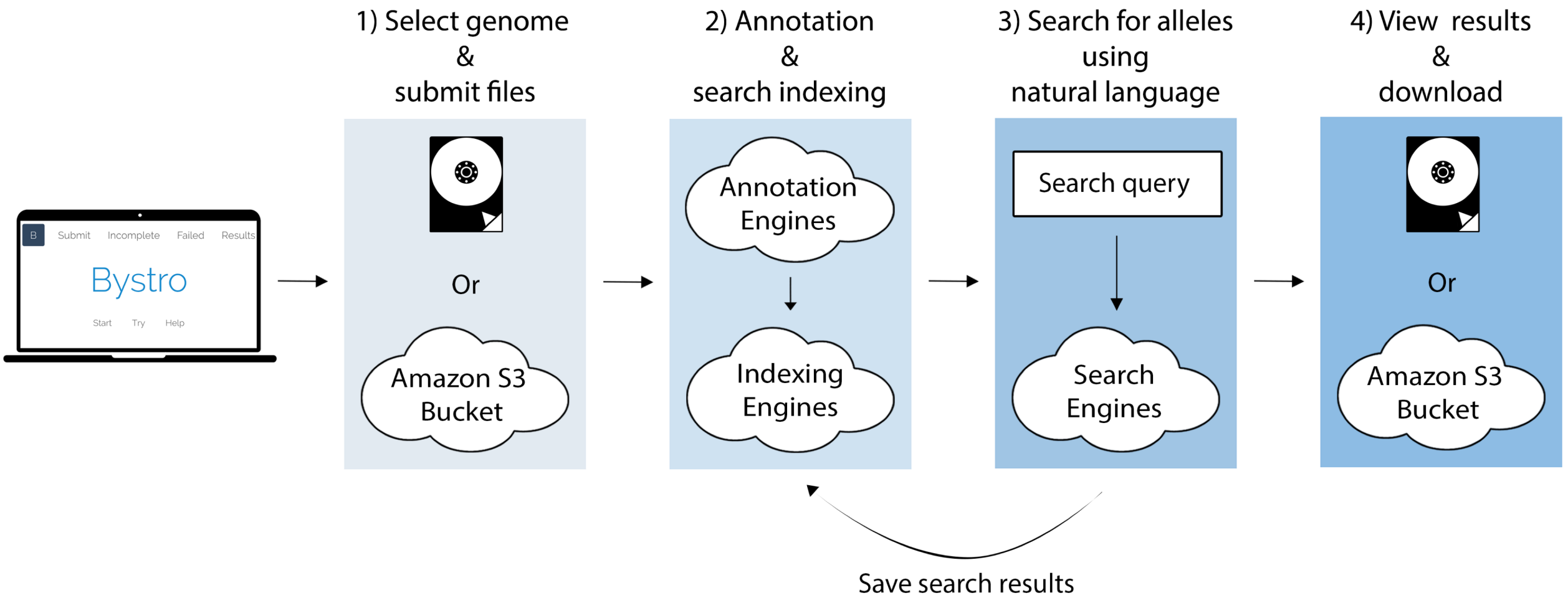
- 1 and only 1 completion guarantee

Real-time state sync (notifications)

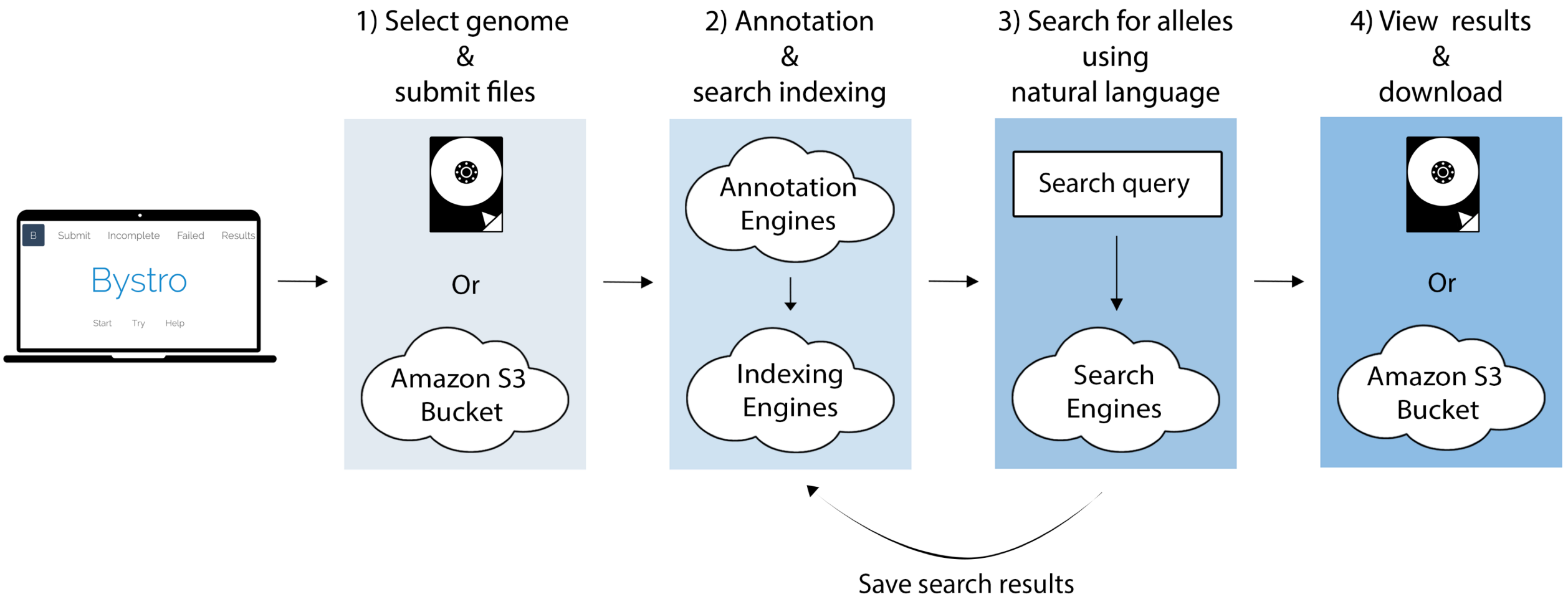
- websocket/Socket.io + beanstalkd + Mongo

Encrypted connection, encrypted data at rest

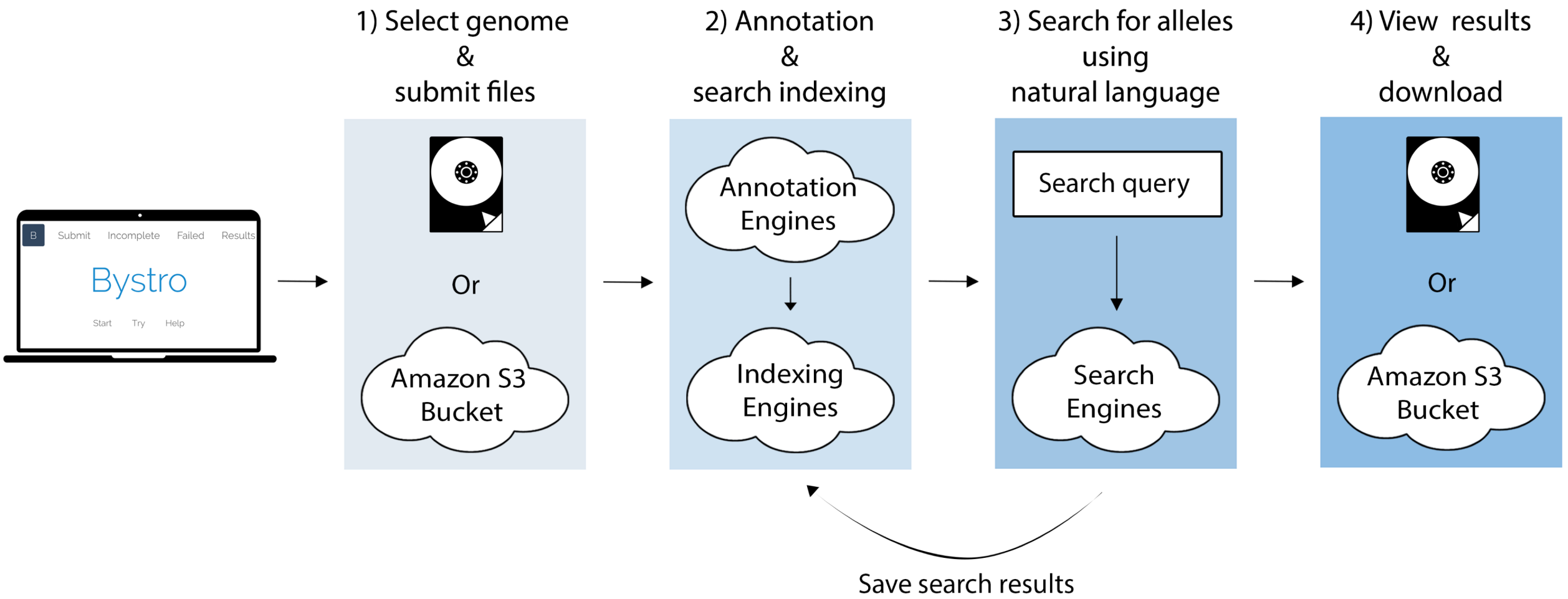
Architecture Overview



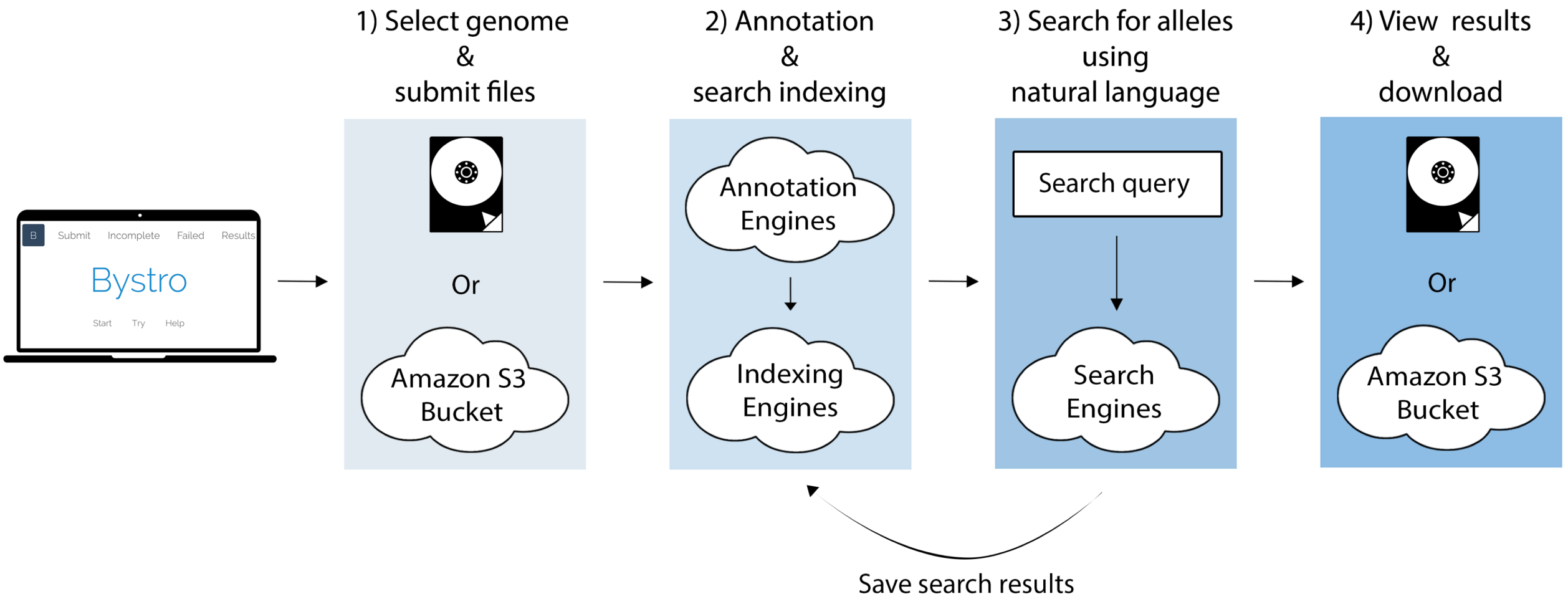
Architecture Overview



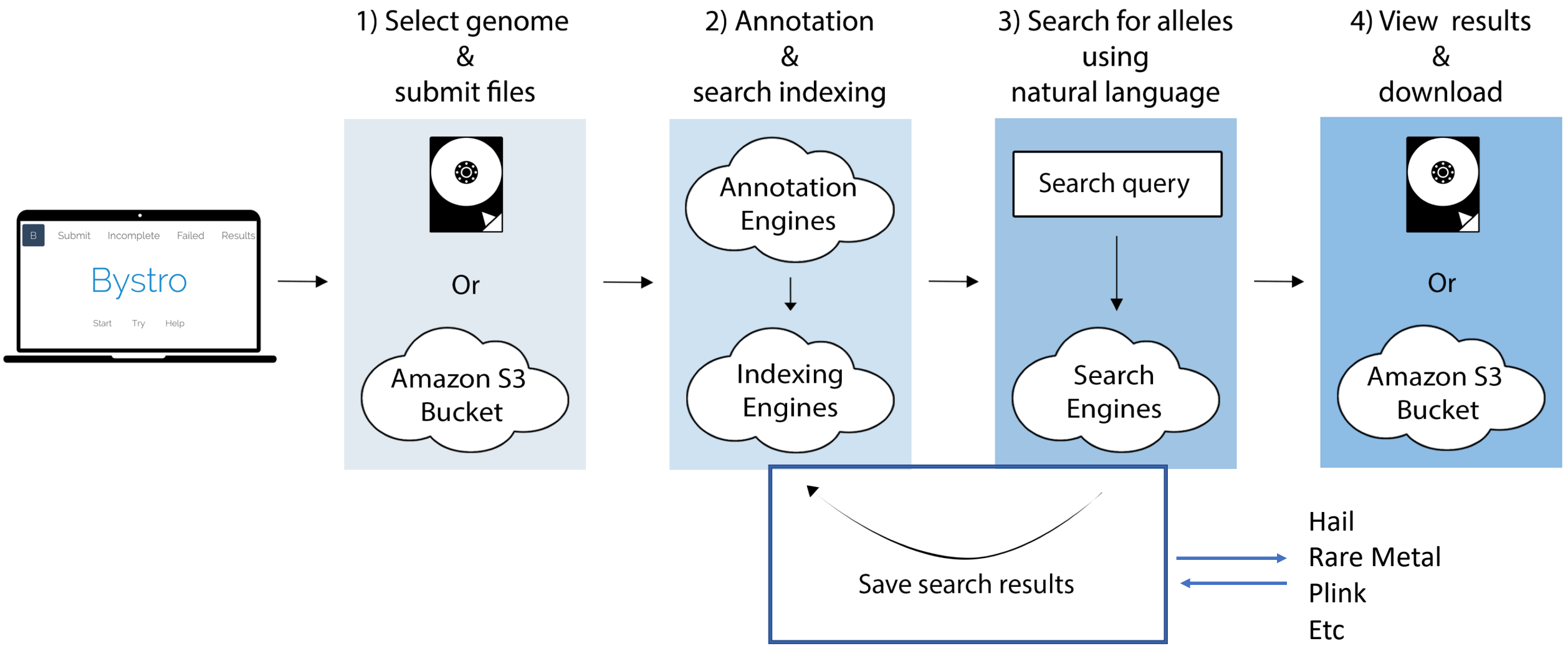
Architecture Overview



Architecture Overview



Pluggable Analysis Interface



The “cloud” necessitates efficiency

How is efficiency achieved?

Multithreaded

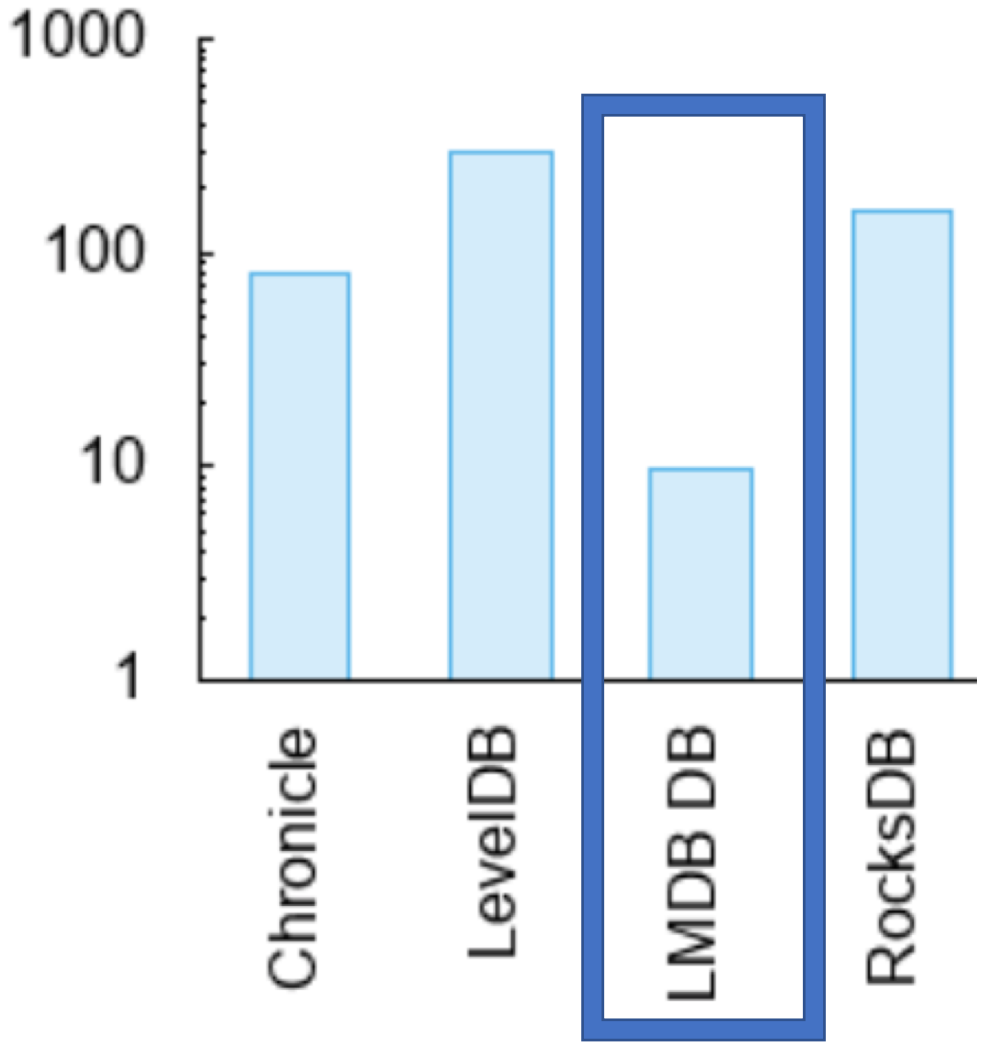
- No in-memory database

LMDB

- Cache: memory-mapped
- Low disk access: B+tree
- Fast: Read w/o copy, lock

1 db read / variant

Go for bottlenecks



Binary database structure

No hashes: 4D arrays

msgpack binary serialization

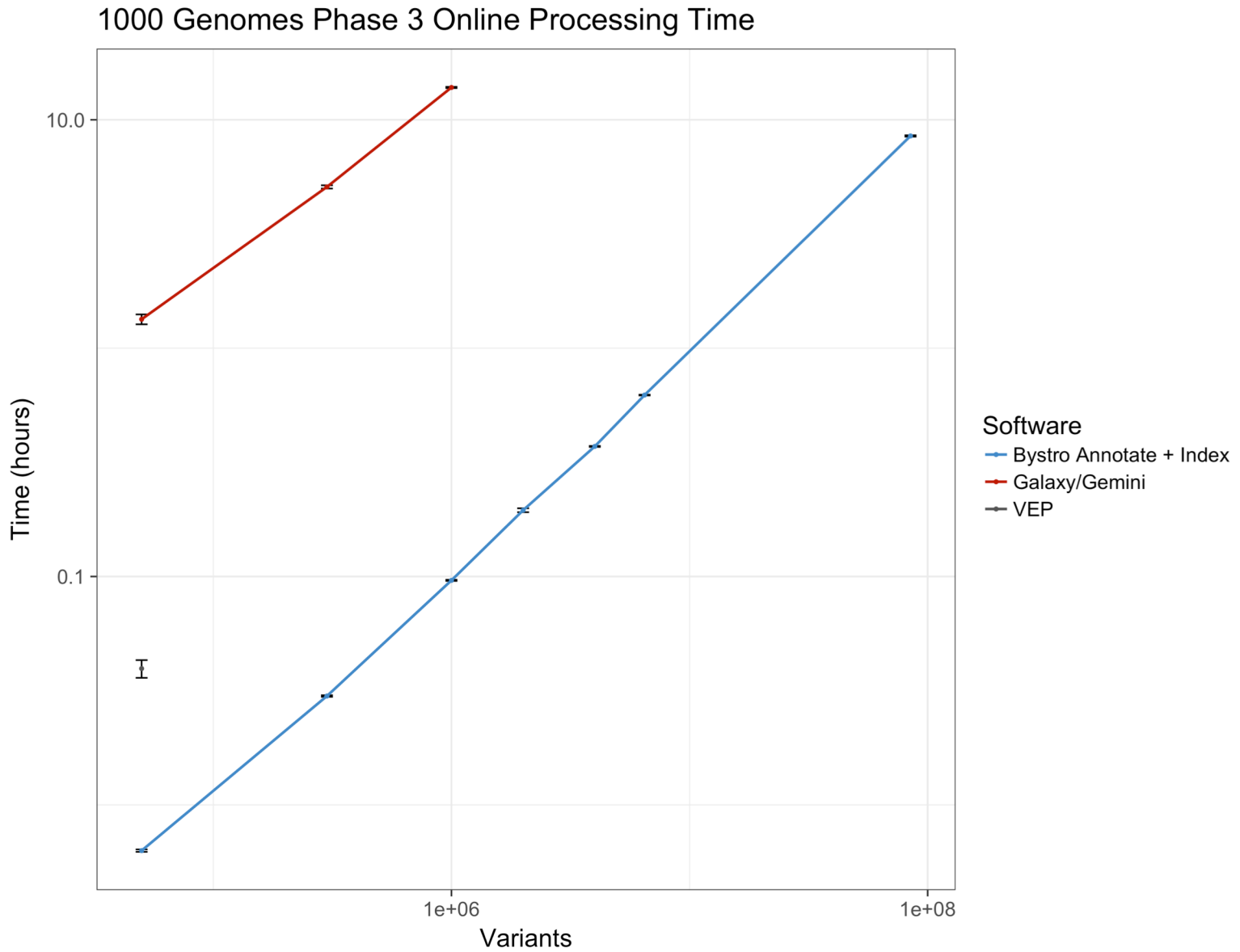
16 bit floats (custom)

1 byte for wigFix values

Implementation	Bytes	Overhead %
(Flat Array)	163,720,000,000	
LevelDB	163,795,005,440	.04
RocksDB	163,795,025,920	.04
Chronicle	163,832,737,792	.06
LMDB DB	164,160,393,216	.26

Performance

Online Performance



Annotation Accuracy

Annotation Accuracy

Variant	POS	REF	ALT
Input	42680000	CA	CAA
Bystro	42680000	C|A	+A
Annovar	42680001-42680001	-	A
VEP	42680001-42680002	-	A

Annotation Accuracy

- Annotation accuracy was performed using the dataset provided here:
  - Yen JL et al., A variant by any name: quantifying annotation discordance across tools and clinical databases. *Genome Med* 2017, 9:7.

- See the Bystro paper – Kotlar et al. 2018 Feb 6;19(1):14

Variant Selection

Selection Accuracy

Query	Variants	% Match Filter	% Match Custom
cadd > 15 alt:(A || C || T || G)	28,099	100%	100%
gnomad.exomes.af < .001	6,840	100%	100%
cadd > 15 missense			
gnomad.exomes.af < .001	6,840	100%	100%
cadd > 15 nonsynonymous			
cadd > 15	29.057	100%	100%
alt:(A || C || T || G)	963,802	100%	100%
gnomad.exomes.af < .001	30,674	100%	100%
missense	16,326	100%	100%
nonsynonymous	16,326	100%	100%

Lets try it - https://bystro.io

We’ve annotated, now what?

Quality Control

Lets try it - https://bystro.io

Sample QC

Sample QC

Bystro flags samples 3 s.d. from mean of:

- heterozygosity/homozygosity
- transition/transversion
- silent/replacement
- theta (Watterson estimator)
- exonic theta (Watterson estimator)

In downloaded archive: **file_name.qc.tsv**

Watterson GA. (1975) On the Number of Segregating Sites in Genetical Models without Recombination. Theoretical Population Biology 7, 256-276

Lets try it - https://bystro.io

Variant QC

Variant processing during annotation

Bystro automatically drops:

1. non-PASS/. variants
2. non-ACTG variants
3. missing-only variants
4. ambiguous multiallelic indel variants*

Normalizes indel representations

Reports key statistics: transition/transversion, silent/replacement

*(very rare)

Variant QC

Filter on any values in annotation

missingness < .05

sampleMaf < 1e-4

Suggest new fields (and fork): github.com/akotlar/bystro/issues

“Filters” add (filtering) abilities

Want to filter on statistics:

1. Variants @ different frequency that gnomAD
   - Bonferroni-corrected alpha
   - Multiple allele frequency estimates (genomes, exomes)

2. Hardy-Weinberg Equilibrium
   - Now : all samples tested
   - Soon: exclude/include samples

Lets try it - https://bystro.io

Statistical Analysis

Analysis Overview

1. Prepare data:
   - Download from Bystro search engine
   - Prepare linkage and covariate data

2. Generate gene sets:
   Each gene independently

3. Run association tests

Step 1: Convert to Plink

Convert to plink

```
bystro_to_vcf hg38.sdx <(pigz -d –c ann.tsv.gz ) \
 sample_list | pigz -c > annotation.vcf.gz

plink --vcf annotation.vcf.gz --keep-allele-order \
--const-fid seq --make-bed --out annotation.plink
```

Add your phenotype information

rm annotation.plink.fam;

cp your.fam annotation.plink.fam;

Step 2: Plink-specific QC

Plink-specific QC

An escape hatch to Bystro

Example:

```
plink --bfile ann.plink --hwe 1e-6 midp --make-bed --out ann.plink.hwe
```

SNP Association

Individual SNP association

Example codes provided online

Generating SKAT models

Generate SKAT models

Query variants

Generating SKAT set file

Generate SKAT set file

A bit more involved

Helper script provided (skat_set_id.pl)

Machine Learning with Bystro

Goal

- Classify variants using bystro annotation and your own external classification
- For our toy example we will use Clinvar since someone has labeled it as “Pathogenic.”
  - Here “pathogenic” refers typically to monogenic illnesses.

Hurdle – genomic data completeness

- One issue with using bystro for this is that the underlying genomic data is sometimes sparse
  - E.g., Allele frequency in dbSNP, gNOMAD, etc.
  - These will need to be changed to numeric data or otherwise dropped.
    - We have functions to help with this.

Lets try it - https://bystro.io

QC with Bystro

Goal: use Bystro to automate more of this

Conversion to VCF/plink will be up Friday. Check back on dev!

Future: Integrate Hail, Plink/Seq.

Future Directions

- Download VCF of filtered data
- Future integration with Hail, Plink/Seq, RAREMETAL

Acknowledgements

Bystro Development

- Alex Kotlar
- David Cutler
- Thomas Wingo

Collaborators

- Michael E. Zwick
- Karen Conneely
- Allan I. Levey
- James J. Lah

Funding:
Veterans Affairs (ORD), IK2 BX001820
National Institutes of Health (NIA), P50 AG025688, R01 AG056533, RF1 AG057470

Questions

Bystro links and citation

- https://bystro.io
- https://github.com/akotlar/bystro
- https://github.com/akotlar/bystro-vcf
- Kotlar et al., Genome Biol. 2018 Feb 6;19(1):14.