KBASE
predictive biology
DOE Systems Biology Knowledgebase

KBase Variation Services

Overview and Demo

Michael Schatz, James Gurtowski
Cold Spring Harbor Laboratory

U.S. DEPARTMENT OF ENERGY

KBASE
predictive biology
DOE Systems Biology Knowledgebase

Agenda

1. Introduction to KBase
2. Resequencing and variation calling theory
3. KBase services for variation calling
4. Live Demo
5. Additional Resources

KBASE
predictive biology
DOE Systems Biology Knowledgebase

Agenda

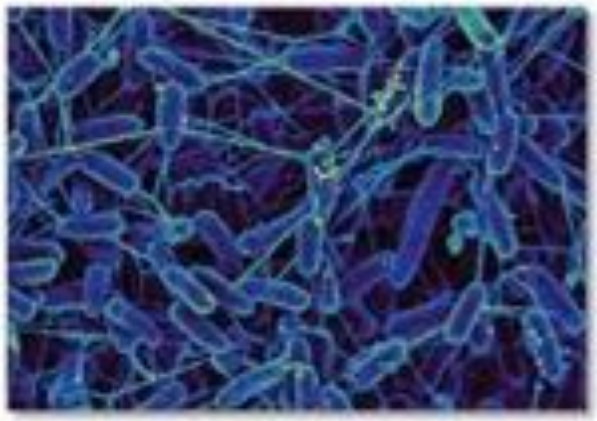
1. Introduction to KBase
2. Resequencing and variation calling theory
3. KBase services for variation calling
4. Live Demo
5. Additional Resources

KBASE predictive biology
DOE Systems Biology Knowledgebase

Overview

Knowledgebase enabling ***predictive*** systems biology.

- Powerful ***modeling*** framework.
- ***Community-driven***, extensible and scalable ***open-source*** software and application system.
- Infrastructure for integration and reconciliation of ***algorithms*** and ***data sources***.
- Framework for standardization, search, and ***association*** of data
- Resources to enable ***experimental design*** and ***interpretation*** of results.



Microbes

Communities

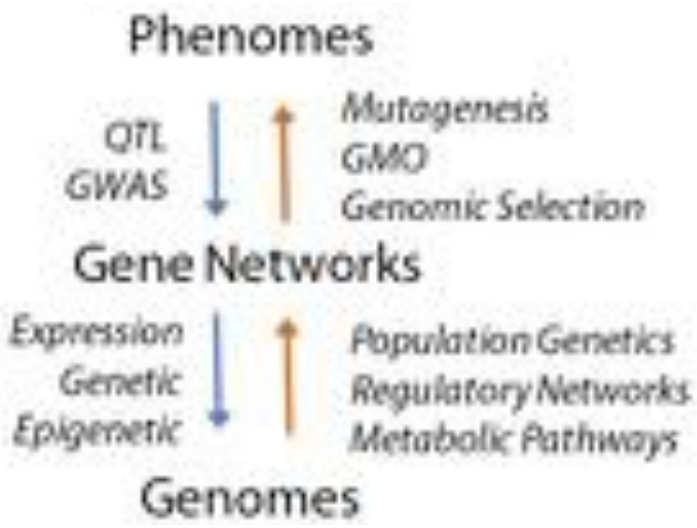
Plants

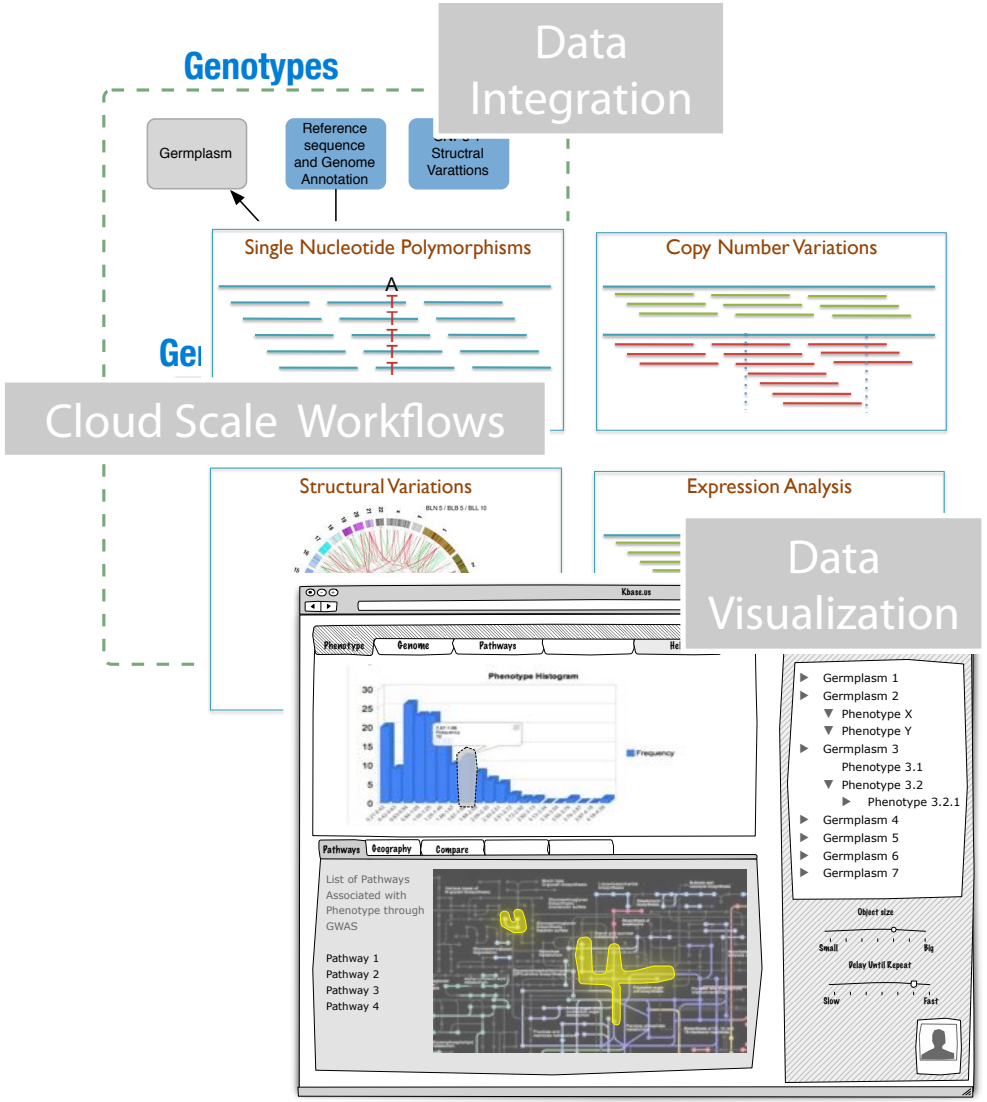
KBASE predictive biology

DOE Systems Biology Knowledgebase

KBase : Plants

Model development
Hypothesis testing
Knowledge Synthesis

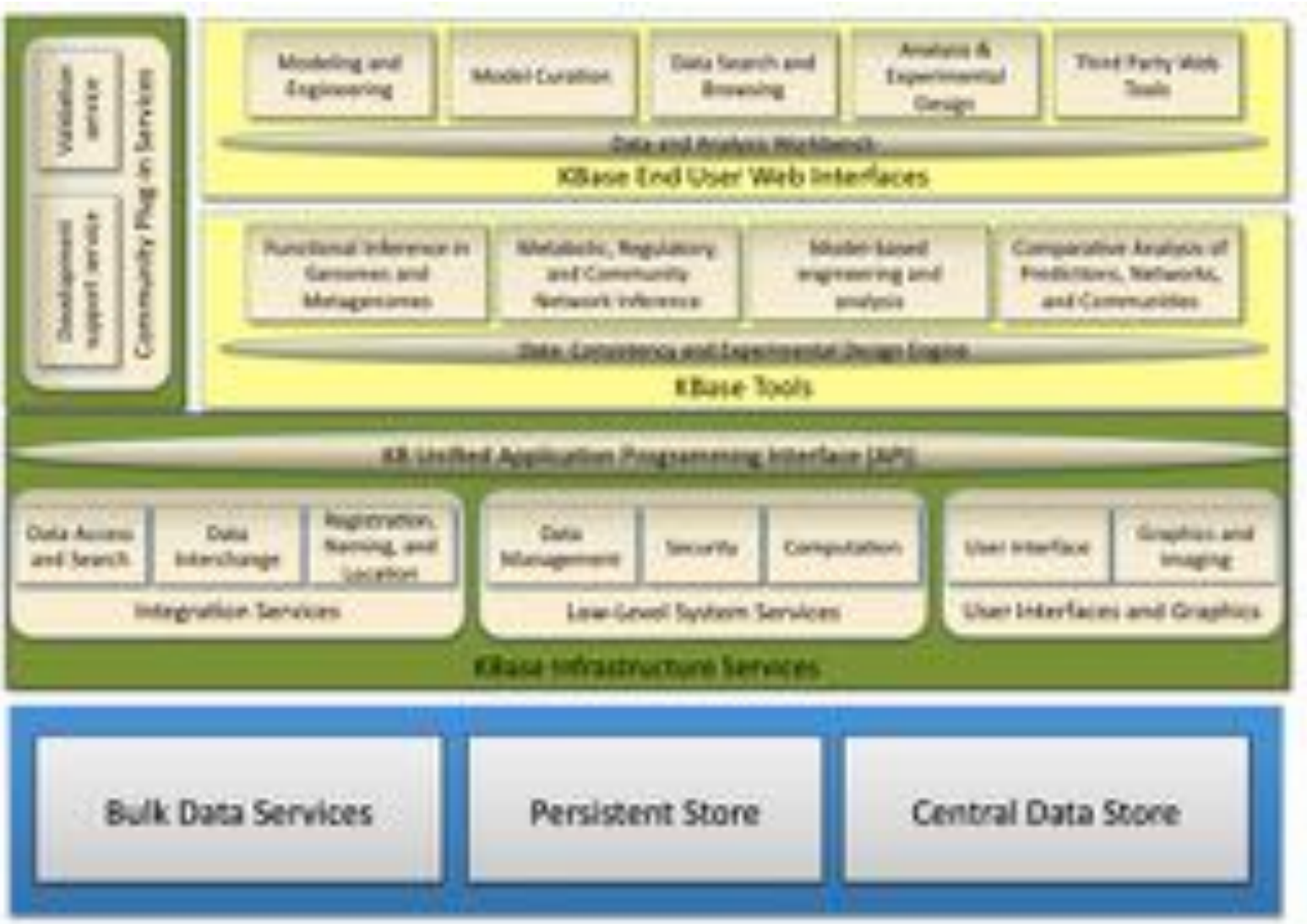




U.S. DEPARTMENT OF ENERGY

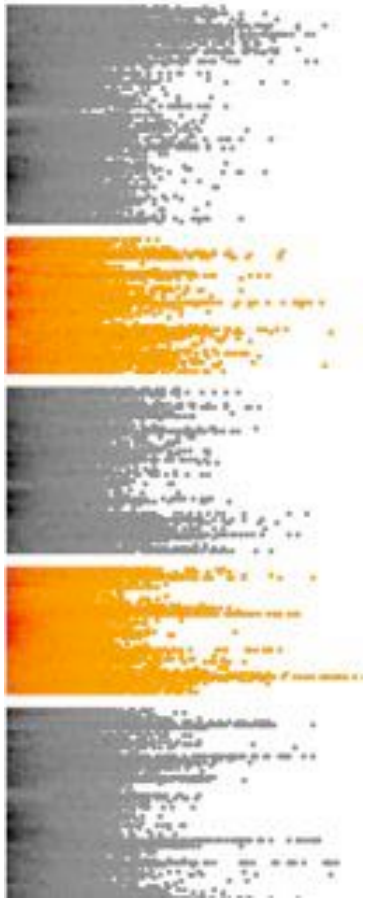
KBASE
predictive biology
DOE Systems Biology Knowledgebase

KBase Infrastructure and Services



KBASE
predictive biology
DOE Systems Biology Knowledgebase

Variation Services: Samples to Discoveries



Powered by KBase

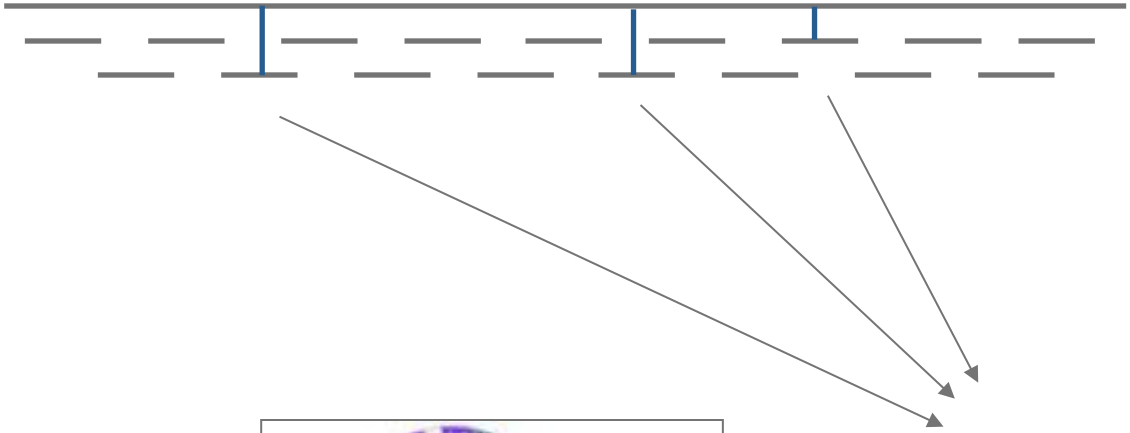
KBASE
predictive biology
DOE Systems Biology Knowledgebase

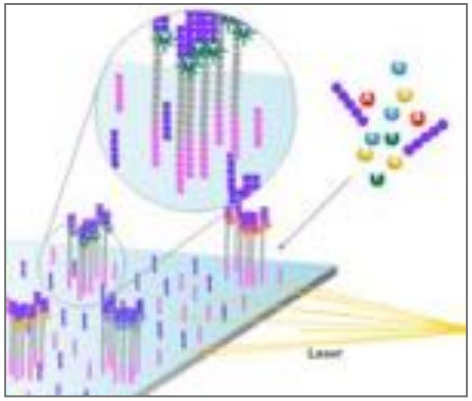
Agenda

1. Introduction to KBase
2. Resequencing and variation calling theory
3. KBase services for variation calling
4. Live Demo
5. Additional Resources

Resequencing & Variations

How does your sample compare to the reference?



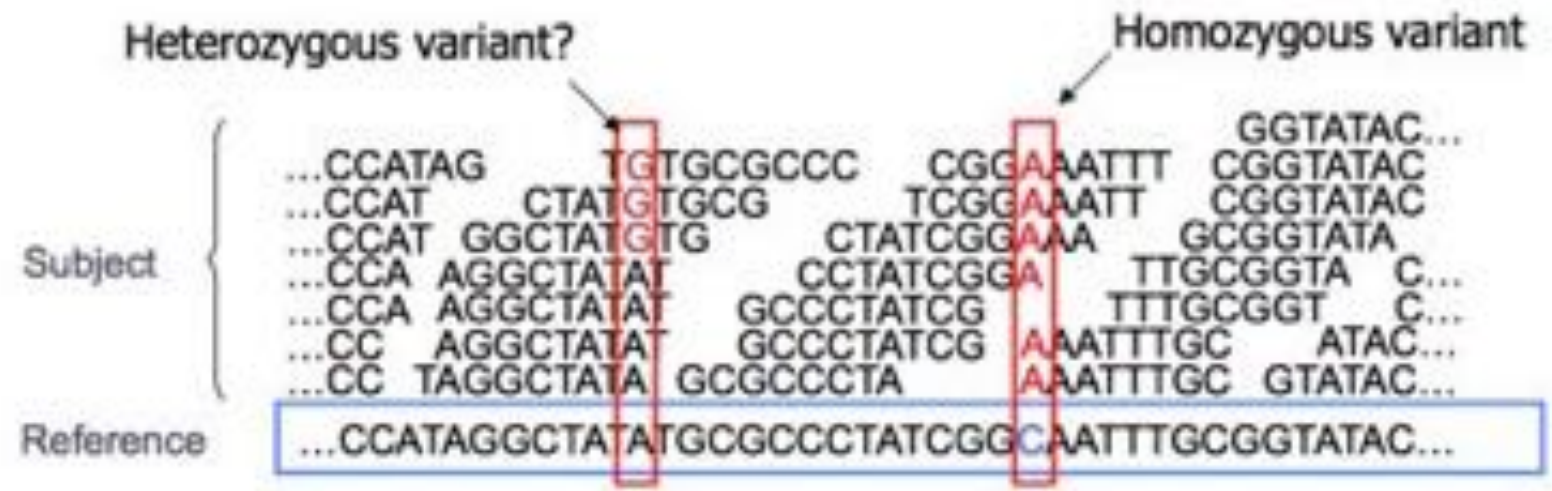


Plant Height

Drought Resistance

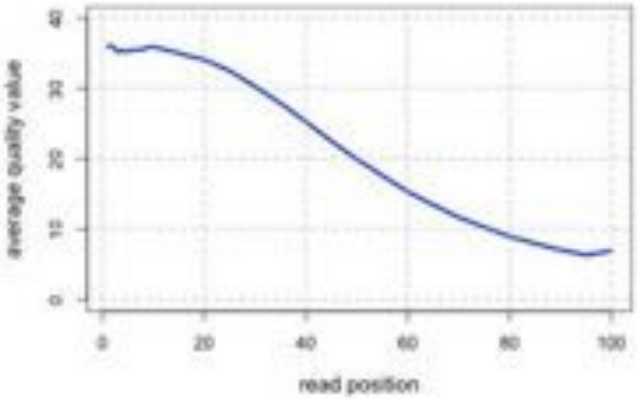
Biomass production

Genotyping



- Sequencing instruments make mistakes
  - Quality of read decreases over the read length

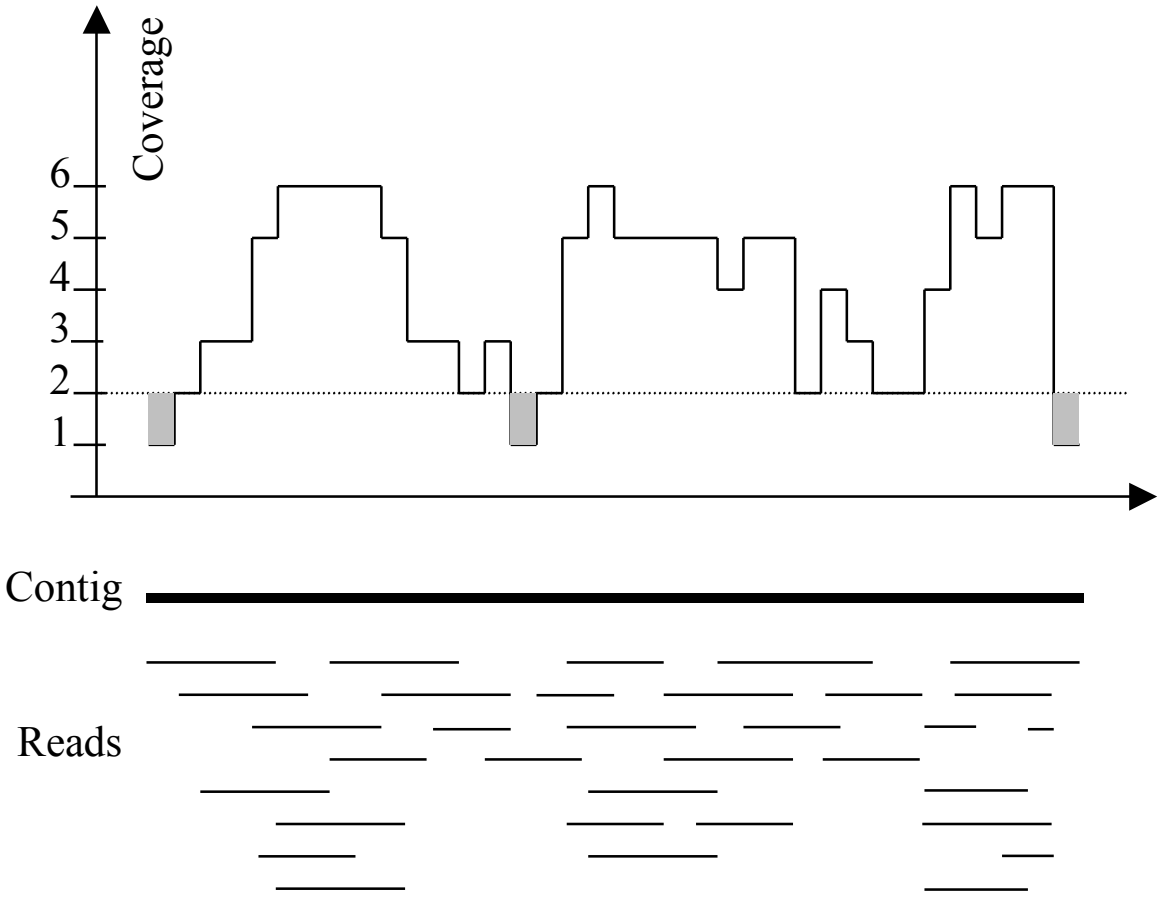
$$Q_{\text{sanger}} = -10 \log_{10} p$$



- A single read differing from the reference is probably just an error, but it becomes more likely to be real as we see it multiple times
  - Often framed as a Bayesian problem of more likely to be a real variant or chance occurrence of N errors
  - Accuracy improves with deeper coverage

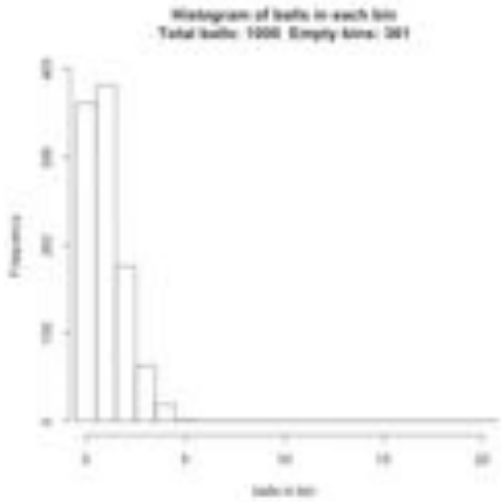
Coverage

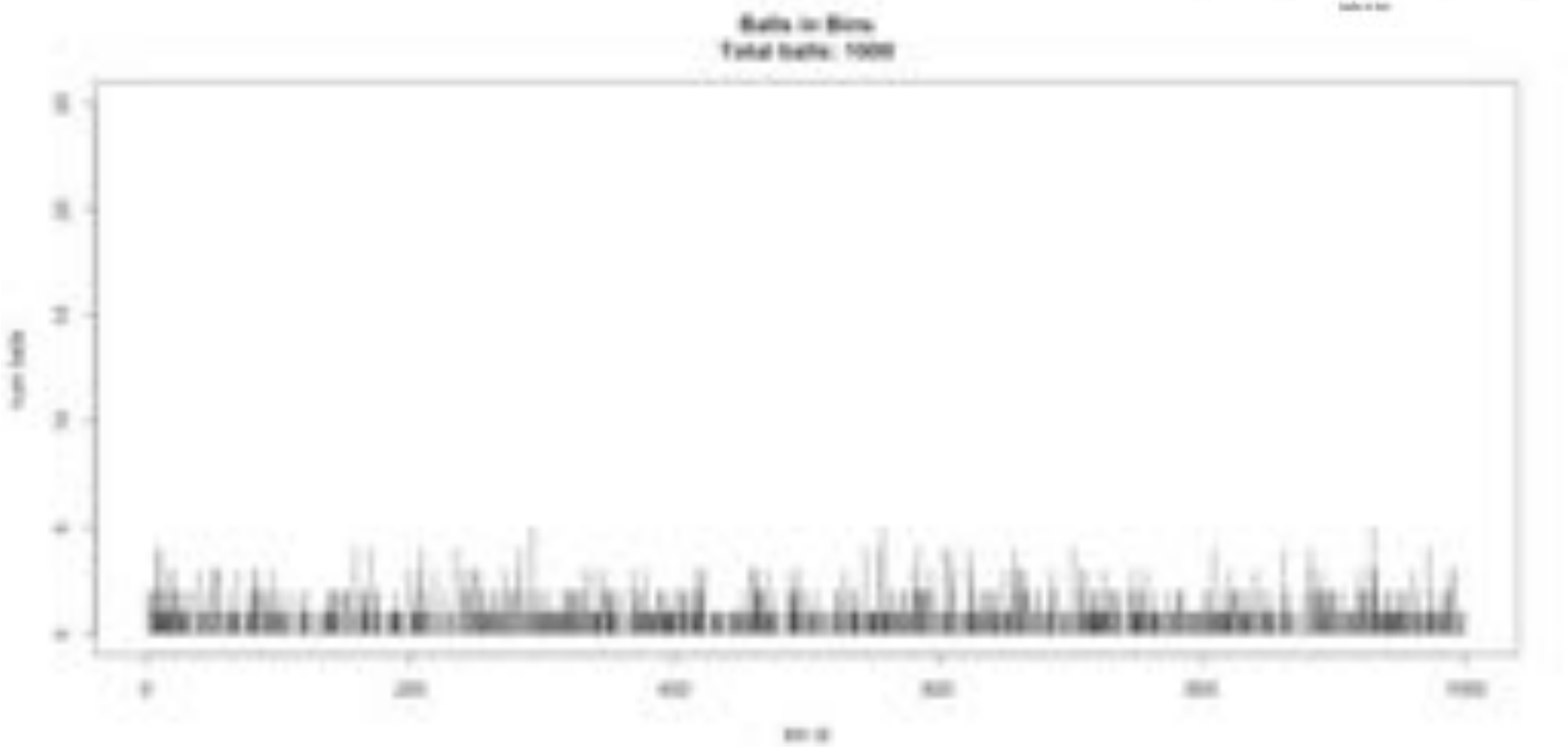
Typical contig coverage



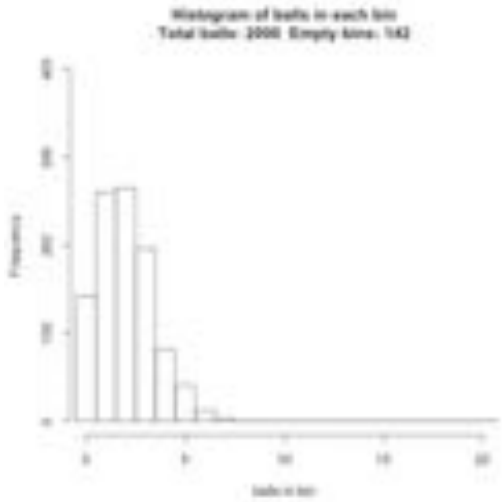
Imagine raindrops on a sidewalk

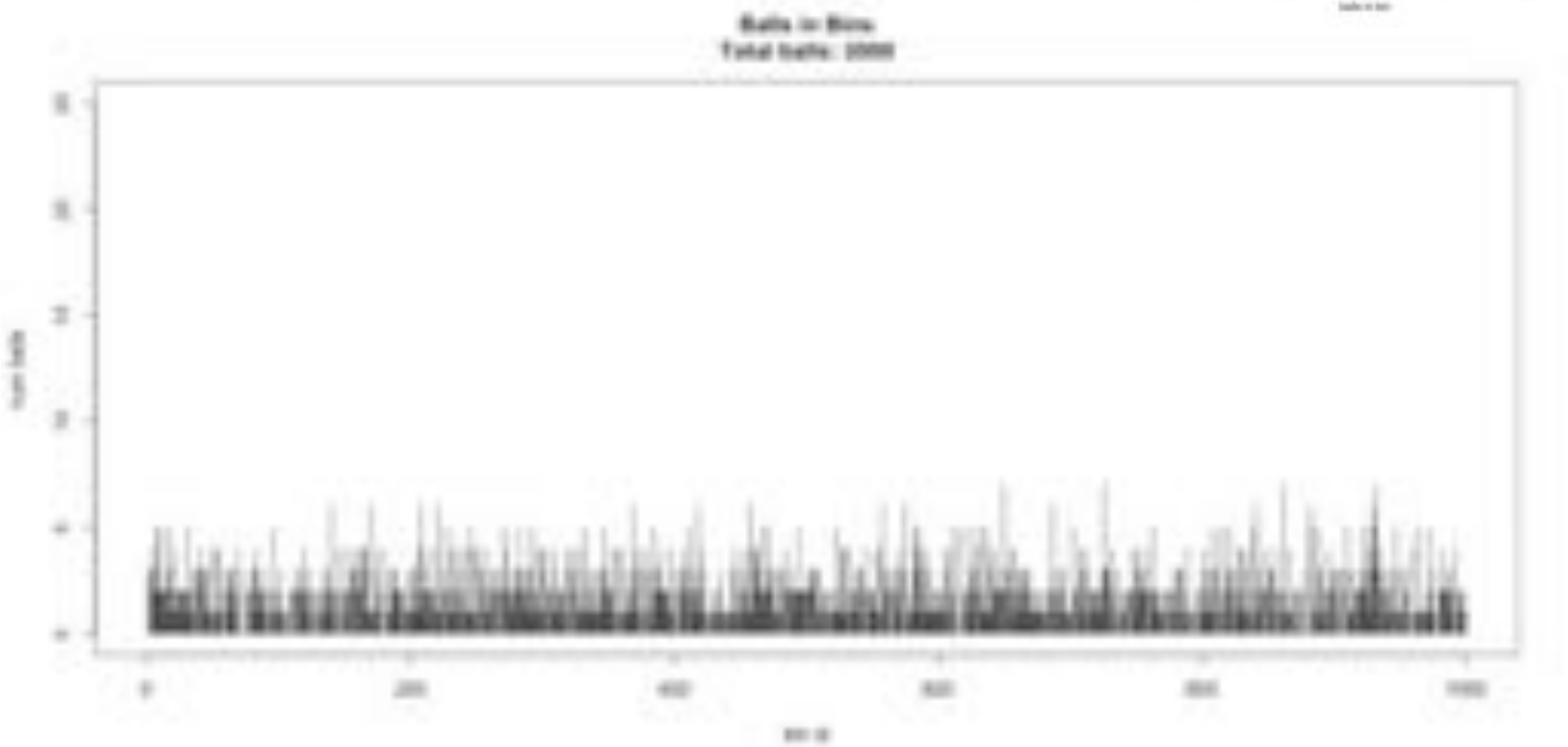
1x Sequencing



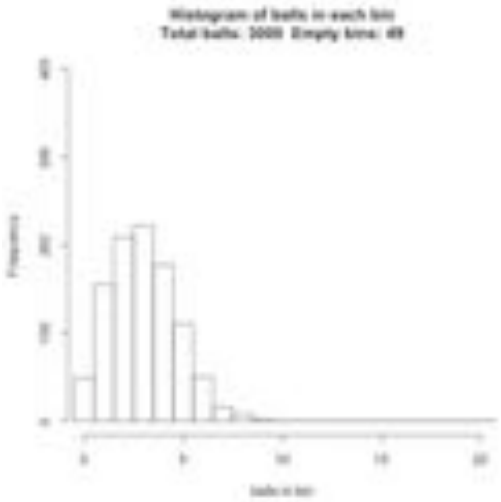


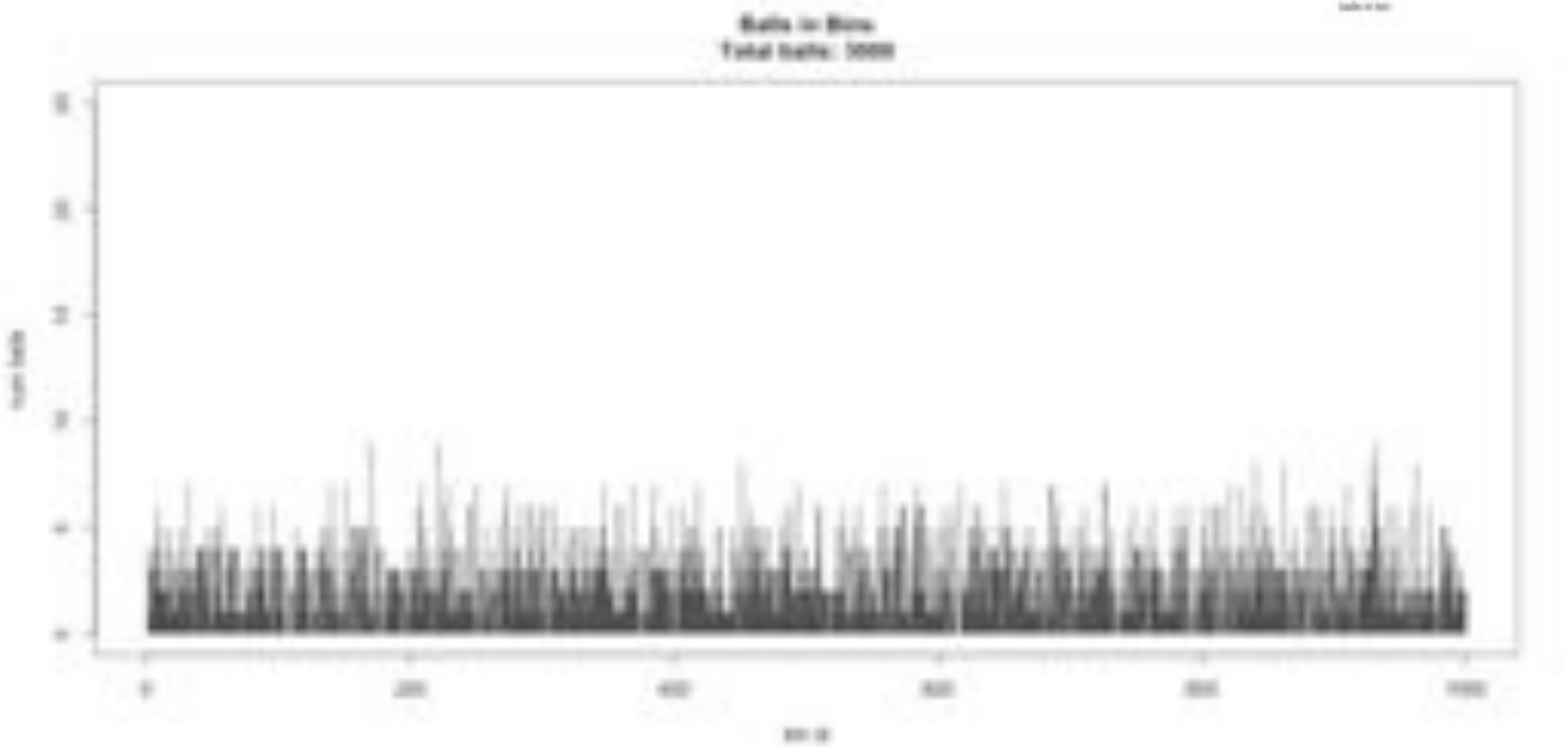
2x Sequencing



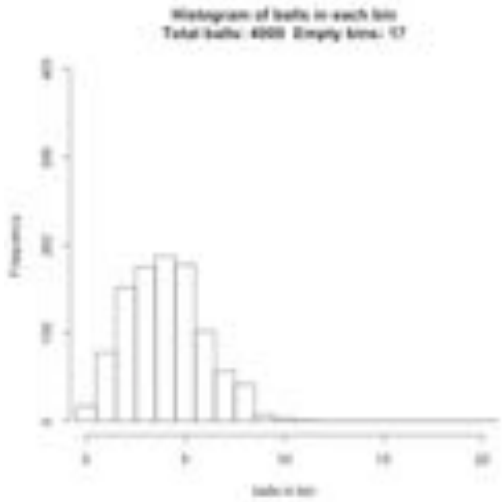


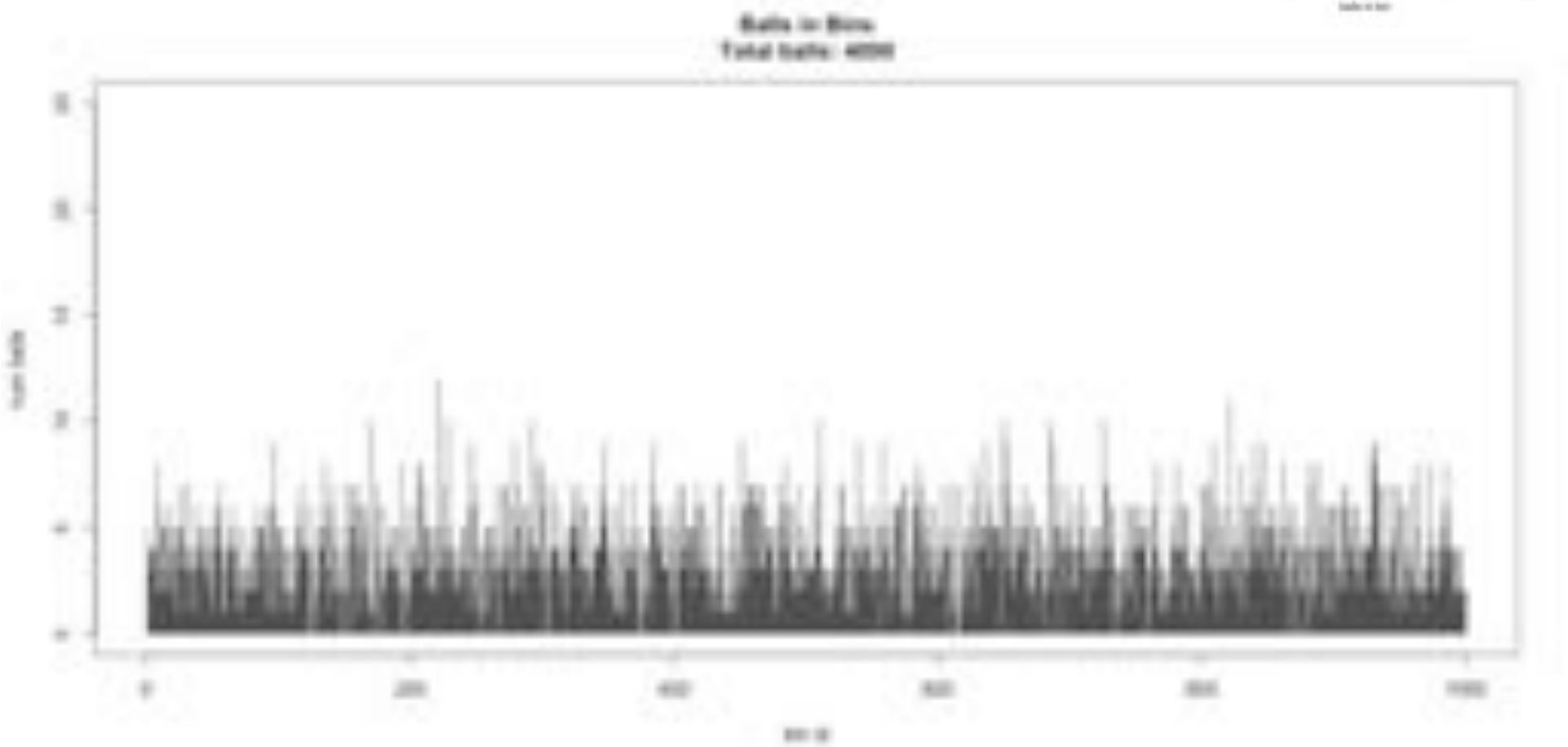
3x Sequencing



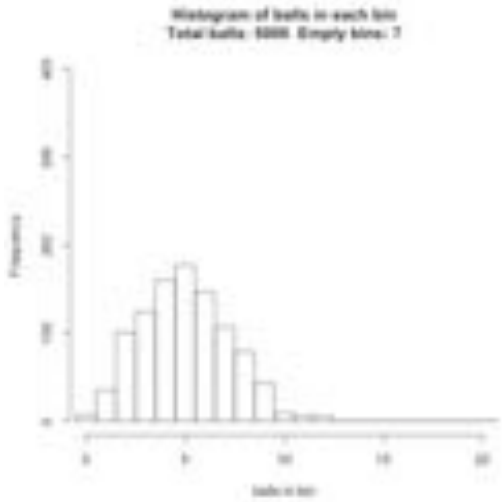


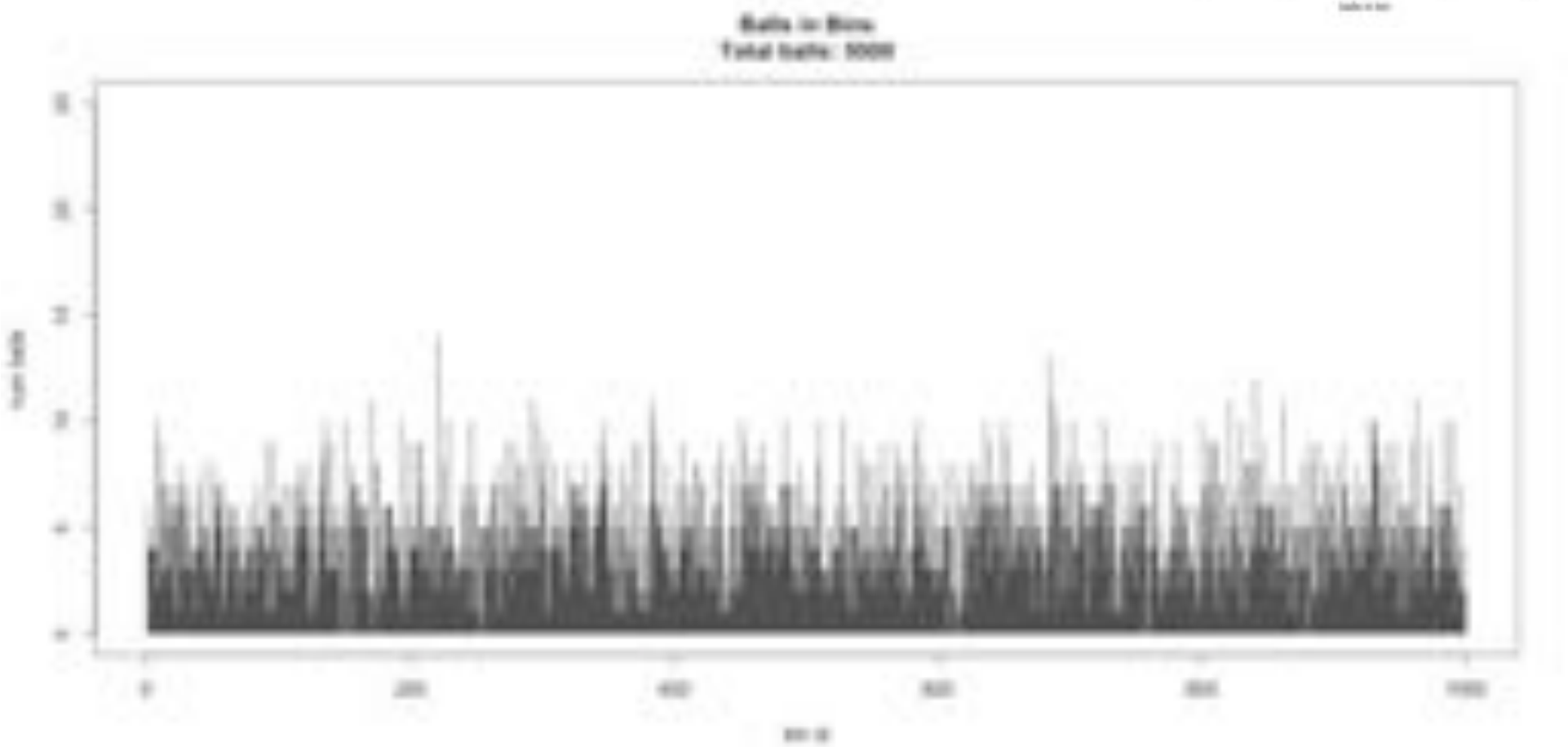
4x Sequencing



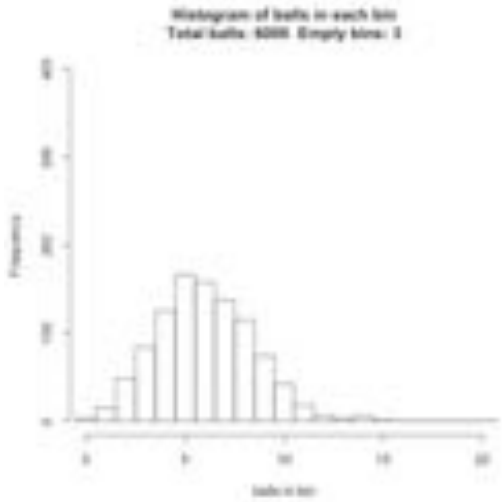


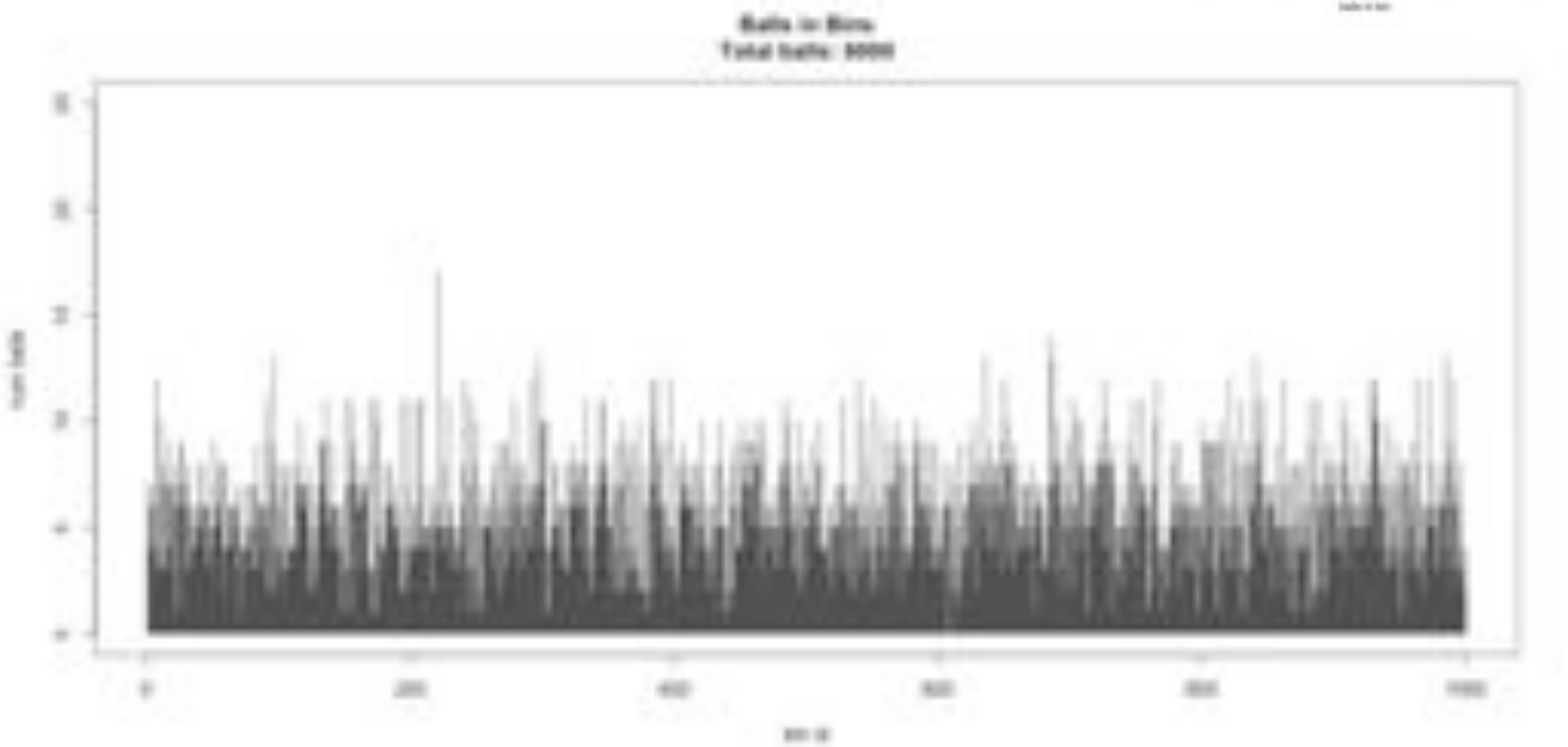
5x Sequencing



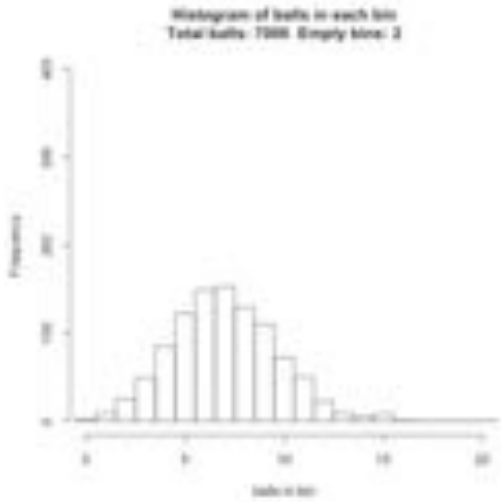


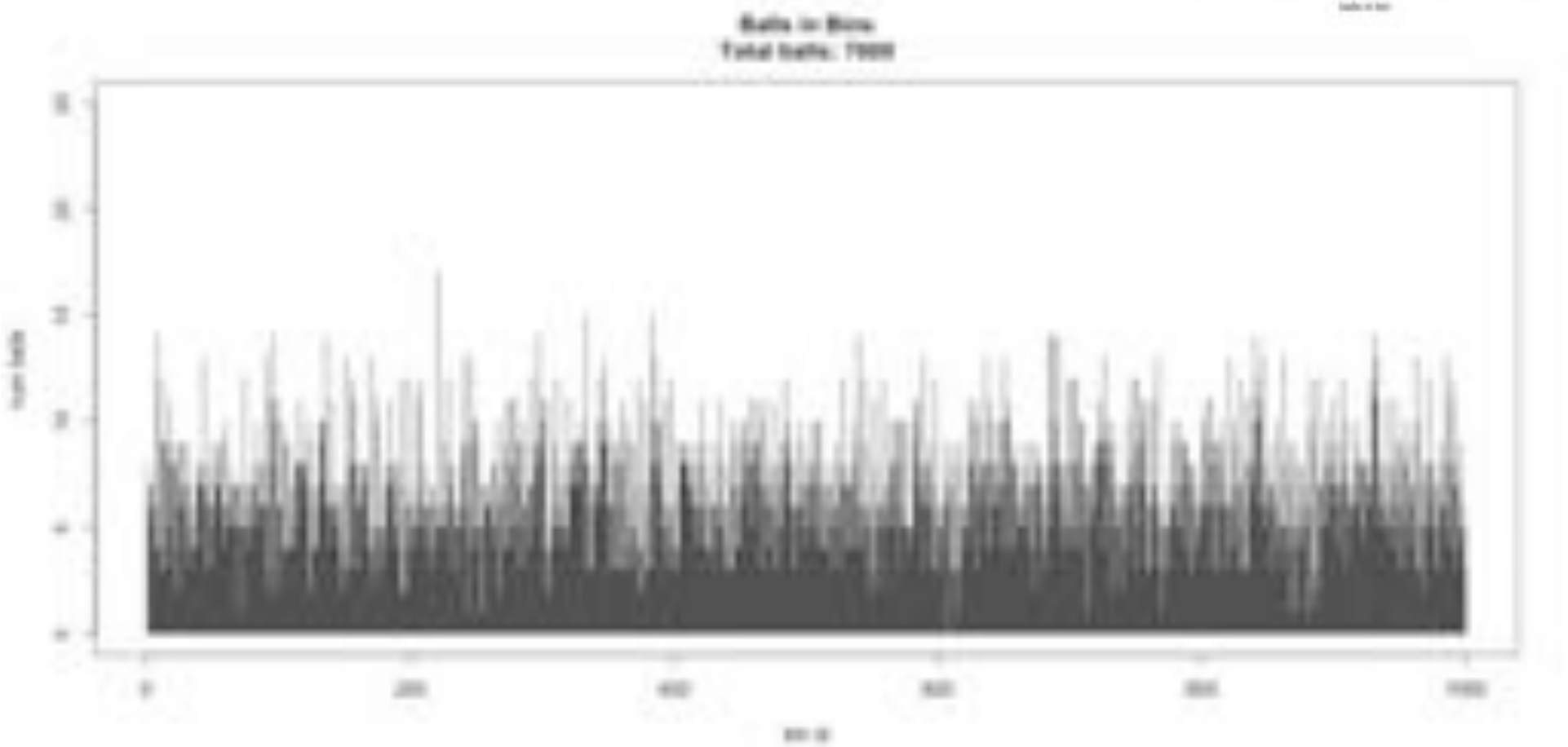
6x Sequencing



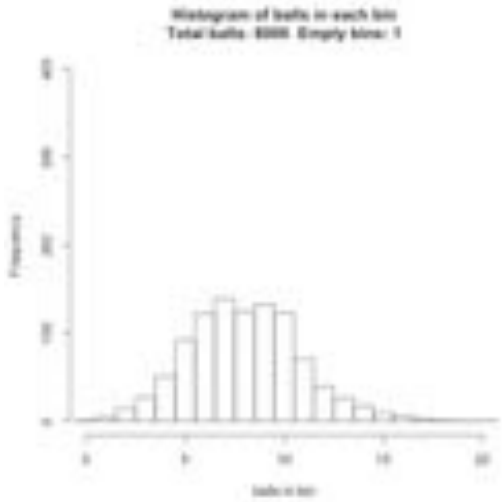


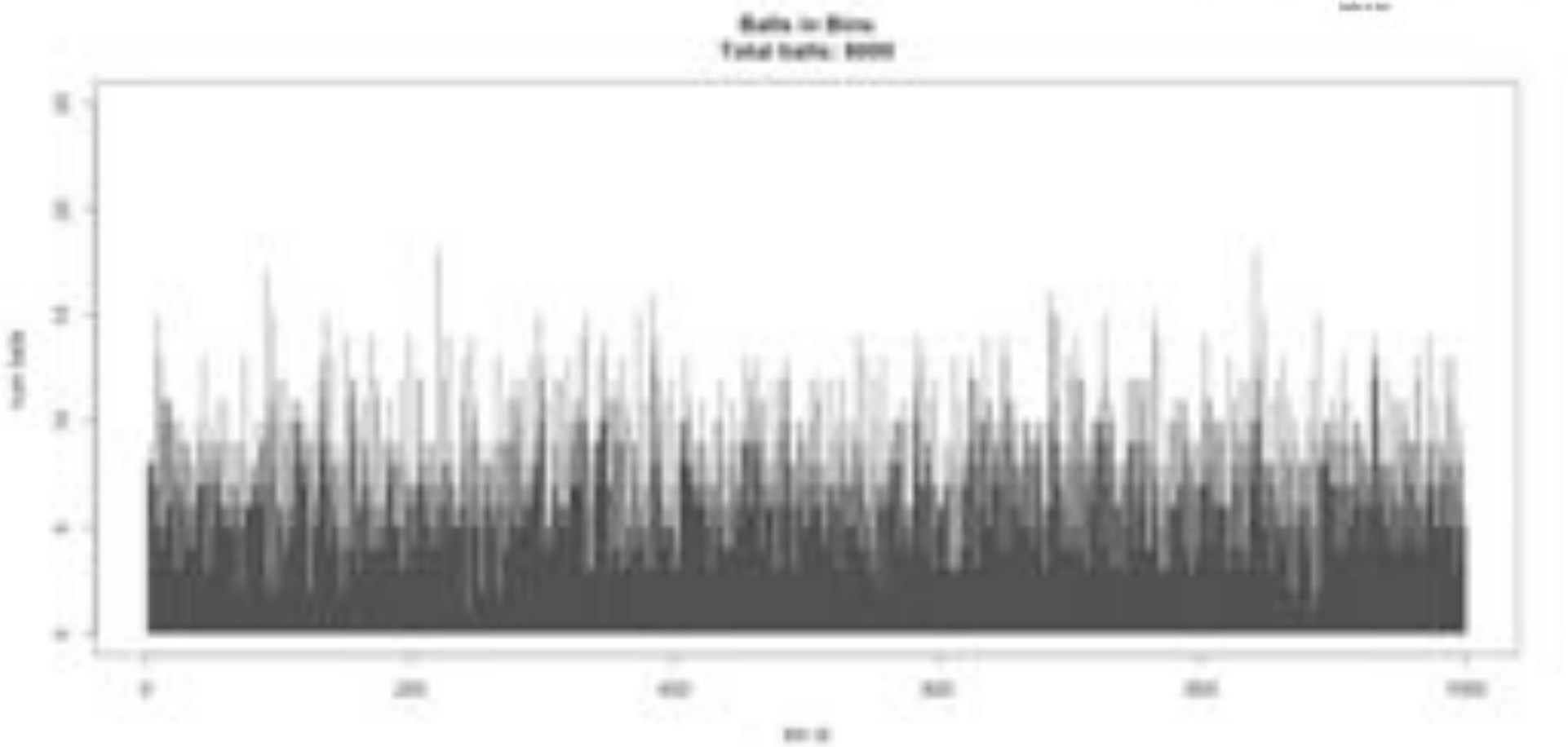
7x Sequencing



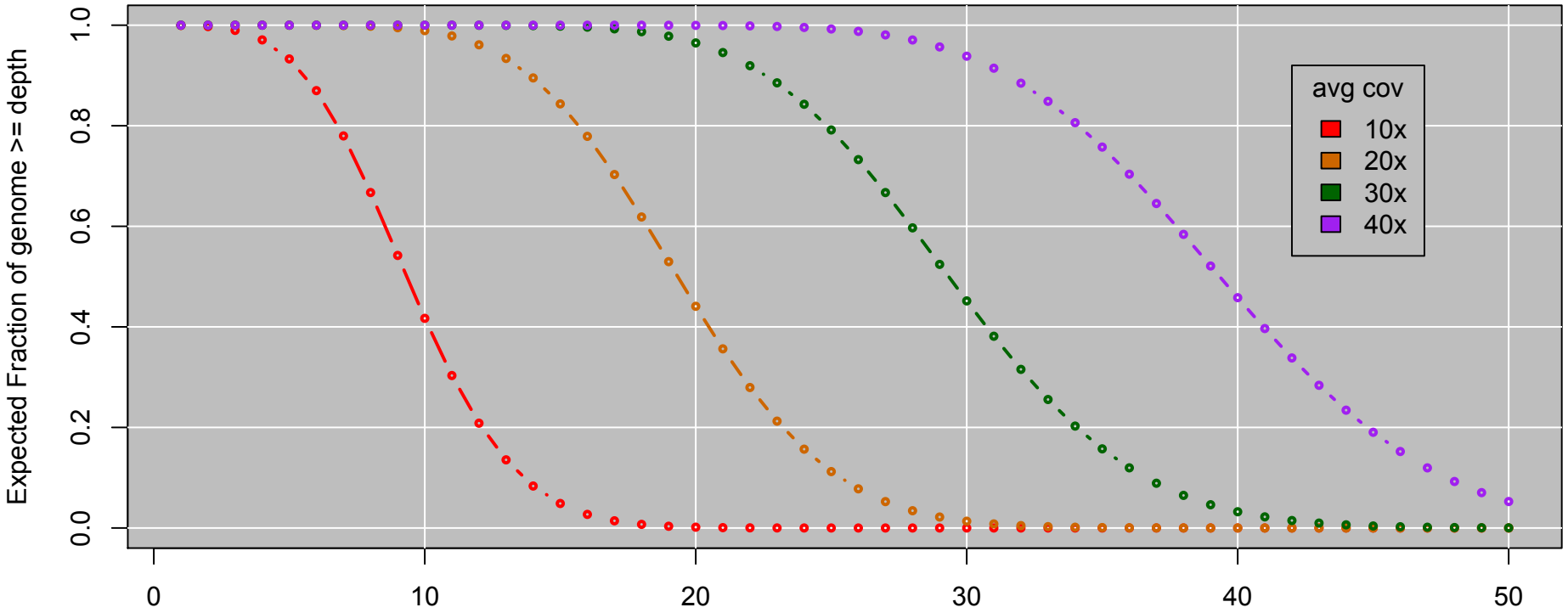


8x Sequencing





Genome Coverage Distribution



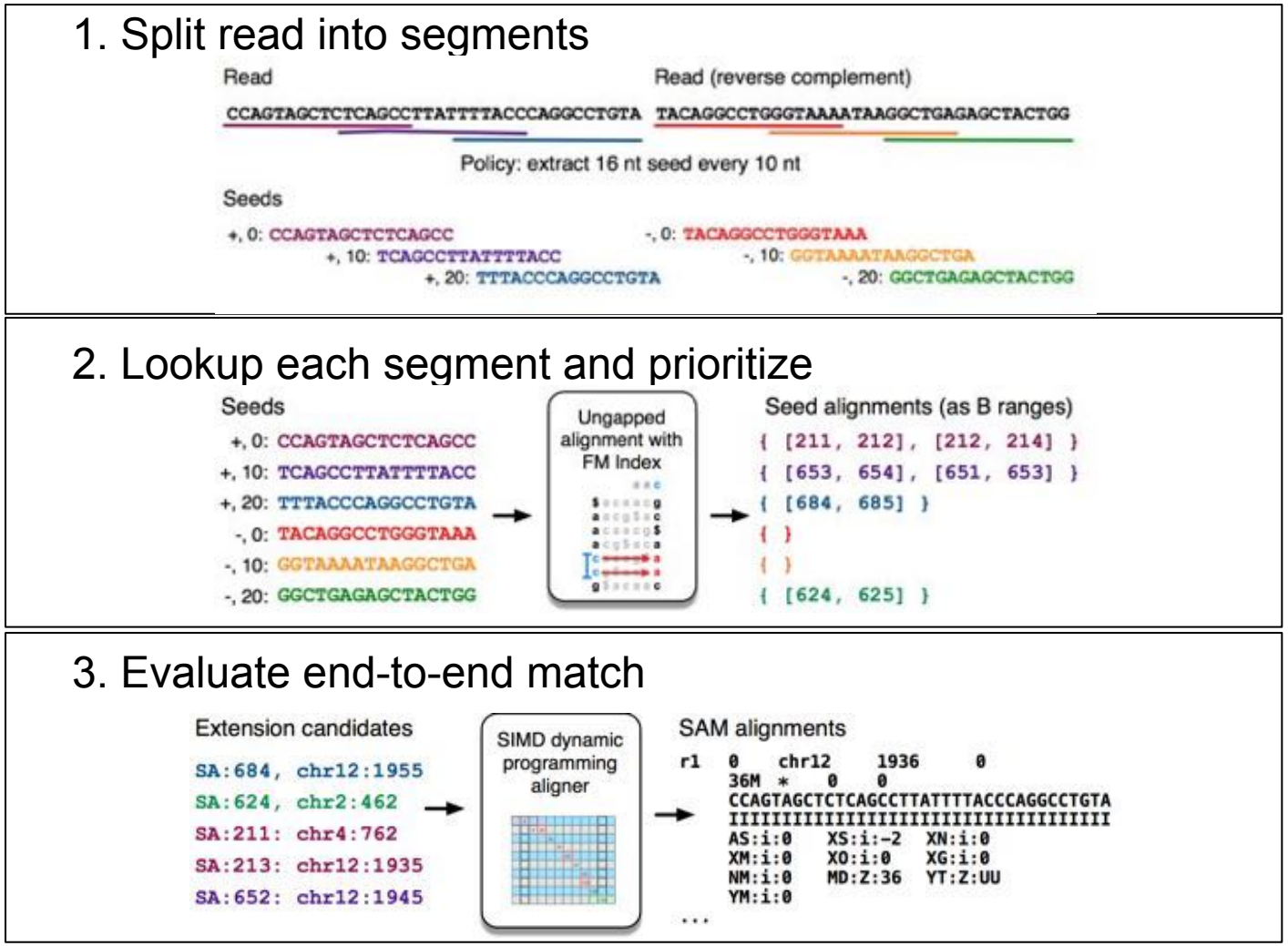
Expect Poisson distribution on depth

Standard Deviation = sqrt(cov)

This is the mathematically model => reality may be much worse

Double your coverage for diploid genomes

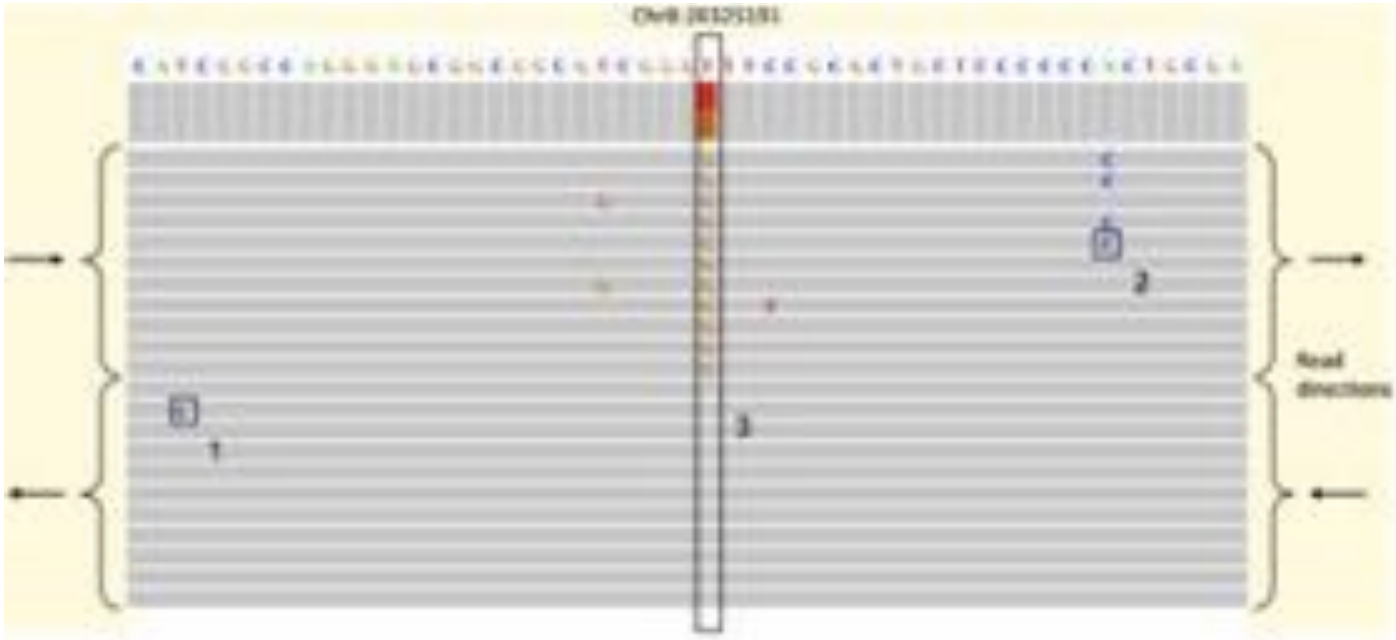
Bowtie2 Overview



Fast gapped-read alignment with Bowtie 2.
Langmead B, Salzberg S. Nature Methods. 2012, 9:357-359.

SNP calling

Beware of (Systematic) Errors



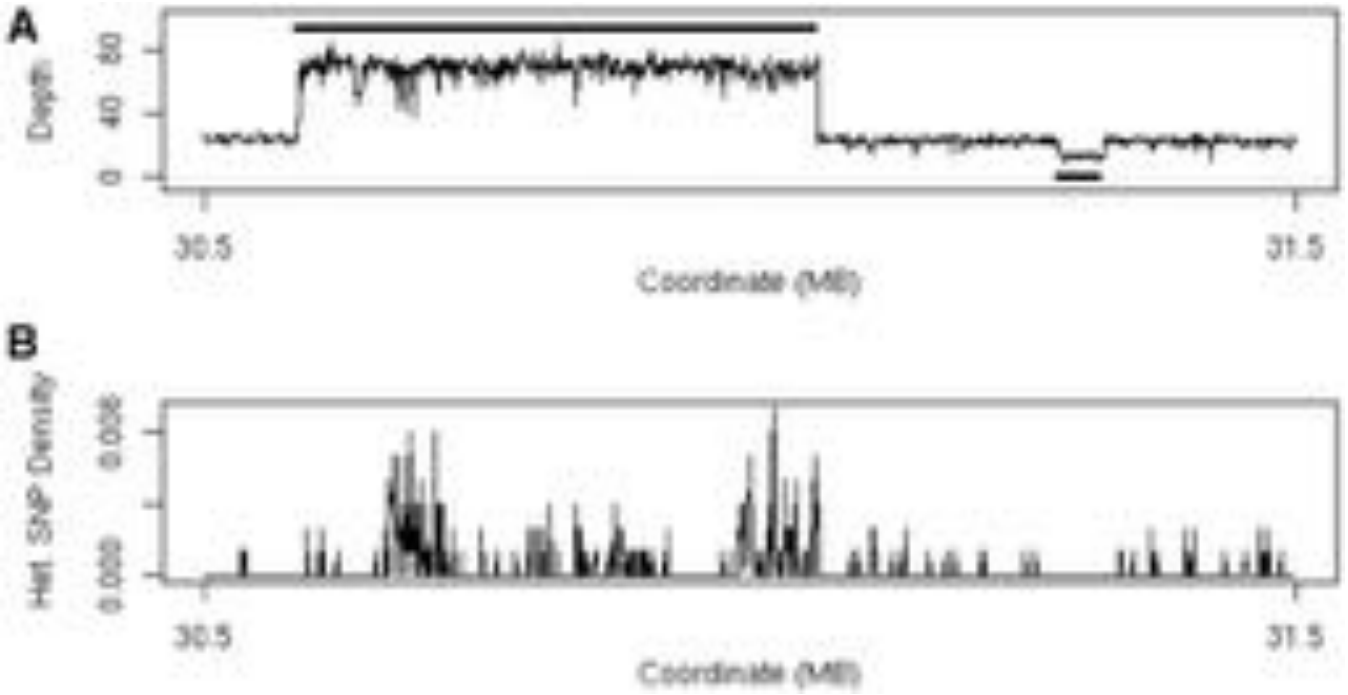
- Distinguishing SNPs from sequencing error typically a likelihood test of the coverage
  - Probability of seeing the data from a heterozygous SNP versus from sequencing error
  - However, some sequencing errors are systematic!

Identification and correction of systematic error in high-throughput sequence data
Meacham et al. (2011) *BMC Bioinformatics*. 12:451

A closer look at RNA editing.
Lior Pachter (2012) *Nature Biotechnology*. 30:246-247

CNV calling

Beware of (Systematic) Errors



(A) Plot of sequencing depth across a one megabase region of A/J chromosome 17 clearly shows both a region of 3-fold increased copy number (30.6–31.1 Mb) and a region of decreased copy number (at 31.3 Mb).

Simpson J T et al. Bioinformatics 2010;26:565-567

- Identify CNVs through increased depth of coverage & increased heterozygosity
  - Segment coverage levels into discrete steps
  - Be careful of GC biases and mapping biases of repeats

KBASE
predictive biology
DOE Systems Biology Knowledgebase

Agenda

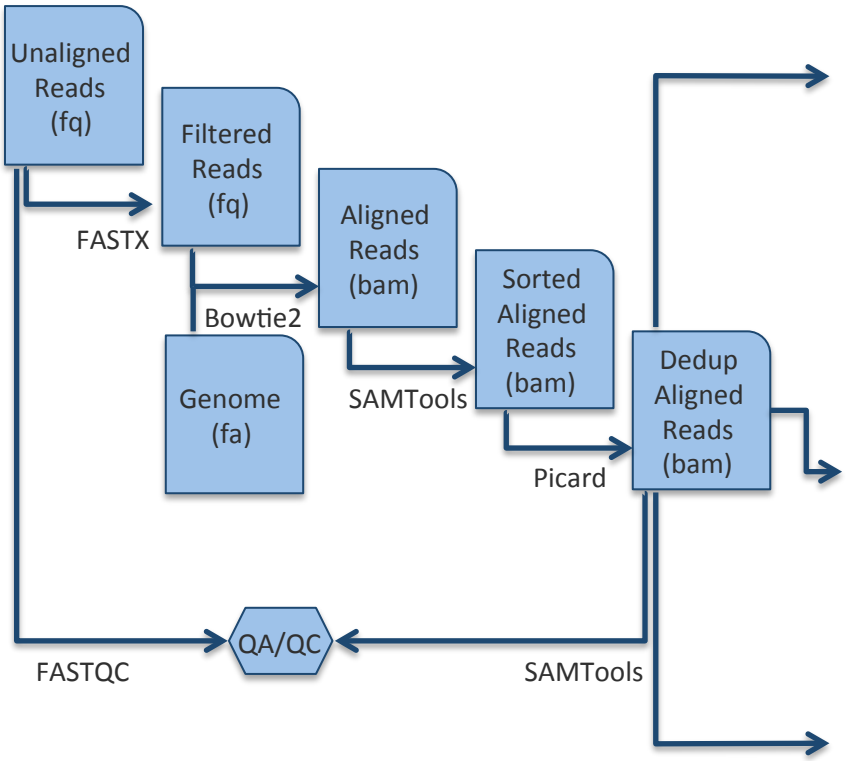
1. Introduction to KBase
2. Resequencing and variation calling theory
3. KBase services for variation calling
4. Live Demo
5. Additional Resources

KBASE predictive biology
DOE Systems Biology Knowledgebase

Sequence to Discovery

Illumina HiSeq 2000
Sequencing by Synthesis

>60Gbp / day



Assays
Read QA/QC
Mapping Stats
SNVs / Indels
CNVs / SVs
RNA-seq
ChIP-seq
DNase-seq
FAIRE-seq
Methyl-seq
ChIA-PET
Hi-C
...

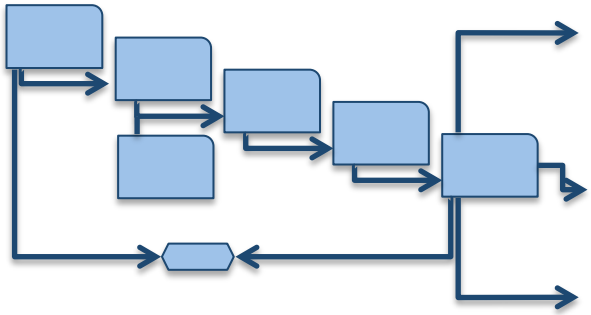
U.S. DEPARTMENT OF ENERGY

KBASE predictive biology
DOE Systems Biology Knowledgebase

Variation Services API 1.0

Genotyping API

- **Bowtie**: Launch alignment task with Bowtie
- **BWA**: Launch alignment task with BWA
- **SNPCalling**: Launch SNPcalling task with SAMTools
- **SortAlignments**: Launch task to sort by chromosome



Job API

- **ClusterStatus**: return basic status of cluster (jobs running, nodes available, etc)
- **JobStatus**: Given a JobID, returns current status
- **ListJobs**: List JobID running with a given username
- **KillJob**: Kills a given JobID

Data API

- **List**: List files in a directory
- **Fetch**: Fetch files from HDFS
- **Put**: Put files into HDFS
- **RM**: Delete files on HDFS
- **FetchBAM**: On-the-fly conversion to BAM
- **PutFastq**: Put reads into HDFS with conversion

Notes:

- All calls are authenticated with KBase username/password

KBASE predictive biology
DOE Systems Biology Knowledgebase

Reads to SNPs in Five Easy Steps

1. Identify reference genome

   $ all_entities_Genome -f scientific_name | grep -i 'Populus'

2. Upload Reads to KBase cloud

   $ jk_fs_put_pe populus.1.fq.gz populus.2.fq.gz populus

3. Align Reads with Bowtie2

   $ jk_compute_bowtie -in=populus.pe -org=populus -out=populus_align

4. Call SNPs with SAMTools

   $ jk_compute_samtools_snp -in=populus_align -org=populus -out=populus_snps

5. Merge and Download VCF files

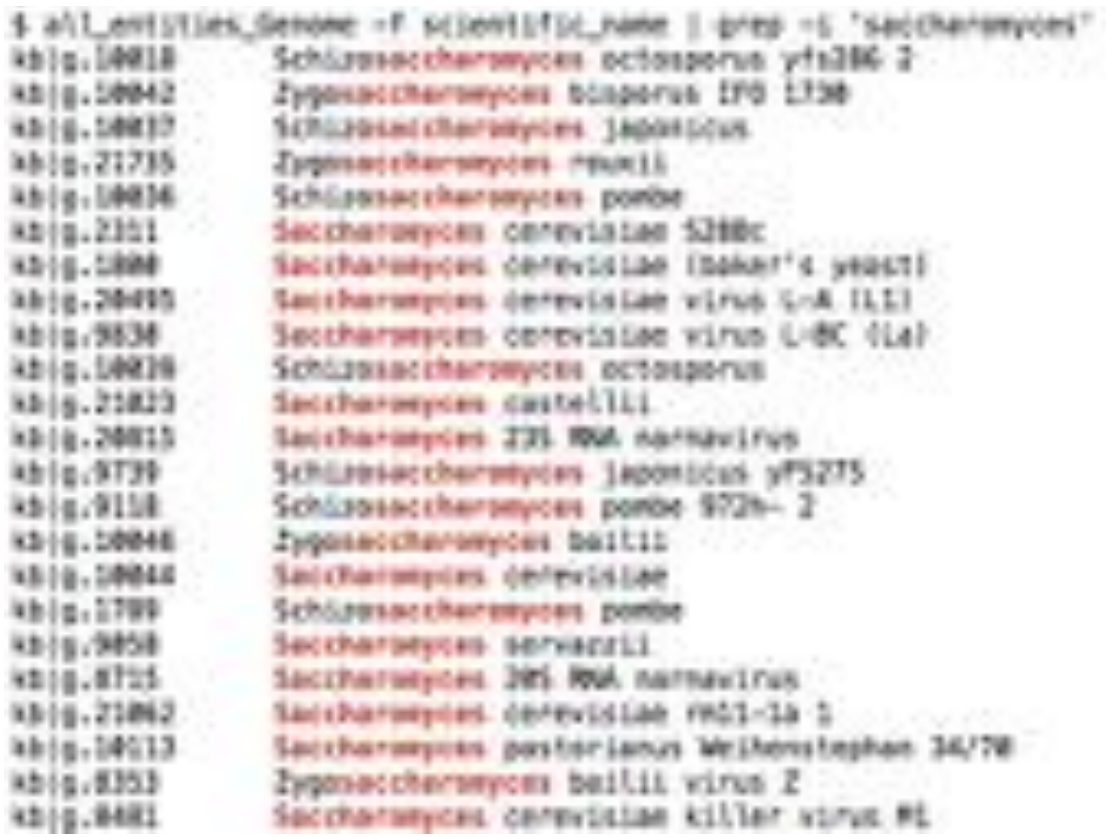
   $ jk_compute_vcf_merge -in=populus_snps --alignments=populus_align -out=populus.vcf

   $ jk_fs_get populus.vcf

KBASE predictive biology
DOE Systems Biology Knowledgebase

Identify a Reference Genome

```
$ all_entities_Genome -f scientific_name | grep -i 'populus'
kb|g.3907       Populus trichocarpa
```



Select the proper KBase ID

Identify reference genome

$ all_entities_Genome -f scientific_name | grep -i 'Populus'

KBASE
predictive biology
DOE Systems Biology Knowledgebase

Upload Reads to KBase Cloud

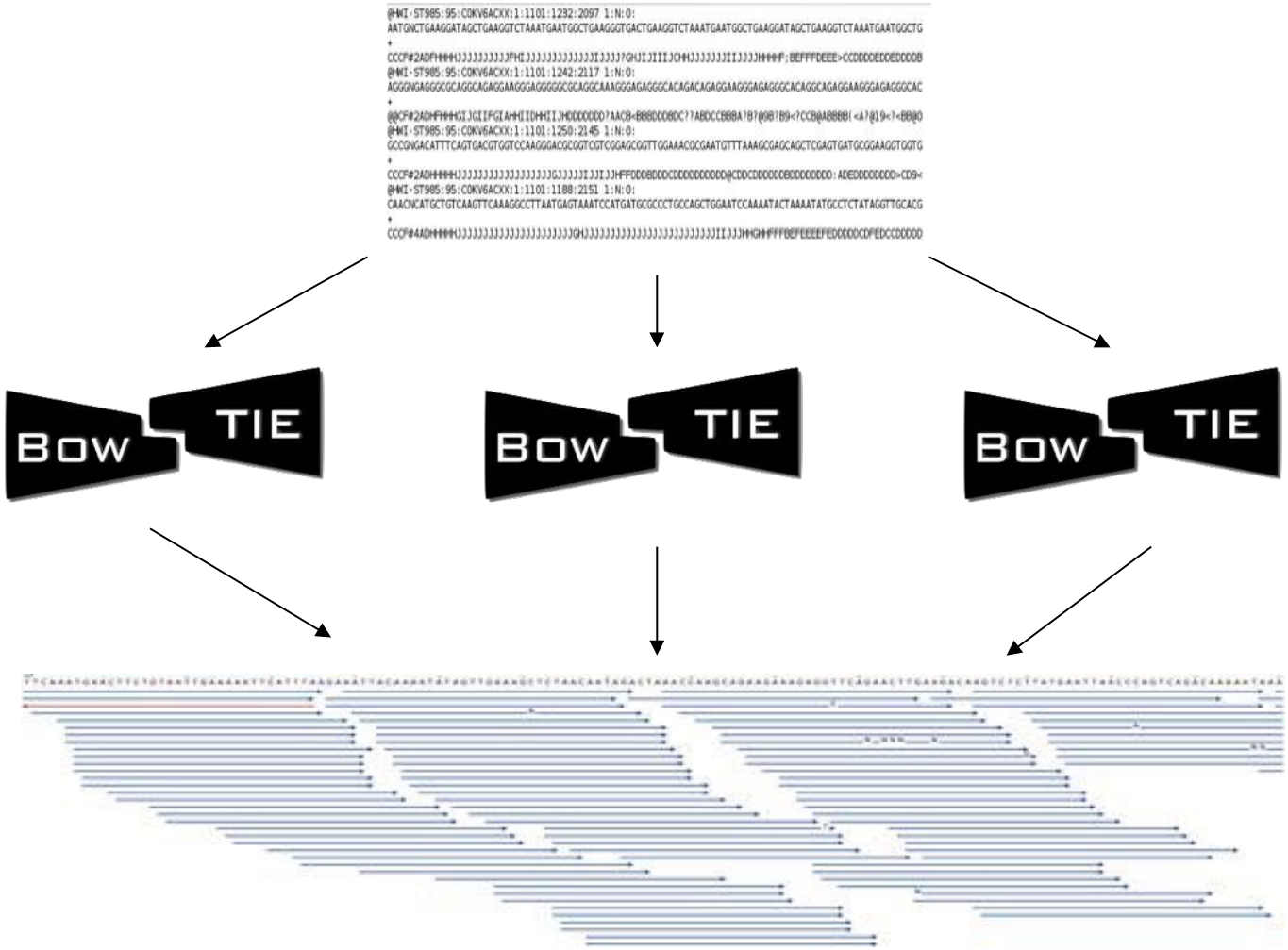
Upload Reads to KBase cloud

$ jk_fs_put_pe populus.1.fq.gz populus.2.fq.gz populus

KBASE predictive biology

DOE Systems Biology Knowledgebase

Align Reads with Bowtie2



Raw Fastq Reads

Bowtie2 Aligner

Alignments

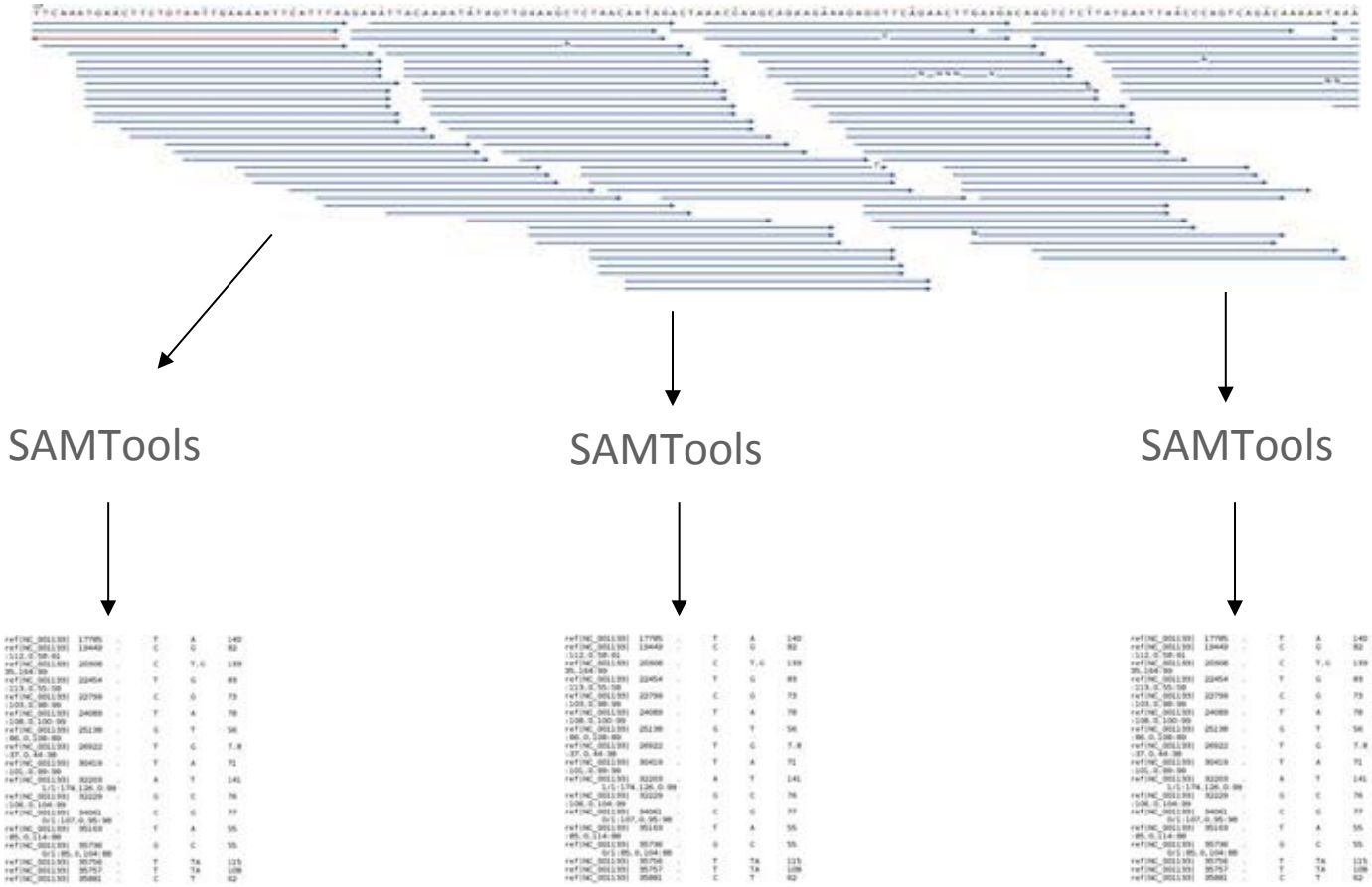
Align Reads with Bowtie2

$ jk_compute_bowtie -in=populus.pe -org='kb|g.3907' -out=populus_align

U.S. DEPARTMENT OF ENERGY

KBASE predictive biology
DOE Systems Biology Knowledgebase

Call SNPs with SAMTools



Alignments

Samtools Variant Detection

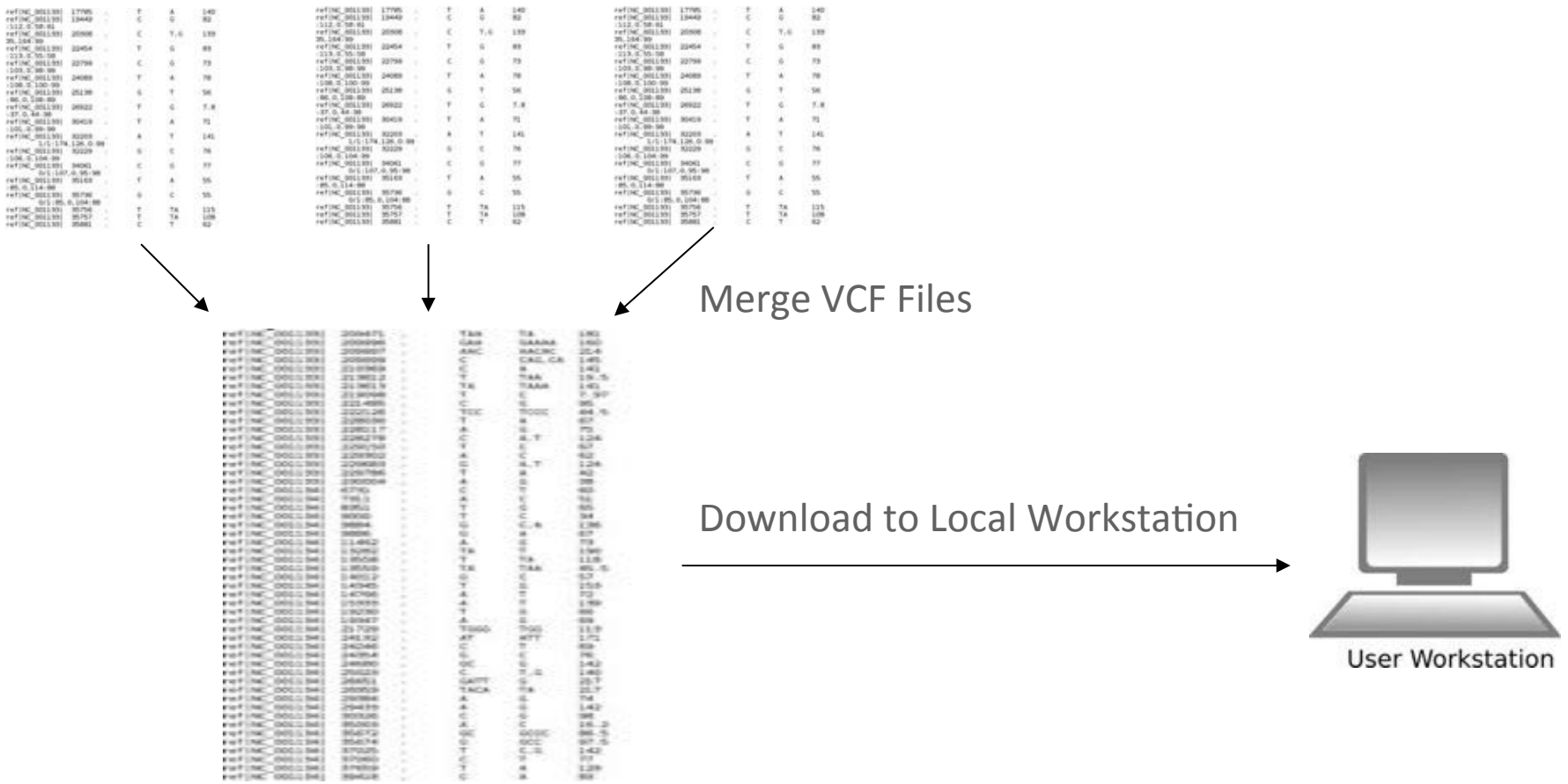
Called Variants (VCF)

Call SNPs with SAMTools

$ jk_compute_samtools_snp -in=populus_align –org='kb|g.3907' -out=populus_snps

KBASE predictive biology
DOE Systems Biology Knowledgebase

Merge and Download VCF Files



Merge and Download

$ jk_compute_vcf_merge -in=populus_snps –alignments=populus_align -out=populus.vcf

$ jk_fs_get populus.vcf

KBASE predictive biology
DOE Systems Biology Knowledgebase

Reads to SNPs in Five Easy Steps

1. Identify reference genome

   $ all_entities_Genome -f scientific_name | grep -i 'Populus'

2. Upload Reads to KBase cloud

   $ jk_fs_put_pe populus.1.fq.gz populus.2.fq.gz populus

3. Align Reads with Bowtie2

   $ jk_compute_bowtie -in=populus.pe -org=populus -out=populus_align

4. Call SNPs with SAMTools

   $ jk_compute_samtools_snp -in=populus_align -org=populus -out=populus_snps

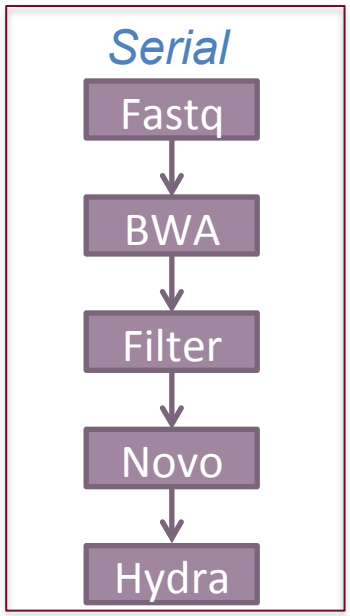
5. Merge and Download VCF files

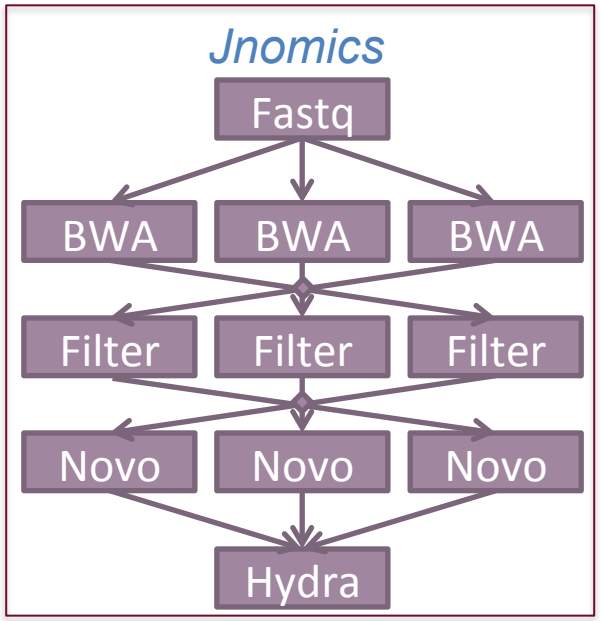
   $ jk_compute_vcf_merge -in=populus_snps --alignments=populus_align -out=populus.vcf

   $ jk_fs_get populus.vcf

KBASE predictive biology
DOE Systems Biology Knowledgebase

Jnomics: Cloud-scale genomics





openstack™ CLOUD SOFTWARE

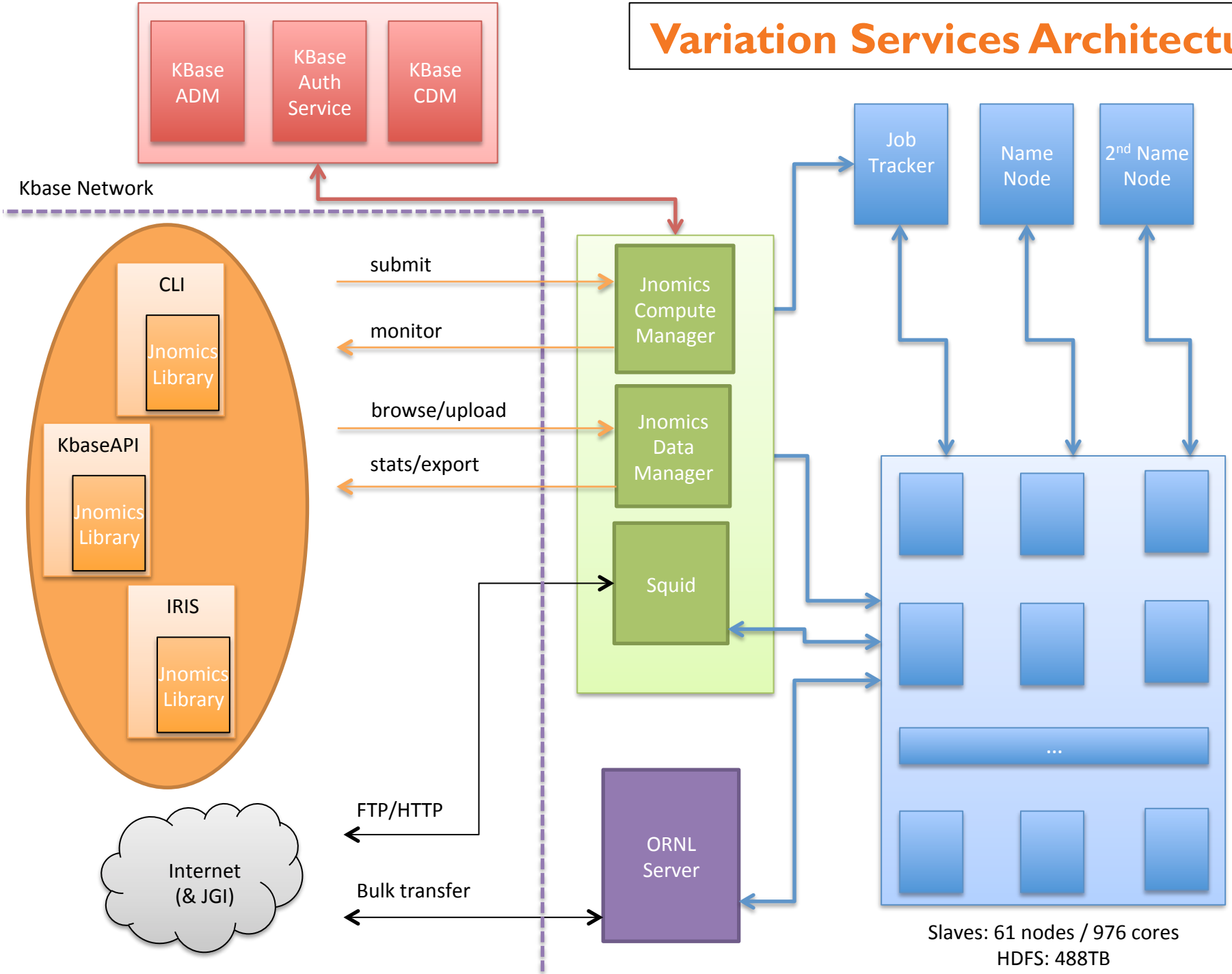
- Rapid parallel execution of data-intensive analysis
  - FASTX, BWA, Bowtie2, Novoalign, SAMTools, Hydra
  - Sorting, merging, filtering, selection, clustering, correlating
  - Supports BAM, SAM, BED, fastq

hadoop

Answering the demands of digital genomics
Titmus, MA, Gurtowski, J, Schatz, MC (2012) *Concurrency & Computation*

U.S. DEPARTMENT OF ENERGY

Variation Services Architecture



KBASE predictive biology
DOE Systems Biology Knowledgebase

Maize Test Run

Align & call SNPs from 35M 80bp (14Gbp) reads with maize genome (zmb73v2)
Identified 372k high confidence SNPs

	Serial	Multitcore	KBase Cloud
Config	1 core (1 node)	44 core (1 node)	118 cores (15 nodes)
Bowtie2	45 h*	1h 10m	23 m
Sort	2 hr	2 hr	N/A
Samtools	2 hr	2 hr	12 m
End-to-End Speedup	50h* 1x	5h 10m 9.6x	35 m 86x

*estimated time

KBASE predictive biology
DOE Systems Biology Knowledgebase

Maize Population Analysis

Align & call SNPs from 131 maize samples
1TB fastq / 408Gbp input data

	Serial	KBase cloud (small)	KBase Cloud (large)
Config	1 core (1 node)	210 cores (15 nodes)	854 cores (61 nodes)
Bowtie2	1311 hr*	19.5 hr	5 hr
Sort	58 hr*	N/A	N/A
Samtools	58 hr*	3.5 hr	1.5 hr
End-to-End Speedup	1427 hr* 1x	23 hr 62x	6.5 hr 219x

*estimated time

KBASE
predictive biology
DOE Systems Biology Knowledgebase

Agenda

1. Introduction to KBase
2. Resequencing and variation calling theory
3. KBase services for variation calling
4. Live Demo
5. Additional Resources

Demo

Online Demo

1. Browse to KBase website: http://kbase.us/

2. Sign up for KBase account: https://gologin.kbase.us/SignUp

3. Download KBase DMG: http://kbase.us/for-users/get-started/
   Or use IRIS: http://kbase.us/services/docs/invocation/Iris/

4. Variation Services Tutorial:
   http://kbase.us/for-users/tutorials/analyzing-data/variation-service/

5. Summarize mutations:
   $ cat yeast.vcf
   $ grep -v '^#' yeast.vcf | cut -f1 | sort | uniq -c
   $ grep -v '^#' yeast.vcf | cut -f 4,5 | sort | uniq -c | sort -nrk1 | head

KBASE
predictive biology
DOE Systems Biology Knowledgebase

Agenda

1. Introduction to KBase
2. Resequencing and variation calling theory
3. KBase services for variation calling
4. Live Demo
5. Additional Resources

KBASE predictive biology
DOE Systems Biology Knowledgebase

Additional Resources

Resource	URL
KBase	http://kbase.us/
Getting Started	http://kbase.us/for-users/user-home/
Variation Services	http://kbase.us/for-users/tutorials/analyzing-data/variation-service/
Bowtie2	http://bowtie-bio.sourceforge.net/bowtie2/index.shtml
BWA	http://bio-bwa.sourceforge.net/
SAMTools	http://samtools.sourceforge.net/
VCF Spec	http://www.1000genomes.org/wiki/Analysis/Variant%20Call%20Format/vcf-variant-call-format-version-40
SNPeff	http://snpeff.sourceforge.net/
KBase Contact	http://kbase.us/contact-us/
Survey	https://www.surveymonkey.com/s/KB-user-info

Questions?

Thank You!

http://schatzlab.cshl.edu
@mike_schatz / @DOEKBase