KBASE predictive biology

DOE Systems Biology Knowledgebase

KBase Variation Services

Overview and Demo

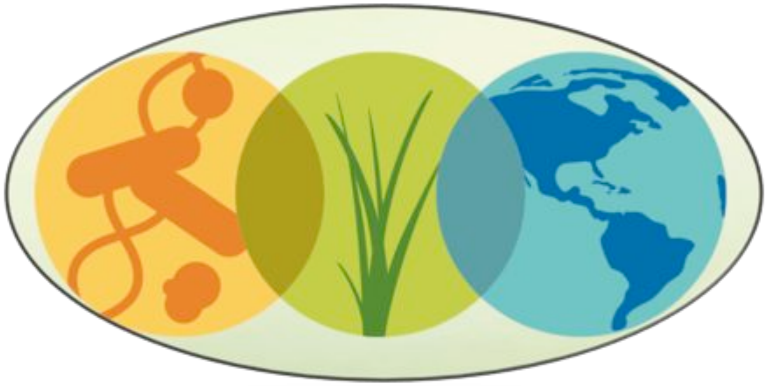
Michael Schatz, James Gurtowski
Cold Spring Harbor Laboratory

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



U.S. DEPARTMENT OF ENERGY

Agenda

1. Introduction to KBase

2. Resequencing and variation calling theory

3. KBase services for variation calling

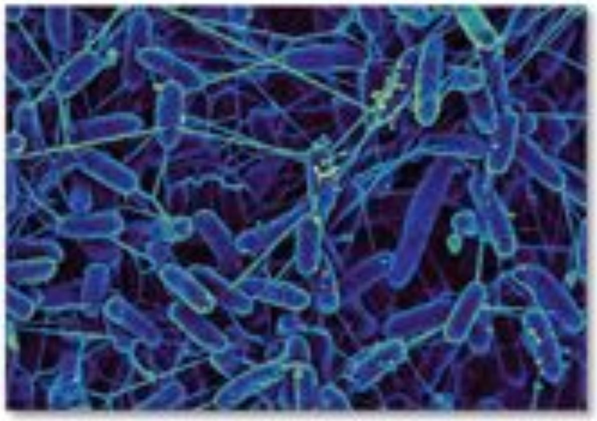
4. Live Demo

5. Additional Resources

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.



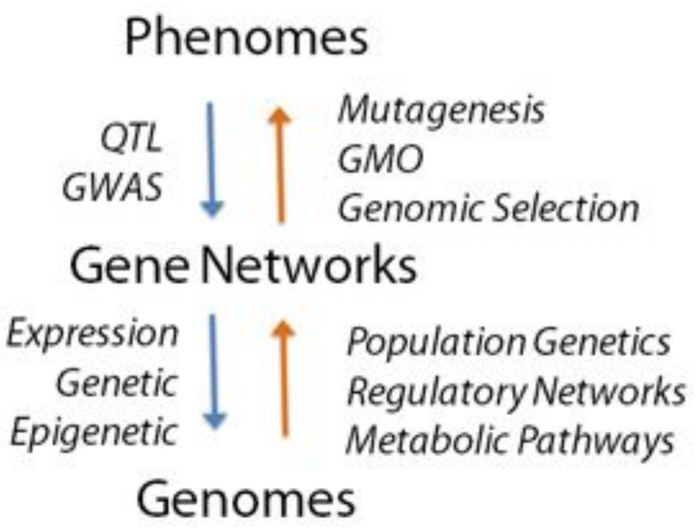
Microbes

Communities

Plants

KBase : Plants

Model development
Hypothesis testing
Knowledge Synthesis



Phenomes

QTL
GWAS

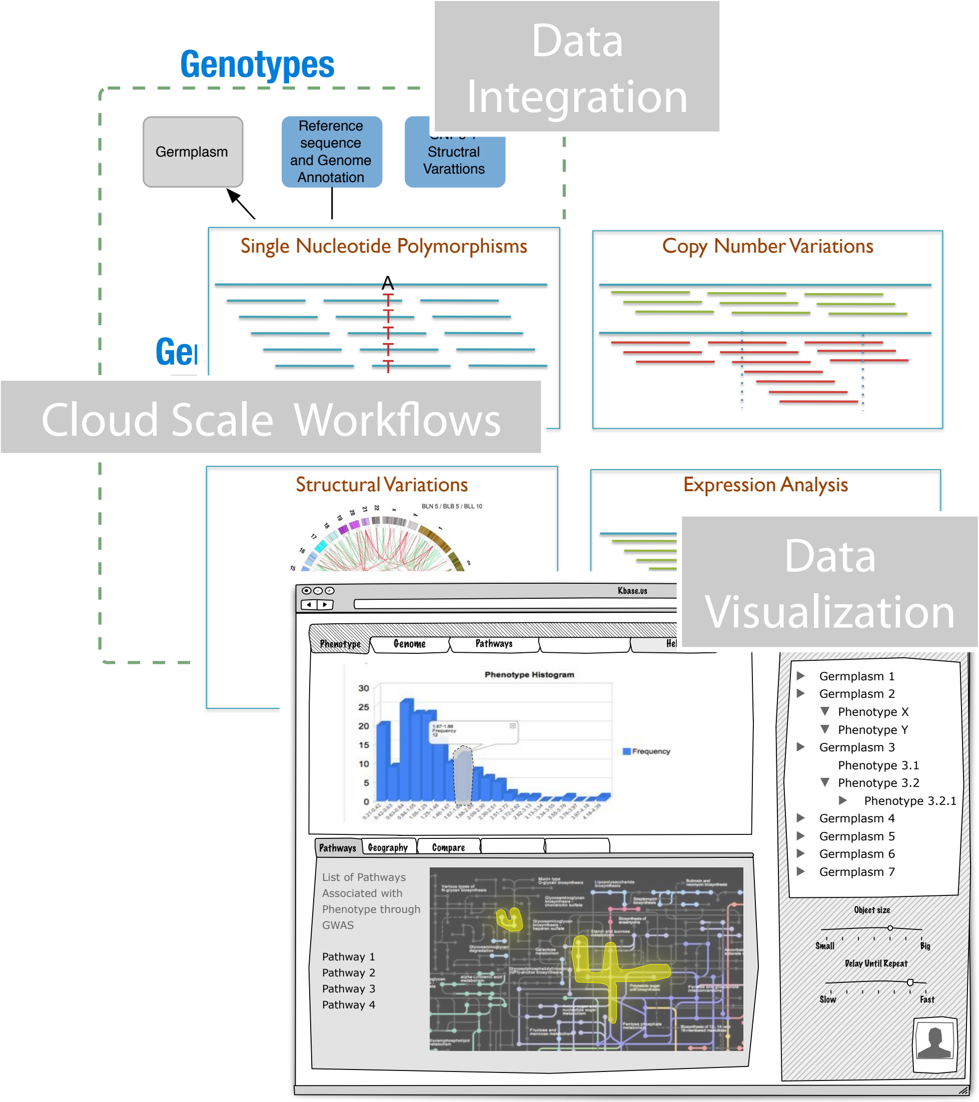
Mutagenesis
GMO
Genomic Selection

Gene Networks

Expression
Genetic
Epigenetic

Population Genetics
Regulatory Networks
Metabolic Pathways

Genomes

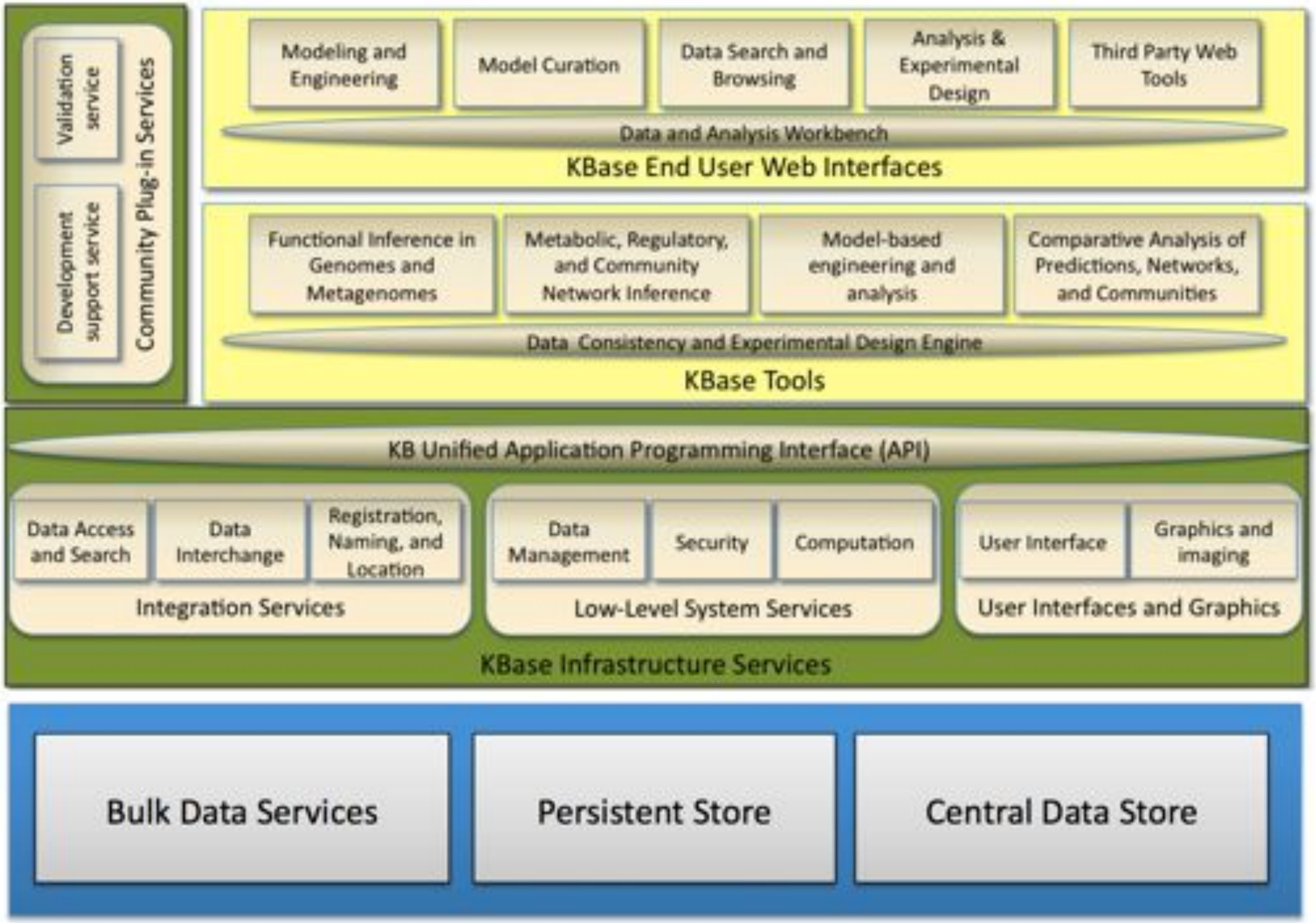


Data Integration

Cloud Scale Workflows

Data Visualization

KBase Infrastructure and Services



Validation service

Development support service

Community Plug-in Services

Modeling and Engineering

Model Curation

Data Search and Browsing

Analysis & Experimental Design

Third Party Web Tools

Data and Analysis Workbench

KBase End User Web Interfaces

Functional Inference in Genomes and Metagenomes

Metabolic, Regulatory, and Community Network Inference

Model-based engineering and analysis

Comparative Analysis of Predictions, Networks, and Communities

Data Consistency and Experimental Design Engine

KBase Tools

KB Unified Application Programming Interface (API)

Data Access and Search

Data Interchange

Registration, Naming, and Location

Integration Services

Data Management

Security

Computation

Low-Level System Services

User Interface

Graphics and imaging

User Interfaces and Graphics

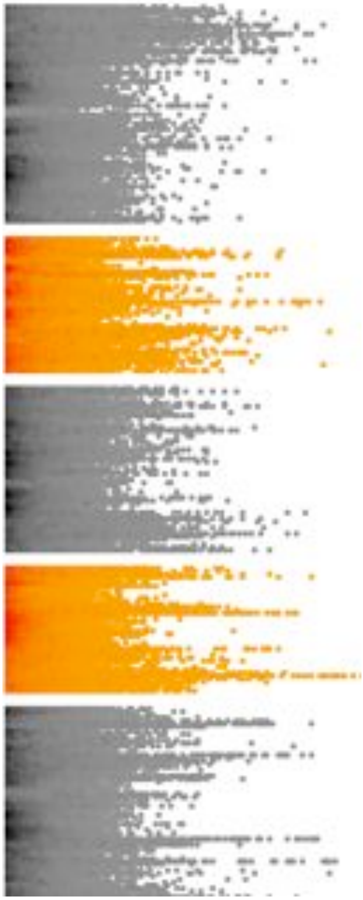
KBase Infrastructure Services

Bulk Data Services

Persistent Store

Central Data Store

Variation Services: Samples to Discoveries



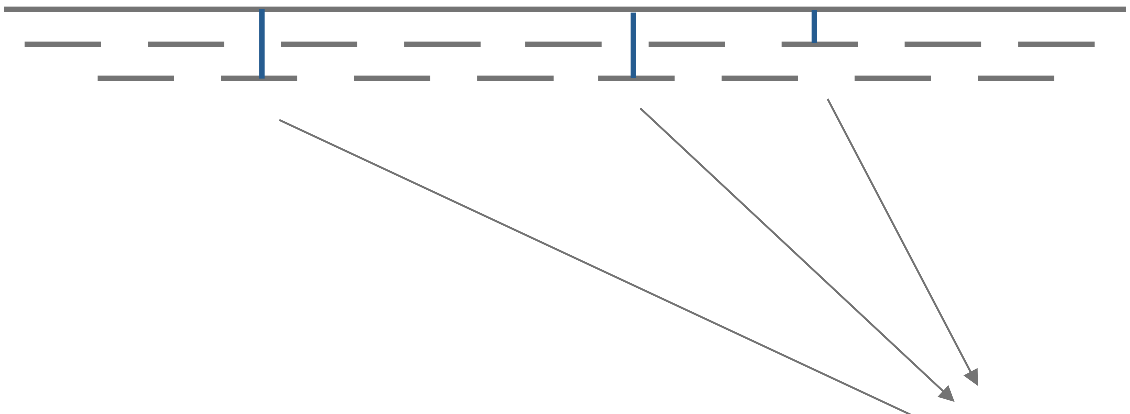
Powered by KBase

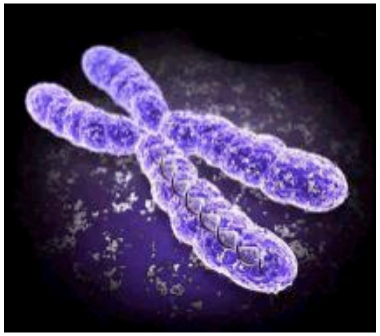
Agenda

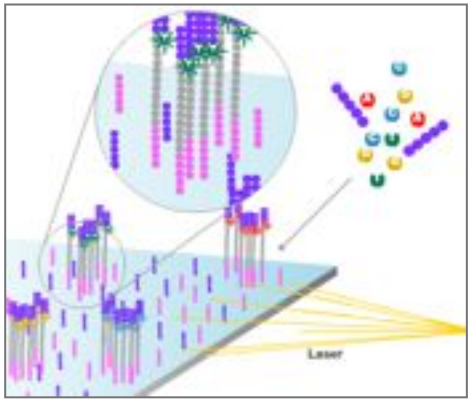
1. Introduction to KBase
2. Resequencing and variation calling theory
3. KBase services for variation calling
4. Live Demo
5. Additional Resources

Resequencing & Variations

How does your sample compare to the reference?





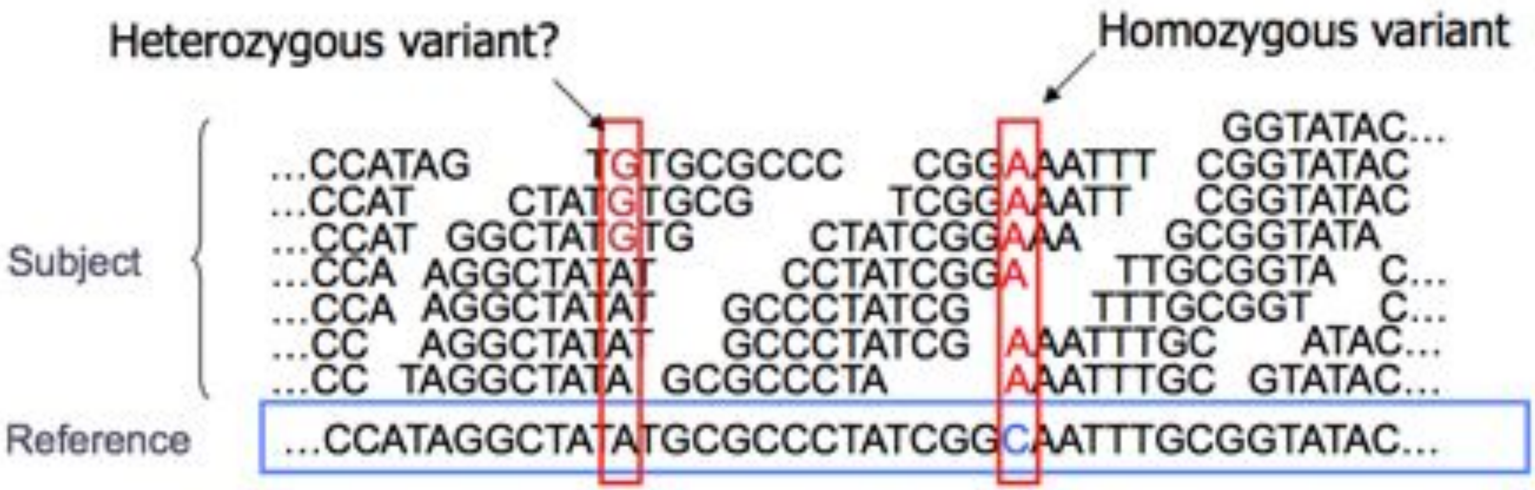


Plant Height

Drought Resistance

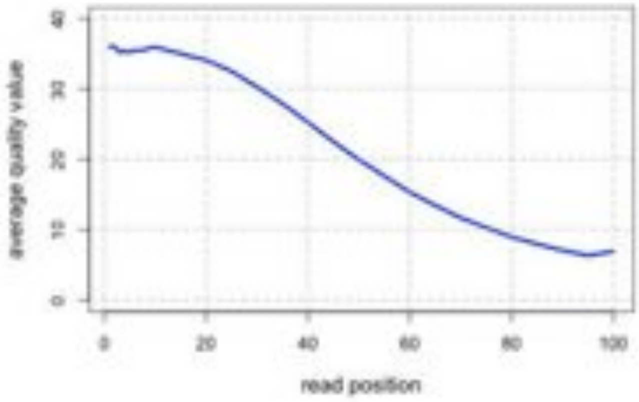
Biomass production

Genotyping


Heterozygous variant?
Homozygous variant
Subject
Reference
GGTATAC...
...CCATAG TGTGCGCCC CGGAAATTT CGGTATAC
...CCAT CTATGTGCG TCGGAAATT CGGTATAC
...CCAT GGCTATGTG CTATCGGAAA GCGGTATA
...CCA AGGCTATAT CCTATCGGA TTGCGGTA C...
...CCA AGGCTATAT GCCCTATCG TTTGCGGT C...
...CC AGGCTATAT GCCCTATCG AAATTTGC ATAC...
...CC TAGGCTATA GCGCCCTA AAATTTGC GTATAC...
...CCATAGGCTATATGCGCCCTATCGGCAATTTGCGGTATAC...

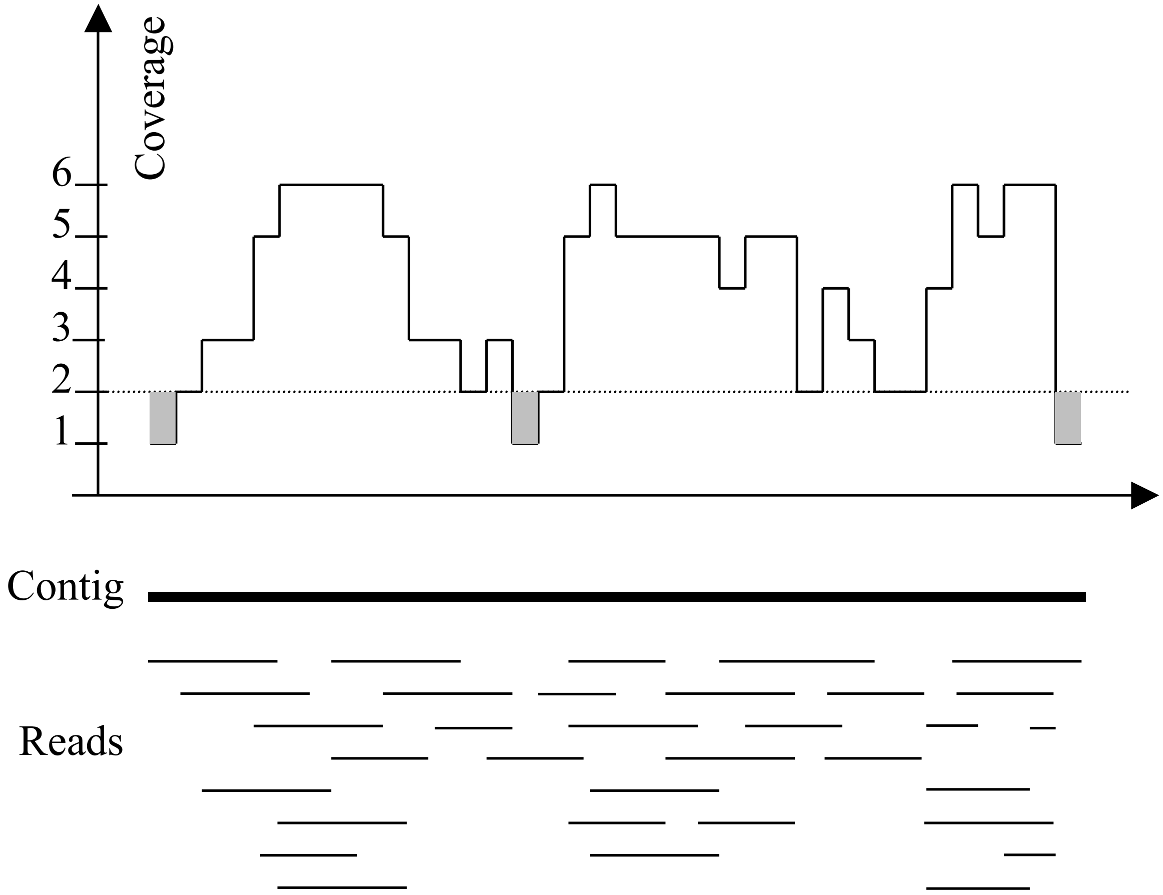
- Sequencing instruments make mistakes
  - Quality of read decreases over the read length
- A single read differing from the reference is probably just an error, but it becomes more likely to be real as we see it multiple times
  - Often framed as a Bayesian problem of more likely to be a real variant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

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



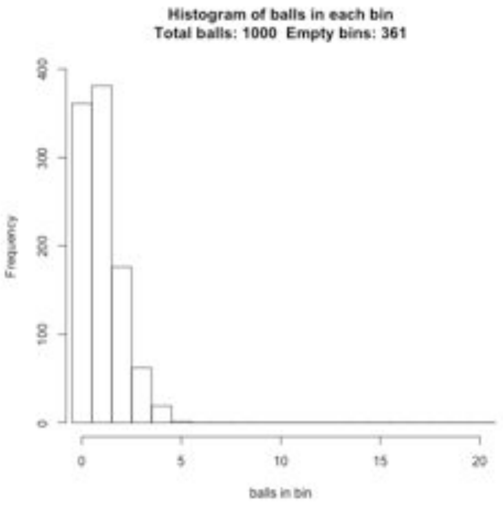
Coverage

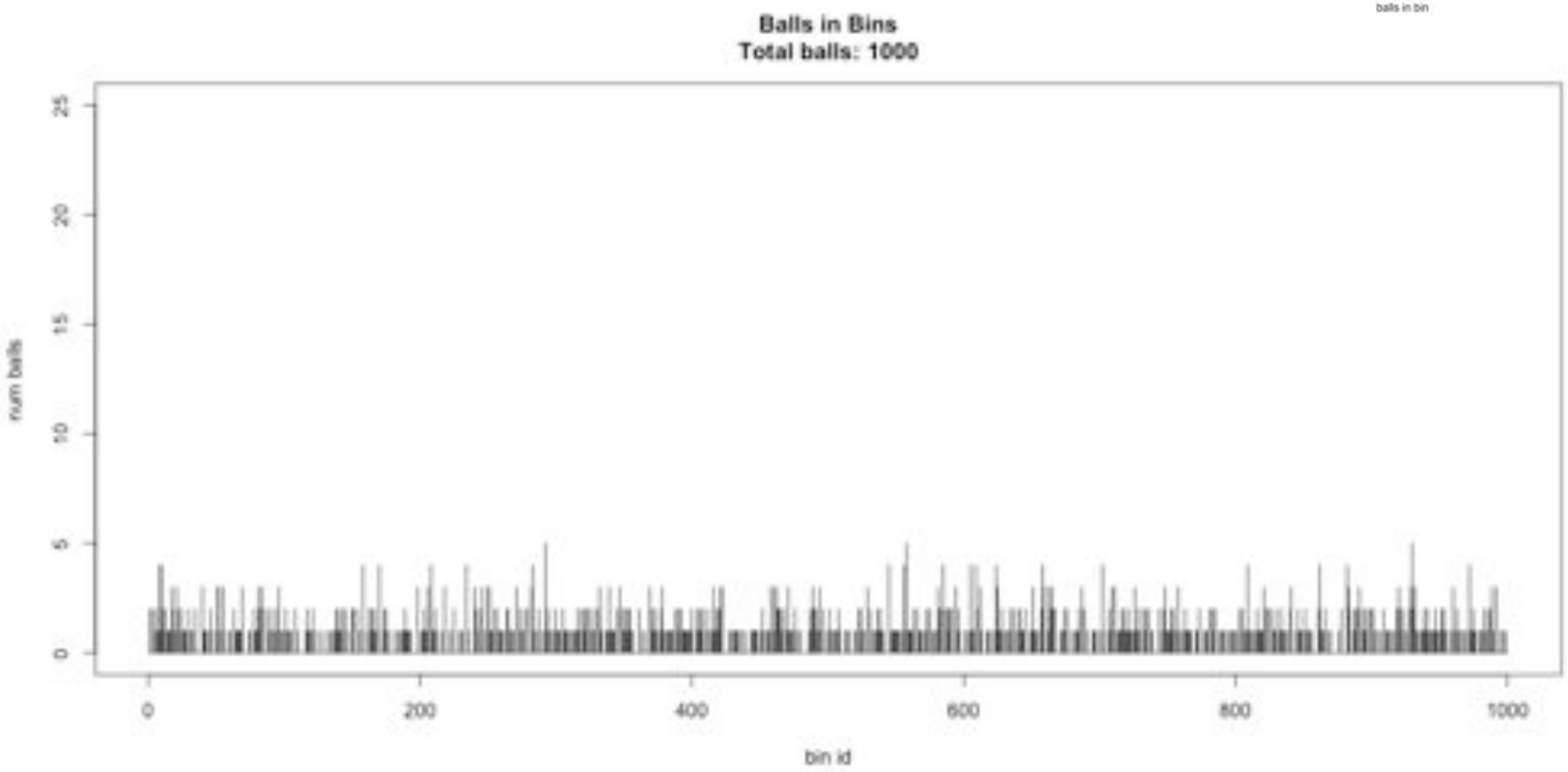
Typical contig coverage



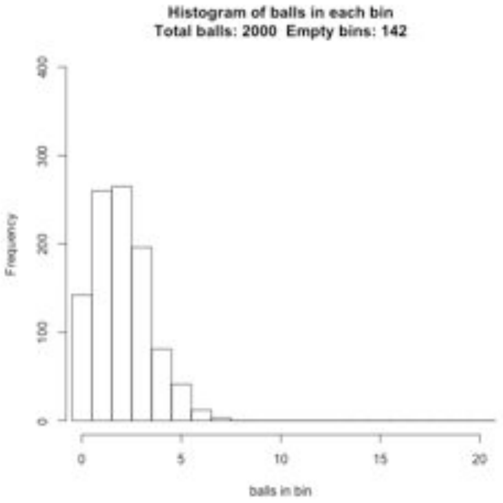
Imagine raindrops on a sidewalk

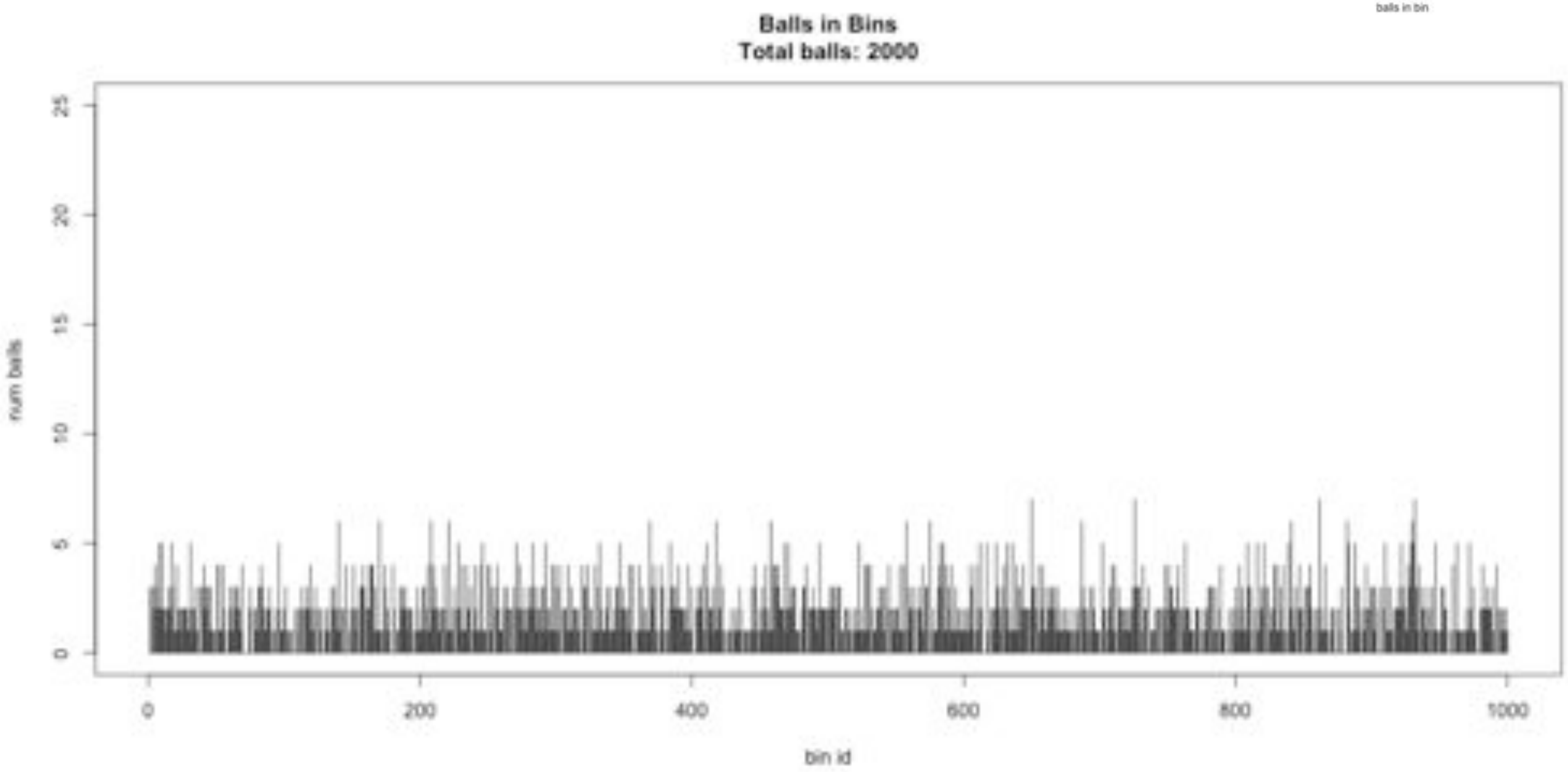
1x Sequencing



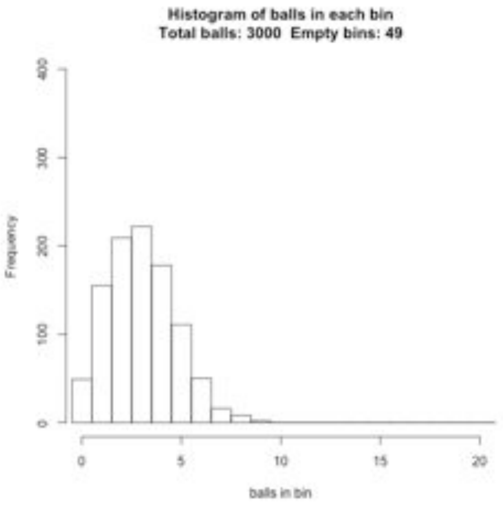


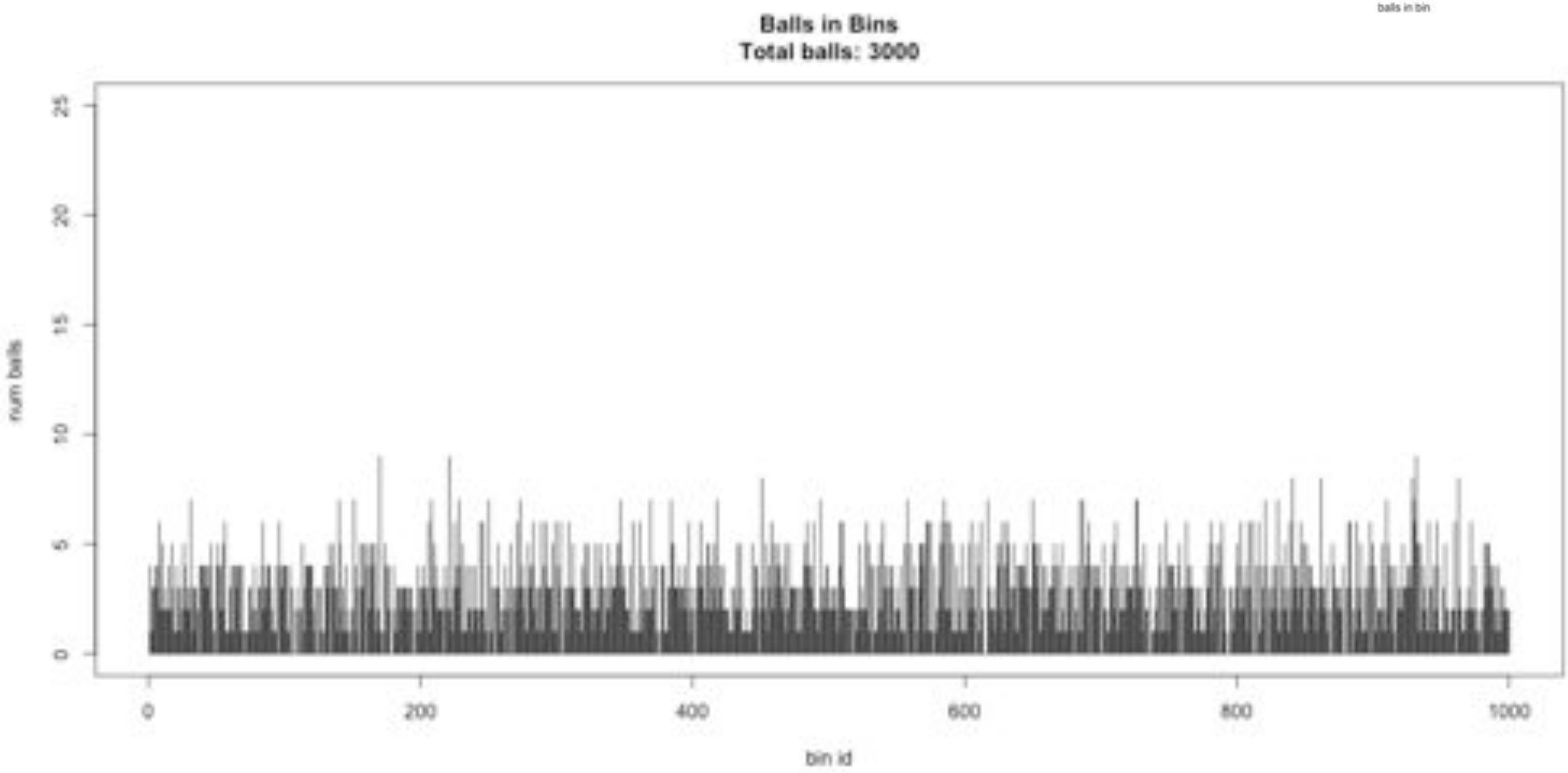
2x Sequencing



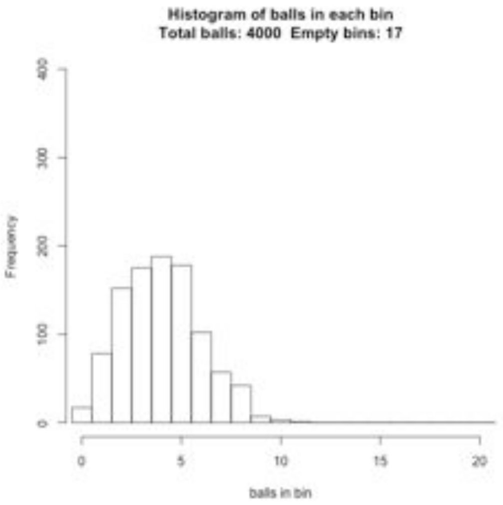


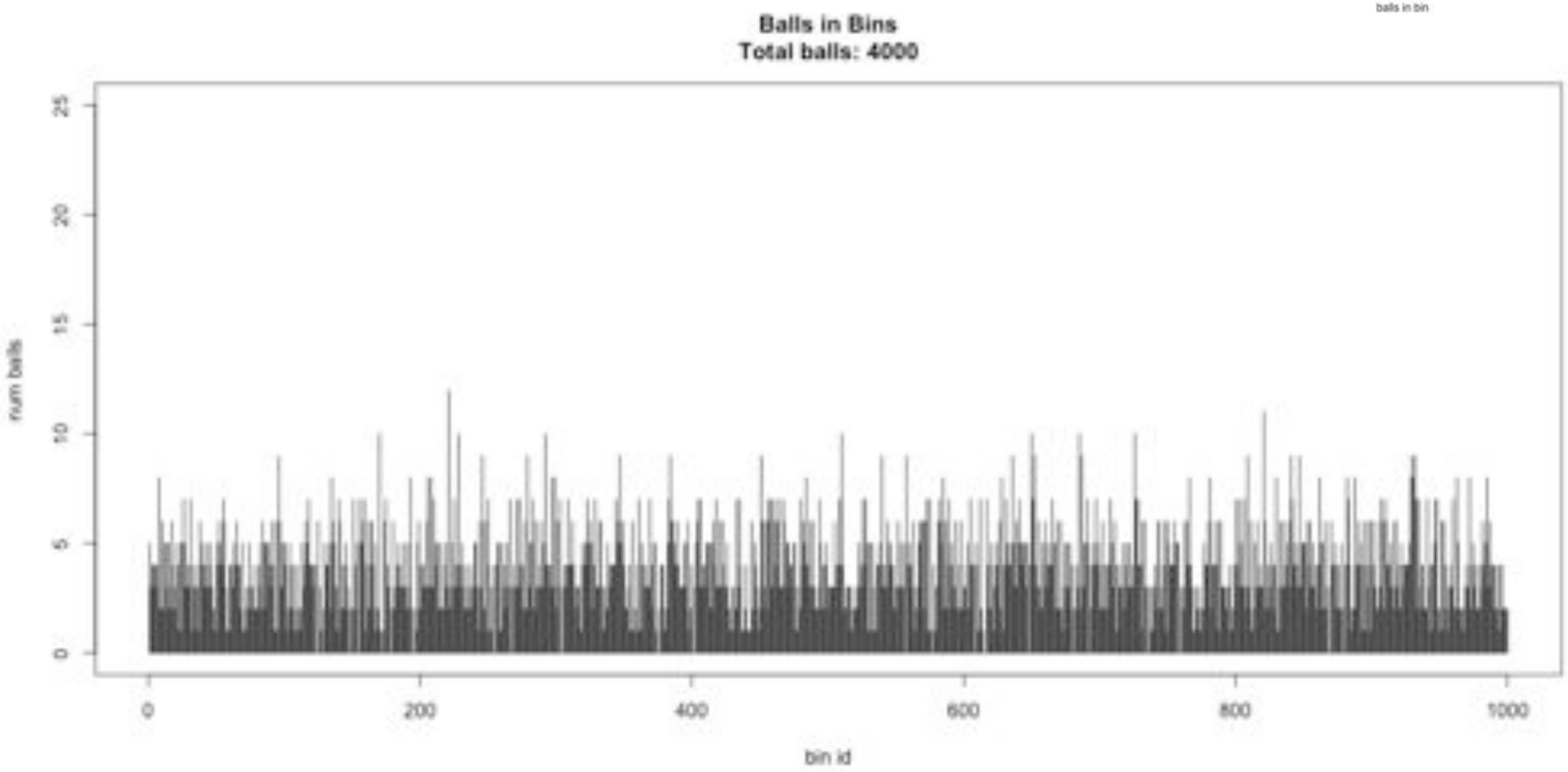
3x Sequencing



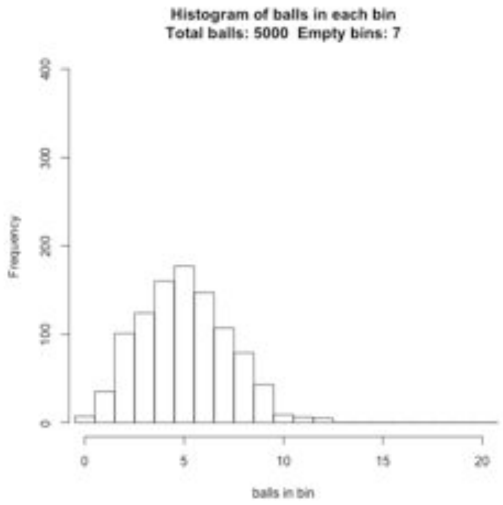


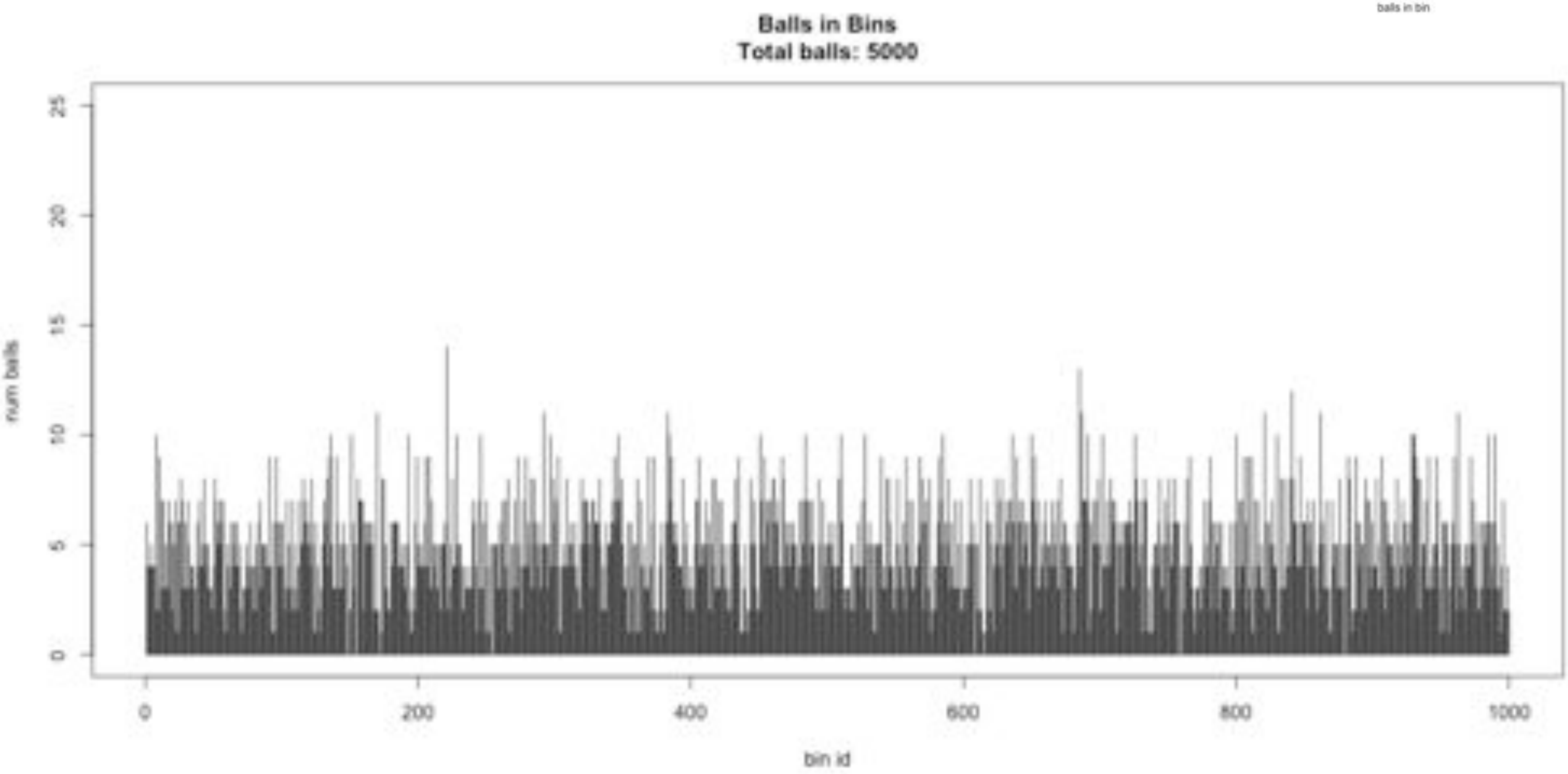
4x Sequencing



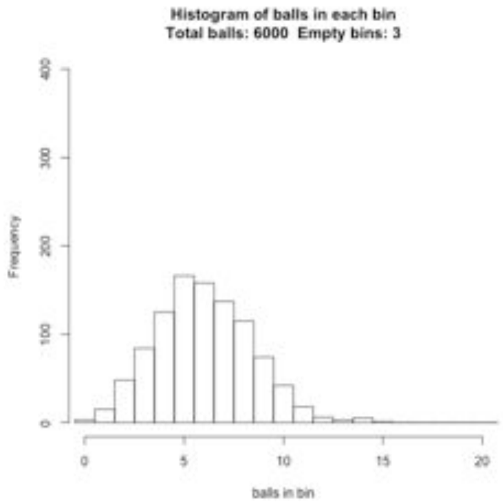


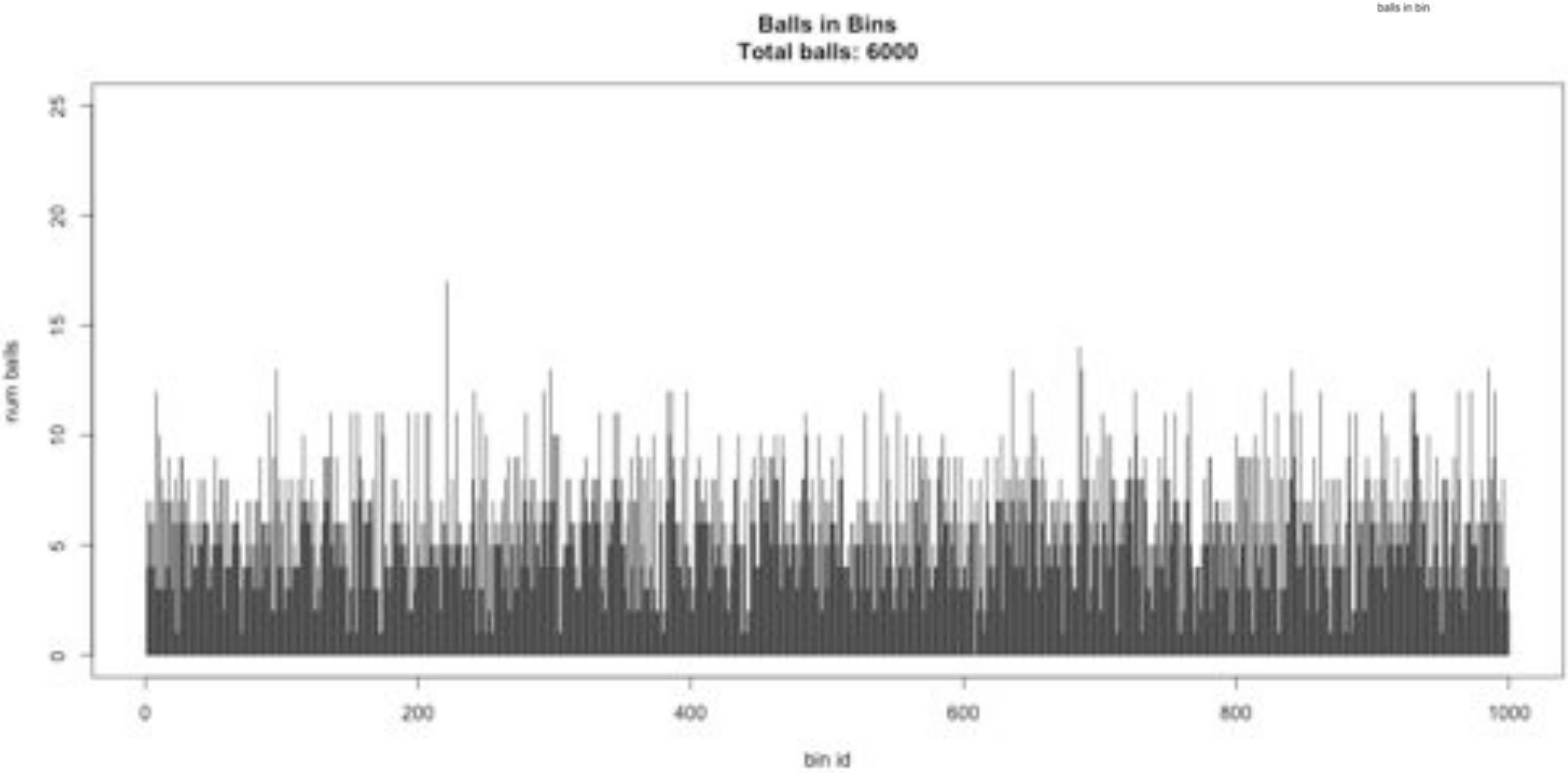
5x Sequencing



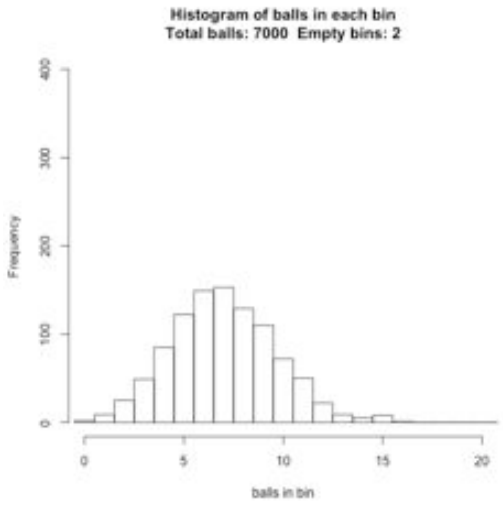


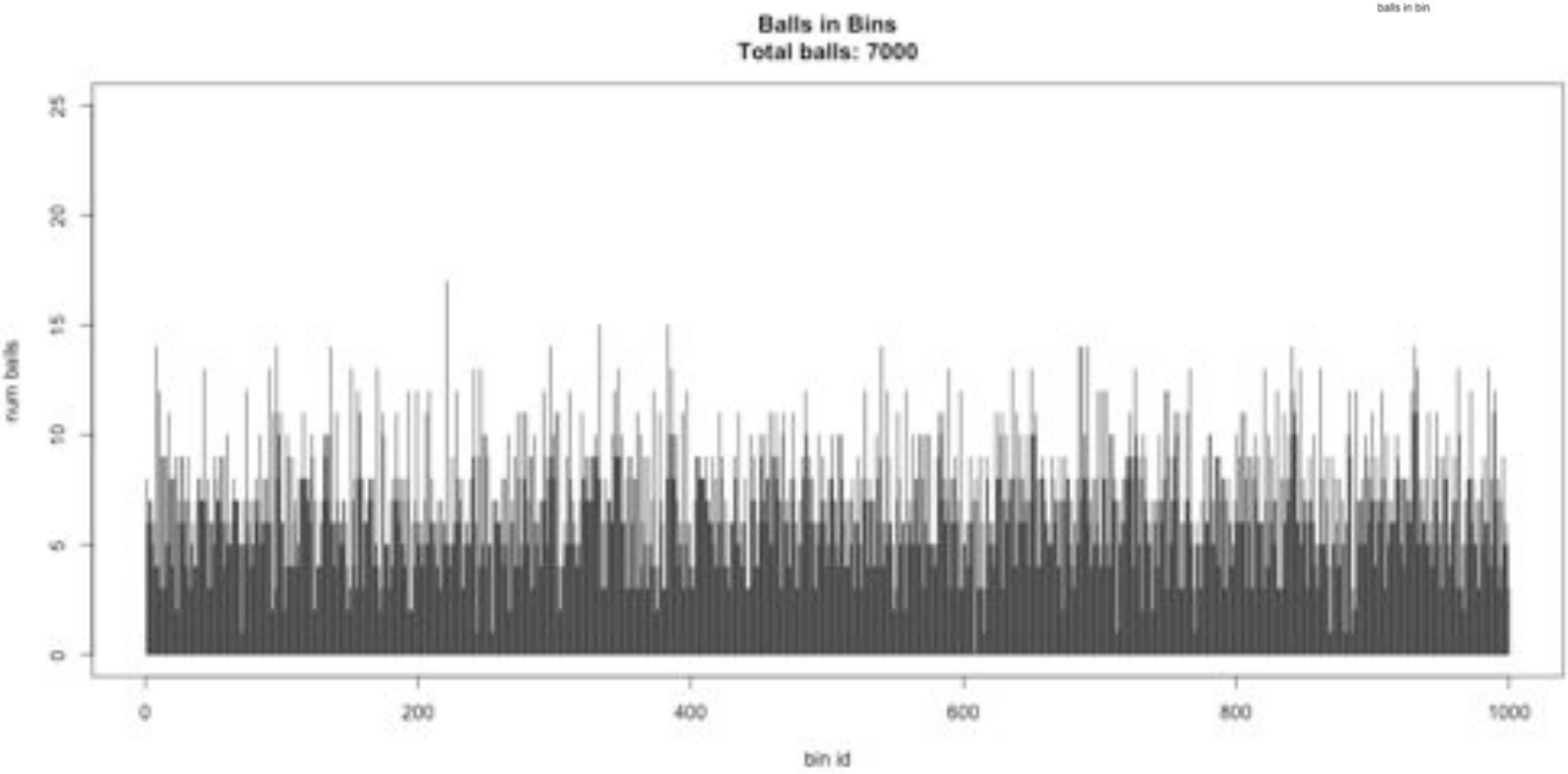
6x Sequencing



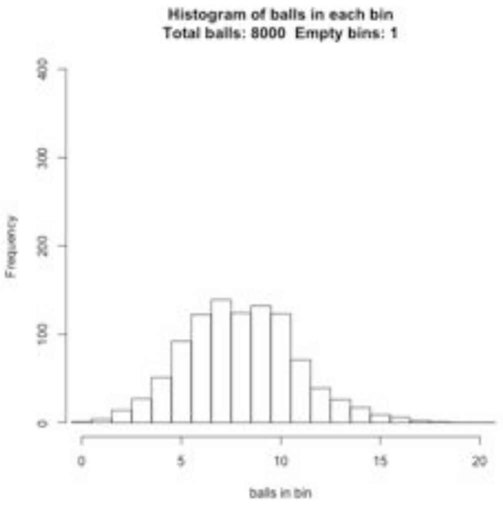


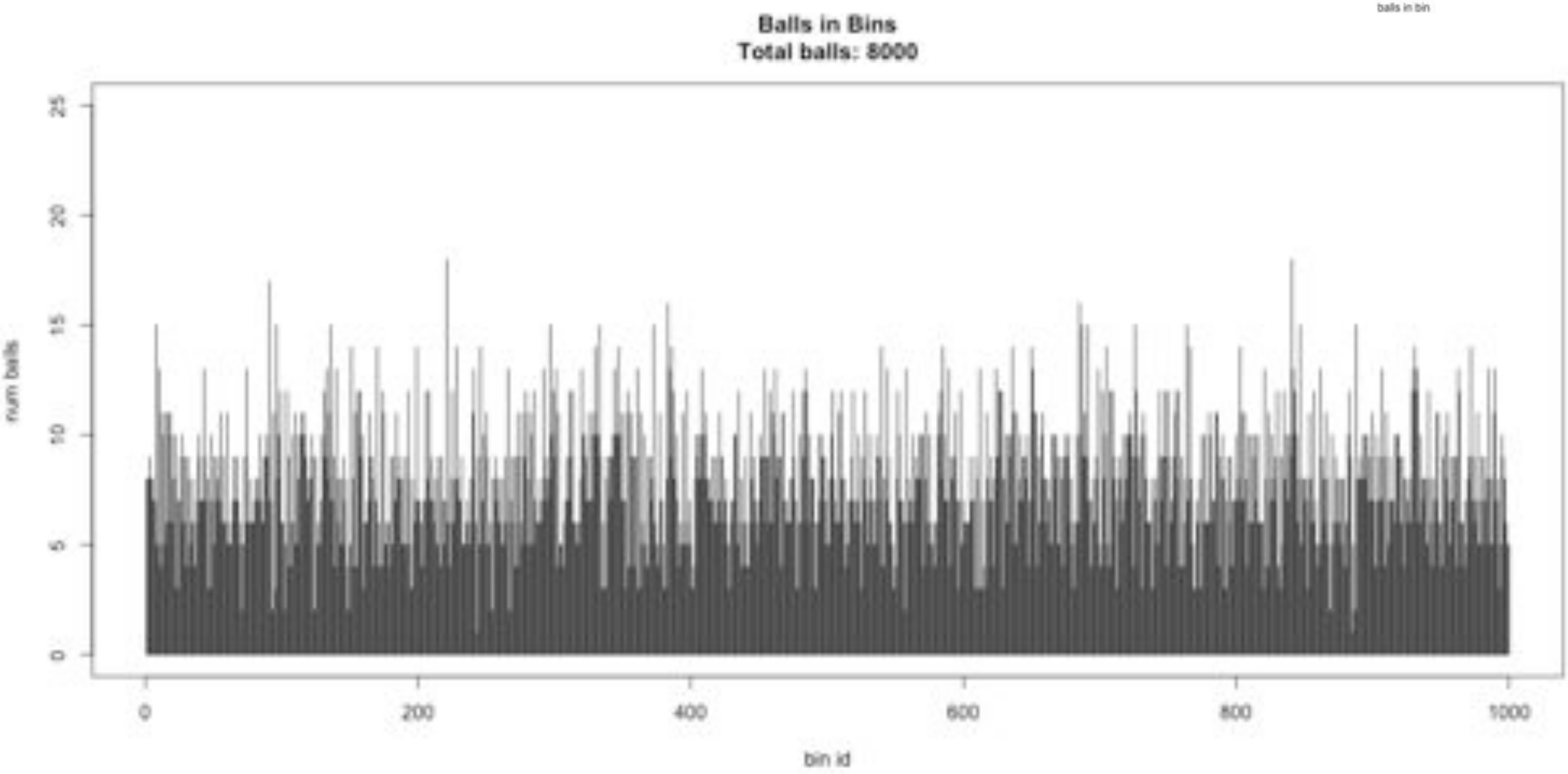
7x Sequencing



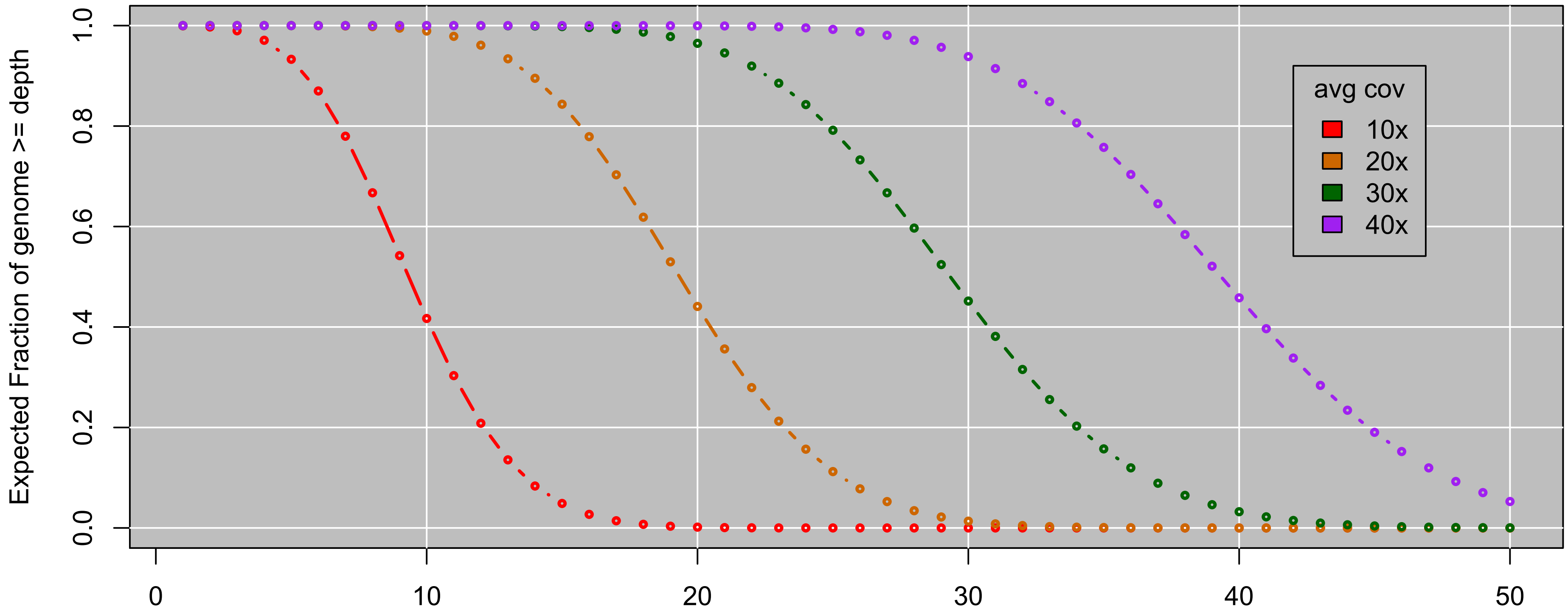


8x Sequencing





Genome Coverage Distribution



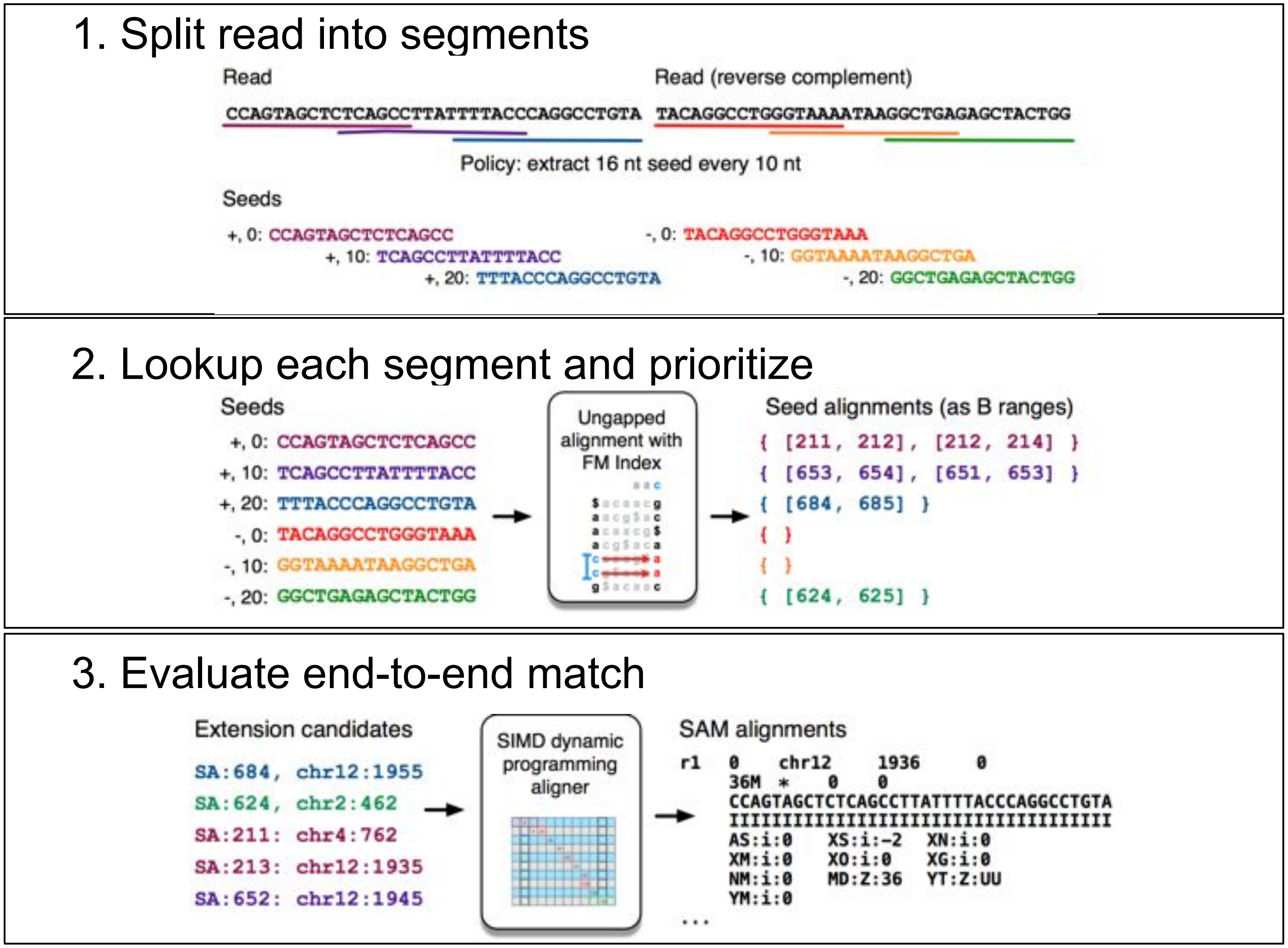
Expect Poisson distribution on depth

Standard Deviation = sqrt(cov)

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

Double your coverage for diploid genomes

Bowtie2 Overview

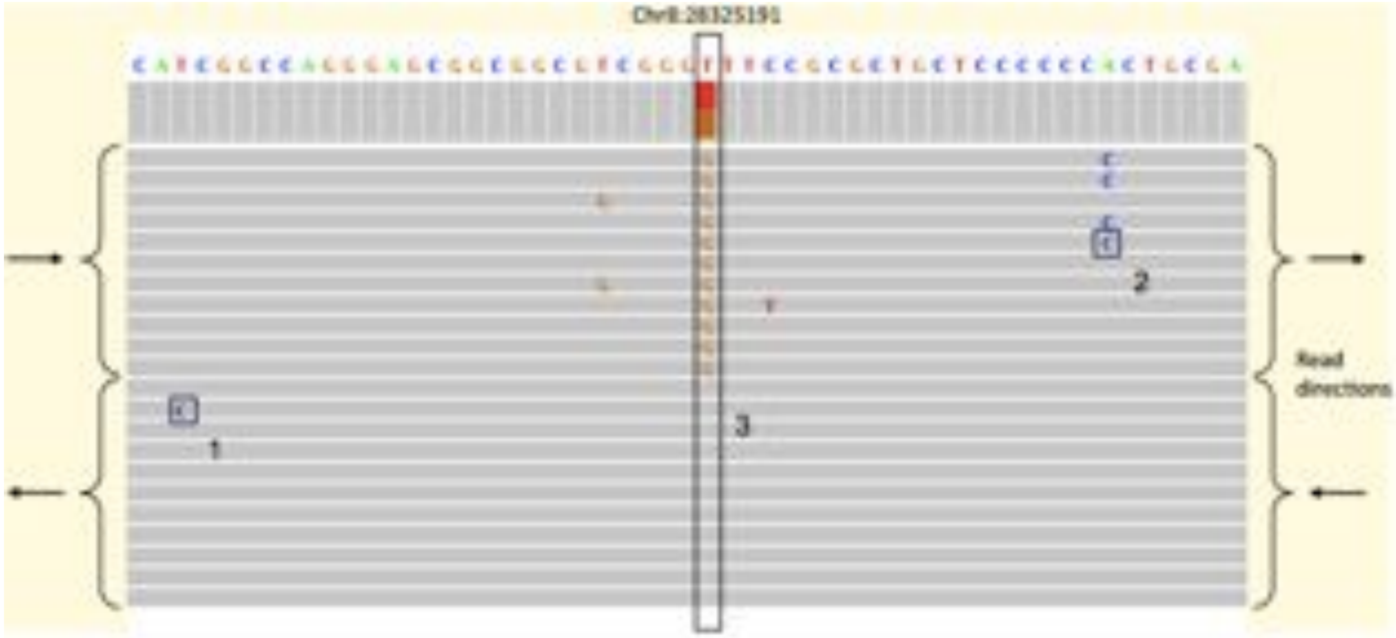


```
r1   0    chr12     1936      0
     36M  *    0    0
     CCAGTAGCTCTCAGCCTTATTTTACCCAGGCCTGTA
     IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
     AS:i:0    XS:i:-2   XN:i:0
     XM:i:0    XO:i:0    XG:i:0
     NM:i:0    MD:Z:36   YT:Z:UU
     YM:i:0
...
```

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

SNP calling

Beware of (Systematic) Errors



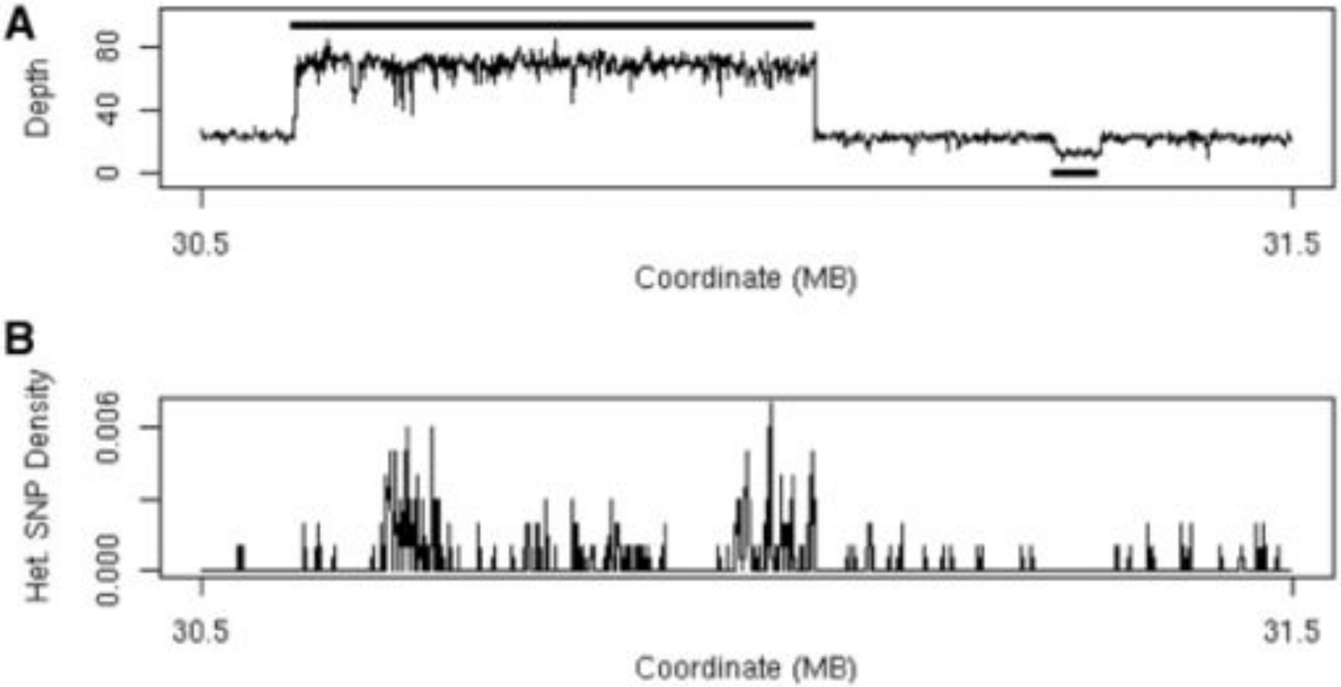
- Distinguishing SNPs from sequencing error typically a likelihood test of the coverage
  - Probability of seeing the data from a heterozygous SNP versus from sequencing error
  - However, some sequencing errors are systematic!

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

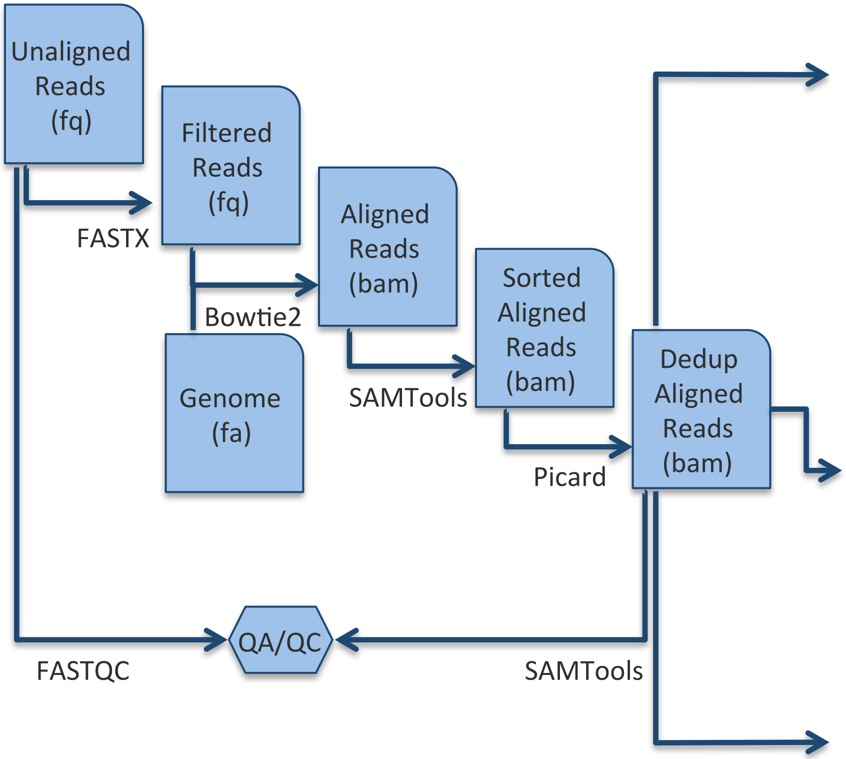
Agenda

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

Sequence to Discovery

Illumina HiSeq 2000
Sequencing by Synthesis

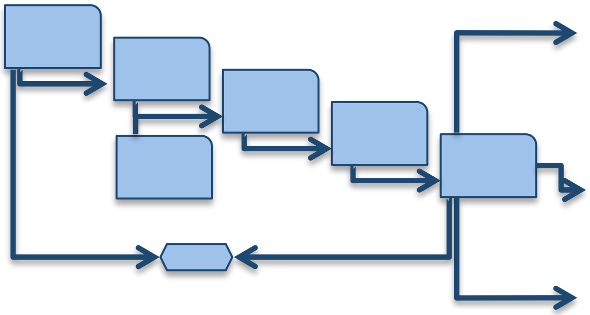
>60Gbp / day



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

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

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
```

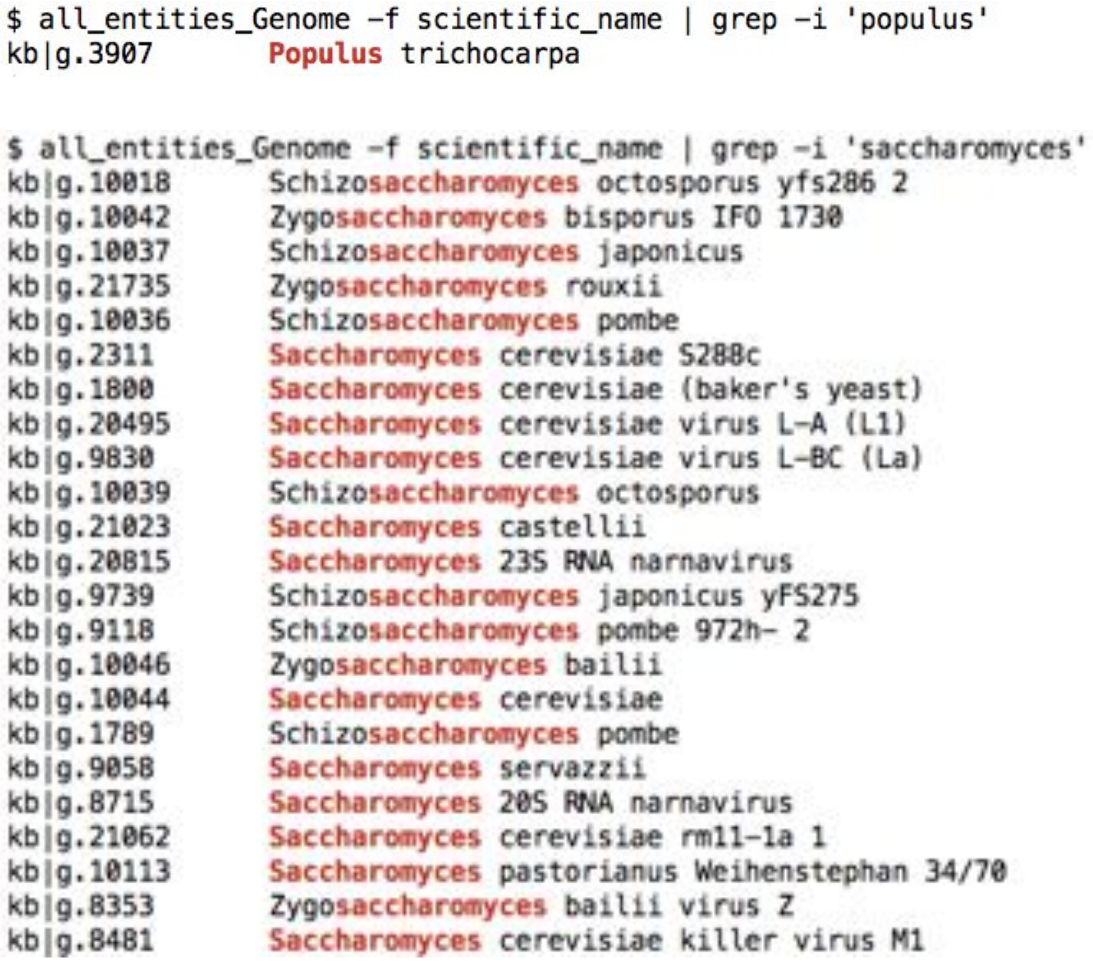
4. Call SNPs with SAMTools

```
$ 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
```

Identify a Reference Genome



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

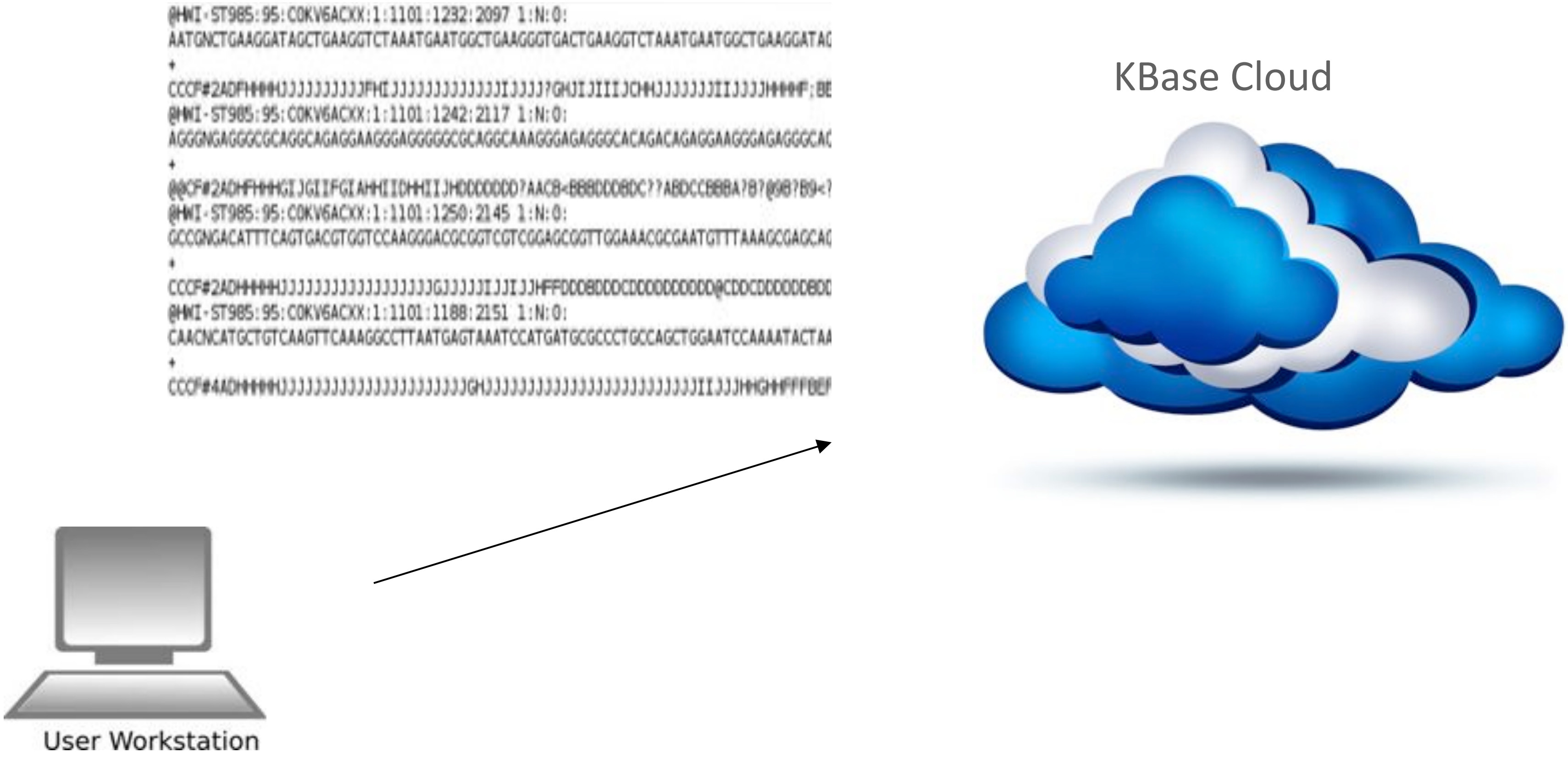
```
$ all_entities_Genome -f scientific_name | grep -i 'saccharomyces'
kb|g.10018      Schizosaccharomyces octosporus yfs286 2
kb|g.10042      Zygosaccharomyces bisporus IFO 1730
kb|g.10037      Schizosaccharomyces japonicus
kb|g.21735      Zygosaccharomyces rouxii
kb|g.10036      Schizosaccharomyces pombe
kb|g.2311       Saccharomyces cerevisiae S288c
kb|g.1800       Saccharomyces cerevisiae (baker's yeast)
kb|g.20495      Saccharomyces cerevisiae virus L-A (L1)
kb|g.9830       Saccharomyces cerevisiae virus L-BC (La)
kb|g.10039      Schizosaccharomyces octosporus
kb|g.21023      Saccharomyces castellii
kb|g.20815      Saccharomyces 23S RNA narnavirus
kb|g.9739       Schizosaccharomyces japonicus yFS275
kb|g.9118       Schizosaccharomyces pombe 972h- 2
kb|g.10046      Zygosaccharomyces bailii
kb|g.10044      Saccharomyces cerevisiae
kb|g.1789       Schizosaccharomyces pombe
kb|g.9058       Saccharomyces servazzii
kb|g.8715       Saccharomyces 20S RNA narnavirus
kb|g.21062      Saccharomyces cerevisiae rm11-1a 1
kb|g.10113      Saccharomyces pastorianus Weihenstephan 34/70
kb|g.8353       Zygosaccharomyces bailii virus Z
kb|g.8481       Saccharomyces cerevisiae killer virus M1
```

Select the proper KBase ID

Identify reference genome

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

Upload Reads to KBase Cloud

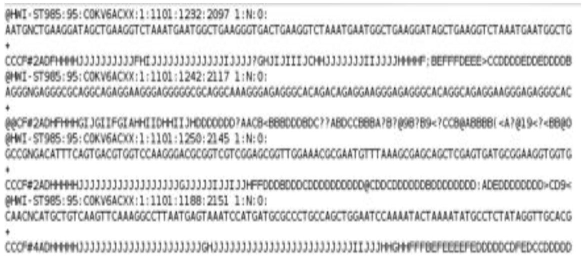


```
@HWI-ST985:95:C0KV6ACXX:1:1101:1232:2097 1:N:0:
AATGNCTGAAGGATAGCTGAAGGTCTAAATGAATGGCTGAAGGGTGACTGAAGGTCTAAATGAATGGCTGAAGGATAG
+
CCCF#2ADFHHHHJJJJJJJJJJFHIJJJJJJJJJJJJJIJJJJ?GHJIJIIIJCHHJJJJJJJIIJJJJHHHHF;BE
@HWI-ST985:95:C0KV6ACXX:1:1101:1242:2117 1:N:0:
AGGGNGAGGGCGCAGGCAGAGGAAGGGAGGGGGCGCAGGCAAAGGGAGAGGGCACAGACAGAGGAAGGGAGAGGGCAC
+
@@CF#2ADHFHHHGIJGIIFGIAHHIIDHHIIJHDDDDDDD?AACB<BBBDDDBDC??ABDCCBBBA?B?@9B?B9<?
@HWI-ST985:95:C0KV6ACXX:1:1101:1250:2145 1:N:0:
GCCGNGACATTTCAGTGACGTGGTCCAAGGGACGCGGTCGTCGGAGCGGTTGGAAACGCGAATGTTTAAAGCGAGCAG
+
CCCF#2ADHHHHHJJJJJJJJJJJJJJJJJJGJJJJJIJJIJJHFFDDDBDDDCDDDDDDDDDD@CDDCDDDDDDBDD
@HWI-ST985:95:C0KV6ACXX:1:1101:1188:2151 1:N:0:
CAACNCATGCTGTCAAGTTCAAAGGCCTTAATGAGTAAATCCATGATGCGCCCTGCCAGCTGGAATCCAAAATACTAA
+
CCCF#4ADHHHHHJJJJJJJJJJJJJJJJJJJJJJJJGHJJJJJJJJJJJJJJJJJJJJJJJJJJIIJJJHHGHHFFFBEF
```

Upload Reads to KBase cloud

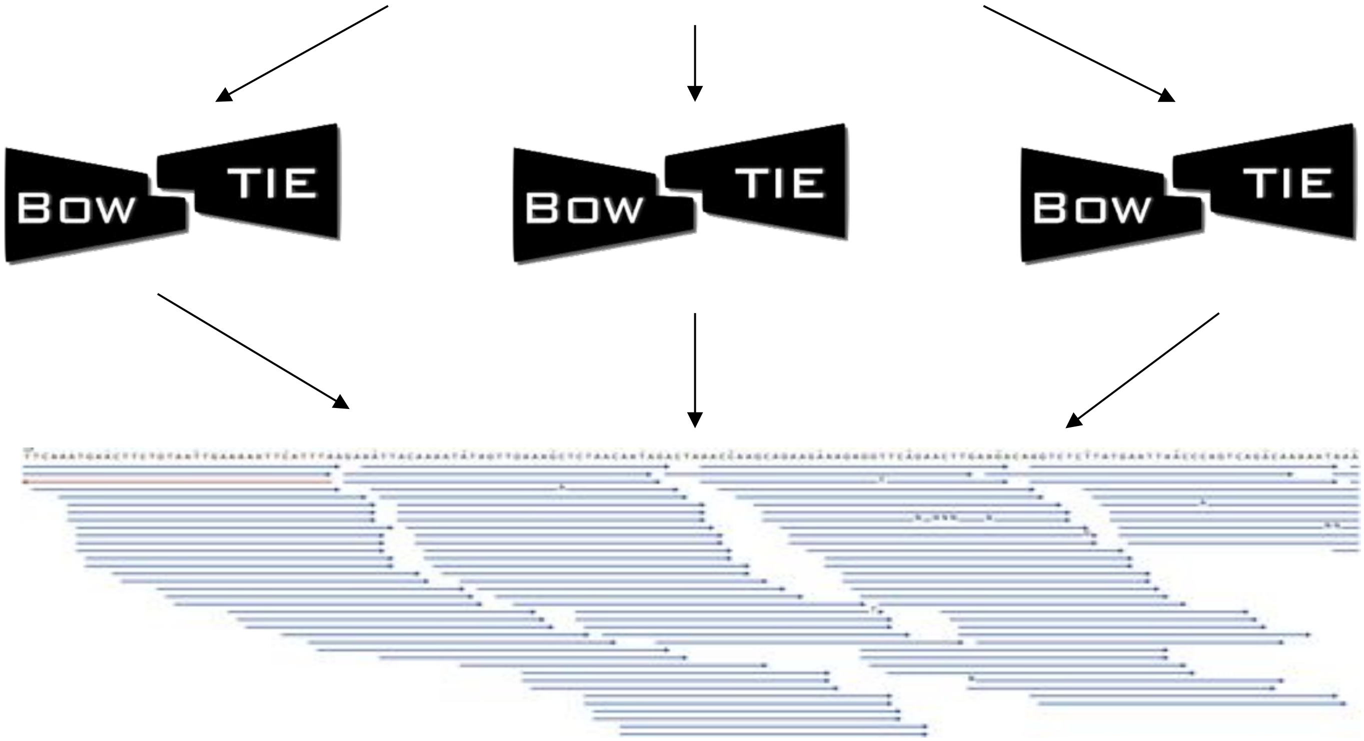
$ jk_fs_put_pe populus.1.fq.gz populus.2.fq.gz populus

Align Reads with Bowtie2



```
@HWI-ST985:95:C0KV6ACXX:1:1101:1232:2097 1:N:0:
AATGNCTGAAGGATAGCTGAAGGTCTAAATGAATGGCTGAAGGGTGACTGAAGGTCTAAATGAATGGCTGAAGGATAGCTGAAGGTCTAAATGAATGGCTG
+
CCCF#2ADFHHHHJJJJJJJJJJFHIJJJJJJJJJJJJJJIJJJJ?GHJIJIIIJCHHJJJJJJJJIIJJJJHHHHF;BEFFFDEEE>CCDDDDEDDEDDDDB
@HWI-ST985:95:C0KV6ACXX:1:1101:1242:2117 1:N:0:
AGGGNGAGGGCGCAGGCAGAGGAAGGGAGGGGGCGCAGGCAAAGGGAGAGGGCACAGACAGAGGAAGGGAGAGGGCACAGGCAGAGGAAGGGAGAGGGCAC
+
@@CF#2ADHFHHHGIJGIIFGIAHHIIDHHIIJHDDDDDDD?AACB<BBBDDDBDC??ABDCCBBBA?B?@9B?B9<?CCB@ABBBB(<A?@19<?<BB@0
@HWI-ST985:95:C0KV6ACXX:1:1101:1250:2145 1:N:0:
GCCGNGACATTTCAGTGACGTGGTCCAAGGGACGCGGTCGTCGGAGCGGTTGGAAACGCGAATGTTTAAAGCGAGCAGCTCGAGTGATGCGGAAGGTGGTG
+
CCCF#2ADHHHHHJJJJJJJJJJJJJJJJJJGJJJJJIJJIJJHFFDDDBDDDCDDDDDDDDDD@CDDCDDDDDDBDDDDDDDD:ADEDDDDDDDD>CD9<
@HWI-ST985:95:C0KV6ACXX:1:1101:1188:2151 1:N:0:
CAACNCATGCTGTCAAGTTCAAAGGCCTTAATGAGTAAATCCATGATGCGCCCTGCCAGCTGGAATCCAAAATACTAAAATATGCCTCTATAGGTTGCACG
+
CCCF#4ADHHHHHJJJJJJJJJJJJJJJJJJJJJJJGHJJJJJJJJJJJJJJJJJJJJJJJJJJJIIJJJHHGHHFFFBEFEEEEFEDDDDDCDFEDCCDDDDD
```

Raw Fastq Reads



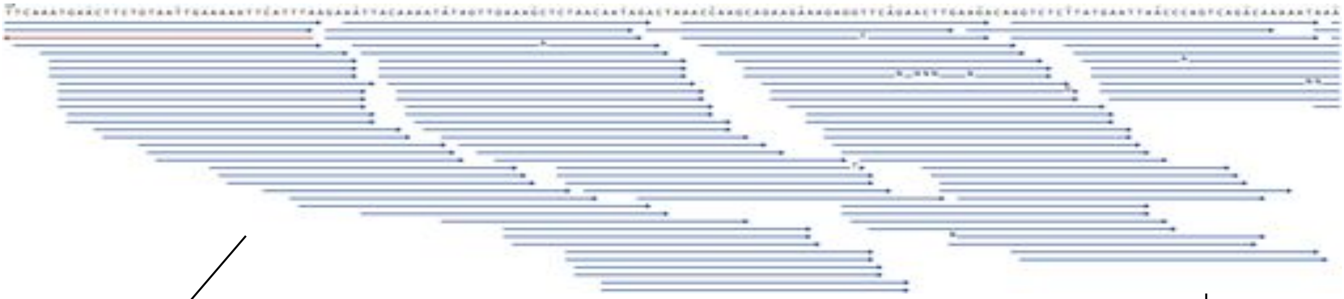
Bowtie2 Aligner

Alignments

Align Reads with Bowtie2

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

Call SNPs with SAMTools



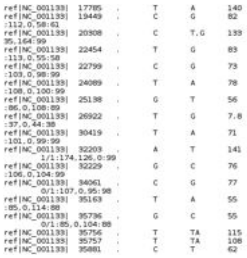
Alignments

SAMTools

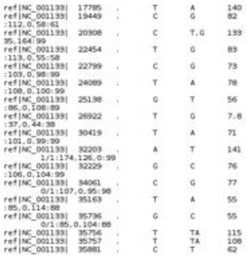
SAMTools

SAMTools

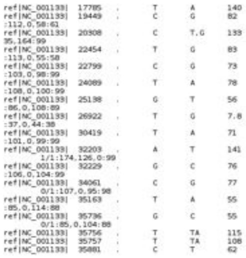
Samtools Variant Detection



```
ref|NC_001133|  17785  .    T    A    140
ref|NC_001133|  19449  .    C    G    82
:112,0,58:61
ref|NC_001133|  20308  .    C    T,G  133
35,164:99
ref|NC_001133|  22454  .    T    G    83
:113,0,55:58
ref|NC_001133|  22799  .    C    G    73
:103,0,98:99
ref|NC_001133|  24089  .    T    A    78
:108,0,100:99
ref|NC_001133|  25138  .    G    T    56
:86,0,108:89
ref|NC_001133|  26922  .    T    G    7.8
:37,0,44:38
ref|NC_001133|  30419  .    T    A    71
:101,0,99:99
ref|NC_001133|  32203  .    A    T    141
          1/1:174,126,0:99
ref|NC_001133|  32229  .    G    C    76
:106,0,104:99
ref|NC_001133|  34061  .    C    G    77
          0/1:107,0,95:98
ref|NC_001133|  35163  .    T    A    55
:85,0,114:88
ref|NC_001133|  35736  .    G    C    55
          0/1:85,0,104:88
ref|NC_001133|  35756  .    T    TA   115
ref|NC_001133|  35757  .    T    TA   108
ref|NC_001133|  35881  .    C    T    62
```



```
ref|NC_001133|  17785  .    T    A    140
ref|NC_001133|  19449  .    C    G    82
:112,0,58:61
ref|NC_001133|  20308  .    C    T,G  133
35,164:99
ref|NC_001133|  22454  .    T    G    83
:113,0,55:58
ref|NC_001133|  22799  .    C    G    73
:103,0,98:99
ref|NC_001133|  24089  .    T    A    78
:108,0,100:99
ref|NC_001133|  25138  .    G    T    56
:86,0,108:89
ref|NC_001133|  26922  .    T    G    7.8
:37,0,44:38
ref|NC_001133|  30419  .    T    A    71
:101,0,99:99
ref|NC_001133|  32203  .    A    T    141
          1/1:174,126,0:99
ref|NC_001133|  32229  .    G    C    76
:106,0,104:99
ref|NC_001133|  34061  .    C    G    77
          0/1:107,0,95:98
ref|NC_001133|  35163  .    T    A    55
:85,0,114:88
ref|NC_001133|  35736  .    G    C    55
          0/1:85,0,104:88
ref|NC_001133|  35756  .    T    TA   115
ref|NC_001133|  35757  .    T    TA   108
ref|NC_001133|  35881  .    C    T    62
```



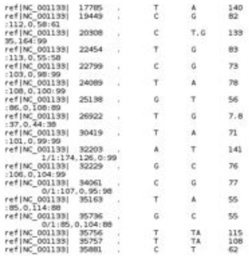
```
ref|NC_001133|  17785  .    T    A    140
ref|NC_001133|  19449  .    C    G    82
:112,0,58:61
ref|NC_001133|  20308  .    C    T,G  133
35,164:99
ref|NC_001133|  22454  .    T    G    83
:113,0,55:58
ref|NC_001133|  22799  .    C    G    73
:103,0,98:99
ref|NC_001133|  24089  .    T    A    78
:108,0,100:99
ref|NC_001133|  25138  .    G    T    56
:86,0,108:89
ref|NC_001133|  26922  .    T    G    7.8
:37,0,44:38
ref|NC_001133|  30419  .    T    A    71
:101,0,99:99
ref|NC_001133|  32203  .    A    T    141
          1/1:174,126,0:99
ref|NC_001133|  32229  .    G    C    76
:106,0,104:99
ref|NC_001133|  34061  .    C    G    77
          0/1:107,0,95:98
ref|NC_001133|  35163  .    T    A    55
:85,0,114:88
ref|NC_001133|  35736  .    G    C    55
          0/1:85,0,104:88
ref|NC_001133|  35756  .    T    TA   115
ref|NC_001133|  35757  .    T    TA   108
ref|NC_001133|  35881  .    C    T    62
```

Called Variants (VCF)

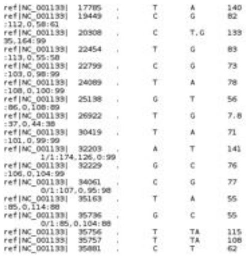
Call SNPs with SAMTools

```
$ jk_compute_samtools_snp -in=populus_align –org=‘kb|g.3907’ -out=populus_snps
```

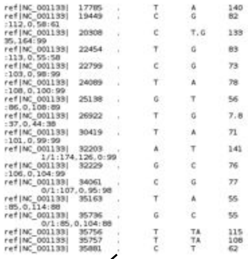
Merge and Download VCF Files



```
ref|NC_001133|  17785   .       T       A       140
ref|NC_001133|  19449   .       C       G       82
:112,0,58:61
ref|NC_001133|  20308   .       C       T,G     133
35,164:99
ref|NC_001133|  22454   .       T       G       83
:113,0,55:58
ref|NC_001133|  22799   .       C       G       73
:103,0,98:99
ref|NC_001133|  24089   .       T       A       78
:108,0,100:99
ref|NC_001133|  25138   .       G       T       56
:86,0,108:89
ref|NC_001133|  26922   .       T       G       7.8
:37,0,44:38
ref|NC_001133|  30419   .       T       A       71
:101,0,99:99
ref|NC_001133|  32203   .       A       T       141
        1/1:174,126,0:99
ref|NC_001133|  32229   .       G       C       76
:106,0,104:99
ref|NC_001133|  34061   .       C       G       77
        0/1:107,0,95:98
ref|NC_001133|  35163   .       T       A       55
:85,0,114:88
ref|NC_001133|  35736   .       G       C       55
        0/1:85,0,104:88
ref|NC_001133|  35756   .       T       TA      115
ref|NC_001133|  35757   .       T       TA      108
ref|NC_001133|  35881   .       C       T       62
```

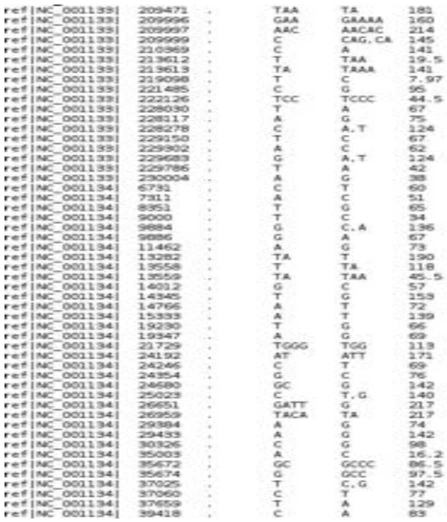


```
ref|NC_001133|  17785   .       T       A       140
ref|NC_001133|  19449   .       C       G       82
:112,0,58:61
ref|NC_001133|  20308   .       C       T,G     133
35,164:99
ref|NC_001133|  22454   .       T       G       83
:113,0,55:58
ref|NC_001133|  22799   .       C       G       73
:103,0,98:99
ref|NC_001133|  24089   .       T       A       78
:108,0,100:99
ref|NC_001133|  25138   .       G       T       56
:86,0,108:89
ref|NC_001133|  26922   .       T       G       7.8
:37,0,44:38
ref|NC_001133|  30419   .       T       A       71
:101,0,99:99
ref|NC_001133|  32203   .       A       T       141
        1/1:174,126,0:99
ref|NC_001133|  32229   .       G       C       76
:106,0,104:99
ref|NC_001133|  34061   .       C       G       77
        0/1:107,0,95:98
ref|NC_001133|  35163   .       T       A       55
:85,0,114:88
ref|NC_001133|  35736   .       G       C       55
        0/1:85,0,104:88
ref|NC_001133|  35756   .       T       TA      115
ref|NC_001133|  35757   .       T       TA      108
ref|NC_001133|  35881   .       C       T       62
```



```
ref|NC_001133|  17785   .       T       A       140
ref|NC_001133|  19449   .       C       G       82
:112,0,58:61
ref|NC_001133|  20308   .       C       T,G     133
35,164:99
ref|NC_001133|  22454   .       T       G       83
:113,0,55:58
ref|NC_001133|  22799   .       C       G       73
:103,0,98:99
ref|NC_001133|  24089   .       T       A       78
:108,0,100:99
ref|NC_001133|  25138   .       G       T       56
:86,0,108:89
ref|NC_001133|  26922   .       T       G       7.8
:37,0,44:38
ref|NC_001133|  30419   .       T       A       71
:101,0,99:99
ref|NC_001133|  32203   .       A       T       141
        1/1:174,126,0:99
ref|NC_001133|  32229   .       G       C       76
:106,0,104:99
ref|NC_001133|  34061   .       C       G       77
        0/1:107,0,95:98
ref|NC_001133|  35163   .       T       A       55
:85,0,114:88
ref|NC_001133|  35736   .       G       C       55
        0/1:85,0,104:88
ref|NC_001133|  35756   .       T       TA      115
ref|NC_001133|  35757   .       T       TA      108
ref|NC_001133|  35881   .       C       T       62
```

Merge VCF Files



```
ref|NC_001133|  209471  .       TAA     TA      181
ref|NC_001133|  209996  .       GAA     GAAAA   160
ref|NC_001133|  209997  .       AAC     AACAC   214
ref|NC_001133|  209999  .       C       CAG,CA  145
ref|NC_001133|  210969  .       C       A       141
ref|NC_001133|  213612  .       T       TAA     19.5
ref|NC_001133|  213613  .       TA      TAAA    141
ref|NC_001133|  219098  .       T       C       7.97
ref|NC_001133|  221485  .       C       G       95
ref|NC_001133|  222126  .       TCC     TCCC    44.5
ref|NC_001133|  228030  .       T       A       67
ref|NC_001133|  228117  .       A       G       75
ref|NC_001133|  228278  .       C       A,T     124
ref|NC_001133|  229150  .       T       C       67
ref|NC_001133|  229302  .       A       C       62
ref|NC_001133|  229683  .       G       A,T     124
ref|NC_001133|  229786  .       T       A       42
ref|NC_001133|  230004  .       A       G       38
ref|NC_001134|  6731    .       C       T       60
ref|NC_001134|  7311    .       A       C       51
ref|NC_001134|  8351    .       T       G       65
ref|NC_001134|  9000    .       T       C       34
ref|NC_001134|  9884    .       G       C,A     136
ref|NC_001134|  9886    .       G       A       67
ref|NC_001134|  11462   .       A       G       73
ref|NC_001134|  13282   .       TA      T       190
ref|NC_001134|  13558   .       T       TA      118
ref|NC_001134|  13559   .       TA      TAA     45.5
ref|NC_001134|  14012   .       G       C       57
ref|NC_001134|  14345   .       T       G       153
ref|NC_001134|  14766   .       A       T       72
ref|NC_001134|  15333   .       A       T       139
ref|NC_001134|  19230   .       T       G       66
ref|NC_001134|  19347   .       A       G       69
ref|NC_001134|  21729   .       TGGG    TGG     113
ref|NC_001134|  24192   .       AT      ATT     171
ref|NC_001134|  24246   .       C       T       69
ref|NC_001134|  24354   .       G       C       76
ref|NC_001134|  24680   .       GC      G       142
ref|NC_001134|  25023   .       C       T,G     140
ref|NC_001134|  26651   .       GATT    G       217
ref|NC_001134|  26959   .       TACA    TA      217
ref|NC_001134|  29384   .       A       G       74
ref|NC_001134|  29433   .       A       G       142
ref|NC_001134|  30326   .       C       G       98
ref|NC_001134|  35003   .       A       C       16.2
ref|NC_001134|  35672   .       GC      GCCC    86.5
ref|NC_001134|  35674   .       G       GCC     97.5
ref|NC_001134|  37025   .       T       C,G     142
ref|NC_001134|  37060   .       C       T       77
ref|NC_001134|  37659   .       T       A       129
ref|NC_001134|  39418   .       C       A       83
```

Download to Local Workstation

Merge and Download

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

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
```

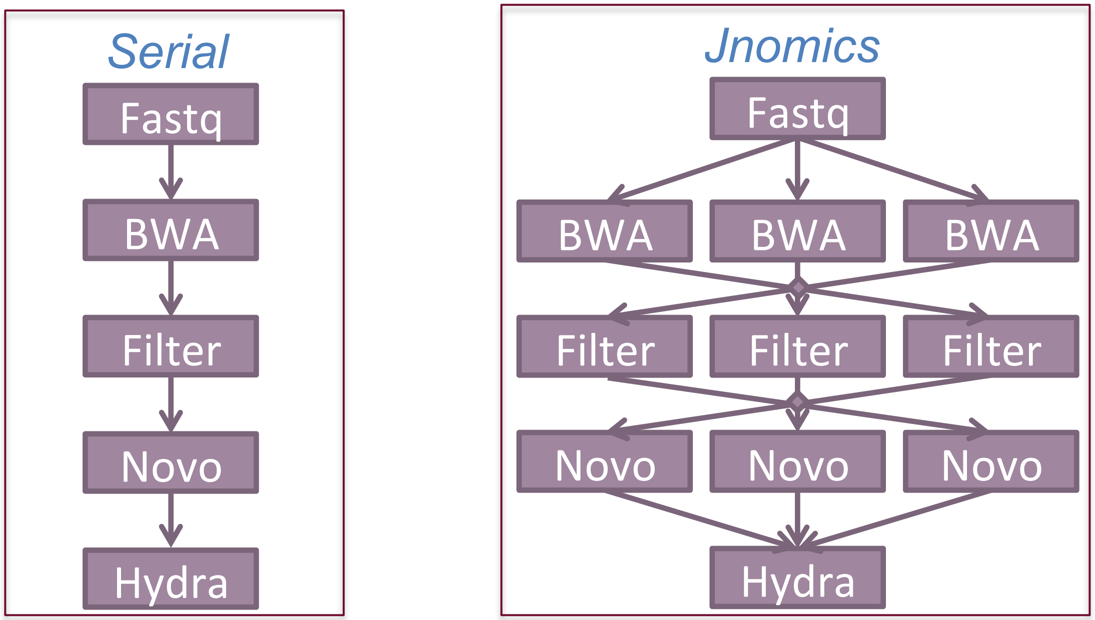
4. Call SNPs with SAMTools

```
$ 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
```

Jnomics: Cloud-scale genomics



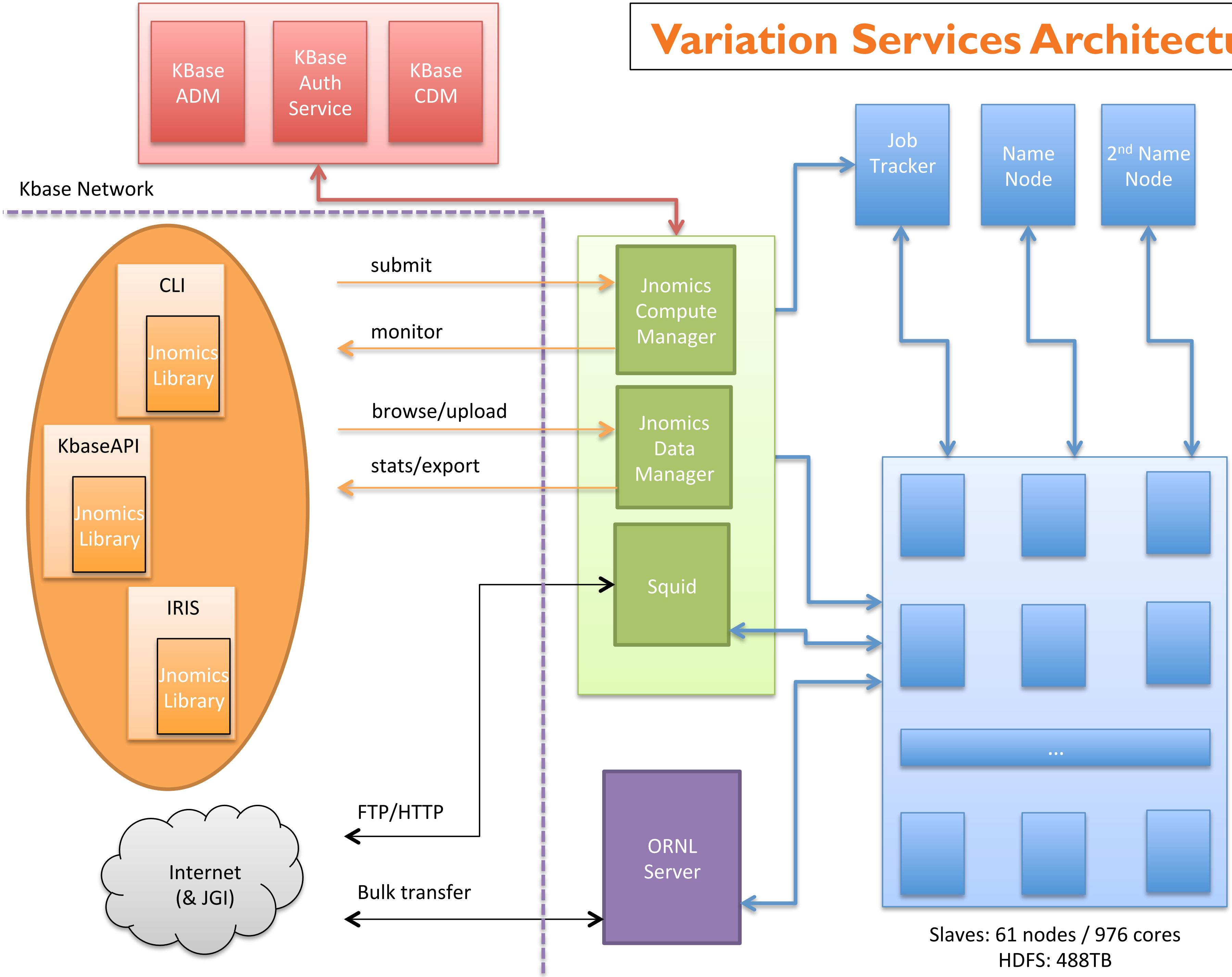
openstack™ CLOUD SOFTWARE

- 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*

Variation Services Architecture



Slaves: 61 nodes / 976 cores
HDFS: 488TB

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

Maize Population Analysis

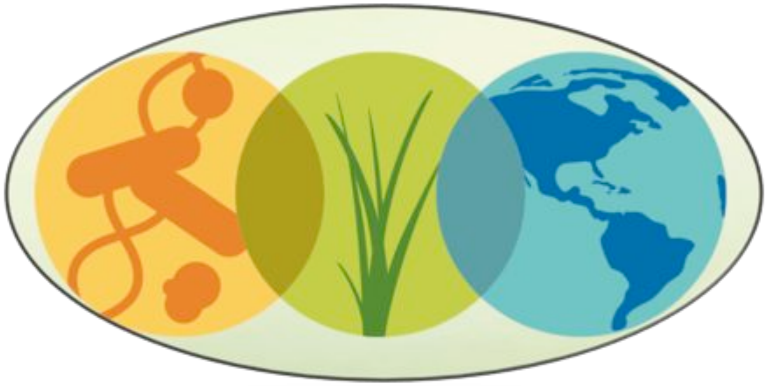
Align & call SNPs from 131 maize samples
1TB fastq / 408Gbp input data

	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

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/

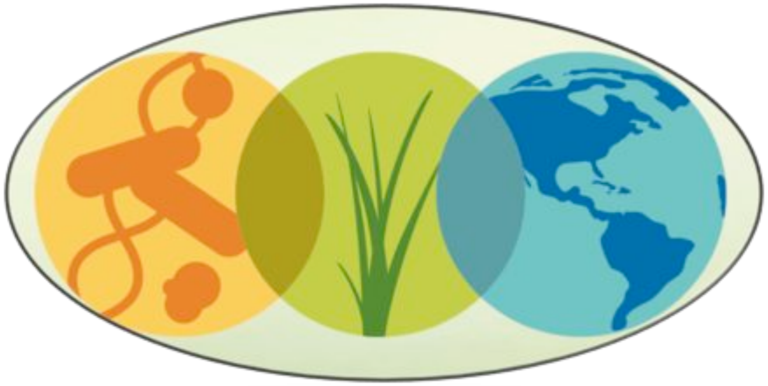
4. Variation Services Tutorial:
   http://kbase.us/for-users/tutorials/analyzing-data/variation-service/

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
```

Agenda

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



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