Genomic Technologies

Michael Schatz

Feb 1, 2018

Lecture 2: Applied Comparative Genomics

Welcome!

The primary goal of the course is for students to be grounded in theory and leave the course empowered to conduct independent genomic analyses.

- We will study the leading computational and quantitative approaches for comparing and analyzing genomes starting from raw sequencing data.
- The course will focus on human genomics and human medical applications, but the techniques will be broadly applicable across the tree of life.
- The topics will include genome assembly & comparative genomics, variant identification & analysis, gene expression & regulation, personal genome analysis, and cancer genomics.

Course Webpage: https://github.com/schatzlab/appliedgenomics2018
Course Discussions: http://piazza.com

Class Hours: Tues + Thurs @ 1:30p – 2:45p, Shaffer 304

Schatz Office Hours: Tues + Thurs @ 3-4p and by appointment
Darby Office Hours: Wed @ 4-5 and by appointment

Please try Piazza first!

Course Webpage

appliedgenomics2018/READM

GitHub, Inc. [US] | https://github.com/schatzlab/appliedgenomics2018/blob/master/README.md

This repository Search Pull requests Issues Marketplace Explore

schatzlab / appliedgenomics2018

Unwatch 1 Star 1 Fork 0

Code Issues 0 Pull requests 0 Projects 0 Wiki Insights Settings

Branch: master appliedgenomics2018 / README.md Find file Copy path

cdarby Update README.md 86b5c57 4 hours ago

2 contributors

73 lines (63 sloc) 15.6 KB Raw Blame History

JHU EN.601.749: Computational Genomics: Applied Comparative Genomics

Prof: Michael Schatz (mschatz @ cs.jhu.edu)
TA: Charlotte Darby (cdarby @ jhu.edu)
Class Hours: Tuesday + Thursday @ 1:30p - 2:45p in Shaffer 304
Schatz Office Hours: Tuesday + Thursday @ 3-4p in Malone 323 and by appointment
Darby Office Hours: Wednesday @ 4pm and by appointment

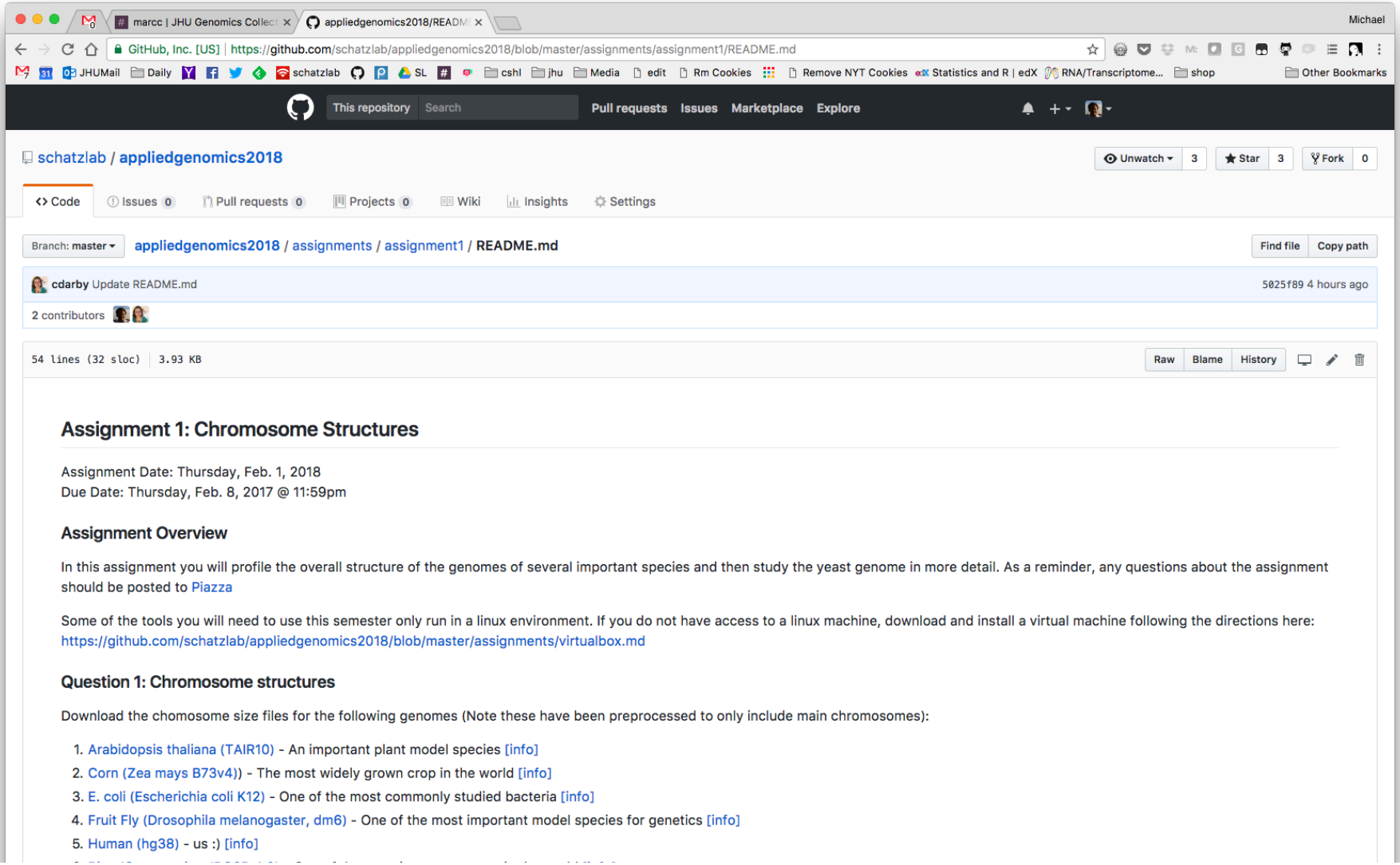
The primary goal of the course is for students to be grounded in theory and leave the course empowered to conduct independent genomic analyses. We will study the leading computational and quantitative approaches for comparing and analyzing genomes starting from raw sequencing data. The course will focus on human genomics and human medical applications, but the techniques will be broadly applicable across the tree of life. The topics will include genome assembly & comparative genomics, variant identification & analysis, gene expression & regulation, personal genome analysis, and cancer genomics. The grading will be based on assignments, a midterm exam, class presentations, and a significant class project. There are no formal course prerequisites, although the course will require familiarity with UNIX scripting and/or programming to complete the assignments and course project.

Prerequisites

- Online introduction to Unix/Linux. Students are strongly recommended to complete one of the following online tutorials (or both) before class begins.
  - Code academy's Intro to Unix
  - Command line bootcamp
  - Rosalind Bioinformatics Programming in Python
  - Minimal Make
- Access to a Linux Machine, and/or Install VirtualBox (Unfortuantely, even Mac will not work correctly for some programs)

https://github.com/schatzlab/appliedgenomics2018

Assignment 1: Chromosome Structures
Due Feb 8 @ 11:59pm



https://github.com/schatzlab/appliedgenomics2018

Piazza

EN. 601.749 | Michael

Secure | https://piazza.com/class/jcumooljtd46p7?cid=6

JHUMail | Daily | schatzlab | SL | cshl | jhu | Media | edit | Rm Cookies | Remove NYT Cookies » | Other Bookmarks

piazza | EN. 601.749 | Q & A | Resources | Statistics | Manage Class | Michael Schatz

hw1 hw2 hw3 hw4 hw5

Unread Updated Unresolved Following

New Post | Search or add a post...

FAVORITES

Instr **Welcome!** 1/29/18
Welcome to JHU EN.601.749: Computational Genomics: Applied Comparative Genomics
Please feel free to ask any questions h

LAST WEEK

Private **Tips & Tricks for a successf...** Thu

note | stop following | 1 views

Welcome!

Welcome to JHU EN.601.749: Computational Genomics: Applied Comparative Genomics

Please feel free to ask any questions here! Also see the course webpage here:
https://github.com/schatzlab/appliedgenomics2018

Good luck!!

Mike & Charlotte

logistics

edit · good note | 0 | Updated 1 minute ago by Michael Schatz

followup discussions *for lingering questions and comments*

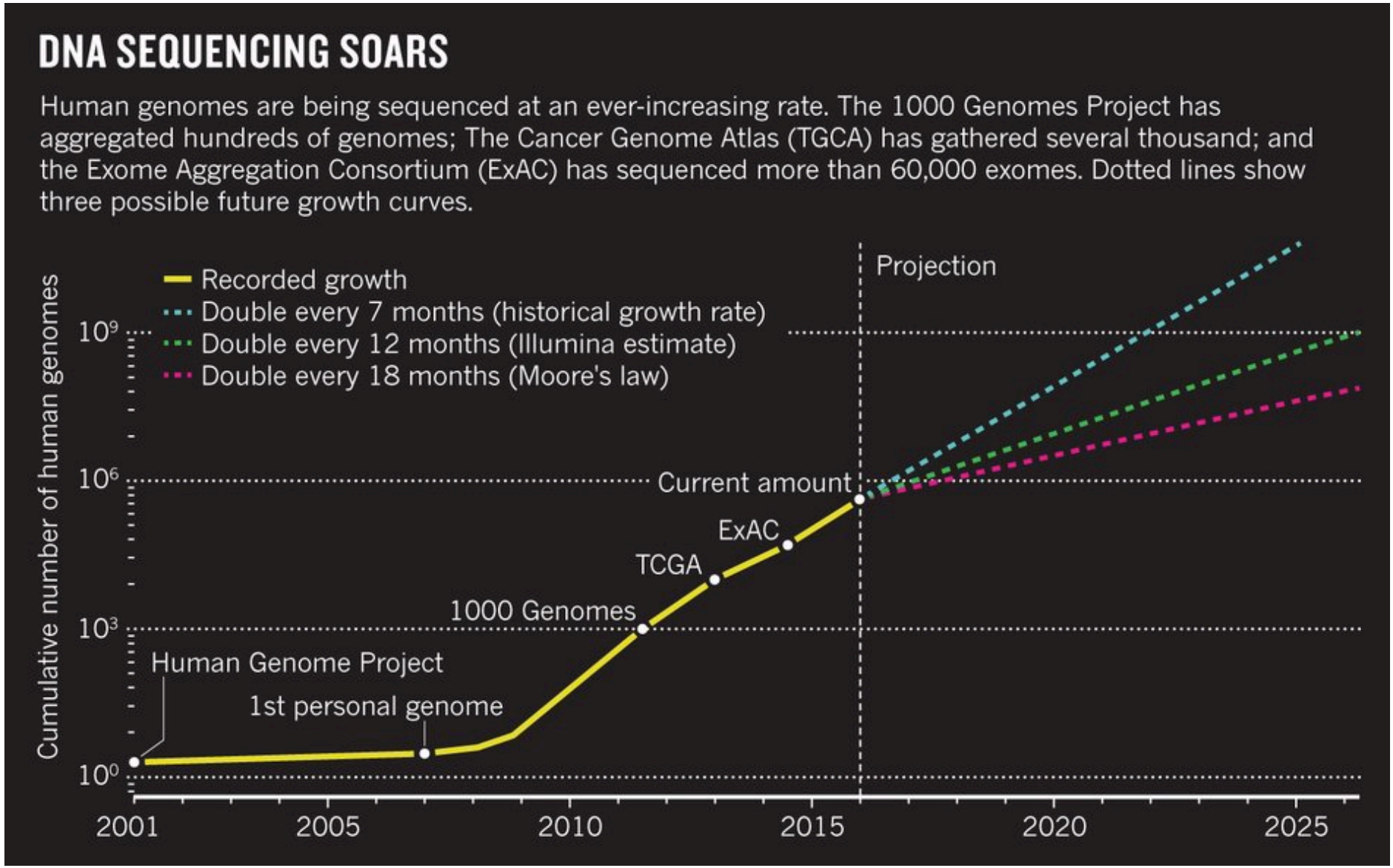
Start a new followup discussion

Compose a new followup discussion

Average Response Time:	Special Mentions:	Online Now	This Week:
N/A	There are no special mentions at this time.	1	4

http://piazza.com/jhu/spring2018/en601749

Sequencing Capacity



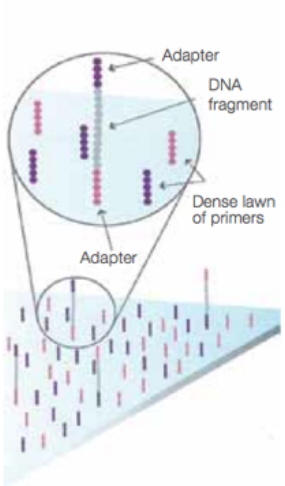
Big Data: Astronomical or Genomical?

Stephens, Z, et al. (2015) PLOS Biology DOI: 10.1371/journal.pbio.1002195

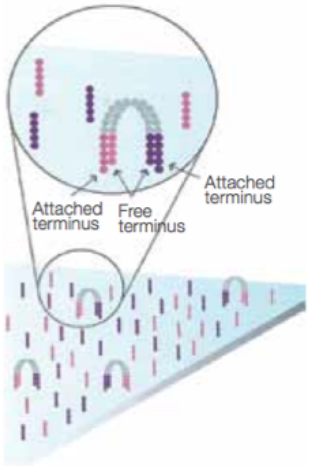
Second Generation Sequencing

Illumina HiSeq 2000
Sequencing by Synthesis

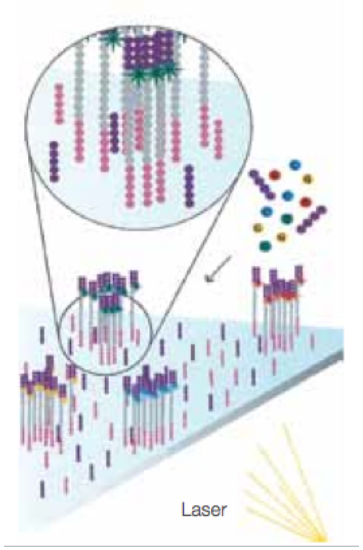
>60Gbp / day



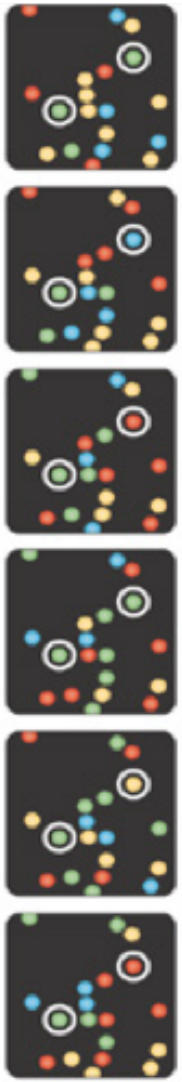
1. Attach



2. Amplify



3. Image



Metzker (2010) Nature Reviews Genetics 11:31-46
https://www.youtube.com/watch?v=fCd6B5HRaZ8

Question?

We would love to generate
longer and longer reads with this technology

What can we do?

Genome Hacking

BIOINFORMATICS ORIGINAL PAPER Vol. 29 no. 12 2013, pages 1492–1497
doi:10.1093/bioinformatics/btt178

Genome analysis Advance Access publication May 22, 2013

Assembling the 20 Gb white spruce (*Picea glauca*) genome from whole-genome shotgun sequencing data

Inanc Birol[1,2,3,*], Anthony Raymond[1], Shaun D. Jackman[1], Stephen Pleasance[1], Robin Coope[1], Greg A. Taylor[1], Macaire Man Saint Yuen[4], Christopher I. Keeling[4], Dana Brand[1], Benjamin P. Vandervalk[1], Heather Kirk[1], Pawan Pandoh[1], Richard A. Moore[1], Yongjun Zhao[1], Andrew J. Mungall[1], Barry Jaquish[5], Alvin Yanchuk[5], Ca
Brian Boyle[7], Jean Bousquet[7,8], Kermit Ritland[6], John MacKay[7,8], Jörg B
Steven J.M. Jones[1,2,9]

[1]Genome Sciences Centre, British Columbia Cancer Agency, Vancouver, BC V5Z 4S6, Canada,
Genetics, University of British Columbia, Vancouver, BC V6H 3N1, Canada, [3]School of Compu
Fraser University, Burnaby, BC V5A 1S6, Canada, [4]Michael Smith Laboratories, University of B
Vancouver, BC V6T 1Z4, Canada, [5]British Columbia Ministry of Forests, Lands and Natural Re
Victoria, BC V8W 9C2, Canada, [6]Department of Forest Sciences, University of British Columbi
1Z4, Canada, [7]Institute for Systems and Integrative Biology, Université Laval, Québec, QC G1
[8]Department of Wood and Forest Sciences, Université Laval, Québec, QC G1V 0A6, Canada
Molecular Biology and Biochemistry, Simon Fraser University, Burnaby, BC V5A 1S6, Canada

Associate Editor: Michael Brudno

ABSTRACT

White spruce (*Picea glauca*) is a dominant conifer of the boreal forests of North America, and providing genomics resources for this commercially valuable tree will help improve forest management and conservation efforts. Sequencing and assembling the large and highly repetitive spruce genome though pushes the boundaries of the current technology. Here, we describe a whole-genome shotgun sequencing strategy using two Illumina sequencing platforms and an assembly approach using the ABySS software. We report a 20.8 giga base pairs draft genome in 4.9 million scaffolds, with a scaffold N50 of 20 356 bp. We demonstrate how recent improvements in the sequencing technology, especially increasing read lengths and paired end reads from longer fragments have a major impact on the assembly contiguity. We also note that scalable bioinformatics tools are instrumental in providing rapid draft assemblies.

Availability: The *Picea glauca* genome sequencing and assembly data are available through NCBI (Accession#: ALWZ01000000000 PID: PRJNA83435). http://www.ncbi.nlm.nih.gov/bioproject/83435.

Contact: ibirol@bcgsc.ca

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on March 20, 2013; revised on April 10, 2013; accepted on April 11, 2013

1 INTRODUCTION

The assembly of short reads to develop genomic resources for non-model species remains an active area of development (Schatz *et al.*, 2012). The feasibility of the approach and its scalability to large genomes was demonstrated b
(Simpson *et al.*, 2009) using human
and was later used to assemble the
SOAPdenovo tool (Li *et al.*, 2010).
high quality results, as demonstrate
et al., 2012; Ladner *et al.*, 2013; Rib
been successfully applied numerous ti
omes (Chan *et al.*, 2011; Chu *et al.*, 2
2011; Godel *et al.*, 2012; Swart *et al.*,
Estimated at 20 giga base pairs (Gi
cing and assembly of the genome of t
the pine (*Pinaceae*) family present uni
generation end, those challenges inclu
whole-genome shotgun sequencing da
ing reduced representation resources
tude of the problem. On the bioin
massive sequencing datasets is extrem
ing cycles, memory usage, storage req
programming implementations on co
We addressed the data representati
and sequencing multiple whole-genom
HiSeq 2000 and MiSeq sequencers f
CA, USA). Compared with localized
as building and sequencing fosmi
approach of isolating ~10 kb DNA s
sequencing fragments in high through
CA, USA), a shotgun only sequencing
sequence data effectively covering th
that can be an order of magnitude le
especially substantial when sequencin
In this work, we demonstrate that
at this scale remains viable and pro

*To whom correspondence should be addressed.

assemble the spruce genome, we used the ABySS algorithm (Simpson *et al.*, 2009), which captures a representation of read-to-read overlaps by a distributed de Bruijn graph and uses parallel computations to build the target genome. The modular nature of the tool allowed us to execute a large number of tests to tune the message passing interface for a successful execution, train the assembly parameters for an optimal assembly and quantify the utility of long reads for large genome assemblies. To the best of our knowledge, the ABySS algorithm is unique in its ability to enable genome assemblies of this scale using whole-genome shotgun sequencing data.

2 METHODS

2.1 Sample collection

Apical shoot tissues were collected in April 2006 from a single white spruce (*Picea glauca*, genotype PG29) tree at the Kalamalka Research Station of the British Columbia Ministry of Forests and Ranges, Vernon, British Columbia, Canada. Genomic DNA was extracted from 60 gm tissue by BioS&T (http://www.biost.com/, Montreal, QC, Canada) using an organelle exclusion method yielding 300 μg of high quality purified nuclear DNA.

2.2 Library preparation and sequencing

DNA quality was assessed by spectrophotometry and gel electrophoresis before library construction. DNA was sheared for 45 s using an E210 sonicator (Covaris) and then analysed on 8% PAGE gels. The 200–300 bp (for libraries with 250 bp insert size) or 450–550 bp (for libraries with 500 bp insert size) DNA size fractions were excised and eluted from the gel slices overnight at 4°C in 300 μl of elution buffer {5:1 [vol/vol] LoTE buffer [3 mM Tris–HCl (pH 7.5), 0.2 mM EDTA]/7.5 M ammonium acetate} and was purified using a Spin-X Filter Tube (Fisher Scientific) and ethanol precipitation. Genome libraries were prepared using a modified paired-end tag (PET) protocol supplied by Illumina Inc. This involved DNA end repair and formation of 3′ adenosine overhangs using the Klenow fragment of DNA polymerase I (3′–5′ exonuclease minus) and ligation to Illumina PE adapters (with 5′ overhangs). Adapter-ligated products were purified on QIAquick spin columns (Qiagen) and amplified using Phusion DNA polymerase (NEB) and 10 PCR cycles with the PE primer 1.0 and 2.0 (Illumina). PCR products of the desired size range were purified from adapter ligation artifacts using 8% PAGE gels. DNA quality was assessed and quantified using an Agilent DNA 1000 series II assay (Agilent) and Nanodrop 7500 spectrophotometer (Nanodrop). DNA was subsequently diluted to 8 nM. The final concentration was confirmed using a Quant-iT dsDNA HS assay kit and Qubit fluorometer (Invitrogen).

The mate pair (MPET, a.k.a. jumping) libraries were constructed using 4 μg of genomic DNA with the Illumina Nextera Mate-Pair library construction protocol and reagent (FC-132-1001). The genomic DNA sample was simultaneously fragmented and tagged with a biotin containing mate pair junction adapter, which left a short sequence gap in the tagmented DNA. The gap was filled by a strand displacement reaction using a polymerase to ensure that all fragments were flush and ready for circularization. After an AMPure Bead cleanup, size selection was done on a 0.6% agarose gel to excise 6–9 kb and 9–13 kb fractions, which were purified using a Zymoclean Large Fragment DNA Recovery Kit. The fragments were circularized by ligation, followed by a digestion to remove any linear molecules and left circularized DNA for shearing. The sheared DNA fragments that contain the biotinylated junction adapter (mate pair fragments) were purified by means of binding to streptavidin magnetic beads, and the unwanted unbiotinylated molecules were washed away. The DNA fragments were then end repaired and A-tailed following the protocol and ligated to indexed TruSeq adapters. The final library was enriched by a 10-cycle PCR and purified by AMPure bead clean-up. Library quality and size were assessed by Agilent DNA 1000 series II assay and KAPA Library Quantification protocol. The two fractions were pooled for sequencing paired end 100 bp using Illumina HiSeq2000.

The construction of the 12 kb mate pair libraries was achieved by a hybrid 454/Illumina procedure. Briefly, 50 μg of genomic DNA was fra
mented for 20 cycles at speed code 12 using a Hydrosh
Marlborough, MA) equipped with a large assembly modu
mented DNA was loaded on a 1% agarose gel, and fragmen
18 kb were extracted. Biotinylated circularization adapters
Titanium Paired-end Adaptor set (454 Life Sciences/Roch
CT) were added to ends of the gel-extracted fragments.
recombination of the ends was performed with Cre recom
England Biolabs, Ipswich, MA), and linear molecules rema
tion were removed with Plasmid Safe (Epicentre, Madison,
molecules were fragmented using GS Rapid Library Nebuli
Sciences/Roche, Branford, CT), and fragment end-repair fo
tailing was performed with the GS Rapid Libray preparation
Sciences/Roche, Branford, CT). TruSeq Adaptors (Illumina
CA) were ligated to the repaired/A-tailed ends. Biotinylat
were enriched using Streptavidin-coupled Dynabeads (Life T
Grand Island, NY) and amplified by PCR using Illu
primers.

Random bacterial artificial chromosome (BAC) seq
performed using DNA from the same genotype on 4
Titanium with 6 kb paired-end libraries at the PlateForm
Génomiques of the Institute for Systems and Integra
(Université Laval, Quebec City, QC). A single paired-end
prepared on a pool of 15 BACs (equimolar concentrations)
earlier in the text with the following modifications: 15 μg
fragmented using a Hydroshear with a standard assembly
at speed code 18, 6–10 kb fragments were extracted from
GS-FLX library adaptors were ligated to the repaired/A
ments. GS-FLX sequencing using the titanium chemistry w
according to manufacturer's instructions (454 Life Sci
Branford, CT). Sanger sequencing method was used to ob
BAC sequencing data as previously described (Hamberger
Keeling *et al.*, 2010).

2.3 MiSeq modification

In sequencing the spruce genome, we generated longer rea
modifying the MiSeq platform. The MiSeq uses a clamshell s
(Supplementary Fig. S1A) to hold reagent tubes in an arra
cessed by the MiSeq's sippers. Most of the reagents are u
length independent steps such as denaturation and cluster ge
three reagents, the Scan, Cleavage and Incorporation mi
sumed at each cycle. Although the MiSeq allows any read
specified in the control software, the reagent cartridge canno
during the run without stopping it. Increasing the read len
requires increasing the quantity of the length-dependent
cartridge. This led to the solution of combining the leng
reagents of two kits into one.

A tool was designed that opens the snap-hook latches
cartridge together (Supplementary Figs S1B and S2), giv
the reagent tubes, yet allowing the cartridge to be put b
without damage to its components (Supplementary Fi
40 ml, the stock length-dependent reagent containers allo
imum of ~650 cycles in total. To maximize the potential of t
kit approach, a new reagent tray with 70 ml wells was d
placed in a modified clamshell base.

Assembling the 20 Gb white spruce (*Picea glauca*) genome from whole-genome shotgun sequencing data

Inanc Birol[1,2,3,*], Anthony Raymond[1], Shaun D Jackman[1], Stephen Pleasance[1], Robin Coope[1], Greg A Taylor[1], Macaire Man Saint Yuen[4], Christopher I Keeling[4], Dana Brand[1], Benjamin P Vandervalk[1], Heather Kirk[1], Pawan Pandoh[1], Richard A Moore[1], Yongjun Zhao[1], Andrew J Mungall[1], Barry Jaquish[5], Alvin Yanchuk[5], Carol Ritland[4,6], Brian Boyle[7], Jean Bousquet[7,8], Kermit Ritland[6], John MacKay[7,8], Jörg Bohlmann[4,6], Steven JM Jones[1,2,9]

[1] British Columbia Cancer Agency, Genome Sciences Centre, Vancouver, BC V5Z 4S6
[2] University of British Columbia, Department of Medical Genetics, Vancouver, BC V6H 3N1
[3] Simon Fraser University, School of Computing Science, Burnaby, BC V5A 1S6
[4] University of British Columbia, Michael Smith Laboratories, Vancouver, BC V6T 1Z4
[5] British Columbia Ministry of Forests, Lands and Natural Resource Operations, Victoria, BC V8W 9C2
[6] University of British Columbia, Department of Forest Sciences, Vancouver, BC V6T 1Z4
[7] Université Laval, Institute for Systems and Integrative Biology, Québec, QC G1V 0A6
[8] Université Laval, Department of Wood and Forest Sciences, Québec, QC G1V 0A6
[9] Simon Fraser University, Department of Molecular Biology and Biochemistry, Burnaby, BC V5A 1S6

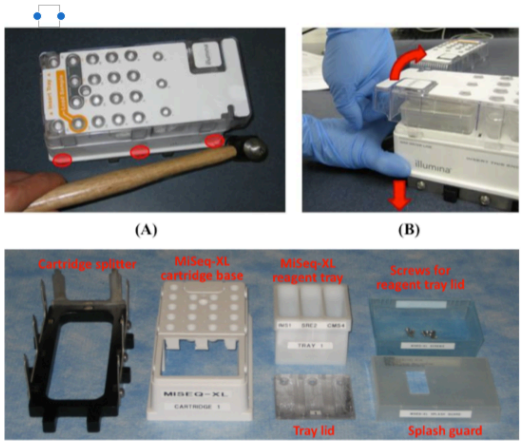
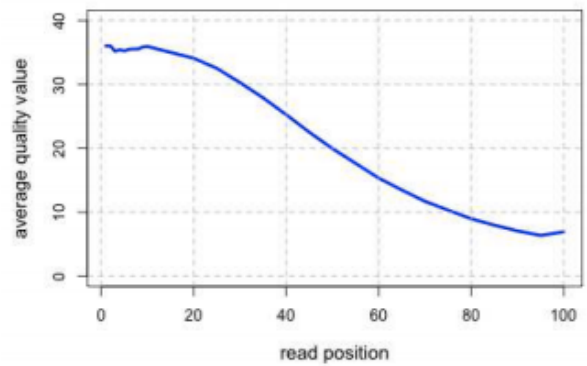


Figure S1. Modification of the MiSeq cartridge. MiSeq reagent cartridge was modified to allow for longer read lengths. **(A, B)** Opening of the clamshell style cartridge. **(C)** Contents of the modified cartridge. This was initially used to combine two PE150 kits for PE300 runs. When Illumina introduced the P250 kit, the same apparatus was used to enable PE500 runs.

Illumina Quality

QV	p_{error}
40	1/10000
30	1/1000
20	1/100
10	1/10

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

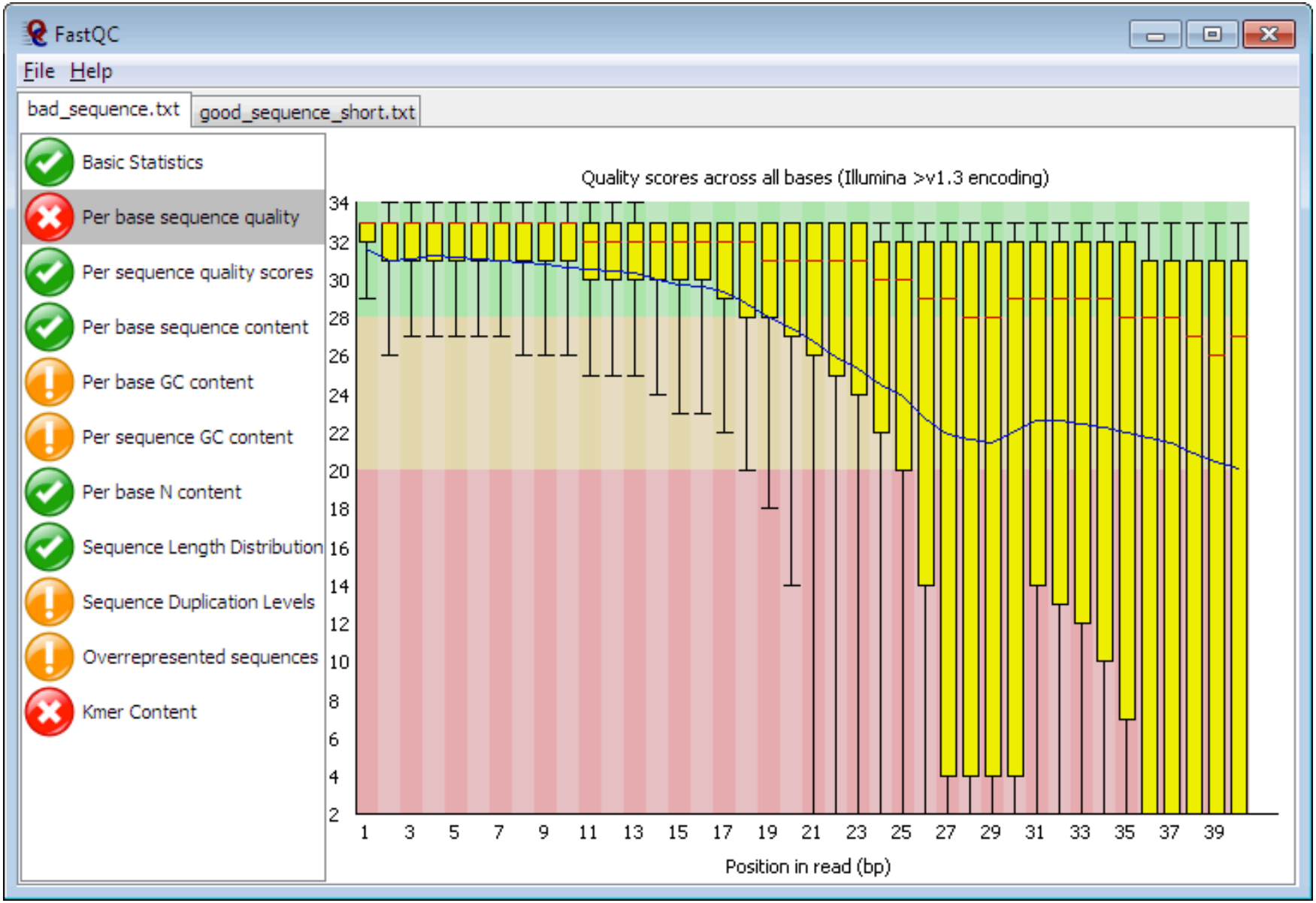



```
 SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.........................................................
 ..........................XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX......................
 ...............................IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII......................
 ..................................JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ......................
 LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL.........................................................
 !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
 |                         |    |        |                              |                     |
33                        59   64       73                            104                   126

S - Sanger        Phred+33,  raw reads typically (0, 40)
X - Solexa        Solexa+64, raw reads typically (-5, 40)
I - Illumina 1.3+ Phred+64,  raw reads typically (0, 40)
J - Illumina 1.5+ Phred+64,  raw reads typically (3, 40)
   with 0=unused, 1=unused, 2=Read Segment Quality Control Indicator (bold)
   (Note: See discussion above).
L - Illumina 1.8+ Phred+33,  raw reads typically (0, 41)
```

http://en.wikipedia.org/wiki/FASTQ_format

FASTQC: Is my data any good?



http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

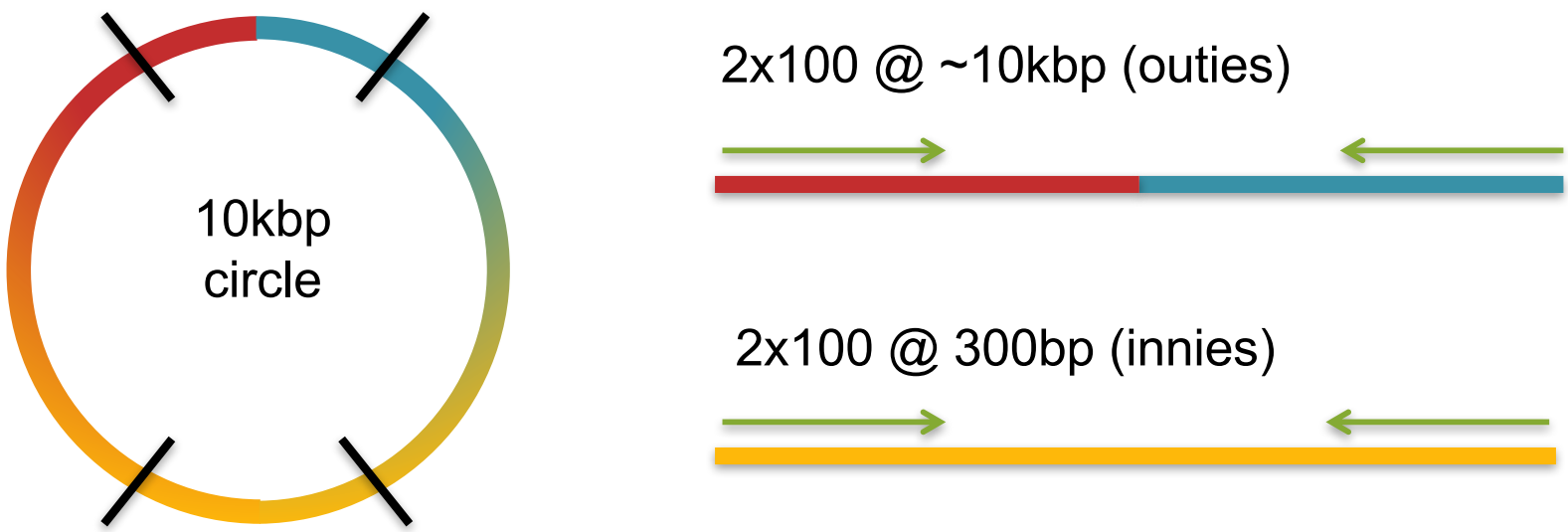
Paired-end and Mate-pairs

Paired-end sequencing

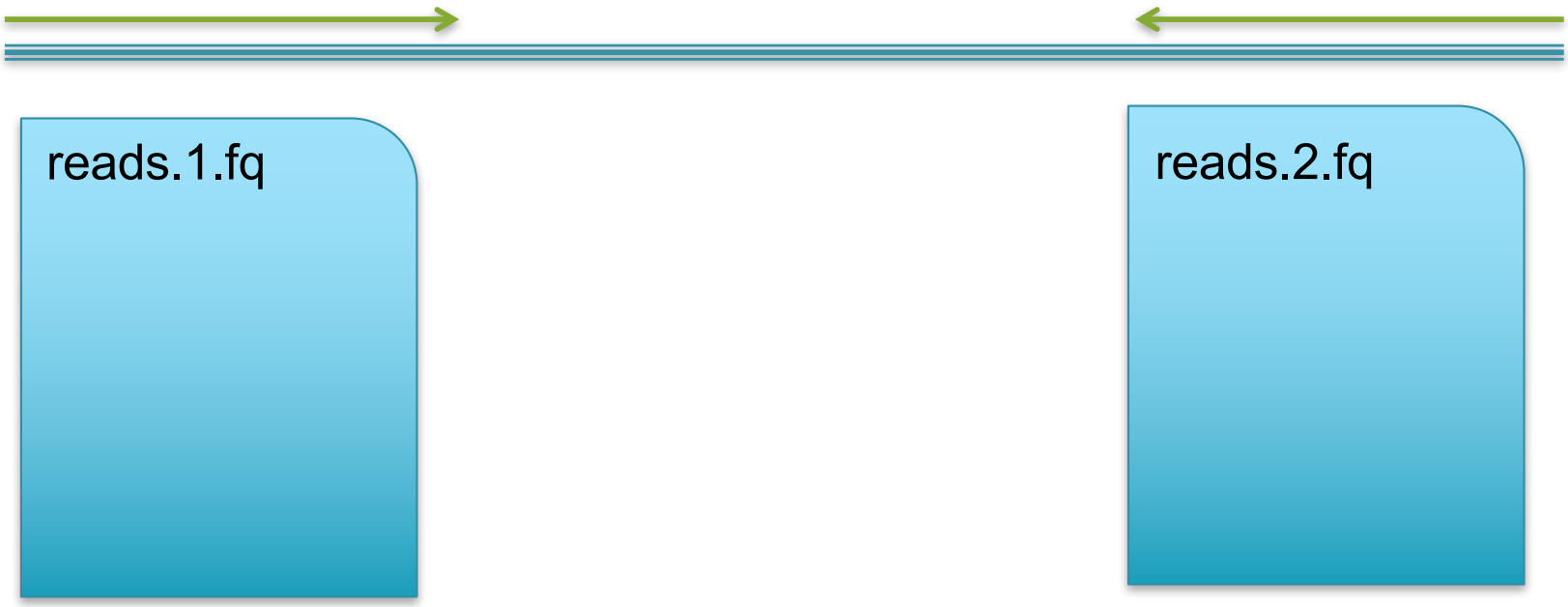
- Read one end of the molecule, flip, and read the other end
- Generate pair of reads separated by up to 500bp with inward orientation

Mate-pair sequencing

- Circularize long molecules (1-10kbp), shear into fragments, & sequence
- Mate failures create short paired-end reads



FASTQ Files



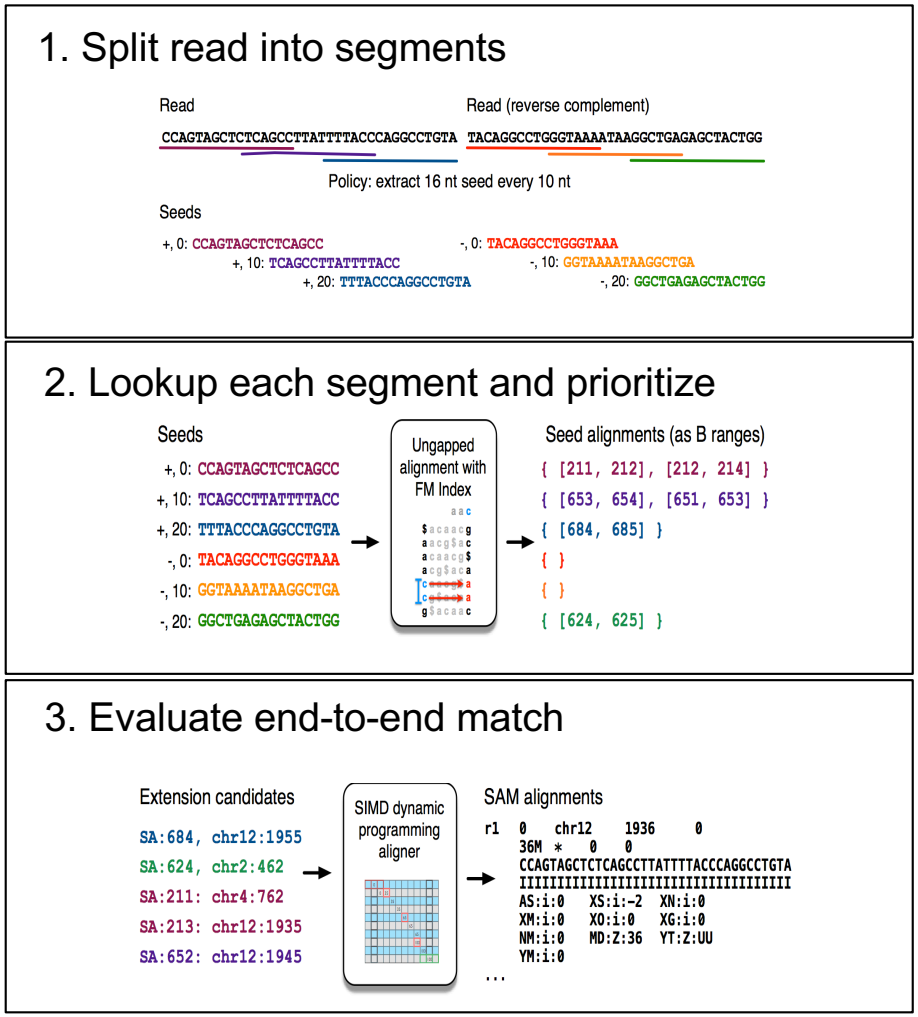

```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

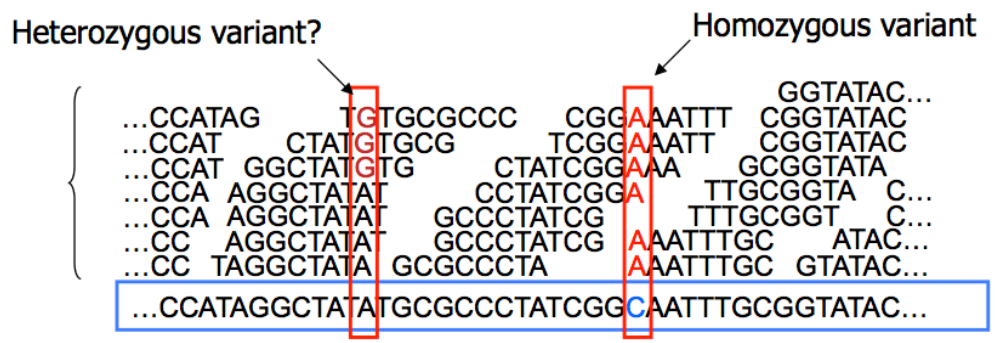
@Identifier
Sequence
+Separator
Quality Values
…

http://en.wikipedia.org/wiki/FASTQ_format

Assembly, Mapping & Genotyping

Week 2/3/4



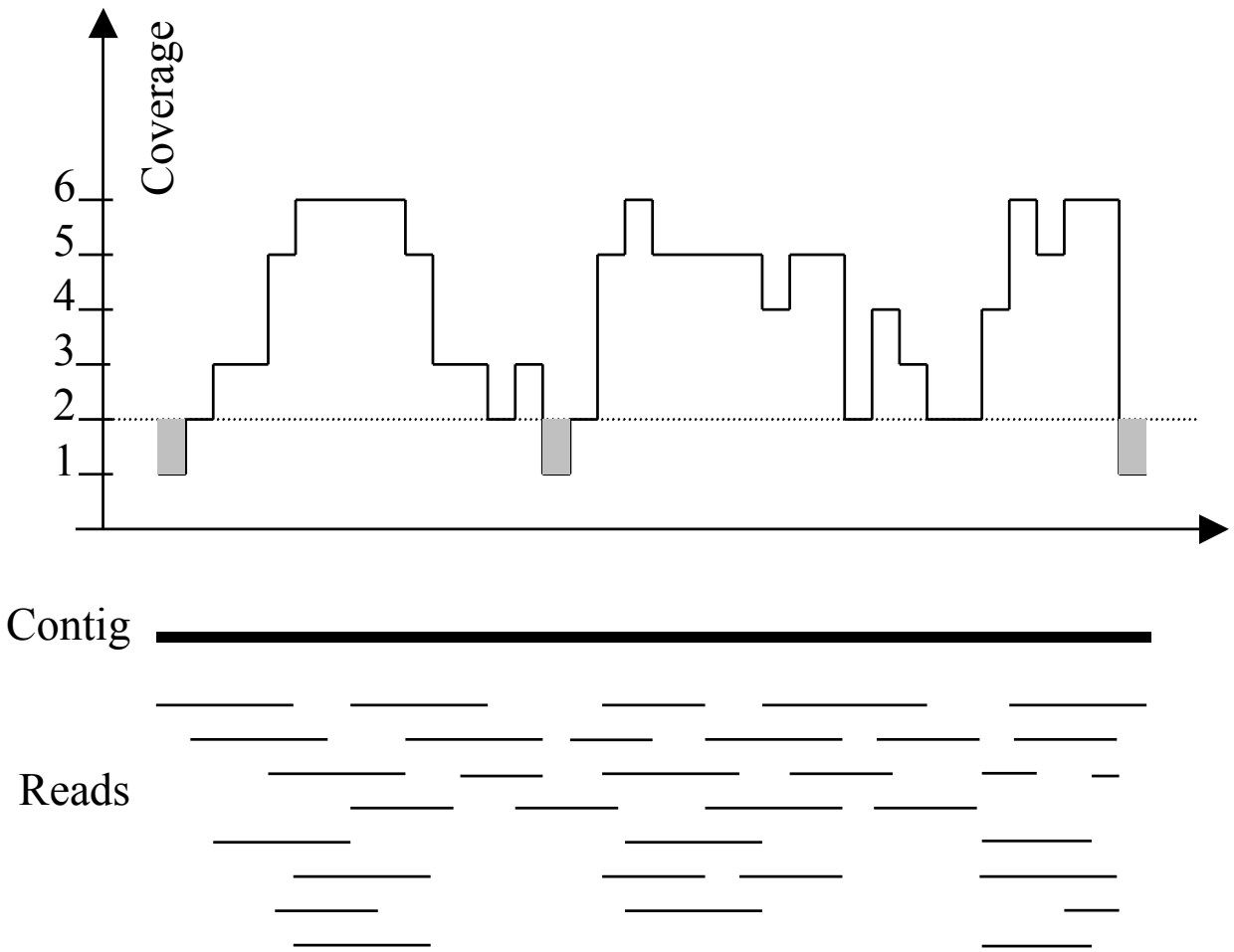


- Distinguishing SNPs from sequencing error typically a likelihood test of the coverage
  - Hardest to distinguish between errors and heterozygous SNP.
  - Coverage is the most important factor!
    - Target at least 10x, 30x more reliable

Fast gapped-read alignment with Bowtie 2
Langmead & Salzberg. (2012) *Nature Methods*. 9:357-359.

The Sequence Alignment/Map format and SAMtools
Li H et al. (2009) *Bioinformatics*. 25:16 2078-9

Typical sequencing coverage

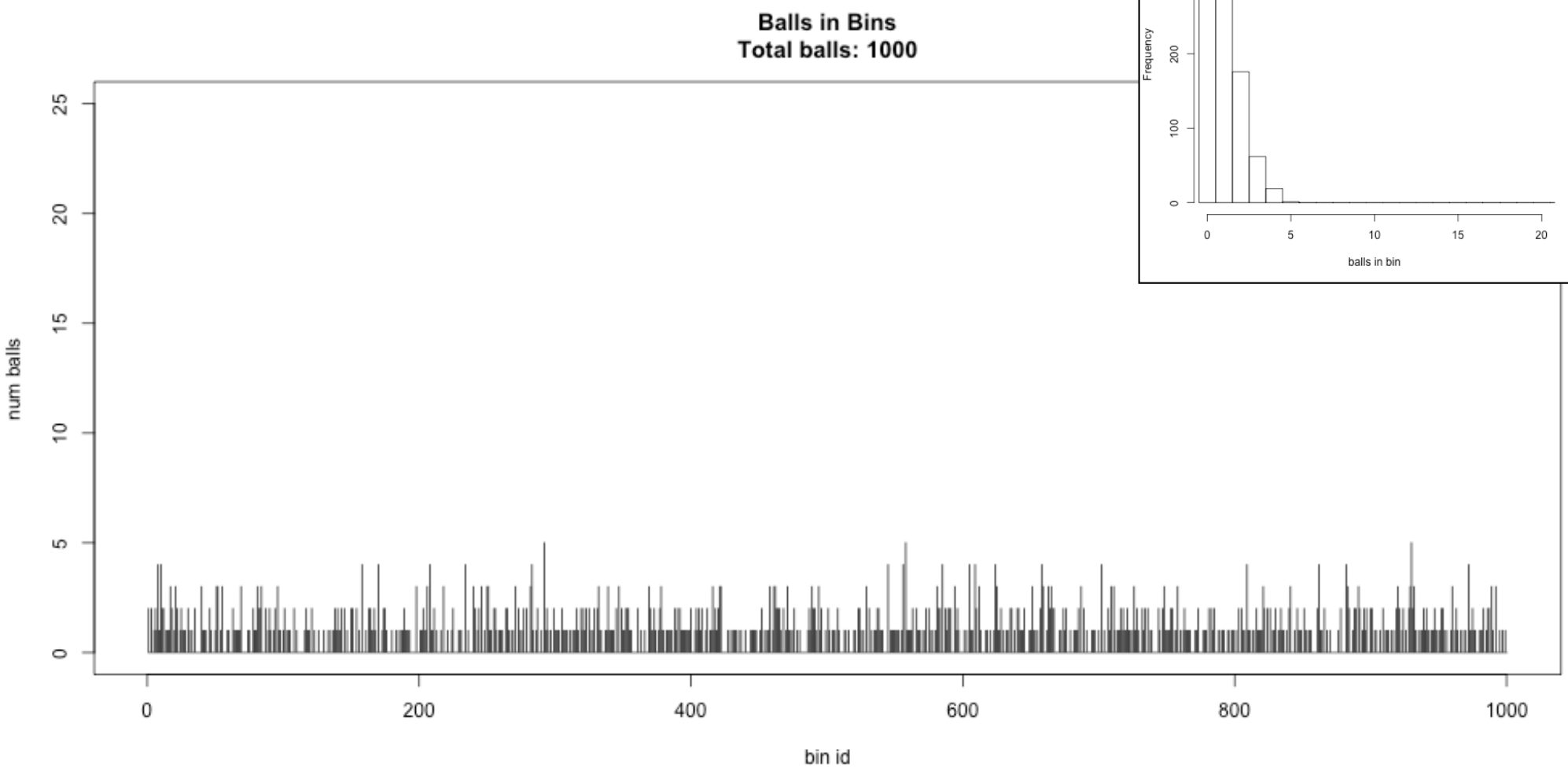


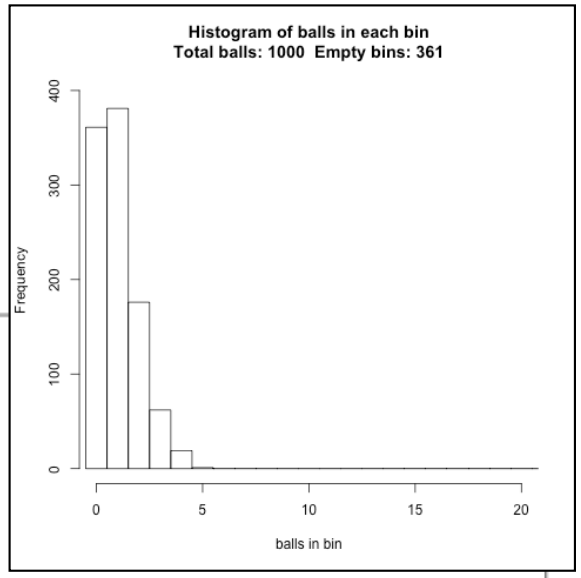
Imagine raindrops on a sidewalk

We want to cover the entire sidewalk but each drop costs $1

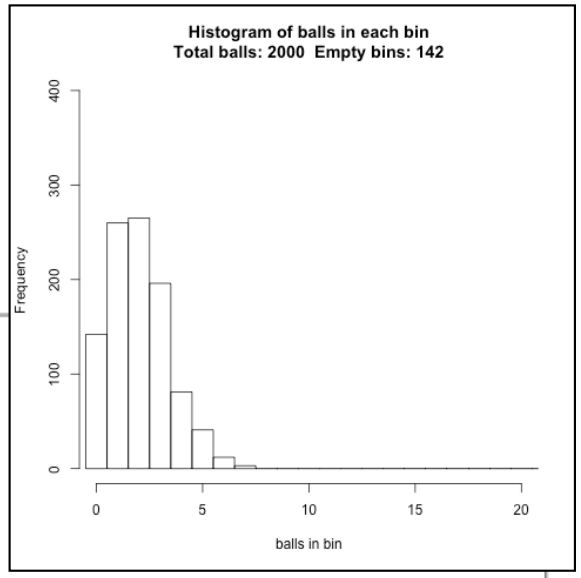
If the genome is 10 Mbp, should we sequence 100k 100bp reads?

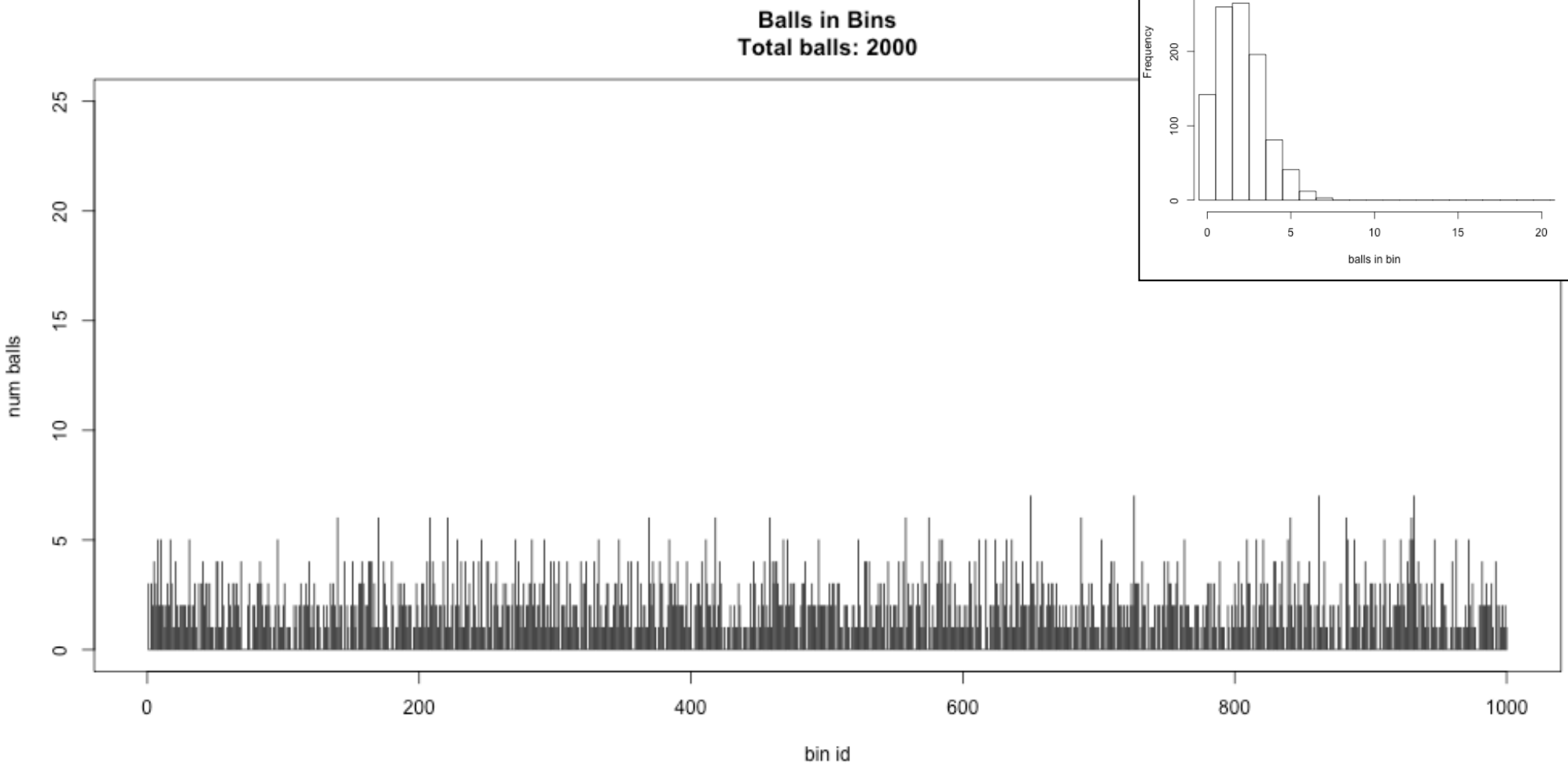
1x sequencing



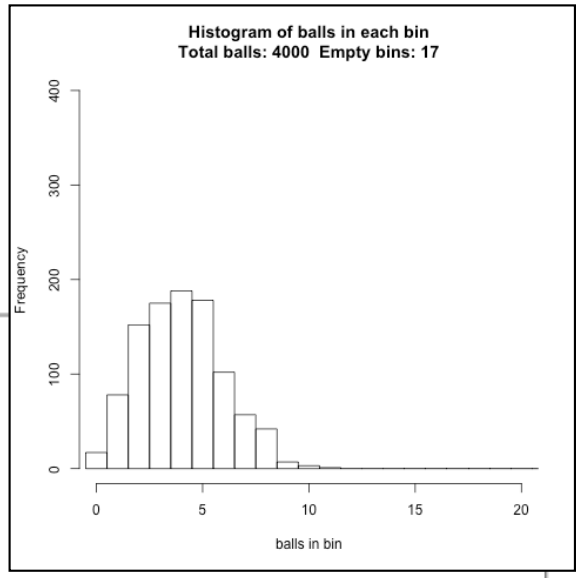


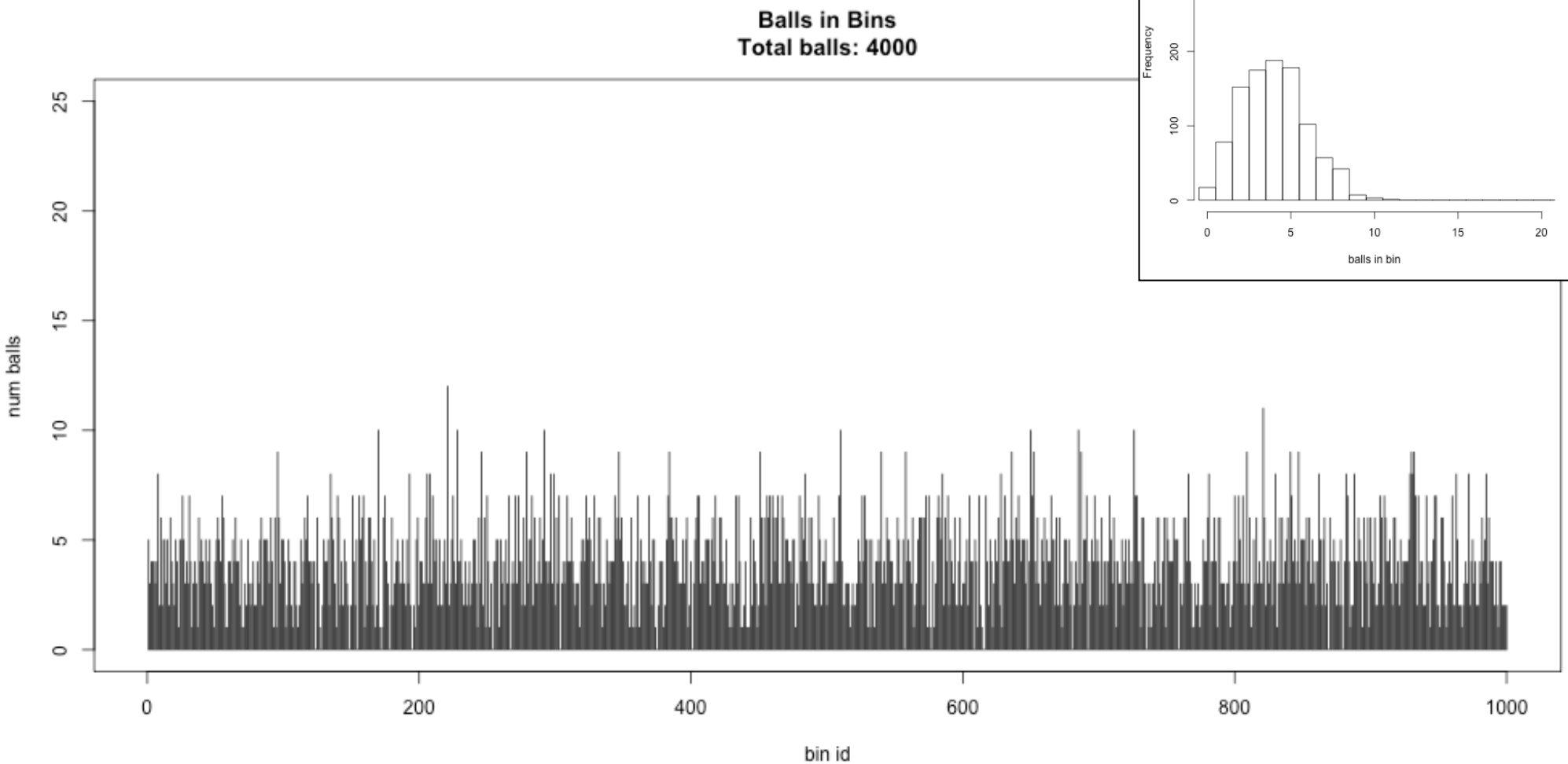
2x sequencing



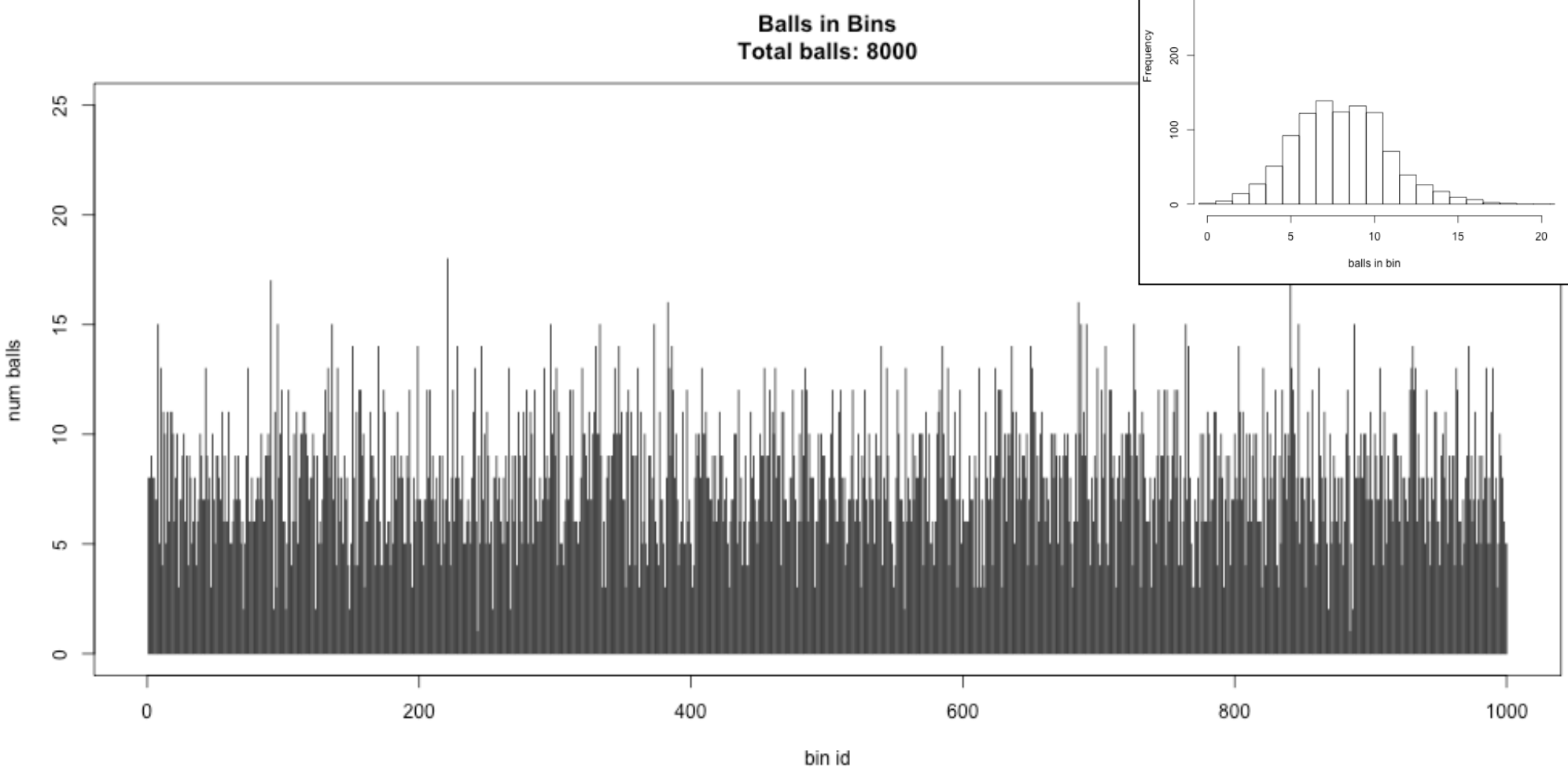


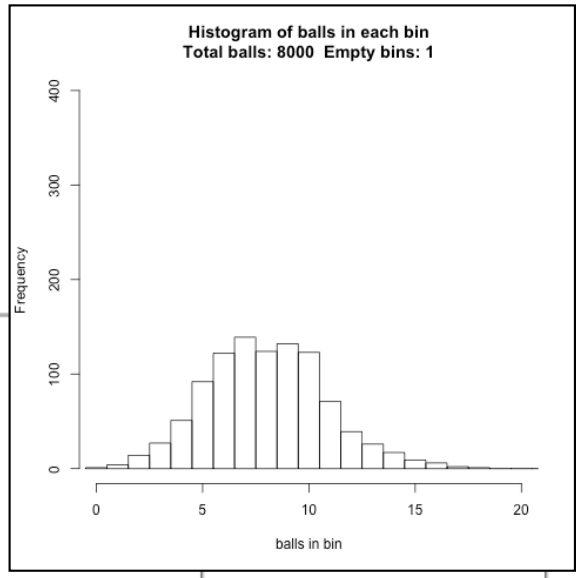
4x sequencing





8x sequencing





Poisson Distribution

The probability of a given number of events occurring in a fixed interval of time and/or space if these events occur with a known average rate and independently of the time since the last event.

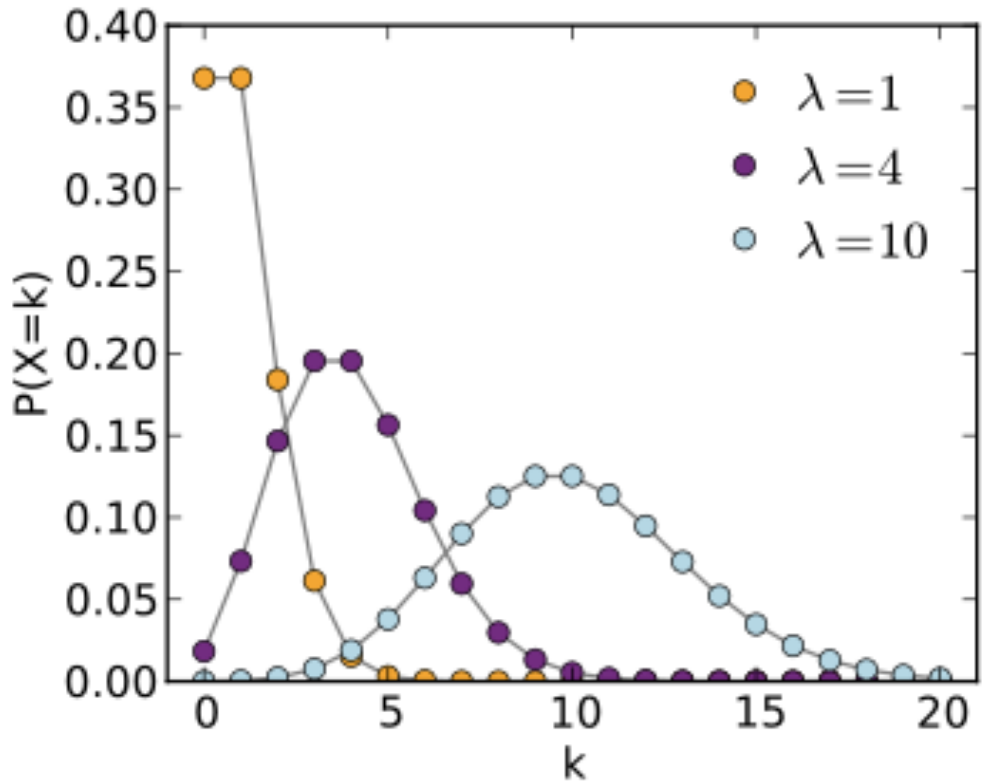
$$P(k) = \frac{\lambda^k}{k!} e^{-\lambda}$$

Formulation comes from the limit of the binomial equation

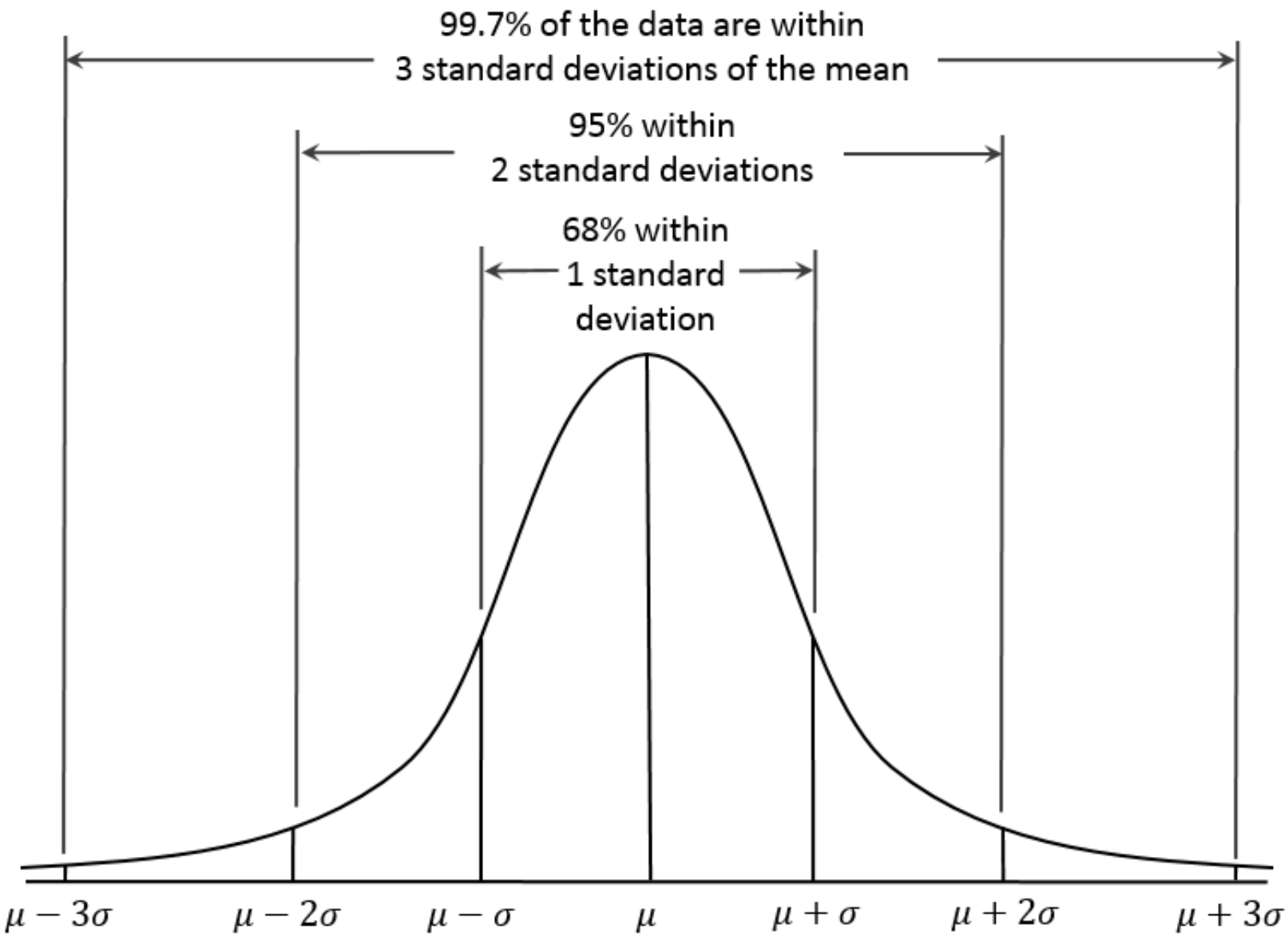
Resembles a normal distribution, but over the positive values, and with only a single parameter.

Key properties:

- ***The standard deviation is the square root of the mean.***
- ***For mean > 5, well approximated by a normal distribution***



Normal Approximation



Can estimate Poisson distribution as a normal distribution when $\lambda > 10$

Pop Quiz!

I want to sequence a 10Mbp genome to 24x coverage.
How many 120bp reads do I need?

I need 10Mbp x 24x = 240Mbp of data
240Mbp / 120bp / read = 2M reads

I want to sequence a 10Mbp genome so that
>97.5% of the genome has at least 24x coverage.
How many 120bp reads do I need?

Find X such that X-2*sqrt(X) = 24

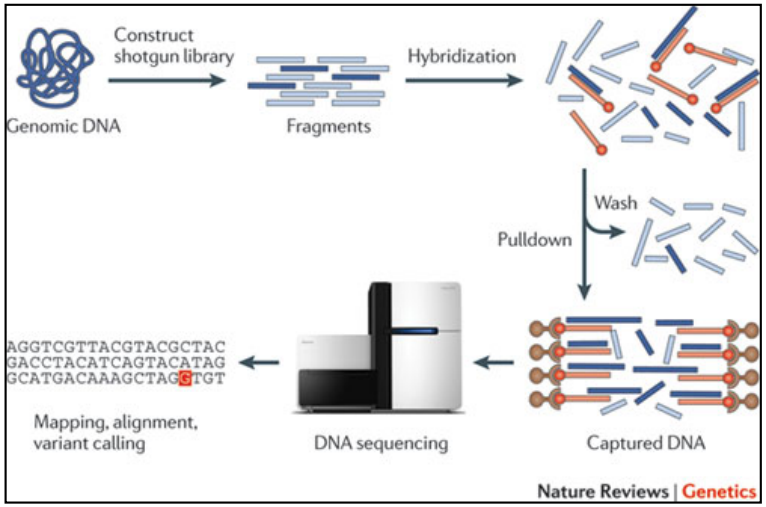
36-2*sqrt(36) = 24

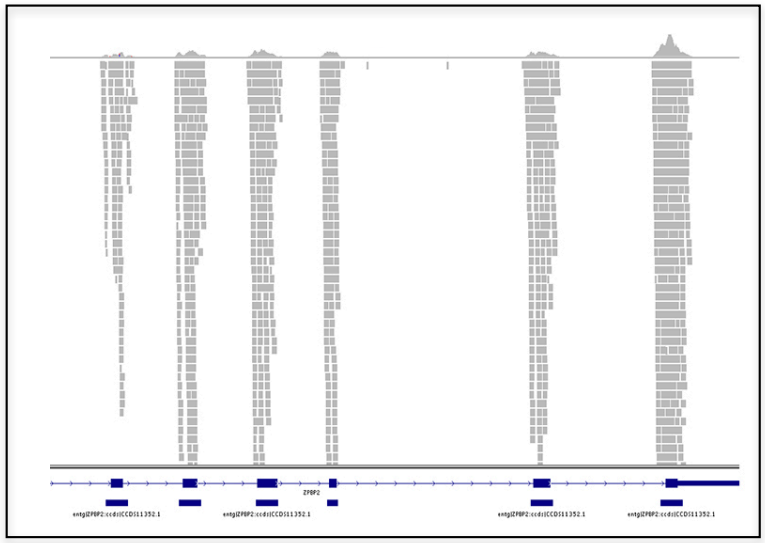
I need 10Mbp x 36x = 360Mbp of data
360Mbp / 120bp / read = 3M reads

Exome-Capture Sequencing

Exome-capture reduces the costs of sequencing

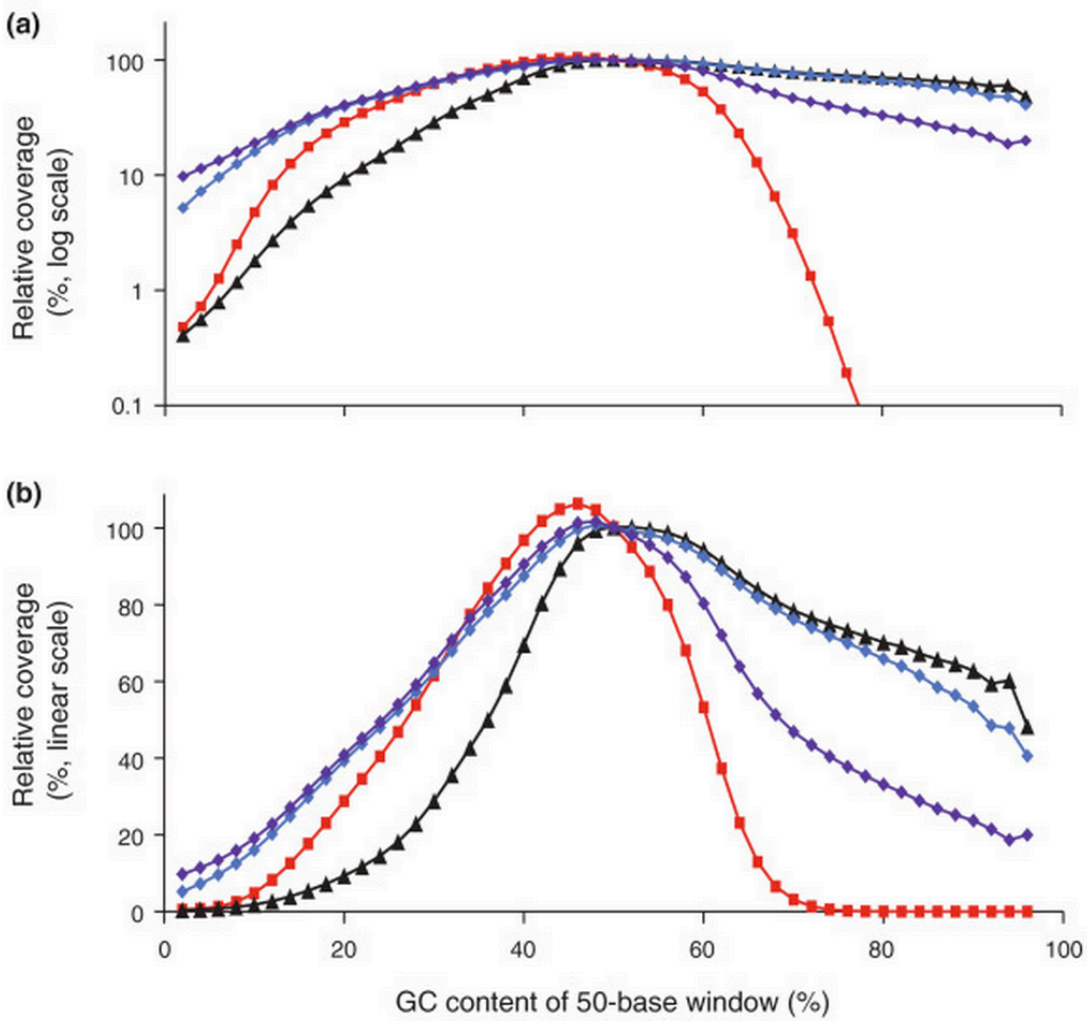
- Currently targets around 50Mbp of sequence: all exons plus flanking regions
- WGS currently costs ~$1200 per sample, while WES currently costs ~$300 per sample
- Coverage is highly localized around genes, although will get sparse coverage throughout rest of genome





Exome sequencing as a tool for Mendelian disease gene discovery
Bamshad et al. (2011) *Nature Reviews Genetics*. 12, 745-755

Beware of GC Biases

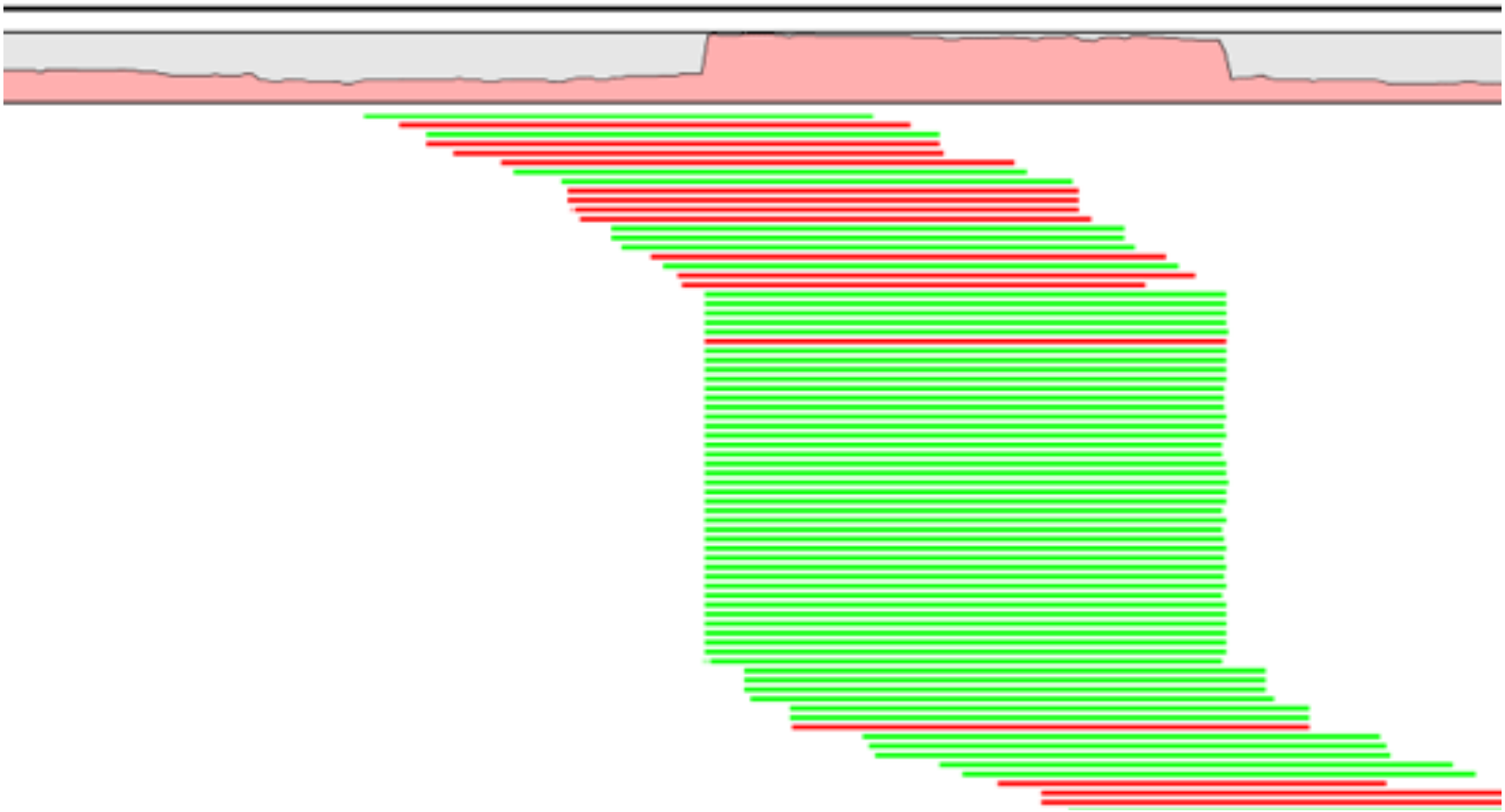


Illumina sequencing does not produce uniform coverage over the genome

- Coverage of extremely high or extremely low GC content will have reduced coverage in Illumina sequencing
- Biases primarily introduced during PCR; lower temperatures, slower heating, and fewer rounds minimize biases
- This makes it very difficult to identify variants (SNPs, CNVs, etc) in certain regions of the genome

Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries.
Aird et al. (2011) *Genome Biology*. 12:R18.

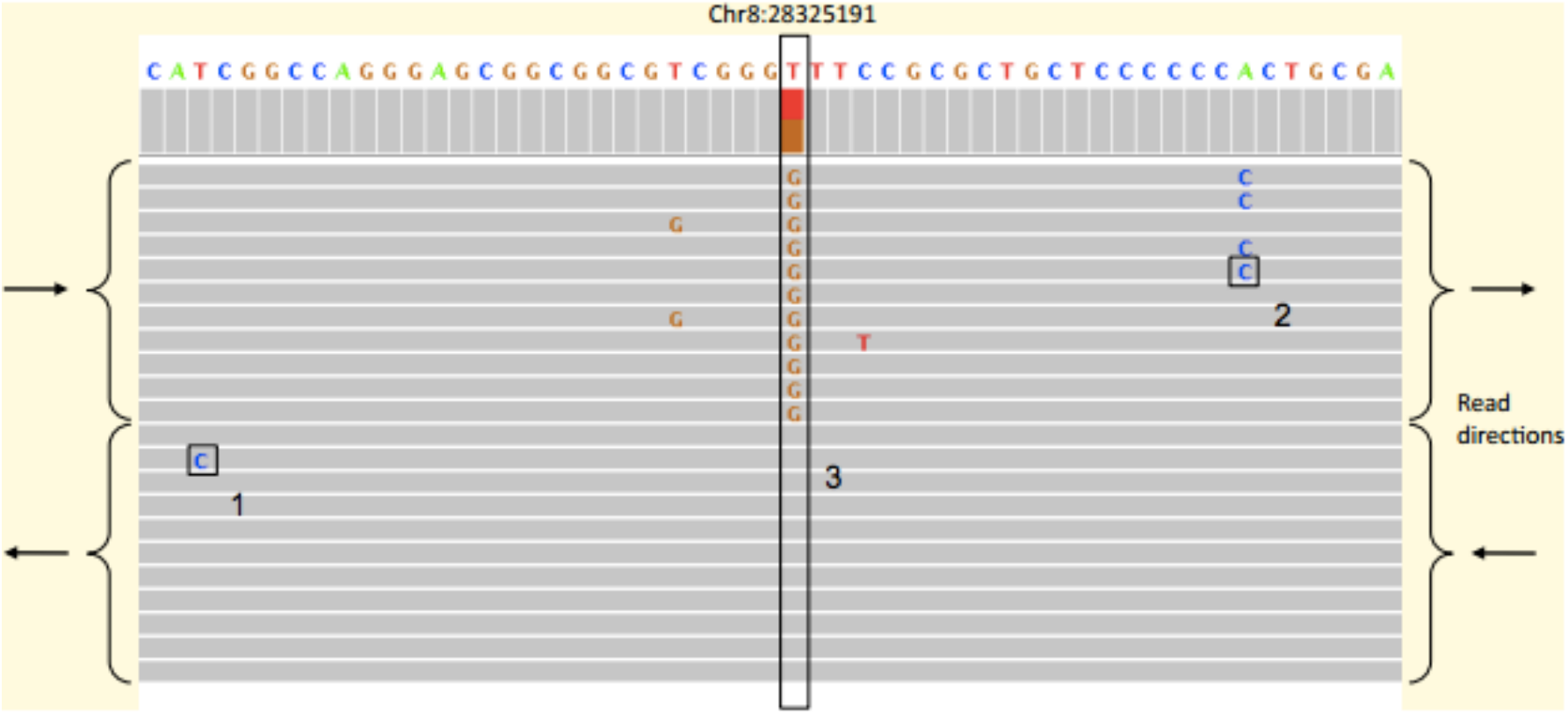
Beware of Duplicate Reads



The Sequence alignment/map (SAM) format and SAMtools.
Li et al. (2009) *Bioinformatics.* 25:2078-9

Picard: http://picard.sourceforge.net

Beware of (Systematic) Errors



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

Illumina Sequencing Summary

Advantages:

- Best throughput, accuracy and read length for any 2nd gen. sequencer
- Fast & robust library preparation

Disadvantages:

- Inherent limits to read length (practically, 150bp)
- Some runs are error prone
- Requires amplification, sequences a population of molecules

Illumina HiSeq
~3 billion paired 100bp reads
~600Gb, $10K, 8 days
(or "rapid run" ~90Gb in 1-2 days)

Illumina X Ten
~6 billion paired 150bp reads
1.8Tb, <3 days, ~1000 / genome($$)
(or "rapid run" ~90Gb in 1-2 days)

Illumina NextSeq
One human genome in **<30 hours**

Ira Hall

Next Steps

1. Reflect on the magic and power of DNA ☺
2. Check out the course webpage
3. Register on Piazza
4. Work on Assignment 1
    1. Set up Linux, set up Virtual Machine
    2. Set up Dropbox for yourself!
    3. Get comfortable on the command line


Welcome to Applied Comparative Genomics
https://github.com/schatzlab/appliedgenomics2018

Questions?