Genome Sequencing

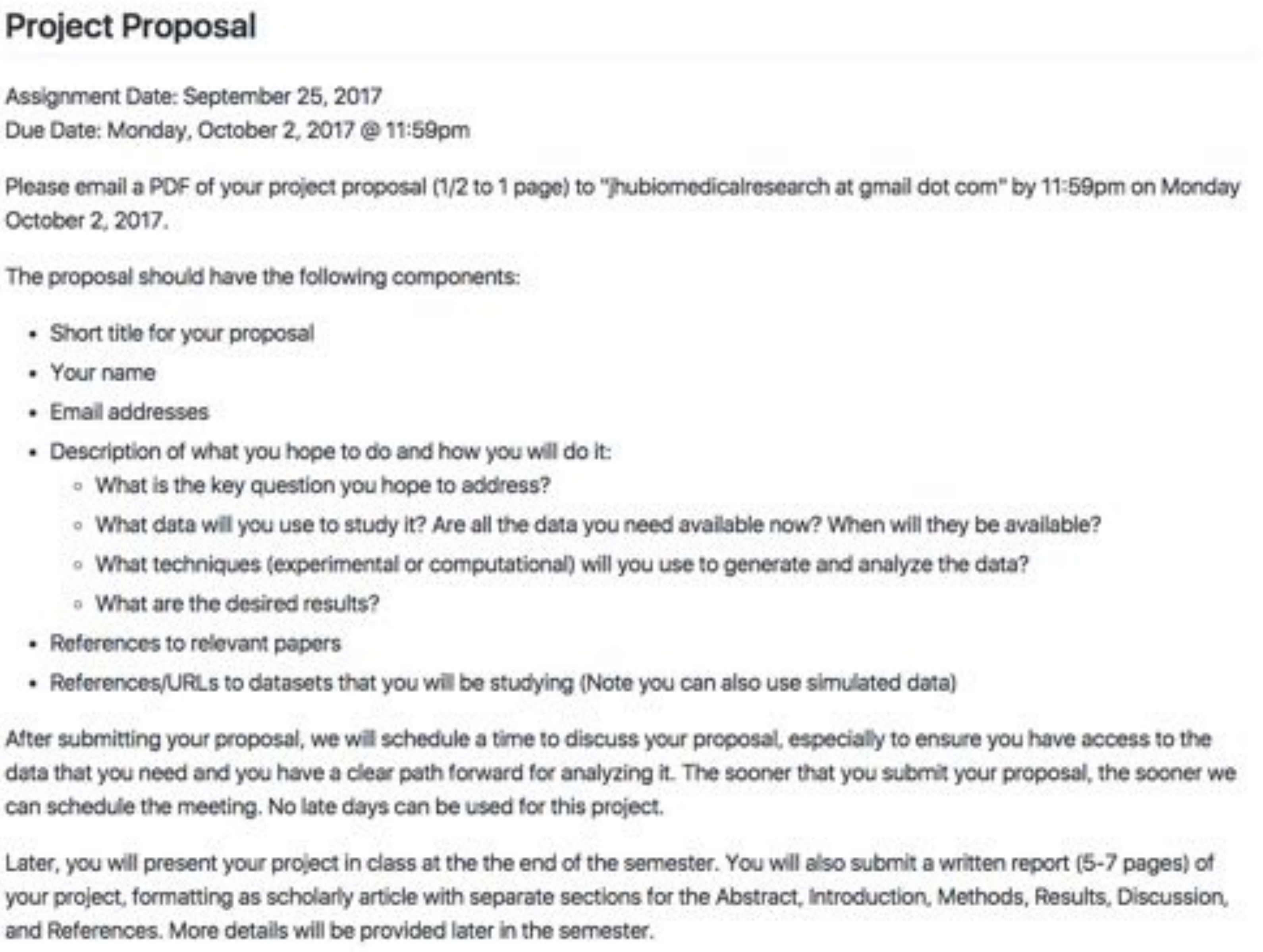
Michael Schatz

October 11 – Lecture 11

EN.601.452 Computational Biomedical Research

AS.020.415 Advanced Biomedical Research

Project Proposals



Project Proposal

Assignment Date: September 25, 2017
Due Date: Monday, October 2, 2017 @ 11:59pm

Please email a PDF of your project proposal (1/2 to 1 page) to "jhubiomedicalresearch at gmail dot com" by 11:59pm on Monday October 2, 2017.

The proposal should have the following components:

- Short title for your proposal
- Your name
- Email addresses
- Description of what you hope to do and how you will do it:
  - What is the key question you hope to address?
  - What data will you use to study it? Are all the data you need available now? When will they be available?
  - What techniques (experimental or computational) will you use to generate and analyze the data?
  - What are the desired results?
- References to relevant papers
- References/URLs to datasets that you will be studying (Note you can also use simulated data)

After submitting your proposal, we will schedule a time to discuss your proposal, especially to ensure you have access to the data that you need and you have a clear path forward for analyzing it. The sooner that you submit your proposal, the sooner we can schedule the meeting. No late days can be used for this project.

Later, you will present your project in class at the the end of the semester. You will also submit a written report (5-7 pages) of your project, formatting as scholarly article with separate sections for the Abstract, Introduction, Methods, Results, Discussion, and References. More details will be provided later in the semester.

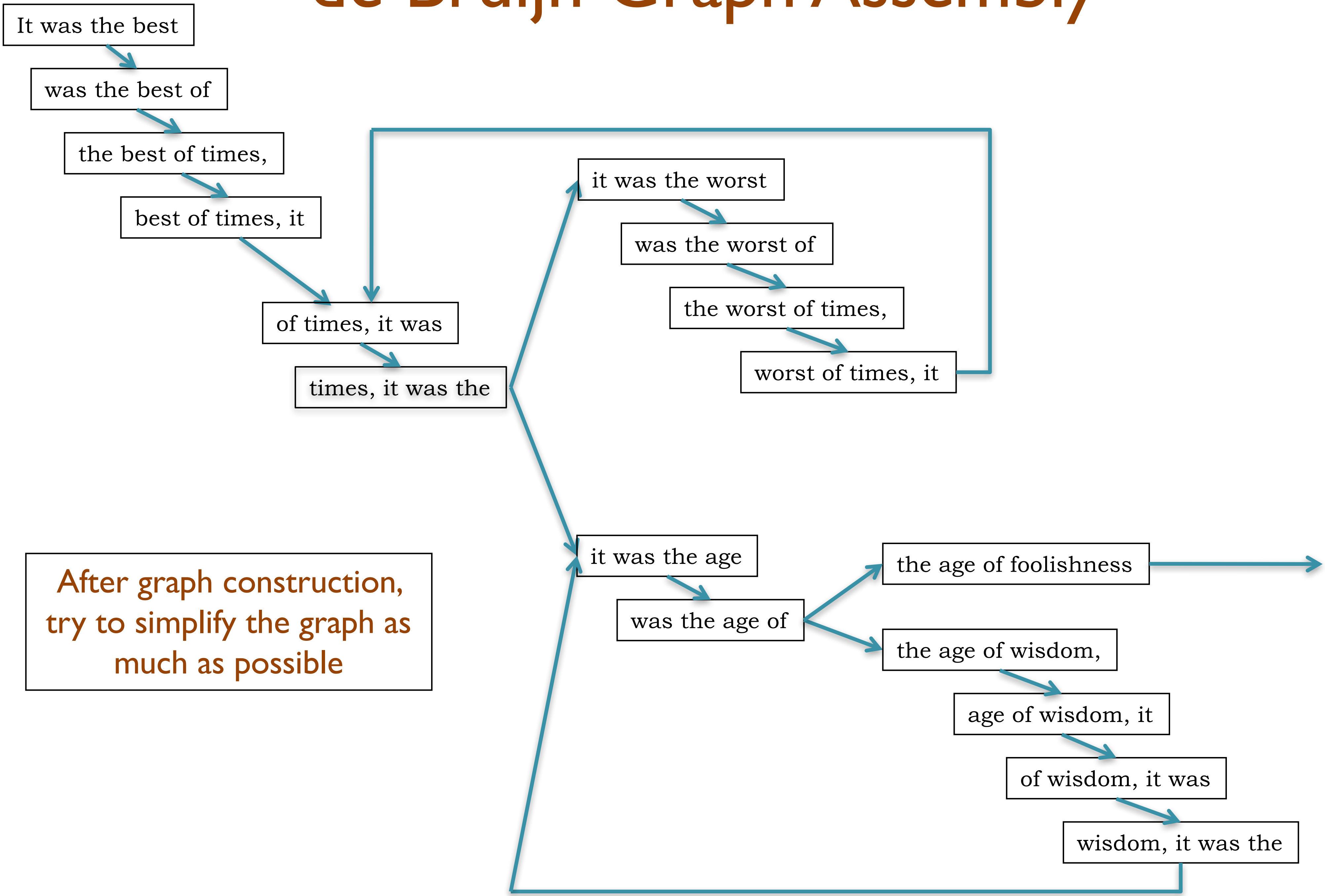
Outline

1. ***Assembly theory***
   – Assembly by analogy

2. ***Practical Issues***
   – Coverage, read length, errors, and repeats

3. ***Next-next-gen Assembly***
   – PacBio/ONT projects

de Bruijn Graph Assembly



It was the best

was the best of

the best of times,

best of times, it

of times, it was

times, it was the

it was the worst

was the worst of

the worst of times,

worst of times, it

it was the age

was the age of

the age of foolishness

the age of wisdom,

age of wisdom, it

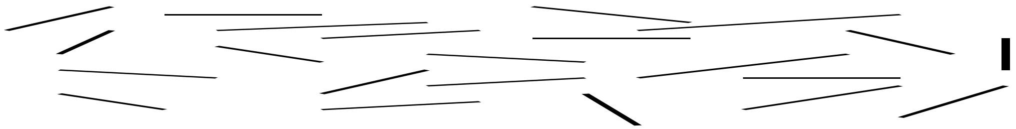
of wisdom, it was

wisdom, it was the

After graph construction, try to simplify the graph as much as possible

Assembling a Genome

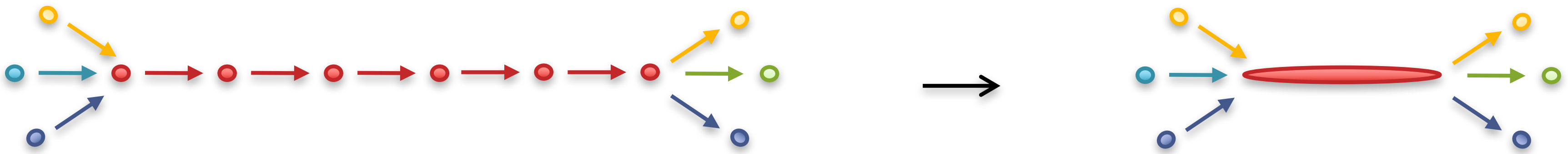
1. Shear & Sequence DNA



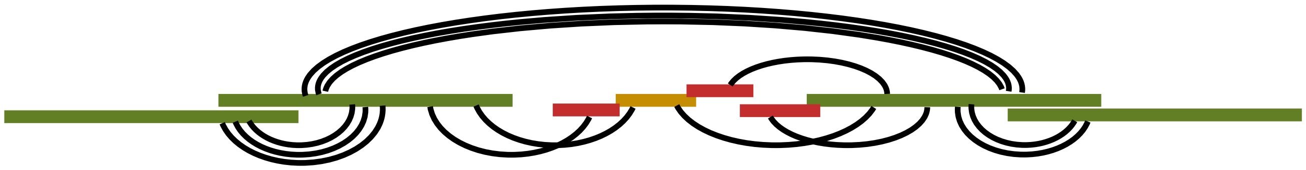
2. Construct assembly graph from reads (de Bruijn / overlap graph)

...AGCCTAGGGATGCGCGACACGT
GGATGCGCGACACGTCGCATATCCGGTTTGGTCAACCTCGGACGGAC
CAACCTCGGACGGACCTCAGCGAA...

3. Simplify assembly graph

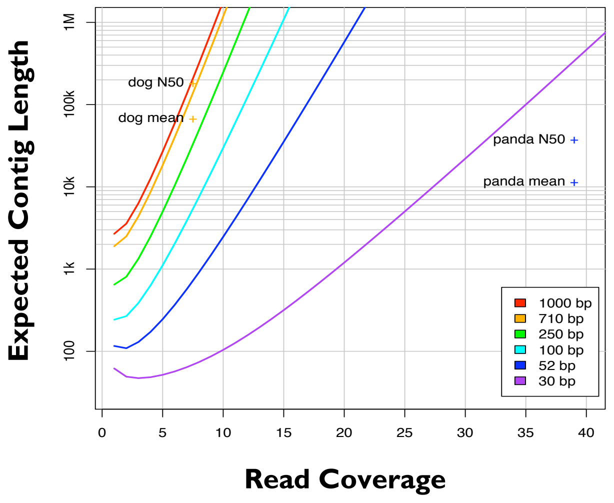


4. Detangle graph with long reads, mates, and other links



Ingredients for a good assembly

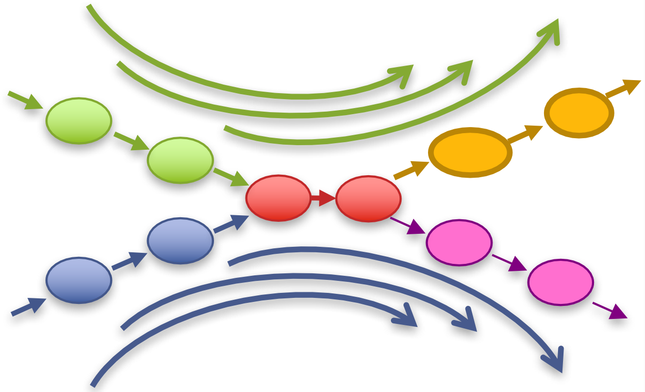
Coverage



High coverage is required

- Oversample the genome to ensure every base is sequenced with long overlaps between reads
- Biased coverage will also fragment assembly

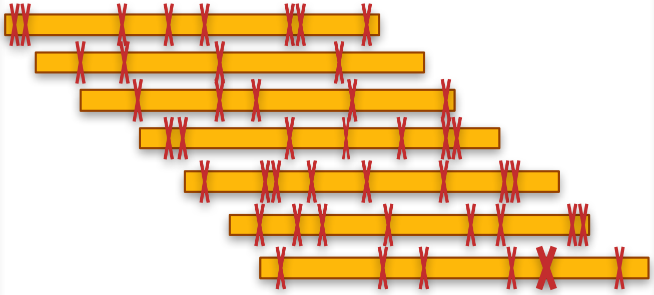
Read Length



Reads & mates must be longer than the repeats

- Short reads will have ***false overlaps*** forming hairball assembly graphs
- With long enough reads, assemble entire chromosomes into contigs

Quality

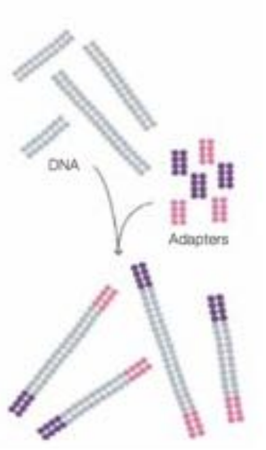


Errors obscure overlaps

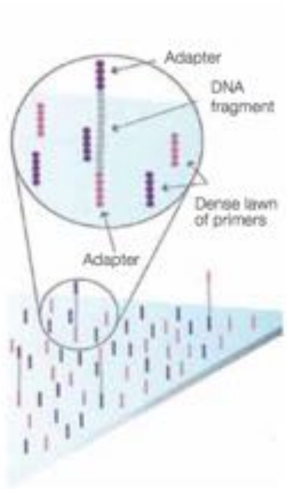
- Reads are assembled by finding kmers shared in pair of reads
- High error rate requires very short seeds, increasing complexity and forming assembly hairballs

Current challenges in *de novo* plant genome sequencing and assembly
Schatz MC, Witkowski, McCombie, WR (2012) *Genome Biology*. 12:243

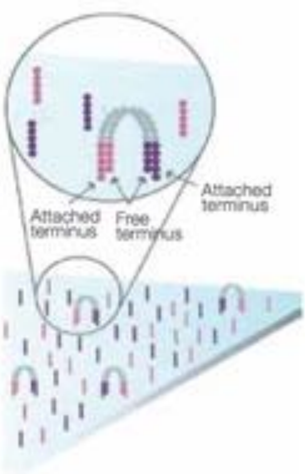
Illumina Sequencing by Synthesis



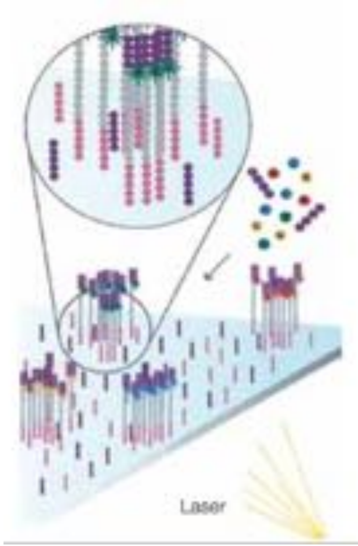
1. Prepare



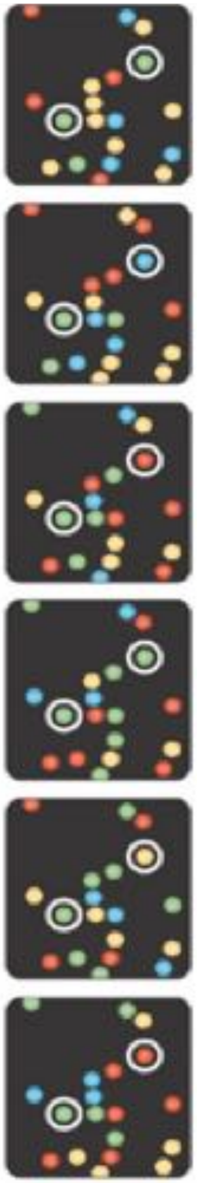
2. Attach



3. Amplify



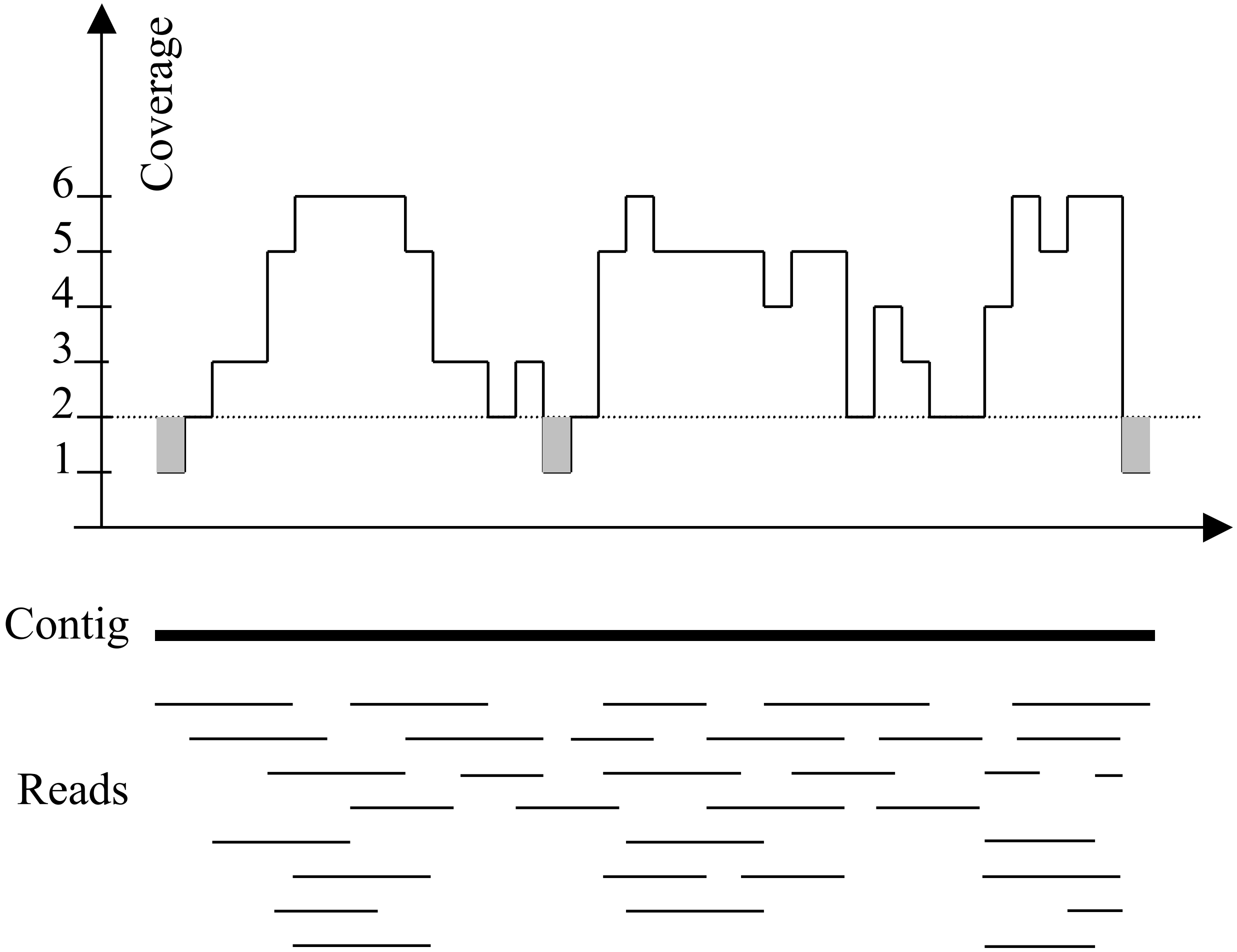
4. Image



5. Basecall

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

Typical sequencing coverage

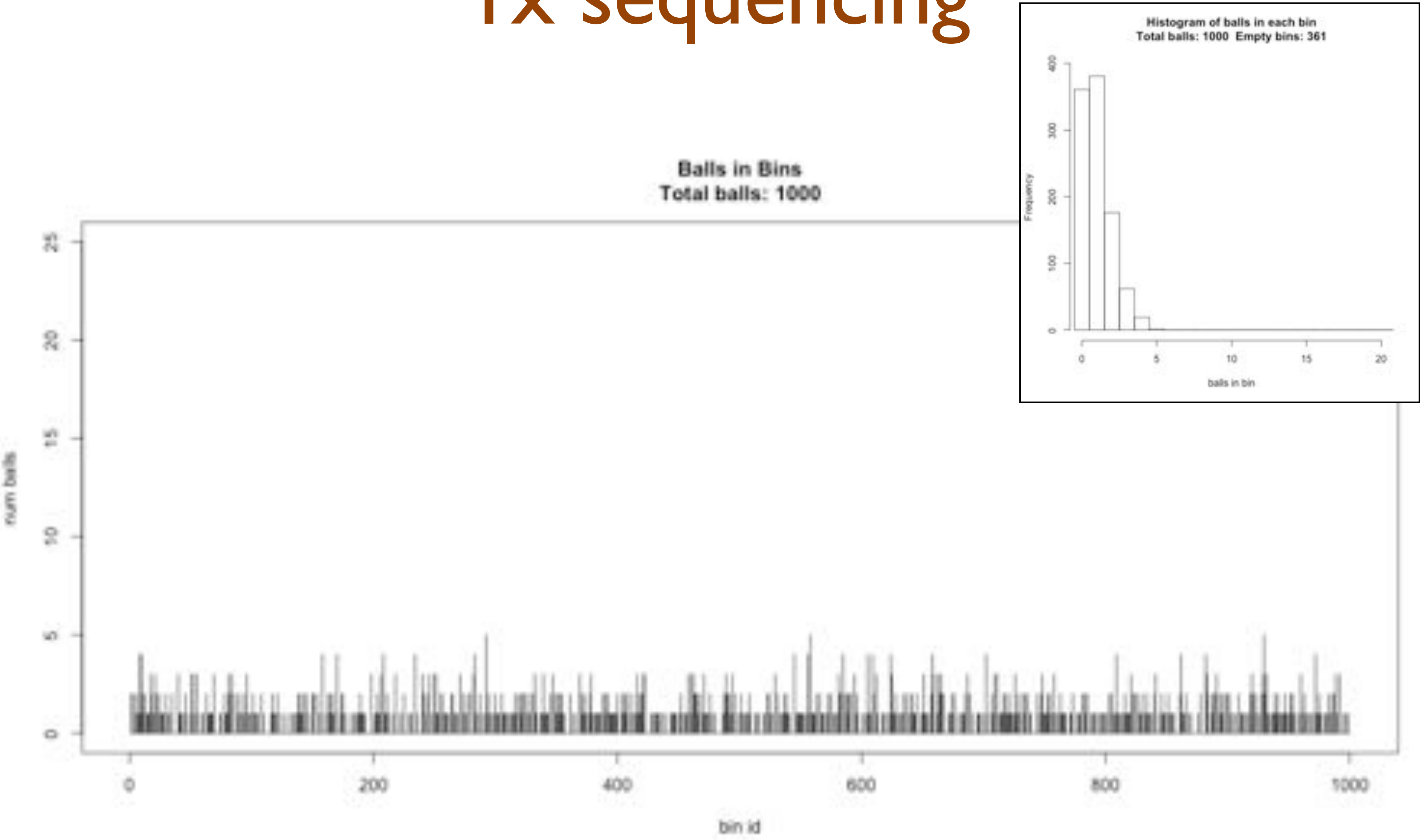


Imagine raindrops on a sidewalk

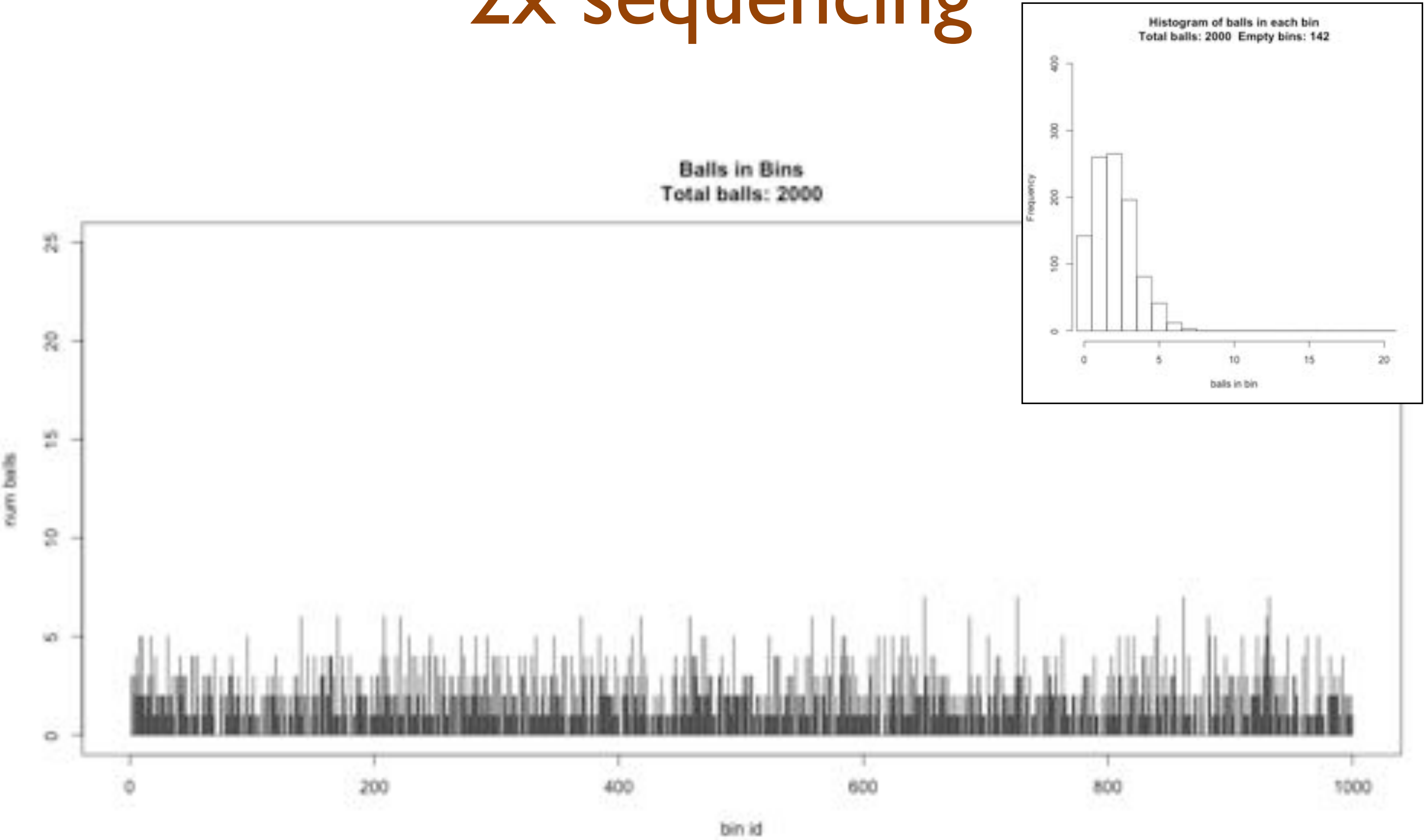
We want to cover the entire sidewalk but each drop costs $1

If the genome is 100 Mbp, should we sequence 1M 100bp reads?

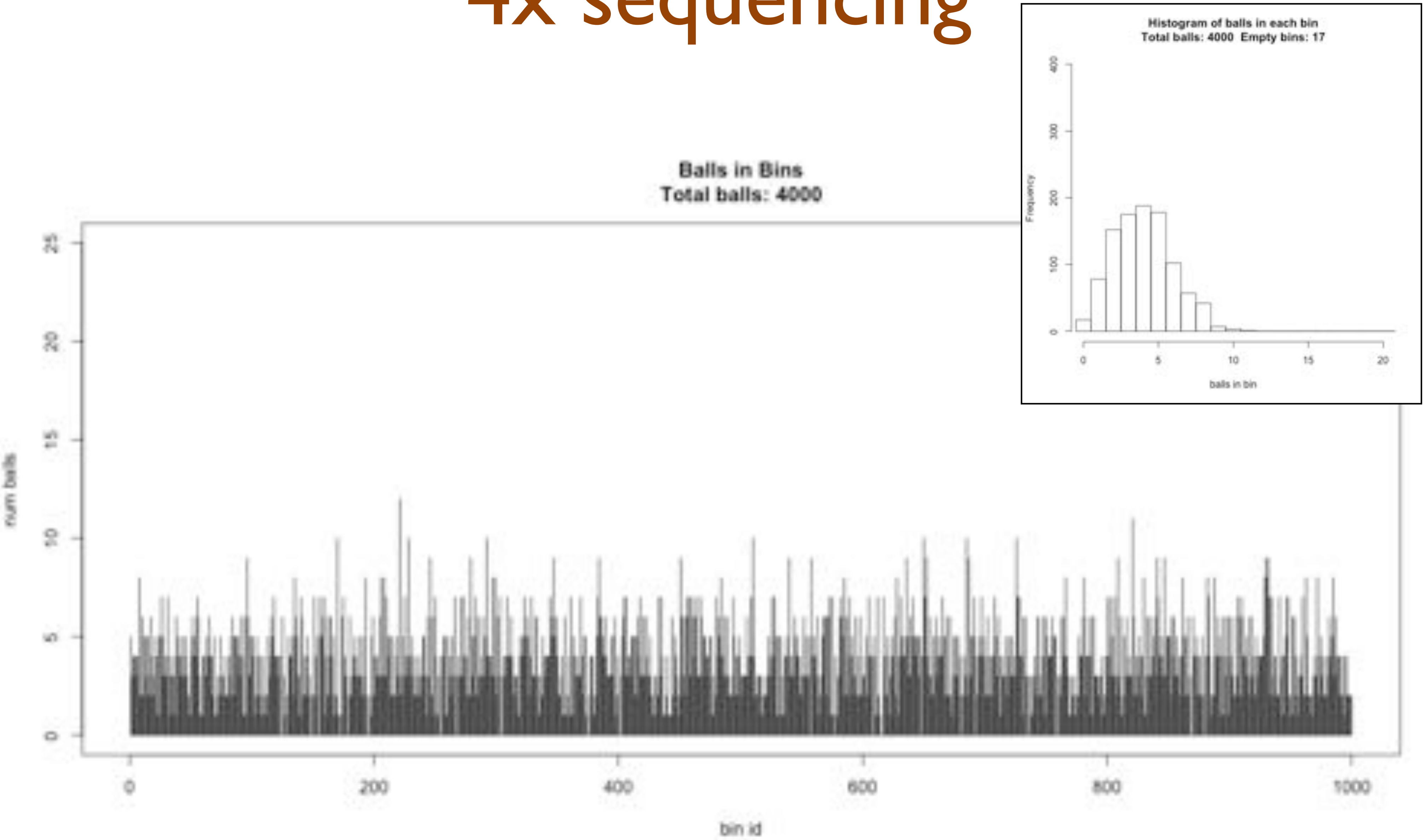
1x sequencing



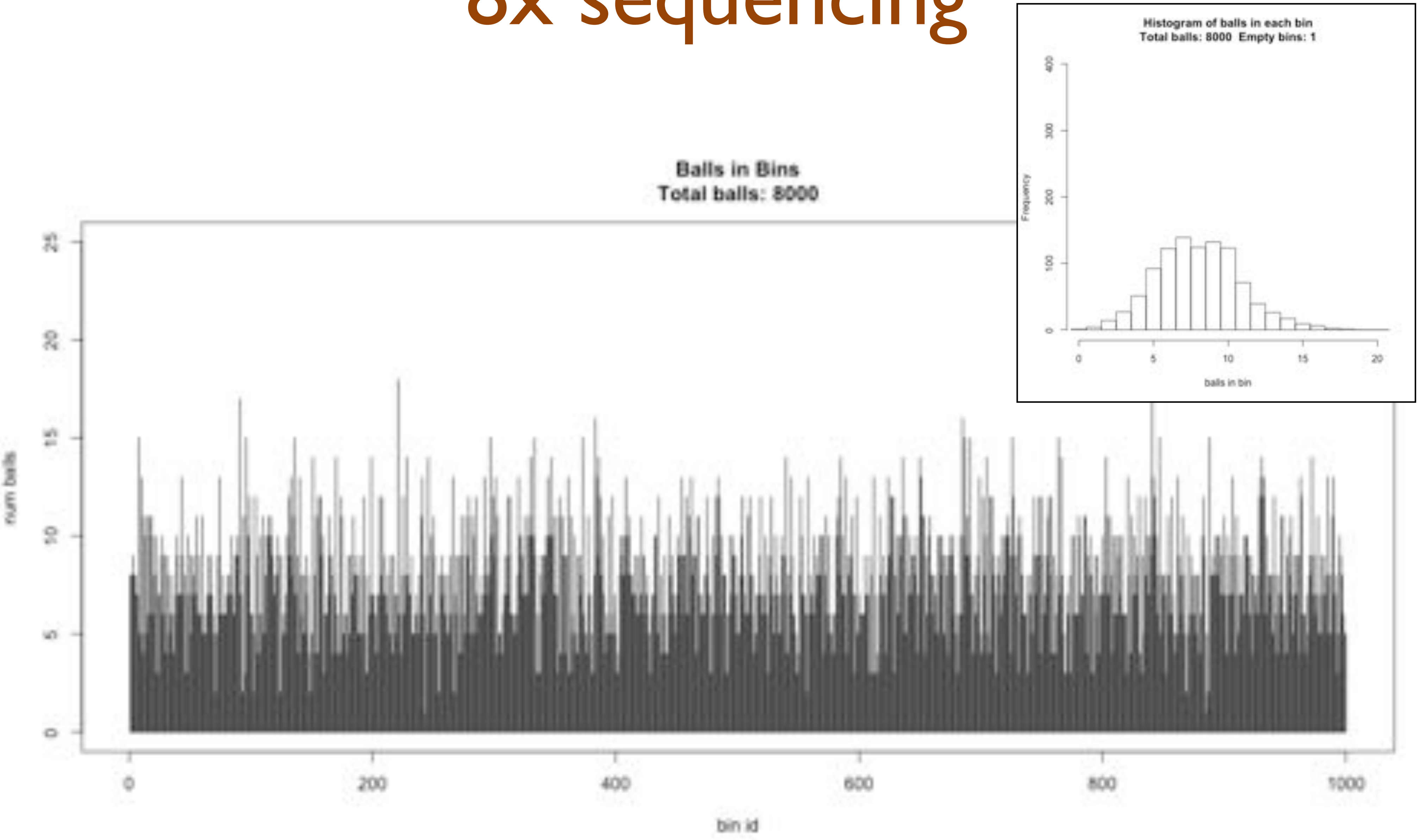
2x sequencing



4x sequencing



8x sequencing



Poisson Distribution

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

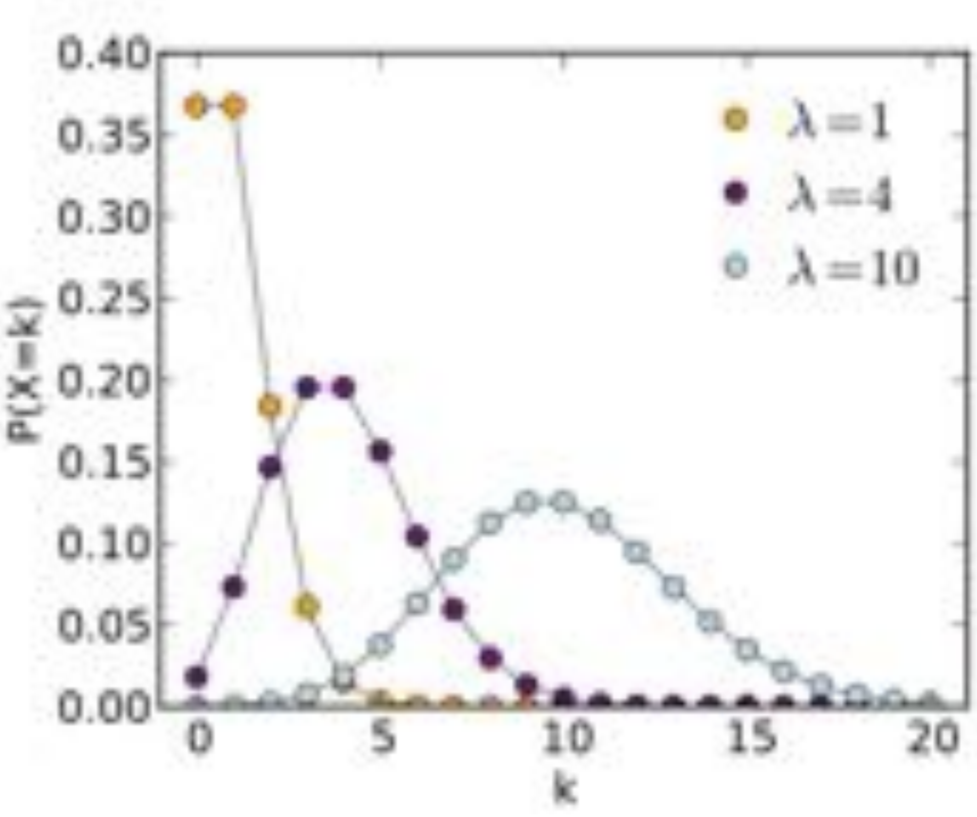
The probability of a given number of events occurring in a fixed interval of time and/or space if these events occur with a known average rate and independently of the time since the last event.

Formulation comes from the limit of the binomial equation

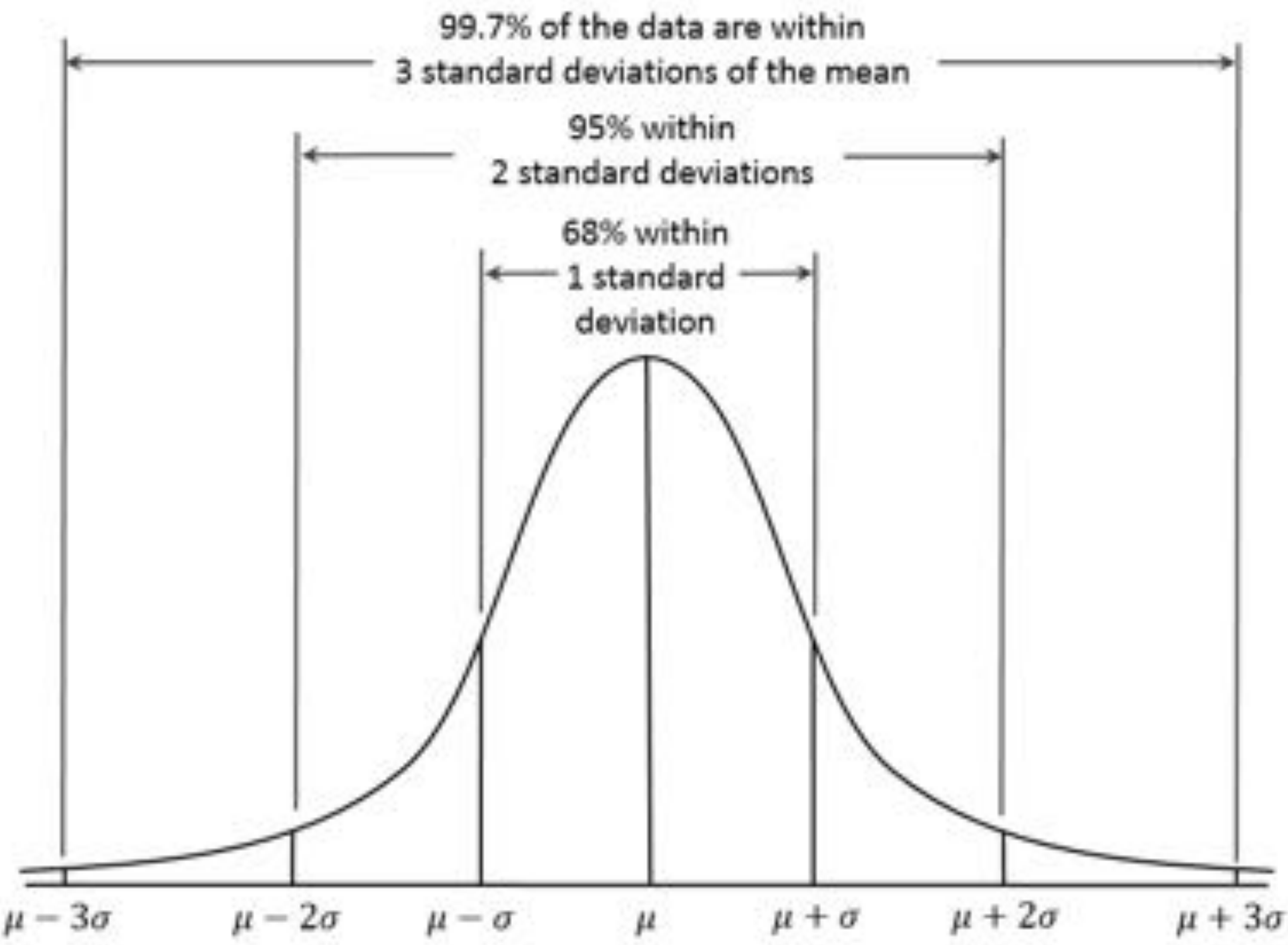
Resembles a normal distribution, but over the positive values, and with only a single parameter.

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 λ > 10

Pop Quiz!

I want to sequence a 10Mbp genome to 24x coverage.
How many 150bp reads do I need?

I need 10Mbp x 24x = 240Mbp of data
240Mbp / 150bp / read = 1.6M reads

I want to sequence a 10Mbp genome so that
>97.5% of the genome has at least 24x coverage.
How many 150bp reads do I need?

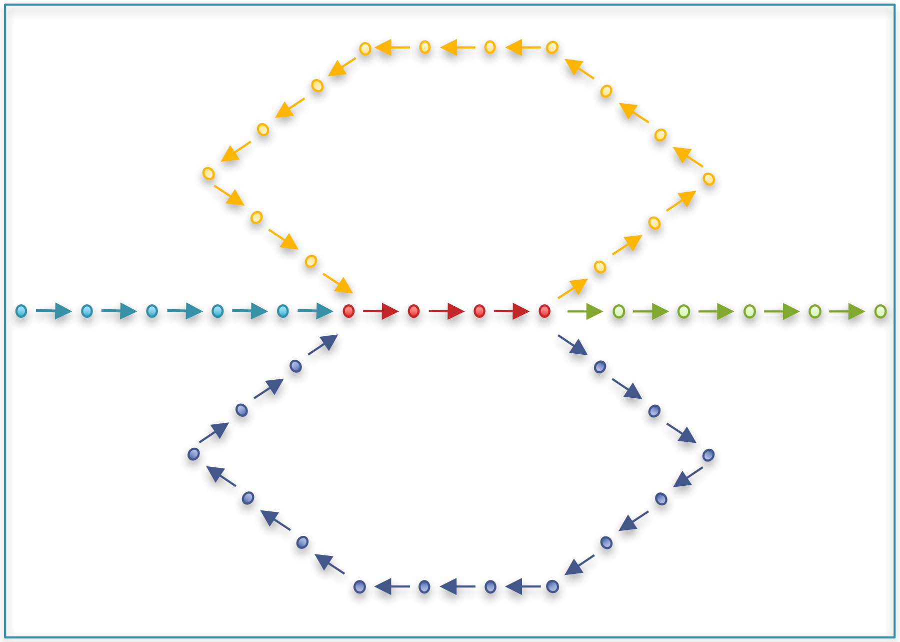
Find X such that X-2*sqrt(X) = 24

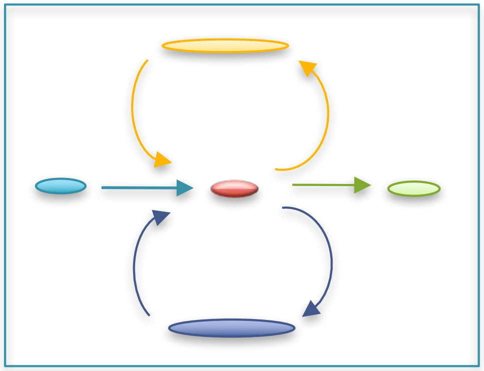
36-2*sqrt(36) = 24

I need 10Mbp x 36x = 360Mbp of data
360Mbp / 150bp / read = 2.4M reads

Unitigging / Unipathing

- After simplification and correction, compress graph down to its non-branching initial contigs
  - Aka "unitigs", "unipaths"
  - Unitigs end because of (1) lack of coverage, (2) errors, (3) heterozygosity and (4) repeats





Repetitive regions

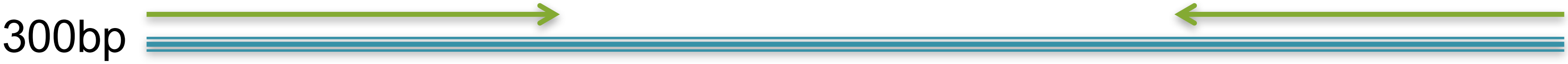
Repeat Type	Definition / Example	Prevalence
Low-complexity DNA / Microsatellites	$(b_1b_2\ldots b_k)^N$ where $1 \le k \le 6$ CACACACACACACACACACA	2%
SINEs (Short Interspersed Nuclear Elements)	*Alu* sequence (~280 bp) Mariner elements (~80 bp)	13%
LINEs (Long Interspersed Nuclear Elements)	~500 – 5,000 bp	21%
LTR (long terminal repeat) retrotransposons	Ty1-copia, Ty3-gypsy, Pao-BEL (~100 – 5,000 bp)	8%
Other DNA transposons		3%
Gene families & segmental duplications		4%

- Over 50% of mammalian genomes are repetitive
  - – Large plant genomes tend to be even worse
  - – Wheat: 16 Gbp; Pine: 24 Gbp

Paired-end and Mate-pairs

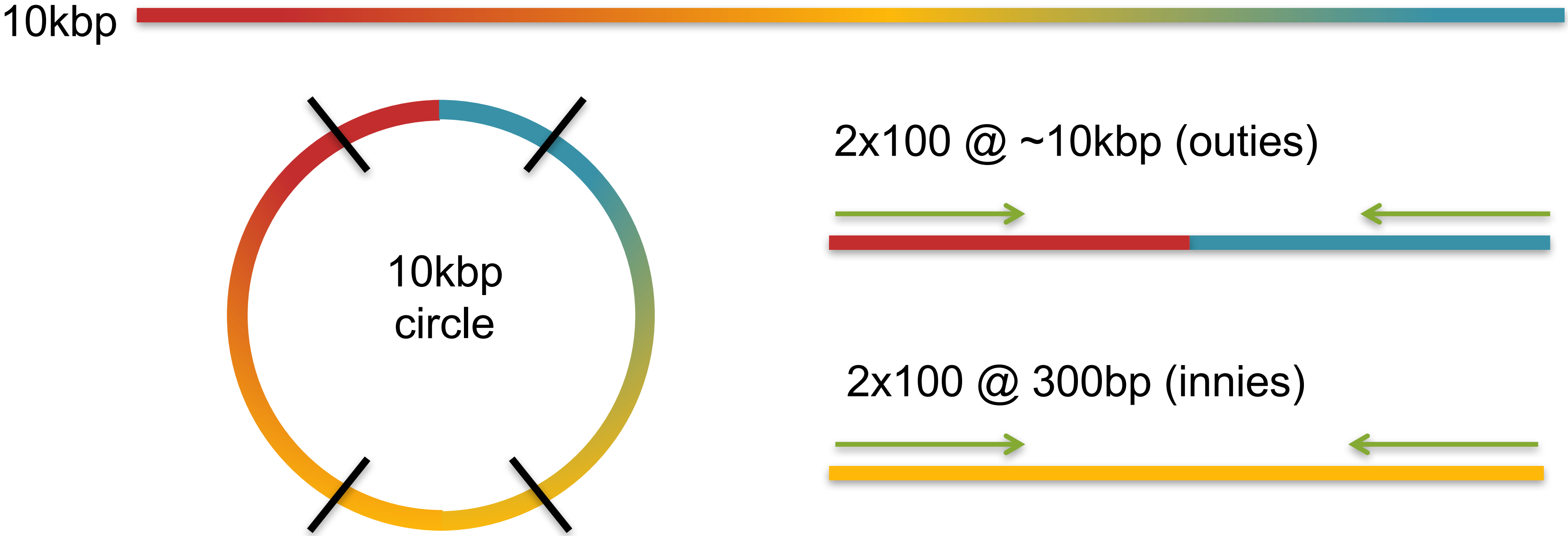
Paired-end sequencing

- 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



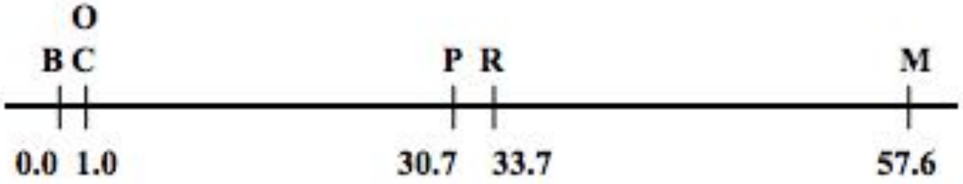
The first genetic map

Mendel's Second Law (The Law of Independent Assortment) states alleles of one gene sort into gametes independently of the alleles of another gene.

However, Morgan and his student Sturtevant noticed that for certain traits the probability of having one trait given another was not 50/50– those traits are genetically linked

http://www.caltech.edu/news/first-genetic-linkage-map-38798

Sturtevant realized the probabilities of co-occurrences could be explained if those alleles were arranged on a linear fashion: traits that are most commonly observed together must be locates closest together

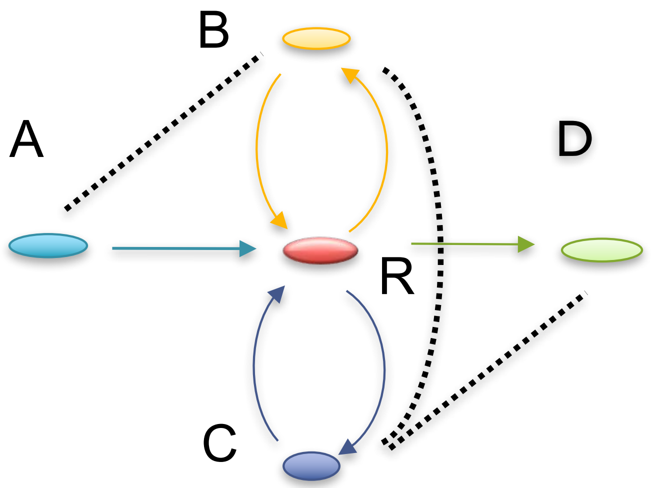


Today genetic maps are routinely generated by measuring the rates of polymorphic markers in large populations of individuals

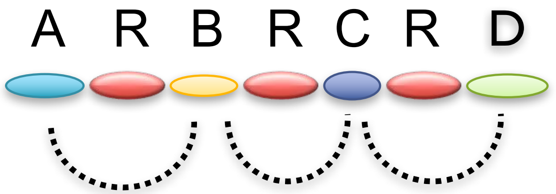
The Linear Arrangement of Six Sex-Linked Factors in Drosophila as shown by their mode of Association
Sturtevant, A. H. (1913) Journal of Experimental Zoology, 14: 43-59

Scaffolding

- Initial contigs (*aka* unipaths, unitigs) terminate at
  - *Coverage gaps*: especially extreme GC
  - *Conflicts*: errors, repeat boundaries



- Use mate-pairs to resolve correct order through assembly graph
  - Place sequence to satisfy the mate constraints
  - Mates through repeat nodes are tangled



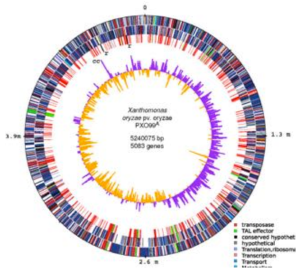
- Final scaffold may have internal gaps called sequencing gaps
  - We know the order, orientation, and spacing, but just not the bases. Fill with Ns instead

The human genome

"Without a doubt, this is the most important, most wondrous map ever produced by humankind."

Bill Clinton
June 26, 2000

Assembly Summary



Assembly quality depends on

1. ***Coverage***: low coverage is mathematically hopeless
2. ***Repeat composition***: high repeat content is challenging
3. ***Read length***: longer reads help resolve repeats
4. ***Error rate***: errors reduce coverage, obscure true overlaps

- Assembly is a hierarchical, starting from individual reads, build high confidence contigs/unitigs, incorporate the mates to build scaffolds
  - Extensive error correction is the key to getting the best assembly possible from a given data set

- Watch out for collapsed repeats & other misassemblies
  - Globally/Locally reassemble data from scratch with better parameters & stitch the 2 assemblies together

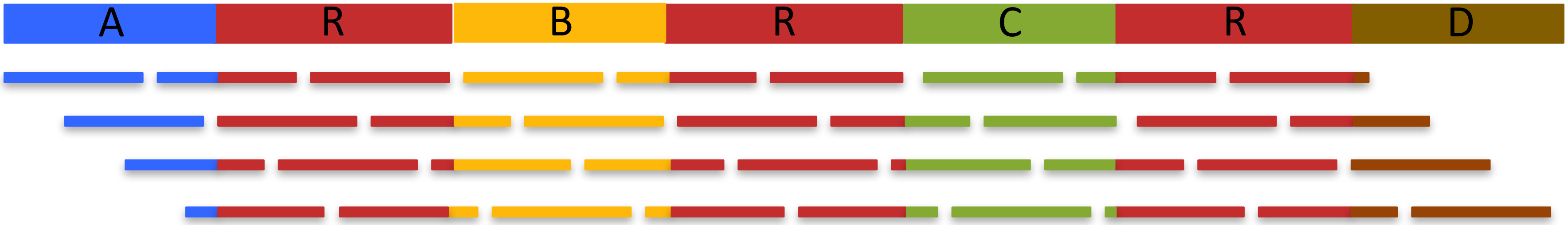
Outline

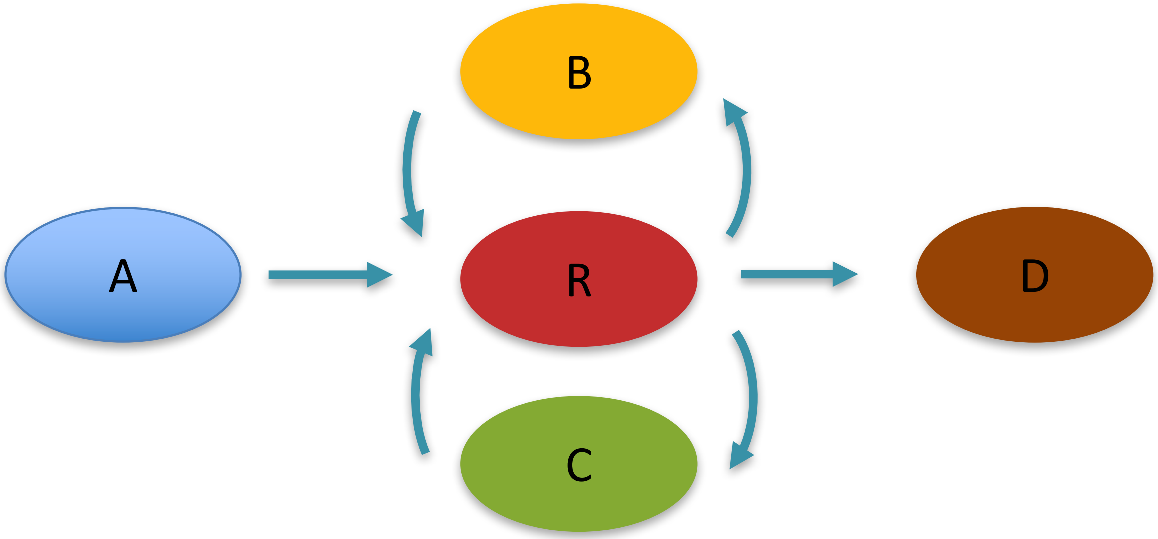
1. ***Assembly theory***
   - Assembly by analogy

2. ***Practical Issues***
   - Coverage, read length, errors, and repeats

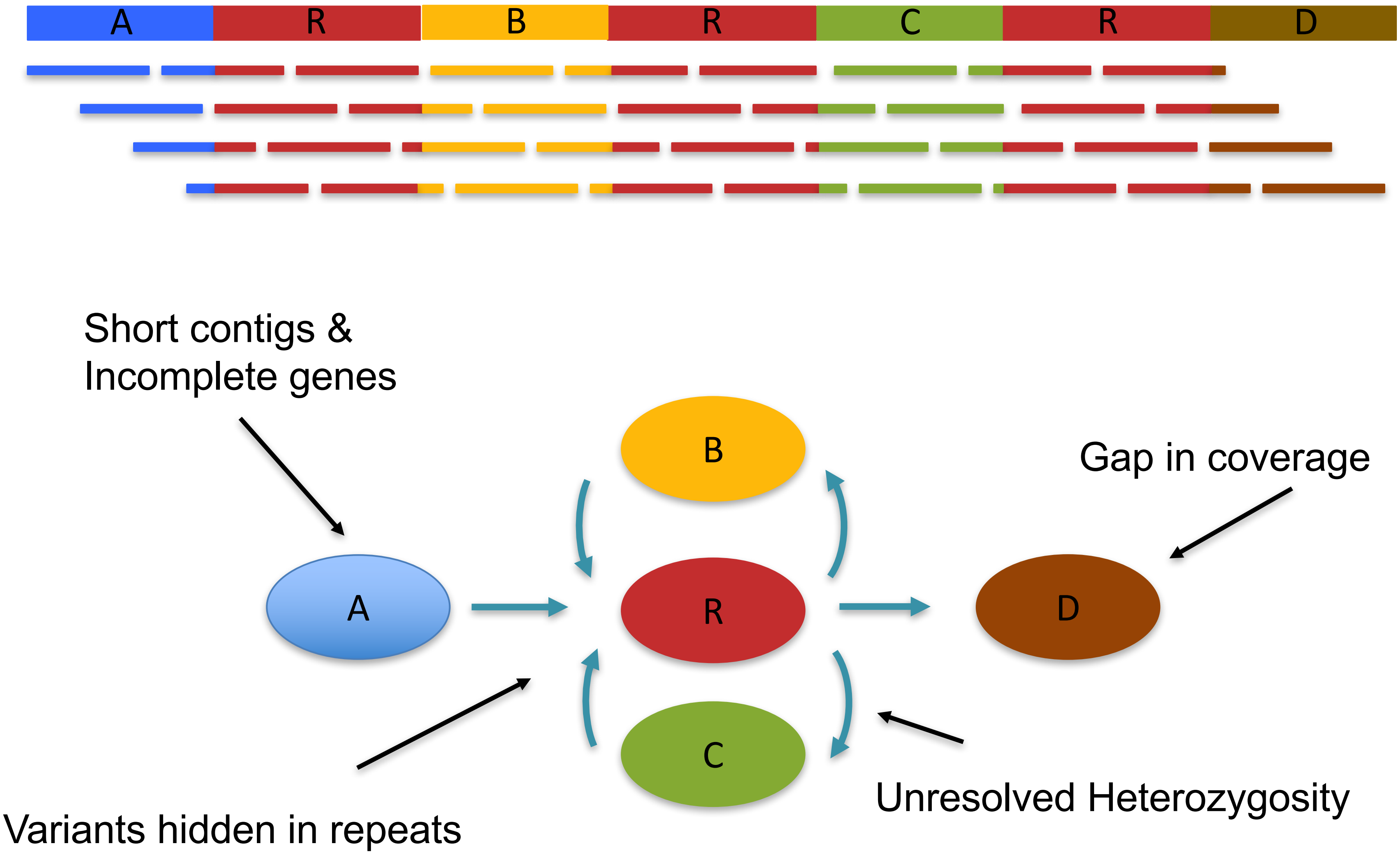
3. ***Next-next-gen Assembly***
   - PacBio/ONT projects

Assembly Complexity





Assembly Complexity



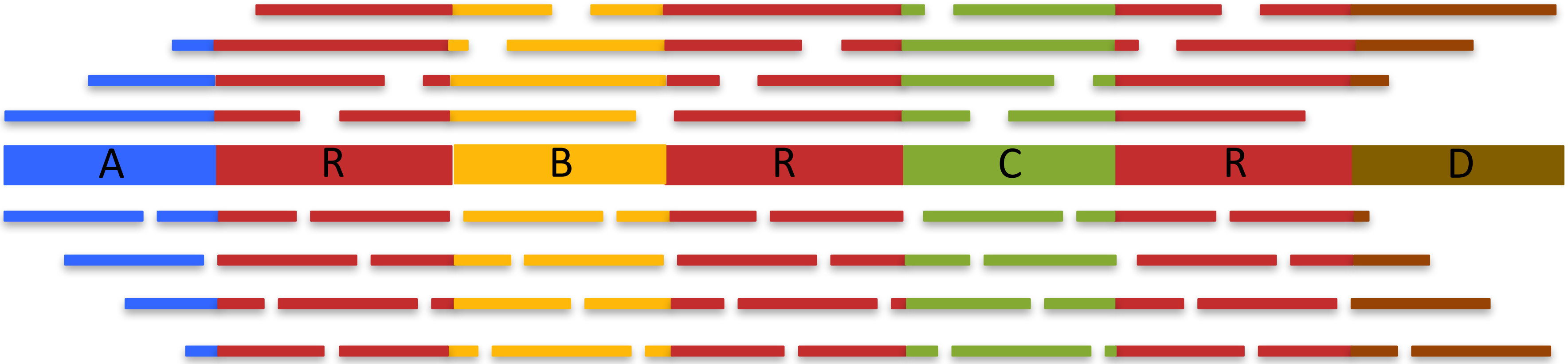
Short contigs & Incomplete genes

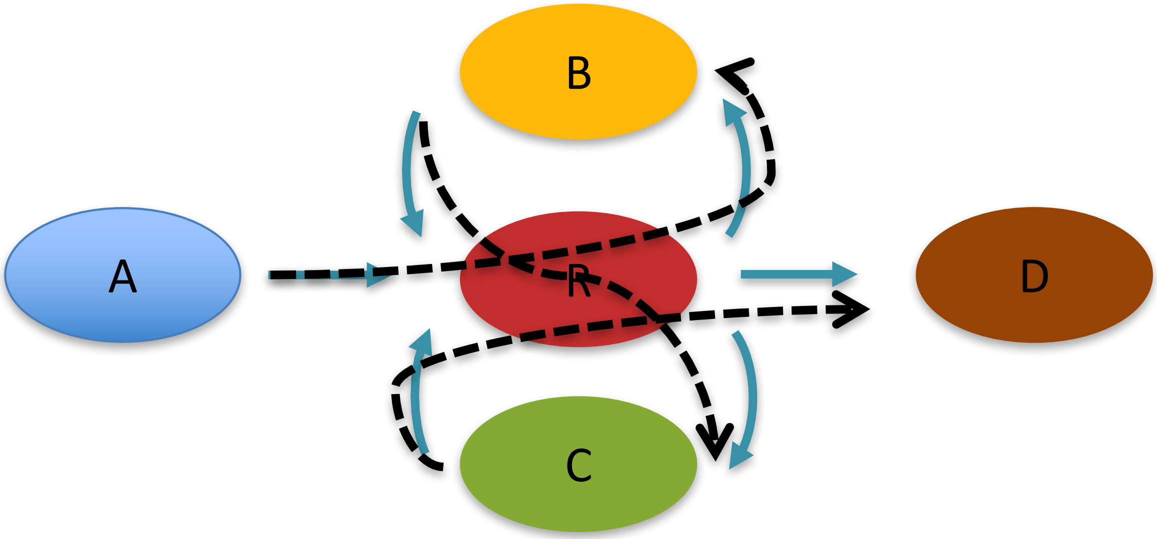
Gap in coverage

Variants hidden in repeats

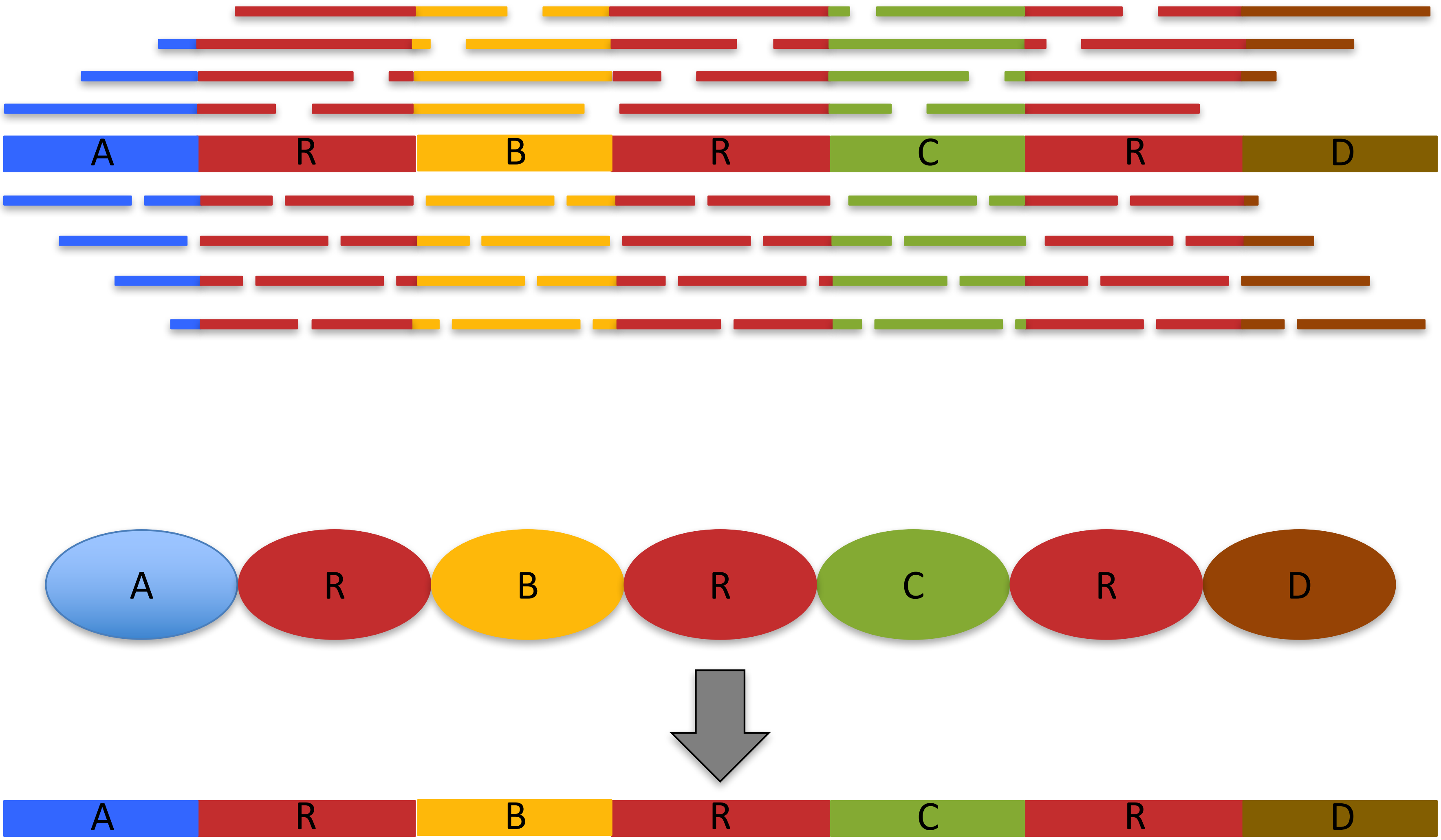
Unresolved Heterozygosity

Assembly Complexity





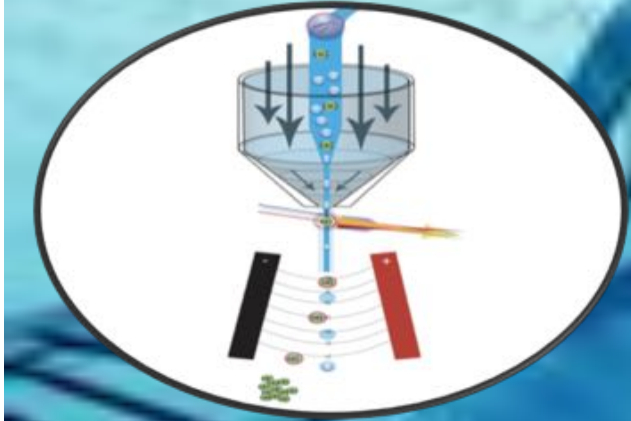
Assembly Complexity



The advantages of SMRT sequencing
Roberts, RJ, Carneiro, MO, Schatz, MC (2013) *Genome Biology.* 14:405

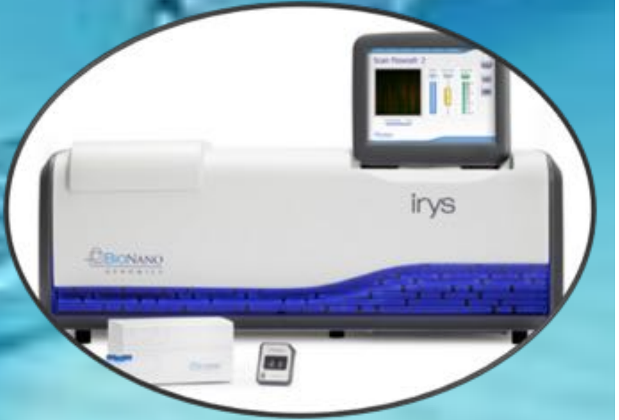
Genomics Arsenal in the year 2017

Sample Preparation



Sequencing

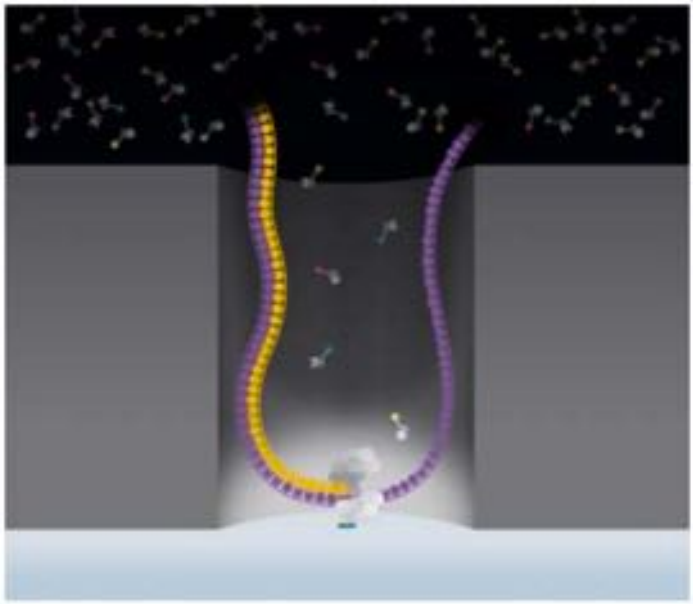
Chromosome Mapping

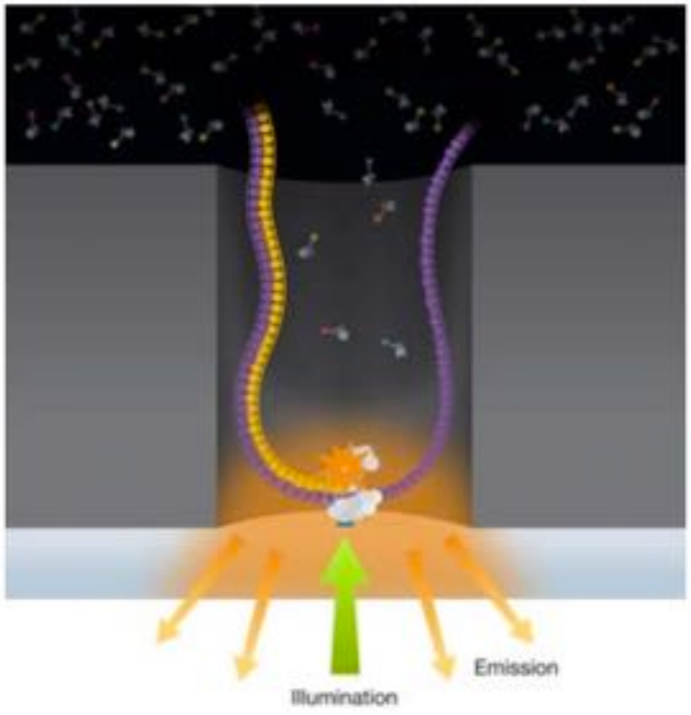


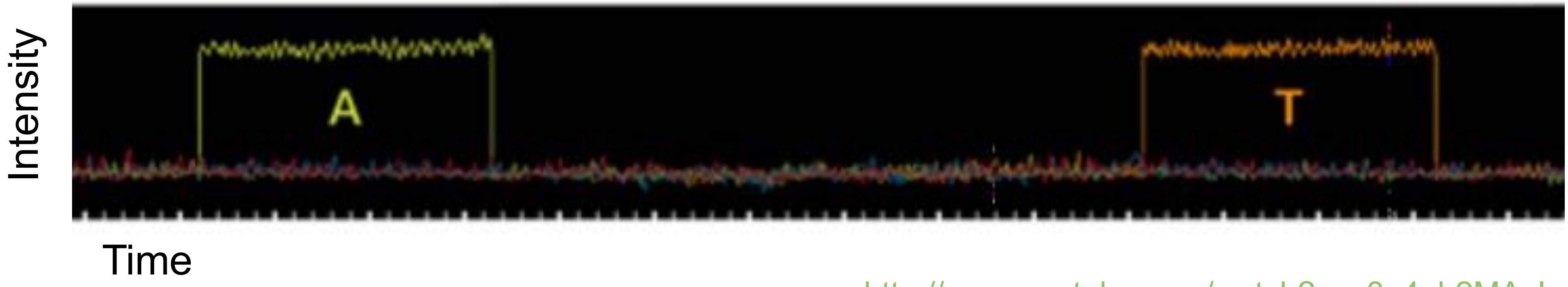
Nhel
AAGCT AGCTT
TTCGA TCGAA

PacBio: SMRT Sequencing

Imaging of florescent phospholinked labeled nucleotides as they are incorporated by a polymerase anchored to a Zero-Mode Waveguide (ZMW).



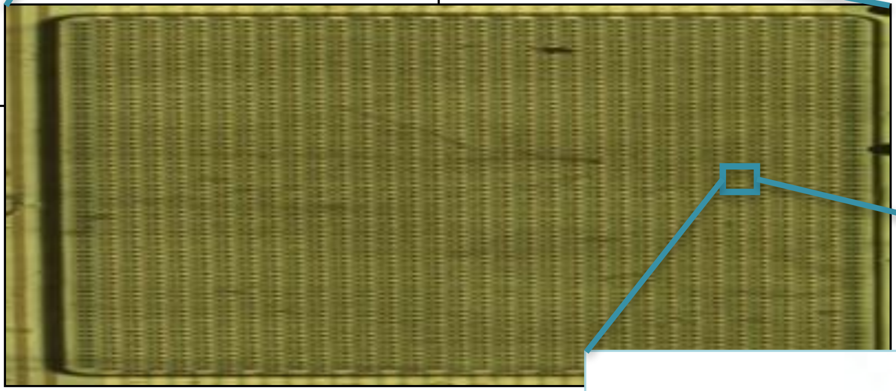


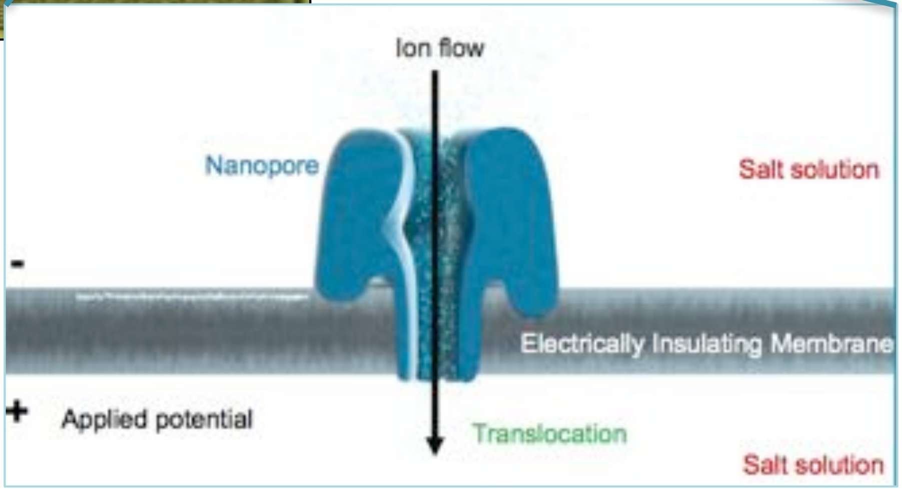


http://www.youtube.com/watch?v=v8p4ph2MAvI

Oxford Nanopore MinION

- Thumb drive sized sequencer powered over USB
- Capacity for 512 reads at once
- Senses DNA by measuring changes to ion flow

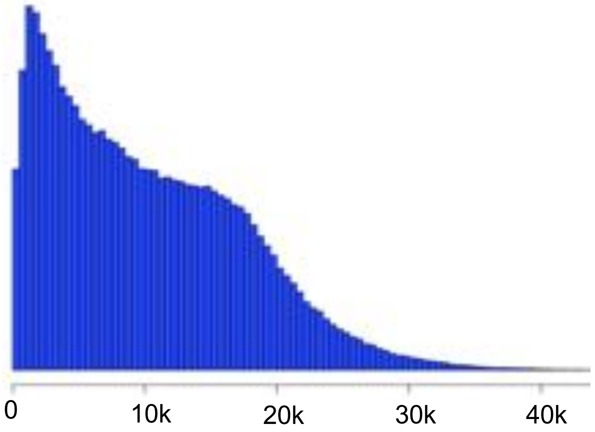




https://www.youtube.com/watch?v=CE4dW64x3Ts

Single Molecule Sequencing

PacBio RS II



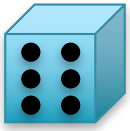
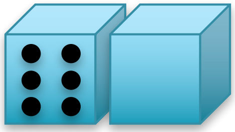
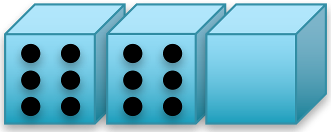
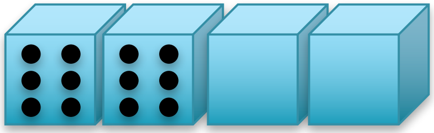
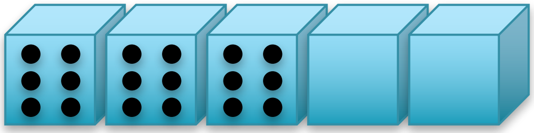
CSHL/PacBio

```
TTGTAAGCAGTTGAAAACTATGTGTGGATTTAGAATAAAGAACATGAAAG
||||||||||||||||||||||||| ||||||| |||||||||||| |||
TTGTAAGCAGTTGAAAACTATGTGT-GATTTAG-ATAAAGAACATGGAAG

ATTATAAA-CAGTTGATCCATT-AGAAGA-AAACGCAAAAGGCGGCTAGG
| |||||| ||||||||||||| |||| | |||||| |||||| ||||||
A-TATAAATCAGTTGATCCATTAAGAA-AGAAACGC-AAAGGC-GCTAGG

CAACCTTGAATGTAATCGCACTTGAAGAACAAGATTTTATTCCGCGCCCG
| |||||| |||| ||  ||||||||||||||||||||||||||||||||
C-ACCTTG-ATGT-AT--CACTTGAAGAACAAGATTTTATTCCGCGCCCG

TAACGAATCAAGATTCTGAAAACACAT-ATAACAACCTCCAAAA-CACAA
| ||||||| |||||||||||||| || ||    |||||||||| |||||
T-ACGAATC-AGATTCTGAAAACA-ATGAT----ACCTCCAAAAGCACAA

-AGGAGGGGAAAGGGGGGAATATCT-ATAAAGATTACAAATTAGA-TGA
 ||||||   ||     |||||||| || |||||||||||||| || |||
GAGGAGG---AA-----GAATATCTGAT-AAAGATTACAAATT-GAGTGA

ACT-AATTCACAATA-AATAACACTTTTA-ACAGAATTGAT-GGAA-GTT
||| ||||||||| | ||||||||||||| ||| ||||||| |||| |||
ACTAAATTCACAA-ATAATAACACTTTTAGACAAAATTGATGGGAAGGTT

TCGGAGAGATCCAAAACAATGGGC-ATCGCCTTTGA-GTTAC-AATCAAA
|| ||||||||| ||||||| ||| |||| |||||| ||||| |||||||
TC-GAGAGATCC-AAACAAT-GGCGATCG-CTTTGACGTTACAAATCAAA
```

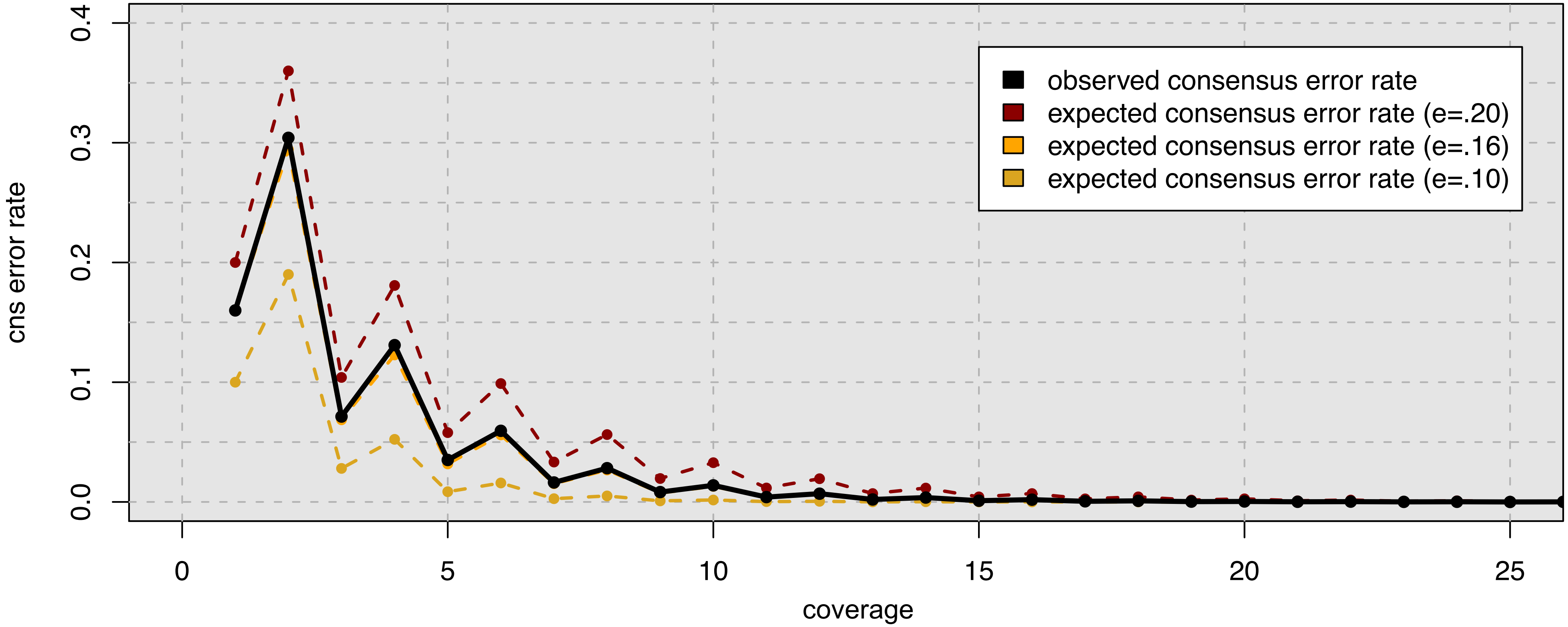
Sample of 100k reads aligned with BLASR requiring >100bp alignment
Average overall accuracy 83.7%: 11.5% insertions, 3.4% deletions, 1.4% mismatch

Consensus Quality: Probability Review

Roll *n* dice => What is the probability that at least half are 6's

n	*Min to Win*	*Winning Events*	*P(Win)*
1		*1/6*	*16.7%*
2		*P(1of 2) + P(2 of 2)*	*30.5%*
3		*P(2 of 3) + P(3 of 3)*	*7.4%*
4		*P(2 of 4) + P(3 of 4) + P(4 of 4)*	*13.2%*
5		*P(3 of 5) + P(4 of 5) + P(5 of 5)*	*3.5%*
n	*ceil(n/2)*	$\sum_{i=\lceil n/2\rceil}^{n} P(i\ \ of\ \ n) = \sum_{i=\lceil n/2\rceil}^{n} \binom{n}{i} (p)^{i} (1-p)^{n-i}$	

Consensus Accuracy and Coverage



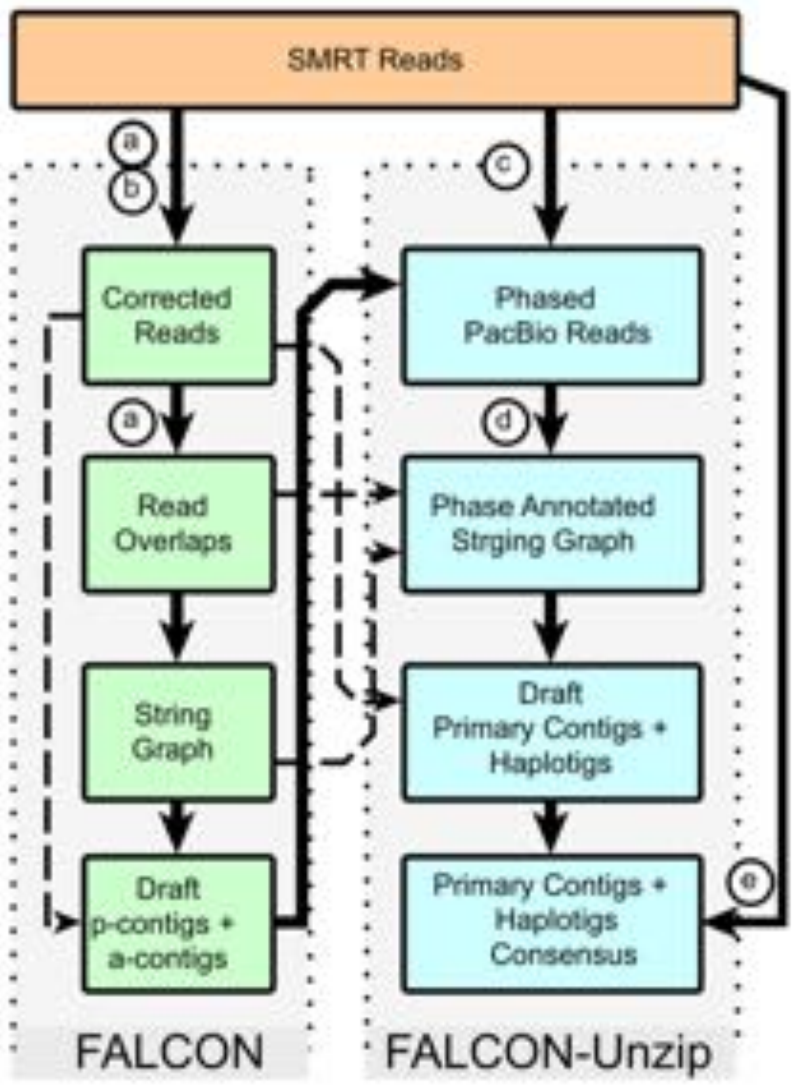
Coverage can overcome random errors

- Dashed: error model from binomial sampling
- Solid: observed accuracy

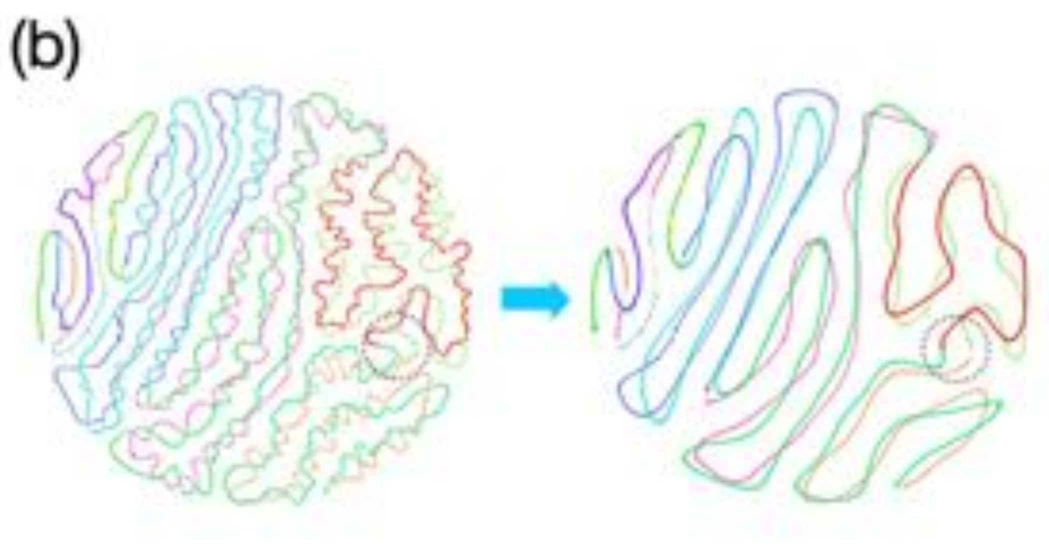
$$CNS\ Error = \sum_{i=\lceil c/2 \rceil}^{c} \binom{c}{i} (e)^{i} (1-e)^{n-i}$$

Hybrid error correction and de novo assembly of single-molecule sequencing reads.
Koren et al (2012) *Nature Biotechnology. doi:10.1038/nbt.2280*

FALCON Accuracy

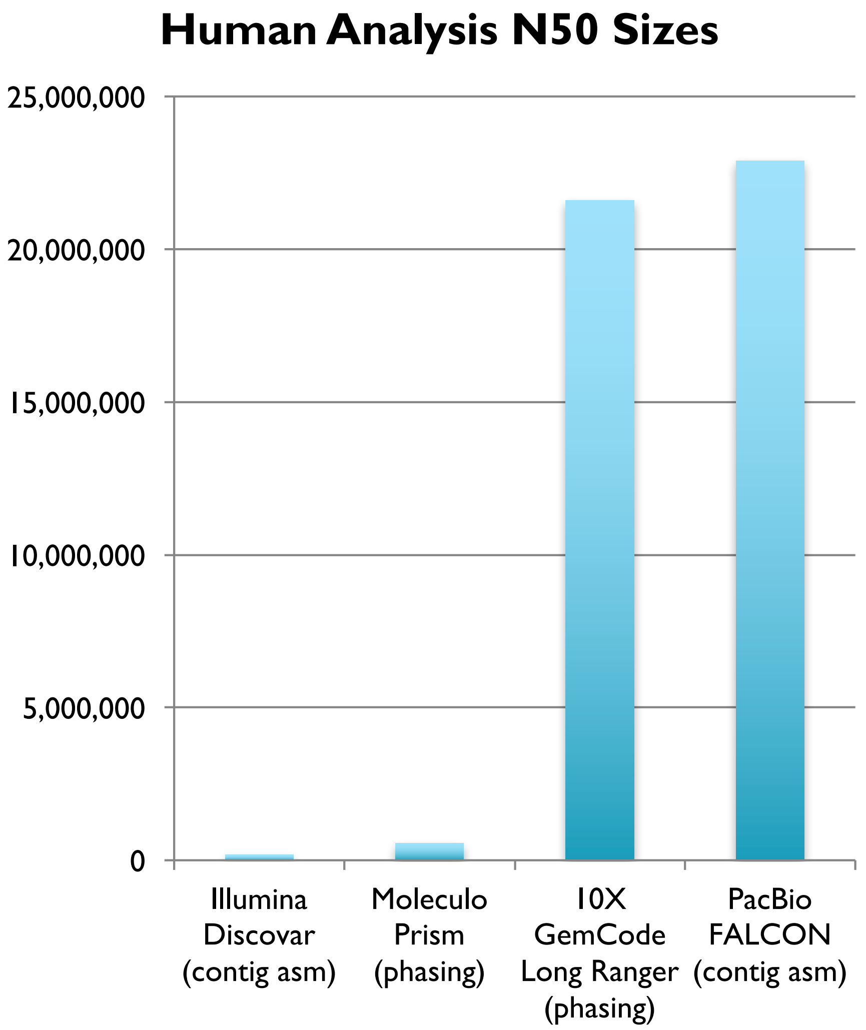

SMRT Reads
a
b
c
Corrected Reads
Phased PacBio Reads
a
d
Read Overlaps
Phase Annotated Strging Graph
String Graph
Draft Primary Contigs + Haplotigs
Draft p-contigs + a-contigs
Primary Contigs + Haplotigs Consensus
e
FALCON
FALCON-Unzip

"***The overall base-to-base concordance rate is about 99.99%*** (QV40 in Phred scale) in the F1 FALCON-Unzip assembly. The insertion and deletion (indel) concordances to the parental lines were lower (about QV40) than the SNP concordance rate (about QV50), with most residual errors concentrated in long homopolymer sequences"

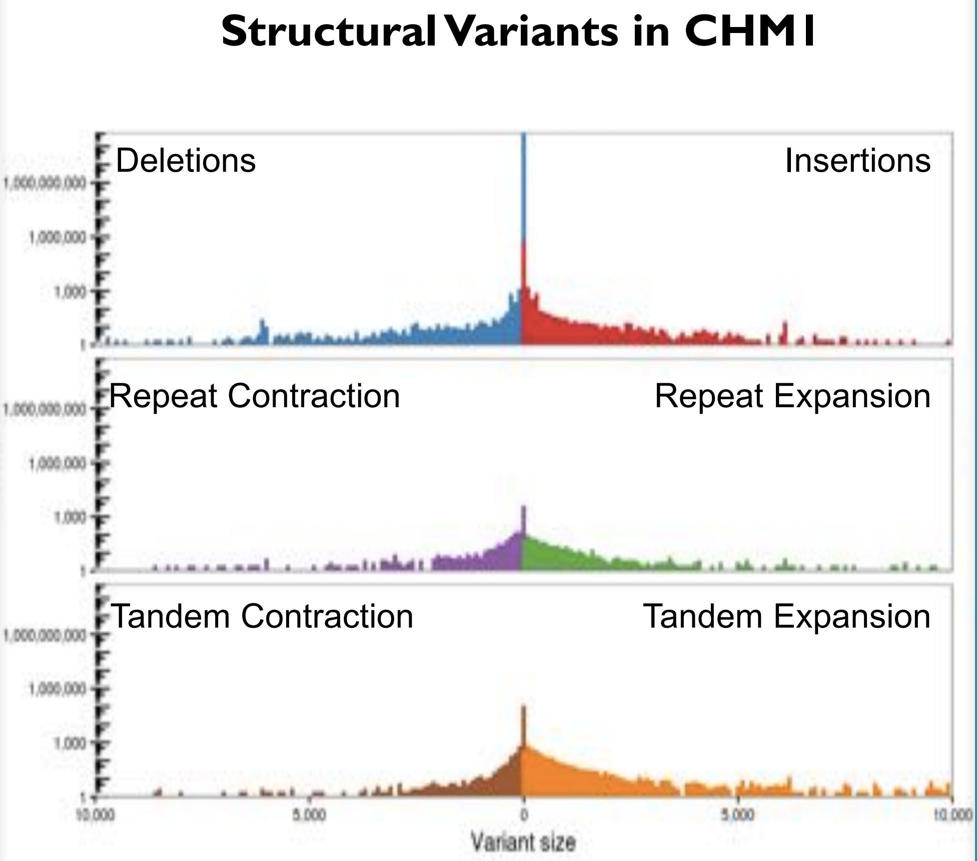


Phased Diploid Genome Assembly with Single Molecule Real-Time Sequencing
Chin et al (2016) *Nature Methods. doi:10.1038/nmeth.4035.*

Recent Long Read Assemblies



Third-generation sequencing and the future of genomics
Lee et al (2016) *bioRxiv*
doi: http://dx.doi.org/10.1101/048603



Assemblytics: a web analytics tool for the detection of variants from an assembly
Nattestad & Schatz (2016) *Bioinformatics.*
doi: 10.1093/bioinformatics/btw369

Illumina Roadmap

Illumina Novaseq

$850k instrument cost
~$1k / human @ 50x
Short reads, high throughput

10X Chromium

$125k instrument costs
~$2k / human
Linked reads, medium throughput

PacBio Roadmap

PacBio Sequel

$350k instrument cost
~$30k / human @ 50x
Long reads, Medium throughput

SMRTcell v2

1M Zero Mode Waveguides
~15kb average read length
~$1000 / SMRTcell

Oxford Nanopore

MinION

$1k / instrument
~$30k / human @ 50x
Long reads, Low throughput

PromethION

$75k / instrument
>>100GB / day
??? / human @ 50x

Oxford Nanopore sequencing, hybrid error correction, and de novo assembly of a eukaryotic genome
Goodwin, S, Gurtowski, J, Ethe-Sayers, S, Deshpande, P, Schatz MC* McCombie, WR* (2015) Genome Research doi: 10.1101/gr.191395.115

Ebola Surveillance

LETTER

doi:10.1038/nature16996

Real-time, portable genome sequencing for Ebola surveillance

Joshua Quick[1]*, Nicholas J. Loman[1]*, Sophie Duraffour[2,3]*, Jared T. Simpson[4,5]*, Ettore Severi[6]*, Lauren Cowley[7]*, Joseph Akoi Bore[2], Raymond Koundouno[2], Gytis Dudas[8], Amy Mikhail[7], Nobila Ouédraogo[9], Babak Afrough[2,10], Amadou Bah[2,11], Jonathan H. J. Baum[2,3], Beate Becker-Ziaja[2,3], Jan Peter Boettcher[2,12], Mar Cabeza-Cabrerizo[2,3], Álvaro Camino-Sánchez[2], Lisa L. Carter[2,13], Juliane Doerrbecker[2,3], Theresa Enkirch[2,14], Isabel García-Dorival[2,15], Nicole Hetzelt[2,12], Julia Hinzmann[2,12], Tobias Holm[2,3], Liana Eleni Kafetzopoulou[2,16], Michel Koropogui[2,17], Abigael Kosgey[2,18], Eeva Kuisma[2,10], Christopher H. Logue[2,10], Antonio Mazzarelli[2,19], Sarah Meisel[2,3], Marc Mertens[2,20], Janine Michel[2,12], Didier Ngabo[2,10], Katja Nitzsche[2,3], Elisa Pallasch[2,3], Livia Victoria Patrono[2,3], Jasmine Portmann[2,21], Johanna Gabriella Repits[2,22], Natasha Y. Rickett[2,15,23], Andreas Sachse[2,12], Katrin Singethan[2,24], Inês Vitoriano[2,10], Rahel L. Yemanaberhan[2,3], Elsa G. Zekeng[2,15,23], Trina Racine[25], Alexander Bello[25], Amadou Alpha Sall[26], Ousmane Faye[26], Oumar Faye[26], N'Faly Magassouba[27], Cecelia V. Williams[28,29], Victoria Amburgey[28,29], Linda Winona[28,29], Emily Davis[29,30], Jon Gerlach[29,30], Frank Washington[29,30], Vanessa Monteil[31], Marine Jourdain[31], Marion Bererd[31], Alimou Camara[31], Hermann Somlare[31], Abdoulaye Camara[31], Marianne Gerard[31], Guillaume Bado[31], Bernard Baillet[31], Déborah Delaune[32,33], Koumpingnin Yacouba Nebie[34], Abdoulaye Diarra[34], Yacouba Savane[34], Raymond Bernard Pallawo[34], Giovanna Jaramillo Gutierrez[35], Natacha Milhano[6,36], Isabelle Roger[34], Christopher J. Williams[6,37], Facinet Yattara[17], Kuiama Lewandowski[10], James Taylor[38], Phillip Rachwal[38], Daniel J. Turner[39], Georgios Pollakis[15,23], Julian A. Hiscox[15,23], David A. Matthews[40], Matthew K. O'Shea[41], Andrew McD. Johnston[41], Duncan Wilson[41], Emma Hutley[42], Erasmus Smit[43], Antonino Di Caro[2,19], Roman Wölfel[2,44], Kilian Stoecker[2,44], Erna Fleischmann[2,44], Martin Gabriel[2,3], Simon A. Weller[38], Lamine Koivogui[45], Boubacar Diallo[34], Sakoba Keïta[17], Andrew Rambaut[8,46,47], Pierre Formenty[34], Stephan Günther[2,3] & Miles W. Carroll[2,10,48,49]

Ebola Surveillance

LETTER

doi:10.1038/nature16996

Real-time, portable genome sequencing for Ebola surveillance

Joshua Quick[1*], Nicholas J. Loman[1*], Sophie Duraffour[2,3*], Jared T. Simpson[4,5*], Ettore Se
Joseph Akoi Bore[2], Raymond Koundouno[2], Cyris Dudas[8], Amy Mikhail[1], Nobila Ouédraog
Amadou Bah[2,11], Jonathan H. J. Baum[2,3], Beate Becker-Ziaja[2,3], Jan Peter Boettcher[2,12], Ma
Álvaro Camino-Sánchez[2], Lisa L. Carter[2,13], Juliane Doerrbecker[2,3], Theresa Enkirch[2,14], Ie
Nicole Hetzelt[2,12], Julia Hinzmann[2,12], Tobias Holm[2,3], Liana Eleni Kafetzopoulou[2,16], Micha
Eeva Kuisma[2,10], Christopher H. Logue[2,10], Antonio Mazzarelli[2,19], Sarah Meisel[2,3], Marc M
Didier Ngabo[2,10], Katja Nitzsche[2,3], Elisa Pallasch[2,3], Livia Victoria Patrono[2,3], Jasmine Port
Natasha Y. Rickett[2,15,23], Andreas Sachse[2,12], Katrin Singethan[2,24], Inês Vitoriano[2,10], Rahel
Elsa G. Zekeng[2,15,22], Trina Racine[25], Alexander Bello[25], Amadou Alpha Sall[26], Ousmane Fa
N'Faly Magassouba[27], Cecelia V. Williams[28,29], Victoria Amburgey[28,29], Linda Winona[28,29], Er
Frank Washington[29,30], Vanessa Monteil[31], Marine Jourdain[31], Marion Berend[31], Alimou Cam
Abdoulaye Camara[31], Marianne Gerard[31], Guillaume Bado[31], Bernard Baillet[31], Déborah Dela
Abdoulaye Diarra[34], Yacouba Savane[34], Raymond Bernard Pallawo[34], Giovanna Jaramillo Gu
Isabelle Roger[34], Christopher J. Williams[6,37], Facinet Yattara[27], Kuiama Lewandowski[20], Jam
Daniel J. Turner[39], Georgios Pollakis[15,23], Julian A. Hiscox[15,23], David A. Matthews[40], Matthe
Andrew McD. Johnston[41], Duncan Wilson[41], Emma Hutley[42], Erasmus Smit[43], Antonino Di
Kilian Stoecker[2,44], Erna Fleischmann[2,44], Martin Gabriel[2,3], Simon A. Weller[18], Lamine Ko
Sakoba Keïta[27], Andrew Rambaut[8,46,47], Pierre Formenty[34], Stephan Günther[2,3] & Miles W. C

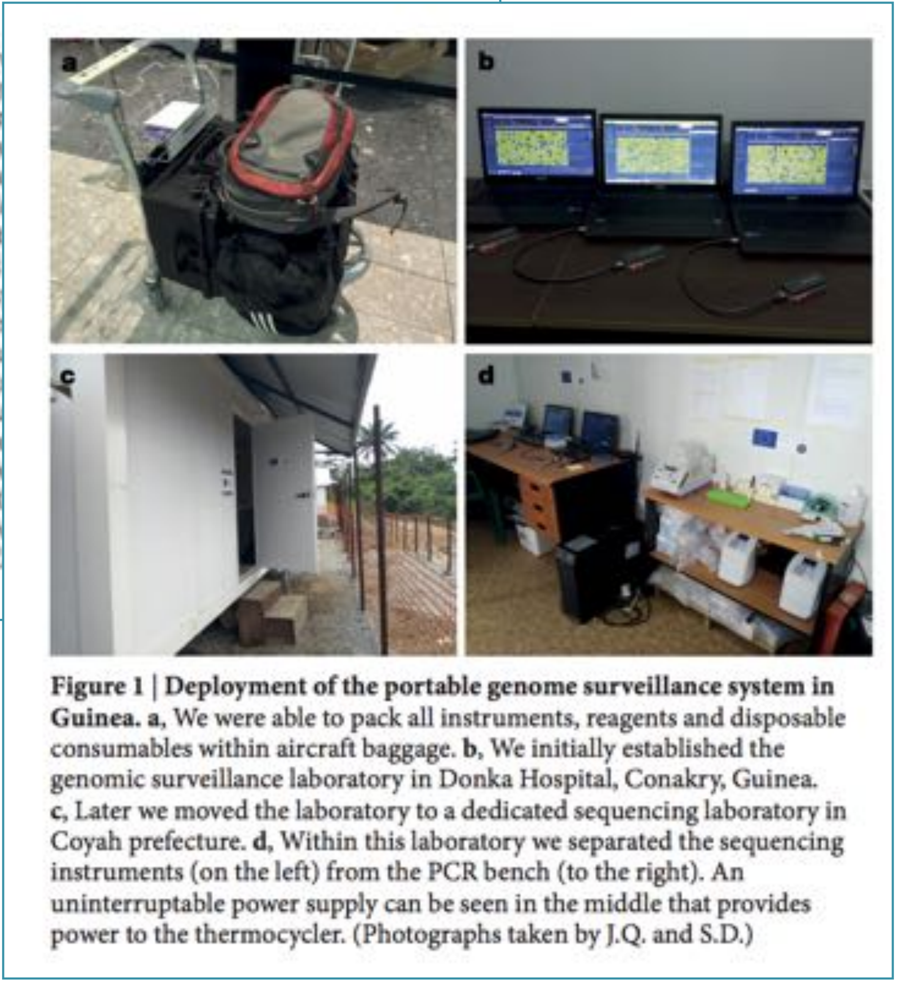


Figure 1 | Deployment of the portable genome surveillance system in Guinea. **a,** We were able to pack all instruments, reagents and disposable consumables within aircraft baggage. **b,** We initially established the genomic surveillance laboratory in Donka Hospital, Conakry, Guinea. **c,** Later we moved the laboratory to a dedicated sequencing laboratory in Coyah prefecture. **d,** Within this laboratory we separated the sequencing instruments (on the left) from the PCR bench (to the right). An uninterruptable power supply can be seen in the middle that provides power to the thermocycler. (Photographs taken by J.Q. and S.D.)

Ebola Surveillance

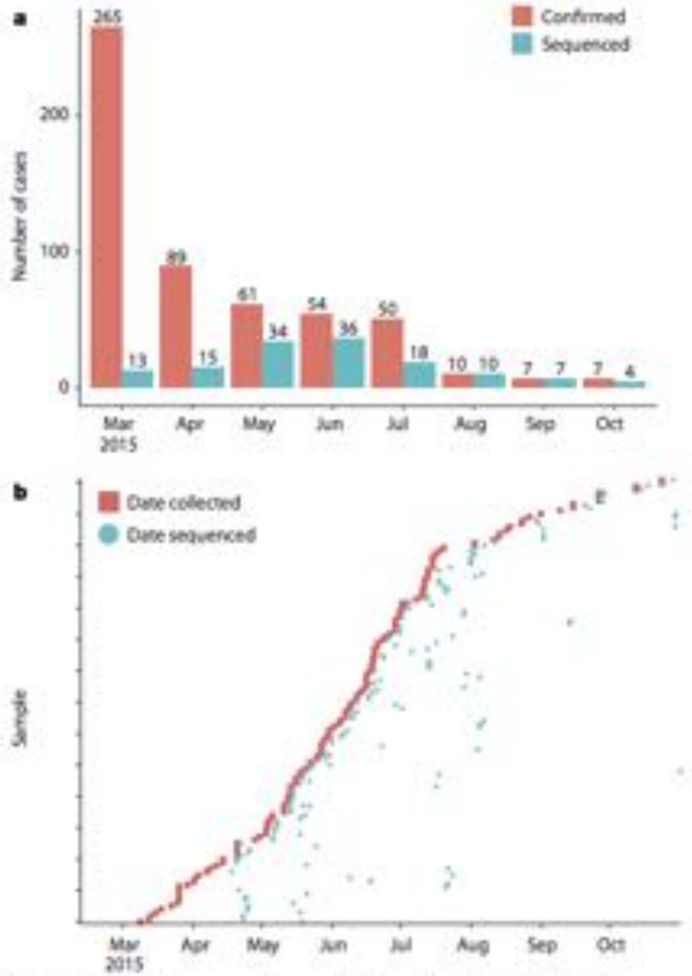


Figure 2 | Real-time genomics surveillance in context of the Guinea Ebola virus disease epidemic. a, Here we show the number of reported cases of Ebola virus disease in Guinea (red) in relation to the number of EBOV new patient samples ($n = 137$, in blue) generated during this study. **b,** For each of the 142 sequenced samples, we show the relationship between sample collection date (red) and the date of sequencing (blue). Twenty-eight samples were sequenced within three days of the sample being taken, and sixty-eight samples within a week. Larger gaps represent retrospective sequencing of cases to provide additional epidemiological context.

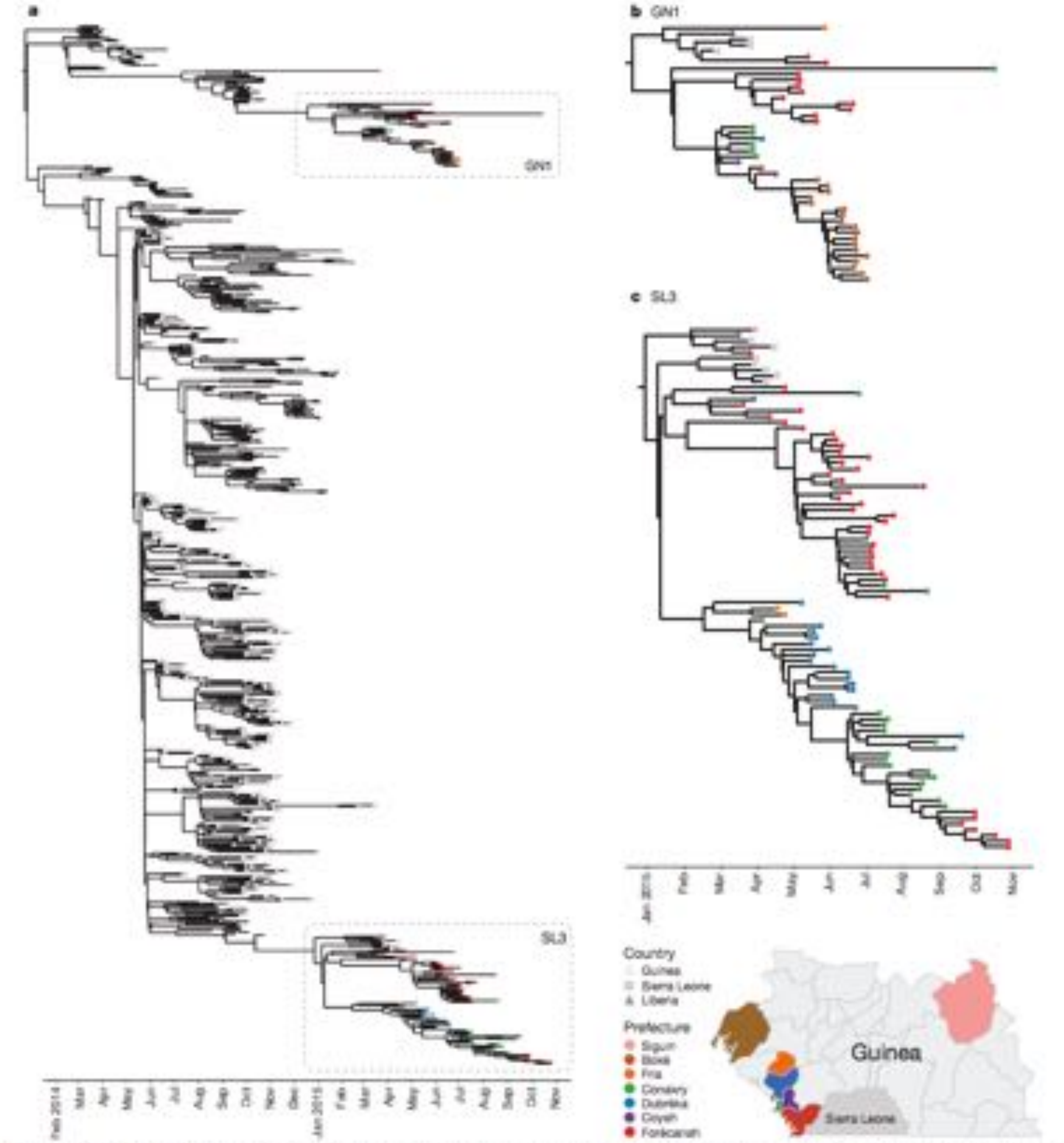
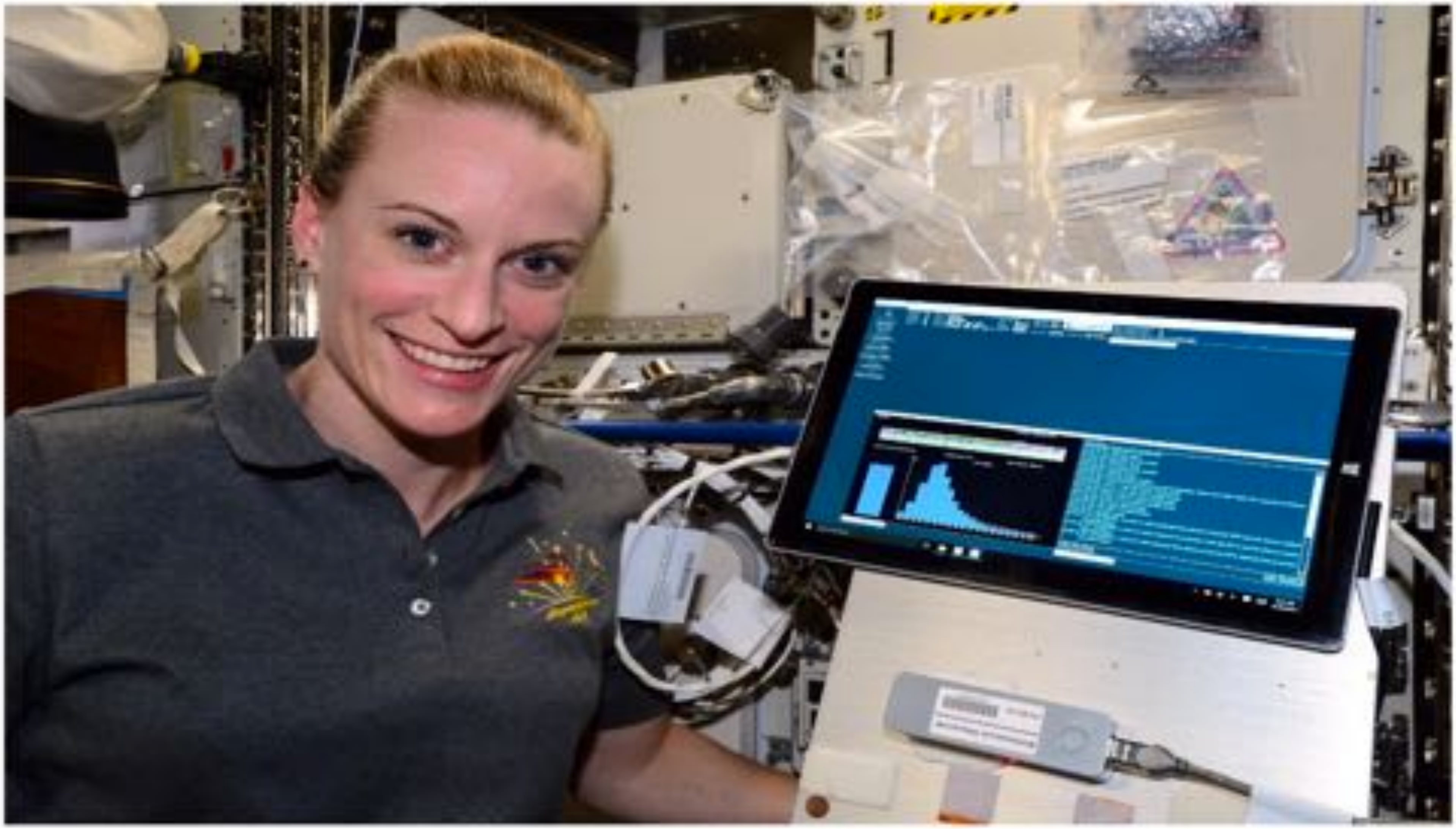


Figure 3 | Evolution of EBOV over the course of the Ebola virus disease epidemic. a, Time-scaled phylogeny of 603 published sequences with 125 high quality sequences from this study. The shape of nodes on the tree demonstrates country of origin. Our results show Guinean samples (coloured circles) belong to two previously identified lineages, GN1 and SL3. **b,** GN1 is deeply branching with early epidemic samples. **c,** SL3 is related to cases identified in Sierra Leone. Samples are frequently clustered by geography (indicated by colour of circle) and this provides information as to origins of new introductions, such as in the Boké epidemic in May 2015. Map figure adapted from SimpleMaps website (http://simplemaps.com/resources/svg-gn).

Extremely Portable Sequencing!



Kate Rubins sequencing DNA on the ISS