Genome Sequencing & Assembly

Michael Schatz

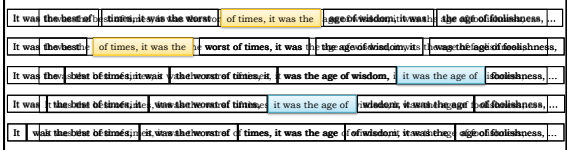
Oct 23, 2014
Programming for Biology

Outline

1. Assembly theory
   1. Assembly by analogy
   2. De Bruijn and Overlap graph
   3. Coverage, read length, errors, and repeats
2. Whole Genome Alignment
   1. Aligning & visualizing with MUMmer
3. Genome assemblers
   1. ALLPATHS-LG: recommended for Illumina-only projects
   2. Celera Assembler: recommended for long read projects
4. Summary & Recommendations

Shredded Book Reconstruction

- Dickens accidentally shreds the first printing of A Tale of Two Cities
  - Text printed on 5 long spools



- How can he reconstruct the text?
  - 5 copies x 138, 656 words / 5 words per fragment = 138k fragments
  - The short fragments from every copy are mixed together
  - Some fragments are identical

Greedy Reconstruction

It was the best of
age of wisdom, it was
best of times, it was
it was the age of
it was the age of
it was the worst of
of times, it was the
of times, it was the
of wisdom, it was the
the age of wisdom, it
the best of times, it
the worst of times, it
times, it was the age
times, it was the worst
was the age of wisdom,
was the age of foolishness,
was the best of times,
was the worst of times,
wisdom, it was the age
worst of times, it was

It was the best of
was the best of times,
the best of times, it
best of times, it was
of times, it was the
of times, it was the
times, it was the worst
times, it was the age

The repeated sequence make the correct reconstruction ambiguous

- It was the best of times, it was the [worst/age]

Model the assembly problem as a graph problem

de Bruijn Graph Construction

- D_k = (V,E)
  - V = All length-k subfragments (k < l)
  - E = Directed edges between consecutive subfragments
    - Nodes overlap by k-1 words

Original Fragment

It was the best of

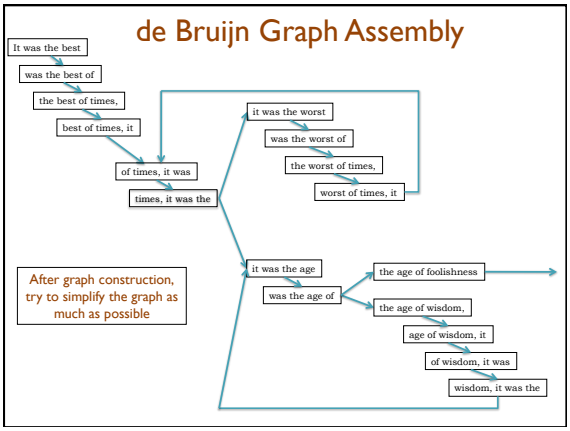
Directed Edge

It was the best → was the best of

- Locally constructed graph reveals the global sequence structure
  - Overlaps between sequences implicitly computed

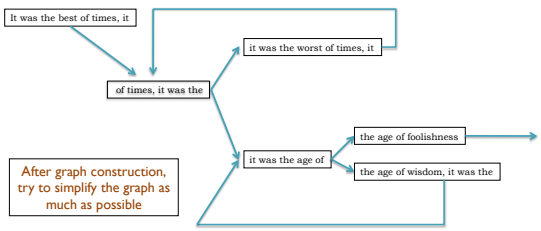
de Bruijn, 1946
Idury and Waterman, 1995
Pevzner, Tang, Waterman, 2001

de Bruijn Graph Assembly



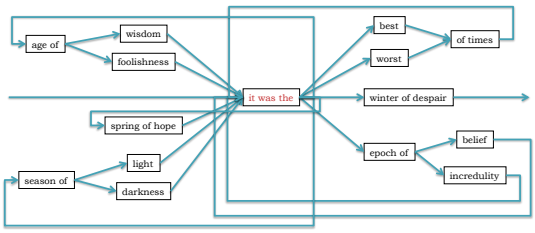
After graph construction, try to simplify the graph as much as possible

de Bruijn Graph Assembly



The full tale

… it was the best of times it was the worst of times …
… it was the age of wisdom it was the age of foolishness …
… it was the epoch of belief it was the epoch of incredulity …
… it was the season of light it was the season of darkness …
… it was the spring of hope it was the winder of despair …



Milestones in Genome Assembly

1977. Sanger *et al.*
1st Complete Organism
5375 bp

1995. Fleischmann *et al.*
1st Free Living Organism
TIGR Assembler. 1.8Mbp

1998. C.elegans SC
1st Multicellular Organism
BAC-by-BAC Phrap. 97Mbp

2000. Myers *et al.*
1st Large WGS Assembly.
Celera Assembler. 116 Mbp

2001. Venter *et al.*, IHGSC
Human Genome
Celera Assembler/GigaAssembler. 2.9 Gbp

2010. Li *et al.*
1st Large SGS Assembly.
SOAPdenovo 2.2 Gbp

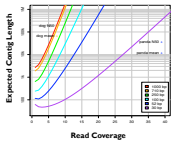
Like Dickens, we must computationally reconstruct a genome from short fragments

Assembly Applications

- Novel genomes
- Metagenomes
- Sequencing assays
  - Structural variations
  - Transcript assembly
  - …

Ingredients for a good assembly

Coverage



High coverage is required

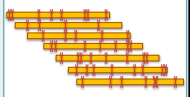
- Oversample the genome to ensure every base is sequenced with long overlaps between reads
- Biased coverage will also fragment assembly

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

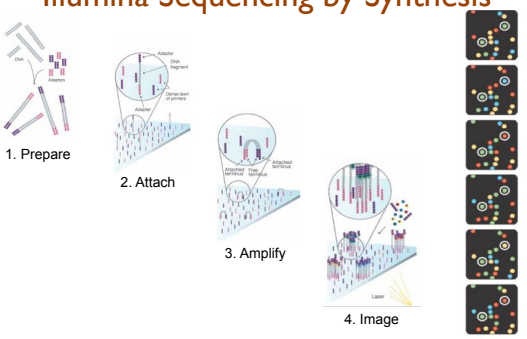


Errors obscure overlaps

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

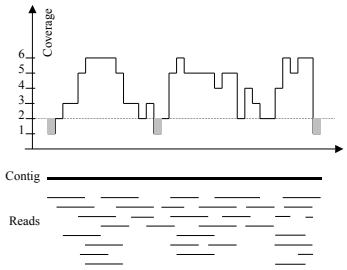
Current challenges in *de novo* plant genome sequencing and assembly
Schatz MC, Witkowski, McCombie, WR (2012) *Genome Biology*. 12:243

Illumina Sequencing by Synthesis



Metzker (2010) Nature Reviews Genetics 11:31-46
http://www.youtube.com/watch?v=I99aKKHcxC4

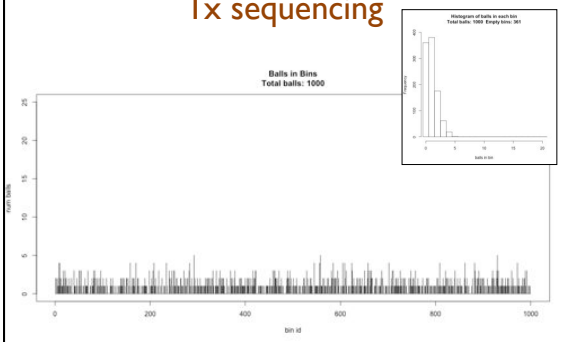
Typical sequencing coverage



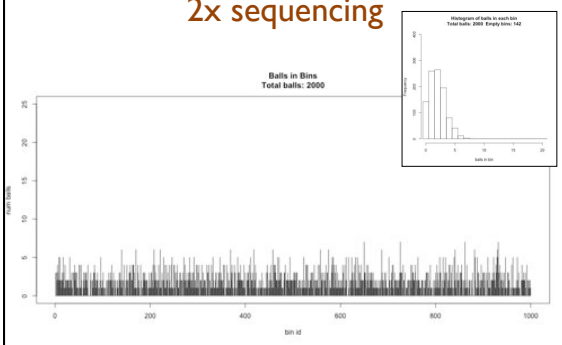
Imagine raindrops on a sidewalk

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

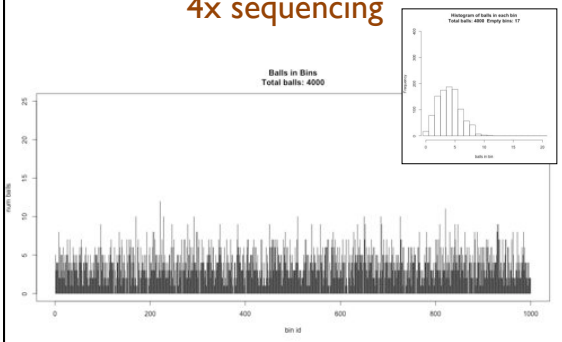
1x sequencing



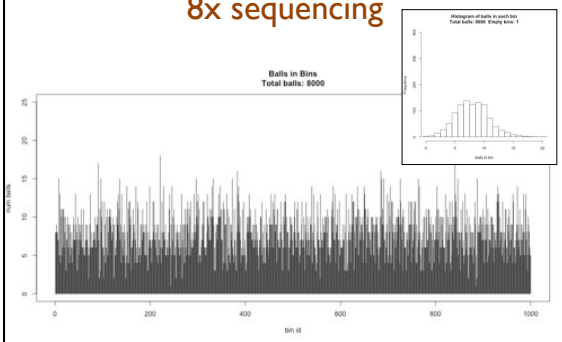
2x sequencing



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.

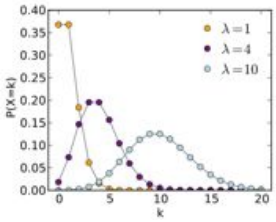
Formulation comes from the limit of the binomial equation

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

Key property:

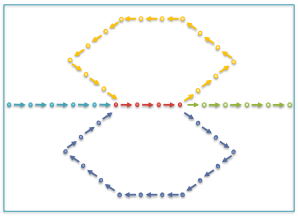
- ***The standard deviation is the square root of the mean.***

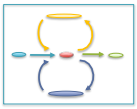
$$P(k) = \frac{\lambda^k}{k!} e^{-\lambda}$$



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



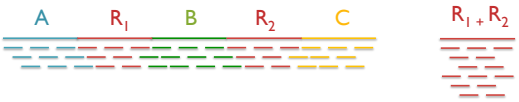


Repetitive regions

Repeat Type	Definition / Example	Prevalence
Low-complexity DNA / Microsatellites	$(b_1b_2\ldots b_k)^N$ where $1 \leq k \leq 6$ 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

Repeats and Coverage Statistics



- If n reads are a uniform random sample of the genome of length G, we expect $k=n\Delta/G$ reads to start in a region of length Δ.
  - If we see many more reads than k (if the arrival rate is > A) , it is likely to be a collapsed repeat

$$\Pr(X-copy) = \binom{n}{k}\left(\frac{X\Delta}{G}\right)^k\left(\frac{G-X\Delta}{G}\right)^{n-k}$$

$$A(\Delta,k) = \ln\left(\frac{\Pr(1-copy)}{\Pr(2-copy)}\right) = \ln\left(\frac{\frac{(\Delta n/G)^k}{k!}e^{\frac{-\Delta n}{G}}}{\frac{(2\Delta n/G)^k}{k!}e^{\frac{-2\Delta n}{G}}}\right) = \frac{n\Delta}{G} - k\ln 2$$

The fragment assembly string graph
Myers, EW (2005) Bioinformatics. 21(suppl 2): ii79-85.

Paired-end and Mate-pairs

Paired-end sequencing

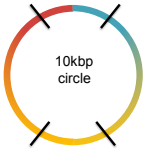
- Read one end of the molecule, flip, and read the other end
- Generate pair of reads separated by up to 500bp with inward orientation

300bp

Mate-pair sequencing

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

10kbp

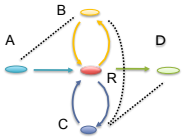


2x100 @ ~10kbp (outies)

2x100 @ 300bp (innies)

Scaffolding

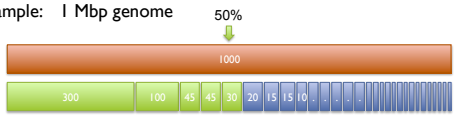
- Initial contigs (*aka* unipaths, unitigs) terminate at
  - *Coverage gaps*: especially extreme GC
  - *Conflicts*: errors, repeat boundaries
- Use mate-pairs to resolve correct order through assembly graph
  - Place sequence to satisfy the mate constraints
  - Mates through repeat nodes are tangled
- 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



N50 size

Def: 50% of the genome is in contigs as large as the N50 value

Example: 1 Mbp genome



N50 size = 30 kbp
(300k+100k+45k+45k+30k = 520k >= 500kbp)

A greater N50 is indicative of improvement in every dimension:

- Better resolution of genes and flanking regulatory regions
- Better resolution of transposons and other complex sequences
- Better resolution of chromosome organization
- Better sequence for all downstream analysis

Outline

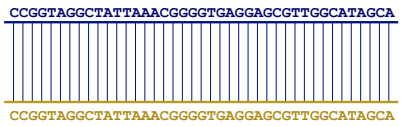
1. Assembly theory
   1. Assembly by analogy
   2. De Bruijn and Overlap graph
   3. Coverage, read length, errors, and repeats
2. Whole Genome Alignment
   1. Aligning & visualizing with MUMmer
3. Genome assemblers
   1. ALLPATHS-LG: recommended for Illumina-only projects
   2. Celera Assembler: recommended for long read projects
4. Summary and Recommendations

Whole Genome Alignment with MUMmer

Slides Courtesy of Adam M. Phillippy
University of Maryland

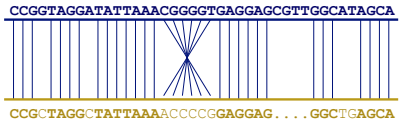
Goal of WGA

- For two genomes, *A* and *B*, find a mapping from each position in *A* to its corresponding position in *B*



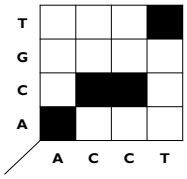
Not so fast...

- Genome *A* may have insertions, deletions, translocations, inversions, duplications or SNPs with respect to *B* (sometimes all of the above)

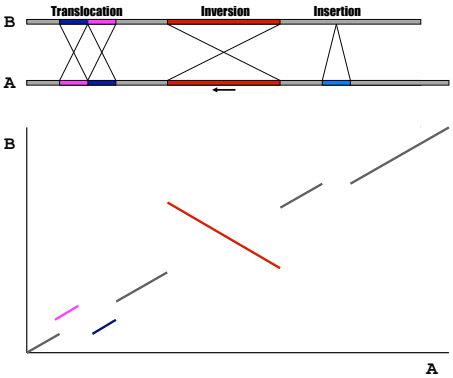


WGA visualization

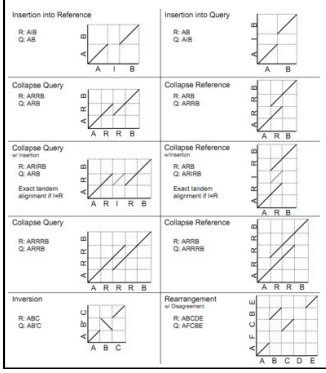
- How can we visualize *whole* genome alignments?
- With an alignment dot plot
  - *N* x *M* matrix
    - Let *i* = position in genome *A*
    - Let *j* = position in genome *B*
    - Fill cell *(i,j)* if A_i shows similarity to B_j



  - A perfect alignment between *A* and *B* would completely fill the positive diagonal

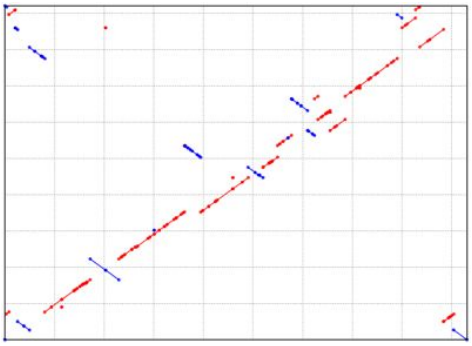


SV Types



- Different structural variation types / misassemblies will be apparent by their pattern of breakpoints
- Most breakpoints will be at or near repeats
- Things quickly get complicated in real genomes

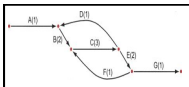
http://mummer.sf.net/manual/AlignmentTypes.pdf



Alignment of 2 strains of *Y. pestis*
http://mummer.sourceforge.net/manual/

Outline

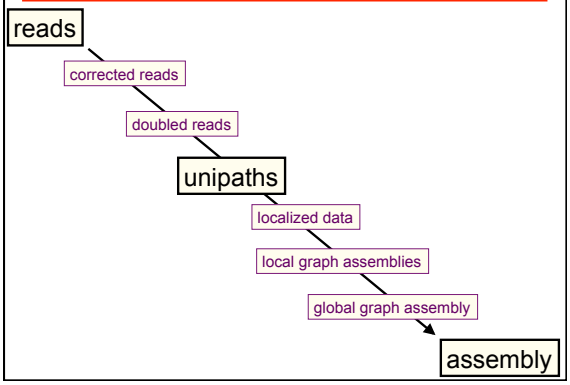
1. Assembly theory
   1. Assembly by analogy
   2. De Bruijn and Overlap graph
   3. Coverage, read length, errors, and repeats
2. Whole Genome Alignment
   1. Aligning & visualizing with MUMmer
3. Genome assemblers
   1. ALLPATHS-LG: recommended for Illumina-only projects
   2. Celera Assembler: recommended for long read projects
4. Summary and Recommendations



Genome assembly with ALLPATHS-LG
Iain MacCallum

BROAD INSTITUTE

How ALLPATHS-LG works



ALLPATHS-LG sequencing model

Libraries (insert types)	Fragment size (bp)	Read length (bases)	Sequence coverage (x)	Required
Fragment	180*	≥ 100	45	yes
Short jump	3,000	≥ 100 preferable	45	yes
Long jump	6,000	≥ 100 preferable	5	no**
Fosmid jump	40,000	≥ 26	1	no**

*See next slide.

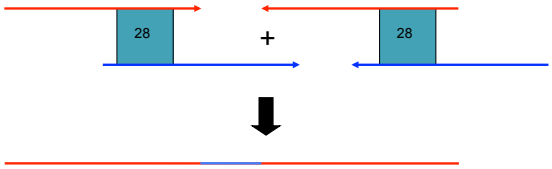
**For best results. Normally not used for small genomes. However essential to assemble long repeats or duplications.

Cutting coverage in half still works, with some reduction in quality of results.

All: protocols are either available, or in progress.

Read doubling

To close a read pair (red), we require the existence of another read pair (blue), overlapping perfectly like this:

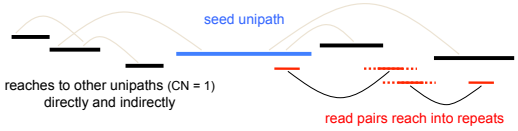


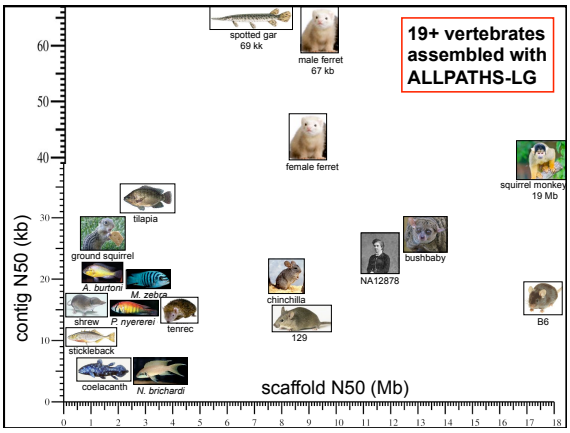
More than one closure allowed (but rare).

Localization

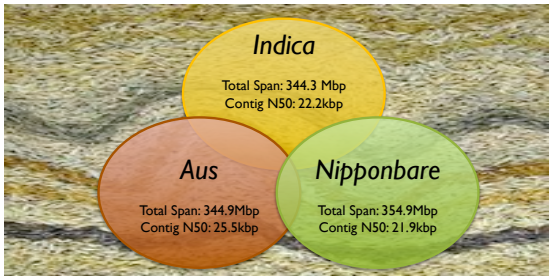
I. Find 'seed' unipaths, evenly spaced across genome
(ideally long, of copy number CN = 1)

II. Form neighborhood around each seed





Population structure of *Oryza sativa*



Whole genome de novo assemblies of three divergent strains of rice (*O. sativa*) documents novel gene space of *aus* and *indica*
Schatz, MC, Maron, L, Stein, *et al* (2014) *In press.*

Strain specific regions

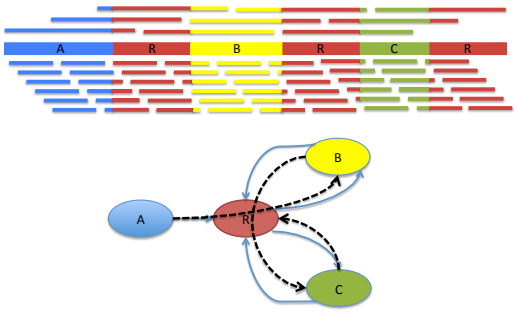
(A) Nipponbare

Conclusions

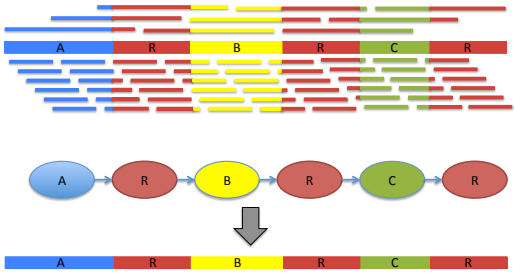
- Very high quality representation of the "gene-space"
  - Overall identity ~99.9%
  - Less than 1% of exonic bases missing
- Genome-specific genes enriched for disease resistance
  - Reflects their geographic and environmental diversity
  - Detailed analysis of agriculturally important loci
- Assemblies fragmented at (high copy) repeats
  - Missing regions have mean k-mer coverage >10,000x
  - Difficult to identify full length gene models and regulatory features

Genome assembly with the Celera Assembler

Assembly Complexity

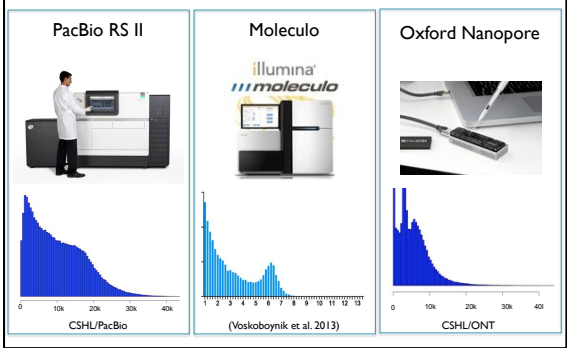


Assembly Complexity

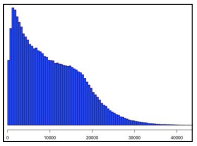


The advantages of SMRT sequencing
Roberts, RJ, Carneiro, MO, Schatz, MC (2013) *Genome Biology*. 14:405

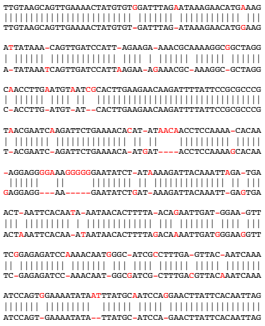
Long Read Sequencing Technology



SMRT Sequencing Data



Match	83.7%
Insertions	11.5%
Deletions	3.4%
Mismatch	1.4%



Sample of 100k reads aligned with BLASR requiring >100bp alignment

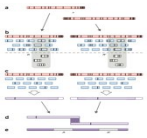
PacBio Assembly Algorithms

PBJelly

Gap Filling and Assembly Upgrade

English *et al* (2012)
PLOS One. 7(11): e47768

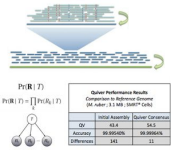
PacBioToCA & ECTools



Hybrid/PB-only Error Correction

Koren, Schatz, *et al* (2012)
Nature Biotechnology. 30:693–700

HGAP & Quiver

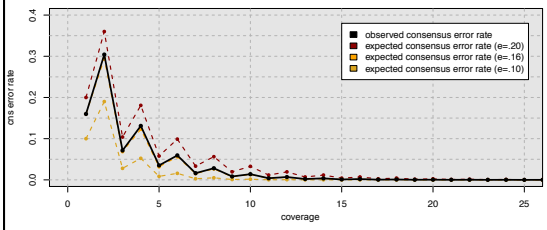


PB-only Correction & Polishing

Chin *et al* (2013)
Nature Methods. 10:563–569

< 5x — PacBio Coverage — > 50x

Consensus Accuracy and Coverage



Coverage can overcome random errors

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

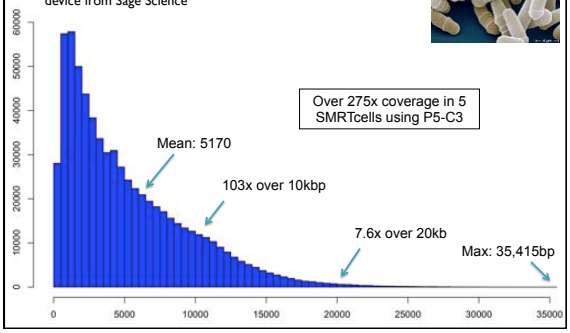
$$CNS\ Error = \sum_{i=\lceil c/2 \rceil}^{c} \binom{c}{i} (e)^{i} (1-e)^{n-i}$$

Koren, Schatz, *et al* (2012)
Nature Biotechnology. 30:693–700

S. pombe dg21

PacBio RS II sequencing at CSHL

- Size selection using an 7 Kb elution window on a BluePippin™ device from Sage Science



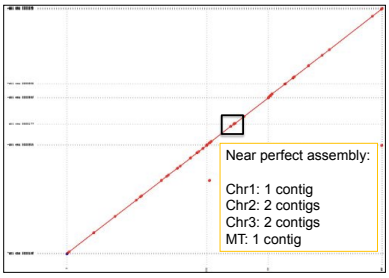
S. pombe dg21

ASM294 Reference sequence

- 12.6Mbp; 3 chromo + mitochondria; N50: 4.53Mbp

PacBio assembly using HGAP + Celera Assembler

- 12.7Mbp; 13 non-redundant contigs; N50: 3.83Mbp; >99.98% id



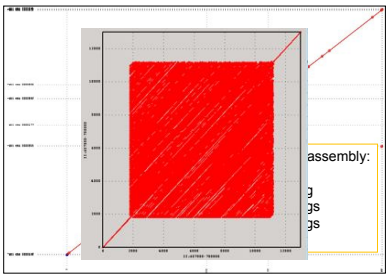
S. pombe dg21

ASM294 Reference sequence

- 12.6Mbp; 3 chromo + mitochondria; N50: 4.53Mbp

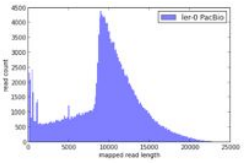
PacBio assembly using HGAP + Celera Assembler

- 12.7Mbp; 13 non-redundant contigs; N50: 3.83Mbp; >99.98% id



A. thaliana Ler-0

http://blog.pacificbiosciences.com/2013/08/new-data-release-arabidopsis-assembly.html



A. thaliana Ler-0 sequenced at PacBio

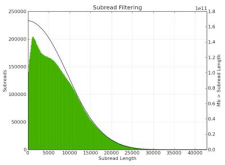
- Sequenced using the previous P4 enzyme and C2 chemistry
- Size selection using an 8 Kb to 50 Kb elution window on a BluePippin™ device from Sage Science
- Total coverage >119x

Genome size:	124.6 Mbp	Sum of Contig Lengths:	149.5Mb
Chromosome N50:	23.0 Mbp	N50 Contig Length:	8.4 Mb
Corrected coverage:	20x over 10kb	Number of Contigs:	1788

High quality assembly of chromosome arms
Assembly Performance: 8.4Mbp/23Mbp = 36%
MiSeq assembly: 63kbp/23Mbp = .2%

Human CHM1

http://blog.pacificbiosciences.com/2014/02/data-release-54x-long-read-coverage-for.html



CHM1hert sequenced at PacBio

- Sequenced using the P5 enzyme and C3 chemistry
- Size selection using an 20kb elution window on a BluePippin™ device from Sage Science
- Total coverage: 54x

Genome size:	3.0 Gb	Sum of Contig Lengths:	3.2 Gb
Chromosome N50:	90.5 Mbp	N50 Contig Length:	4.38 Mbp
Average read length:	7,680 bp	Max Contig:	44 Mbp

High quality draft assembly
Assembly Performance: 4.38Mbp/90.5Mbp = 4.5%
Sanger HuRef assembly: 107kbp / 90.5Mbp = .1%

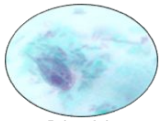
Current Collaborations

Indica & Aus Rice
McCombie/Ware/McCouch

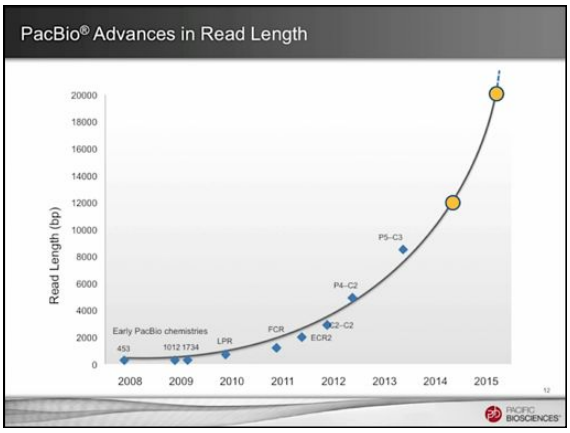
Pinapple
UIUC

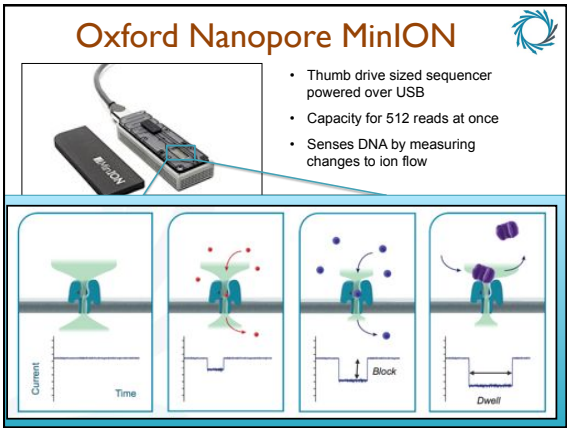
M. ligano
Hannon

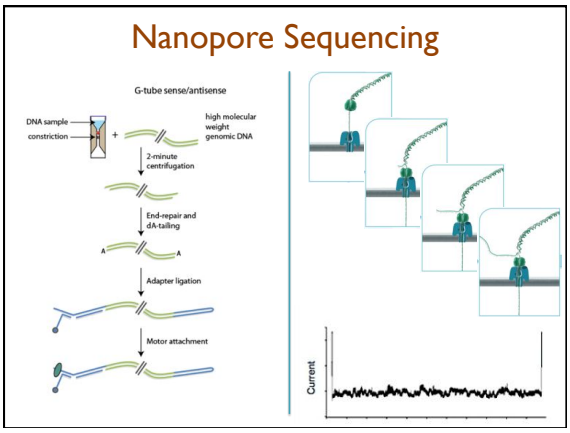
Asian Sea Bass
Temasek Life Sciences Laboratory

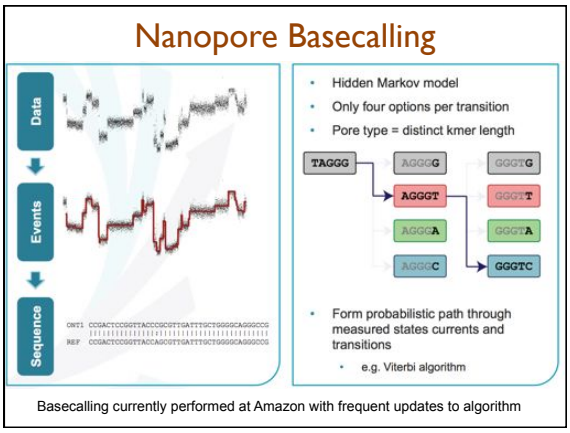


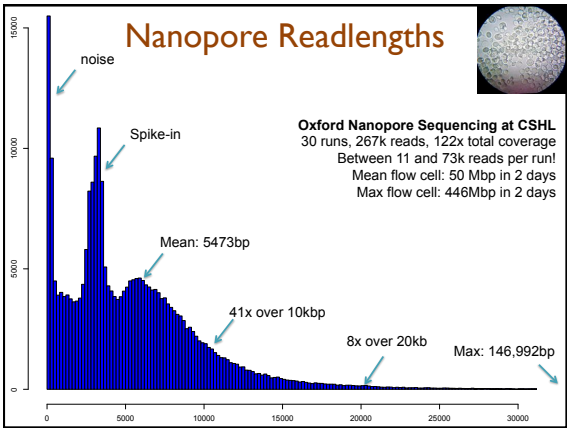
P. hominis
NYU

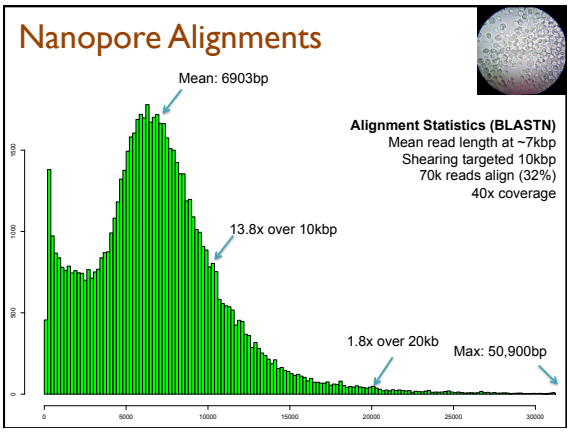

PacBio® Advances in Read Length
Read Length (bp)
20000
18000
16000
14000
12000
10000
8000
6000
4000
2000
0
Early PacBio chemistries
453
1012 1734
LPR
FCR
ECR2
C2-C2
P4-C2
P5-C3
2008
2009
2010
2011
2012
2013
2014
2015
PACIFIC BIOSCIENCES


Oxford Nanopore MinION
Thumb drive sized sequencer powered over USB
Capacity for 512 reads at once
Senses DNA by measuring changes to ion flow
Current
Time
Block
Dwell

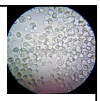

Nanopore Sequencing
G-tube sense/antisense
DNA sample
constriction
high molecular weight genomic DNA
2-minute centrifugation
End-repair and dA-tailing
Adapter ligation
Motor attachment
Current


Nanopore Basecalling
Data
Events
Sequence
Hidden Markov model
Only four options per transition
Pore type = distinct kmer length
TAGGG
AGGGG
GGGTG
AGGGT
GGGTT
AGGGA
GGGTA
AGGGC
GGGTC
Form probabilistic path through measured states currents and transitions
e.g. Viterbi algorithm
Basecalling currently performed at Amazon with frequent updates to algorithm

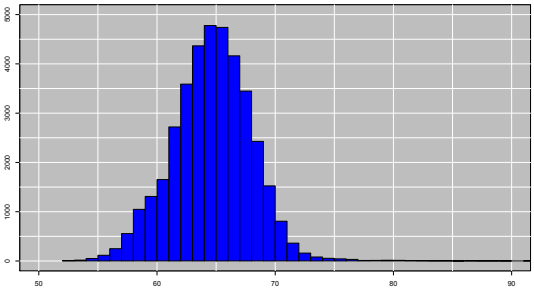

Nanopore Readlengths
noise
Spike-in
Oxford Nanopore Sequencing at CSHL
30 runs, 267k reads, 122x total coverage
Between 11 and 73k reads per run!
Mean flow cell: 50 Mbp in 2 days
Max flow cell: 446Mbp in 2 days
Mean: 5473bp
41x over 10kbp
8x over 20kb
Max: 146,992bp
0
5000
10000
15000
20000
25000
30000


Nanopore Alignments
Mean: 6903bp
Alignment Statistics (BLASTN)
Mean read length at ~7kbp
Shearing targeted 10kbp
70k reads align (32%)
40x coverage
13.8x over 10kbp
1.8x over 20kb
Max: 50,900bp
0
5000
10000
15000
20000
25000
30000

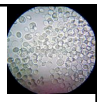
Nanopore Accuracy



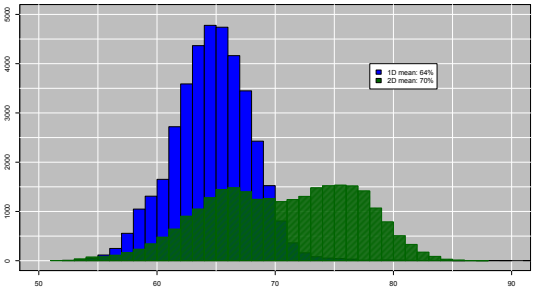
Alignment Quality (BLASTN)
Of reads that align, average ~64% identity



Nanopore Accuracy



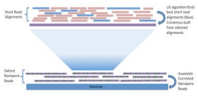
Alignment Quality (BLASTN)
Of reads that align, average ~64% identity
"2D base-calling" improves to ~70% identity

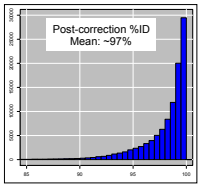


NanoCorr: Nanopore-Illumina Hybrid Error Correction

https://github.com/jgurtowski/nanocorr

1. BLAST Miseq reads to all raw Oxford Nanopore reads
2. Select non-repetitive alignments
   - First pass scans to remove "contained" alignments
   - Second pass uses Dynamic Programming (LIS) to select set of high-identity alignments with minimal overlaps
3. Compute consensus of each Oxford Nanopore read
   - Currently using Pacbio's pbdagcon

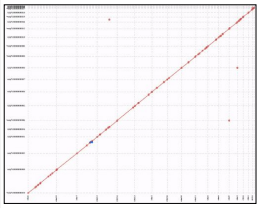




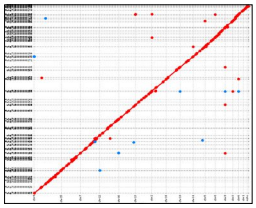
Long Read Assembly

S288C Reference sequence
- 12.1Mbp; 16 chromo + mitochondria; N50: 924kbp

Pacific Biosciences
HGAP + Celera Assembler
- 21 non-redundant contigs
- N50: 811kbp >99.8% id



Oxford Nanopore
NanoCorr + Celera Assembler
- 234 non-redundant contigs
- N50: 362kbp >99.78% id



Platform Technology:

Instruments for scaled nanopore analysis

What should we expect from an assembly?

Analysis of dozens of genomes from across the tree of life with real and simulated data

Summary & Recommendations

< 100 Mbp: HGAP/PacBio2CA @ 100x PB C3-P5
expect near perfect chromosome arms

< 1GB: HGAP/PacBio2CA @ 100x PB C3-P5
high quality assembly: contig N50 over 1Mbp

> 1GB: hybrid/gap filling
expect contig N50 to be 100kbp – 1Mbp

> 5GB: Email mschatz@cshl.edu

Error correction and assembly complexity of single molecule sequencing reads.
Lee, H*, Gurtowski, J*, Yoo, S, Marcus, S, McCombie, WR, Schatz, MC
http://www.biorxiv.org/content/early/2014/06/18/006395

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

Acknowledgements

Schatz Lab
Rahul Amin
Tyler Gavin
James Gurtowski
Han Fang
Hayan Lee
Maria Nattestad
Aspyn Palatnick
Srividya Ramakrishnan

Eric Biggers
Ke Jiang
Shoshana Marcus
Giuseppe Narzisi
Rachel Sherman
Greg Vurture
Alejandro Wences

CSHL
Hannon Lab
Gingeras Lab
Jackson Lab
Hicks Lab
Iossifov Lab
Levy Lab
Lippman Lab
Lyon Lab
Martienssen Lab
McCombie Lab
Tuveson Lab
Ware Lab
Wigler Lab

Pacific Biosciences
Oxford Nanopore

NSF
National Human Genome Research Institute
U.S. DEPARTMENT OF ENERGY
SFARI
SIMONS FOUNDATION AUTISM RESEARCH INITIATIVE

Thank you

http://schatzlab.cshl.edu
@mike_schatz