de novo Genome Assembly

Presented by Peter R. Hoyt

Originally Authored by Haibao Tang

J. Craig Venter Institute

OSU Bioinformatics Workshop
(Aug-14-2012: Modified April-2014)

J. Craig Venter™
INSTITUTE

Who am I and how did I get here?

Peter R. Hoyt, Ph.D., Graduate Program Director, Bioinformatics Certificate
Oklahoma State University, Department of Biochemistry and Molecular Biology

Marine Microbiology (70's)
- Antibiotics/plasmids
- Gut flora

Work (70's)
- Protein production
- P2 cloning facility

Human Genetics & Cell Biology (80's)
- Protein-DNA
- Gene expression
- Chromatin Structure
- Worlds first Genome Browser

ORNL (90's)
- Virology in mice
- Transgenic mice
- Targeted Mutations In Mice/ES cells
- Developmental Biology

Univ. of TN (00's)
- Protein-DNA imaging
- Gene mapping by AFM
- (Human Genome Project)
- Automated sample prep
- Microarrays

OSU (10's)
- Director of Bioinformatics!
- Director of Microarrays
- Bioinformatics Certificate
- Stem Cell Commitment
- Workshops
- New curricula

OKLAHOMA STATE UNIVERSITY

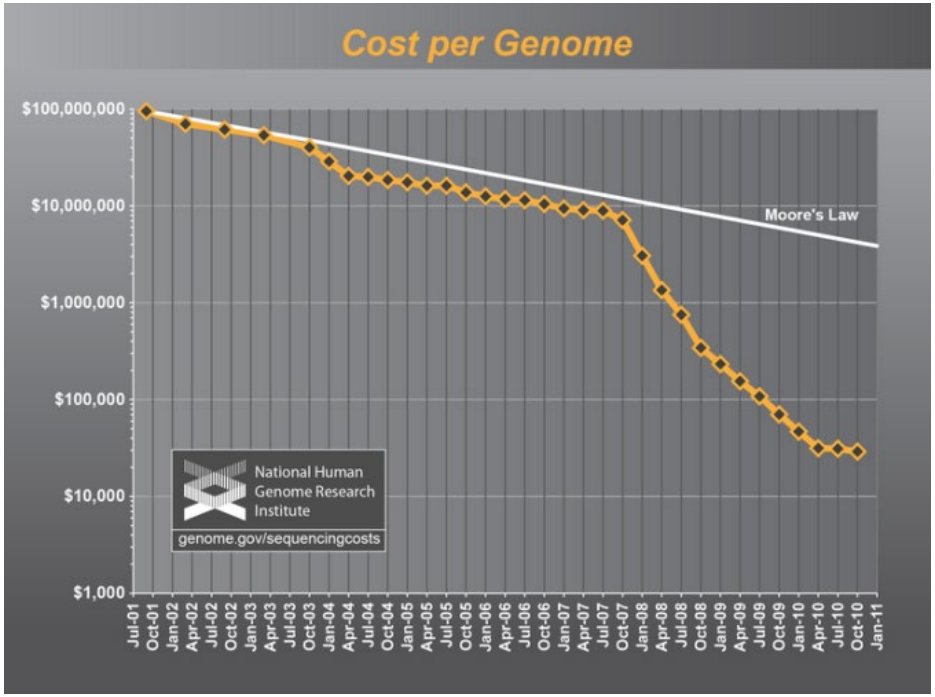
J. Craig Venter INSTITUTE

Sequencing Gets Cheaper and Faster

(How we all got here)

Cost of one human genome

- HGP: $3 billion*
- 2004: $30,000,000
- 2008: $100,000
- 2010: $10,000
- 2011: $4,000
- **2013: $1,000**
- 2014-: $800- ??



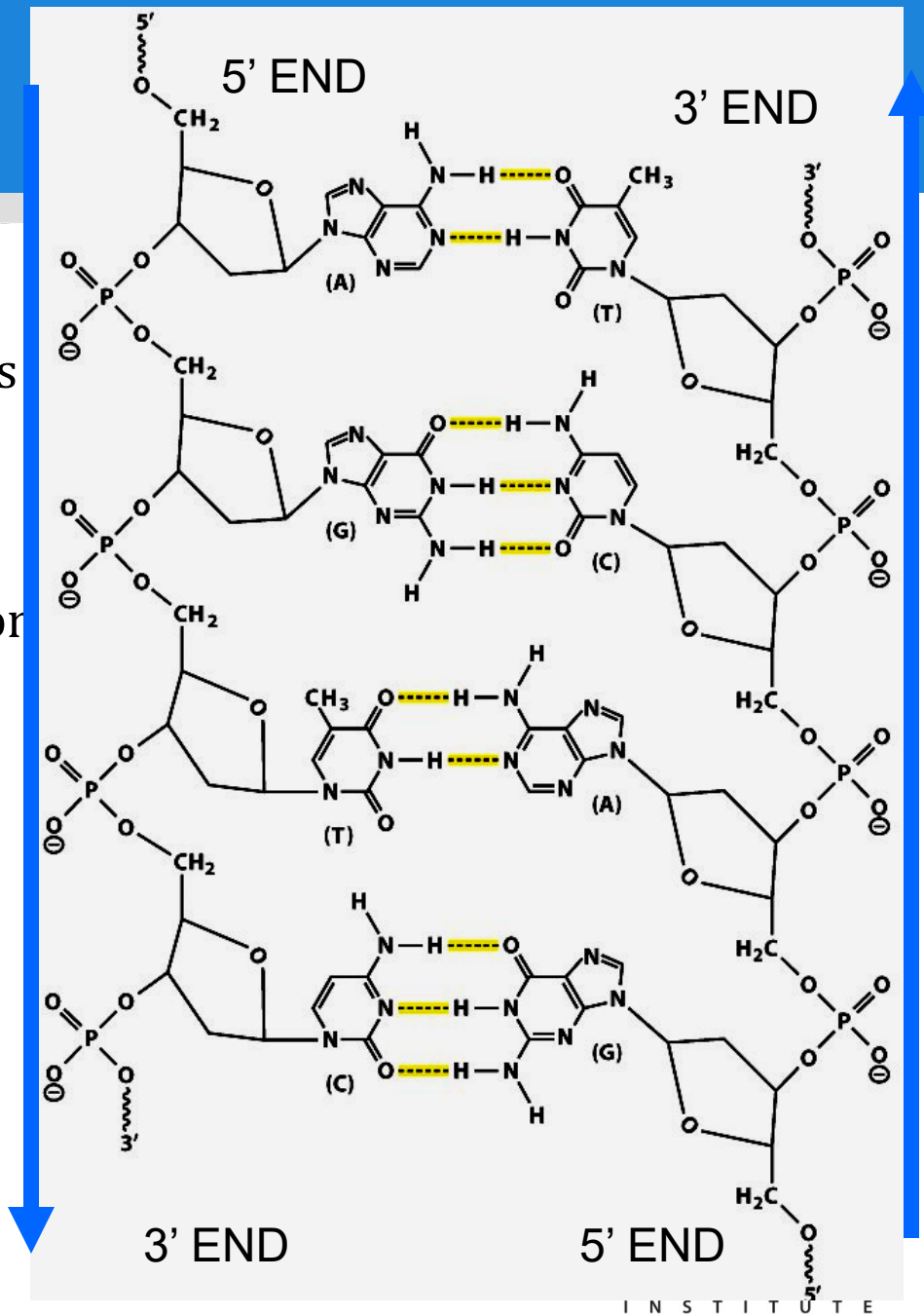
Time to sequence one genome: years/months → hours/days

Massive parallelization.

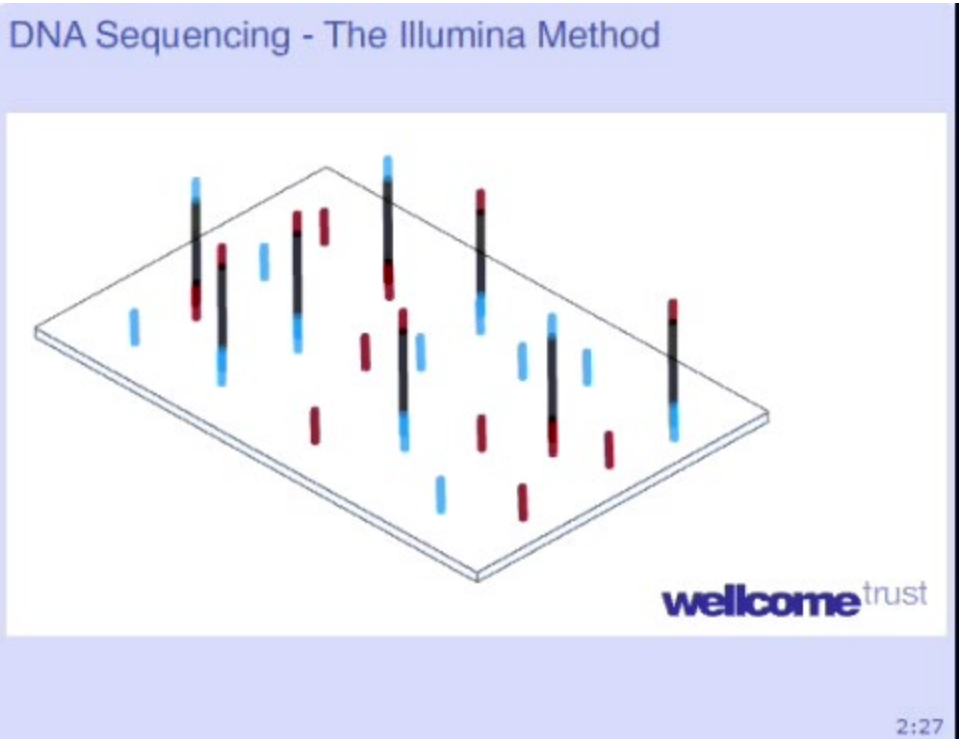
J. Craig Venter INSTITUTE

DNA

- Most DNA's consist of two strands of phosphate-backbone-linked nucleotides
- Strands are in opposite orientation (reverse complementation)
- Held together by hydrogen bonding between bases
  - Adenine with Thymine
  - Guanine with Cytosine

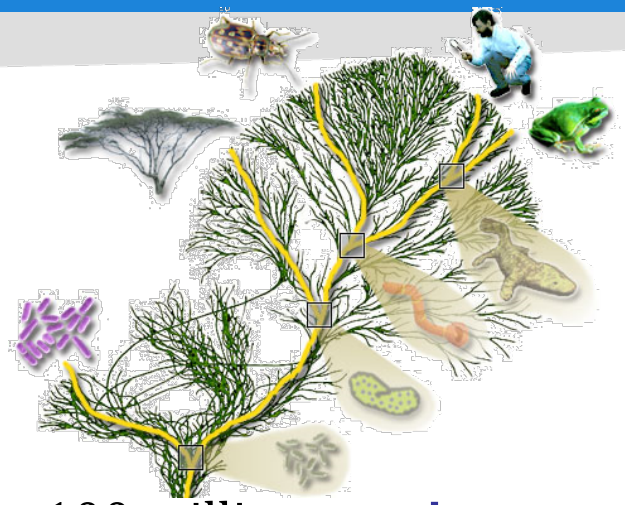


1. DNA is broken to small fragments and different linkers are attached on each end
2. The linker-ed DNA is hybridized to PRIMERS on a surface and PCR amplified
3. The fragments are sequenced by synthesis using fluorescent nucleotides.
4. As each nucleotide is added, a picture is taken and the fluorescent images identify which base was added to each position on the surface



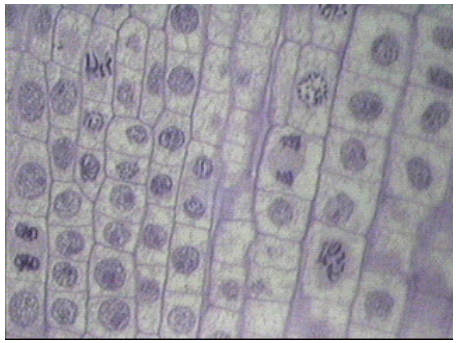
J. Craig Venter INSTITUTE

Many genomes to sequence



100 million **species**
(e.g. phylogeny)

7 billion **individuals**
(SNP, personal genomics)



10^{13} **cells** in a human
(e.g. somatic mutations, cancer)

Genome assembly = JIGSAW puzzle

Unknown Genome: AGCTATAGCGCTATCGTAGCTAGCGCTAGCT

↓ Next-generation sequencing machine

AGCTATAG CTATAGCG
GCTAGCGC CGCTAGCT
TCTAGCGC CGCTATCG
AGCTAGCG ATCGTAGG

↓ Genome assembly software

AGCTATAG GCTAGCGC
TCTAGCGC AGCTAGCG
CTATAGCG ATCGTAGG CGCTAGCT
CGCTATCG

↓

Reconstructed genome : AGCTATAGCGCTATCGTAGCTAGCGCTAGCT

J. Craig Venter INSTITUTE

A difficult JIGSAW puzzle!

- Millions/Billions of pieces
- Lots of malformed pieces
- Often missing pieces
- Pieces mixed from another puzzle
- Lots of similar blue sky pieces…
- If *de novo* you… don't even know the final picture

Credits: Jason Miller, JCVI

Todays Outline

- Assembly preparation – reads, libraries, QC, etc.
- Assembly – OLC assemblers vs. de Bruijn (K-mer) graph assemblers
- Assembly QC - Identify data or assembly issues
- Assembly curation - Further scaffolding, build chromosomes

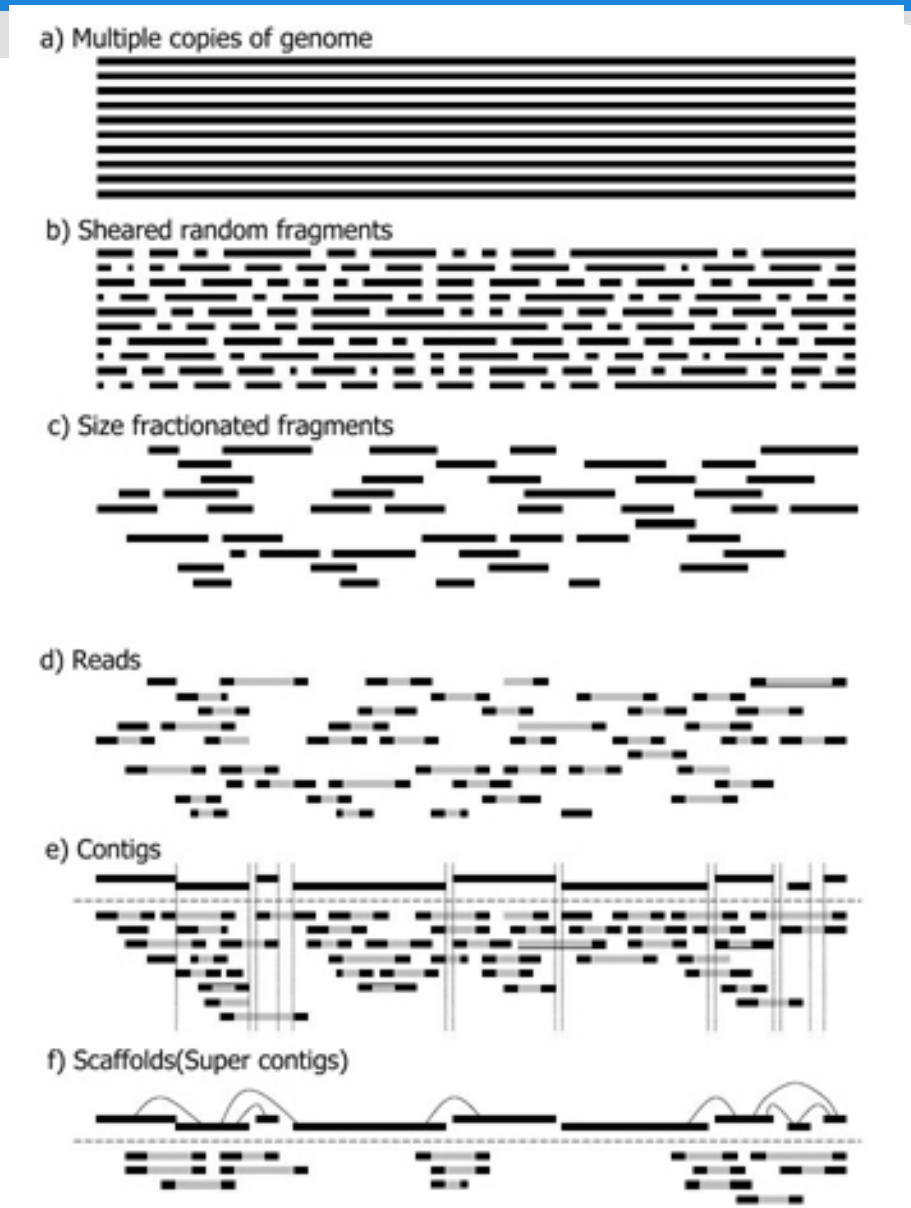
Assembly Seminar: iPlant Workshop (2013) at CSHL
http://www.youtube.com/watch?v=USlTWhmw0oQ

OKLAHOMA STATE UNIVERSITY

J. Craig Venter INSTITUTE

1. Preparing Reads

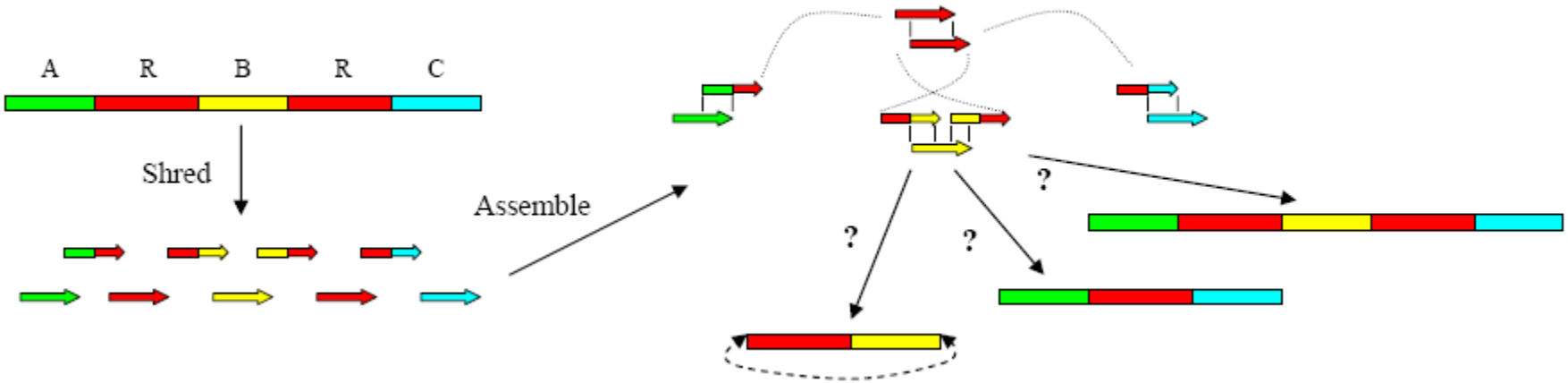
Reads sampled from genomes



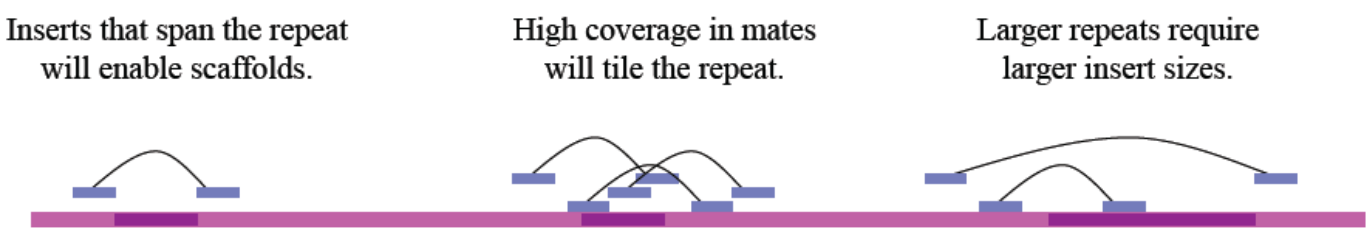
J. Craig Venter INSTITUTE

Repeats are major problems for assembly

- **Short reads harder to assemble**



- **Paired reads are needed to span the repeats**



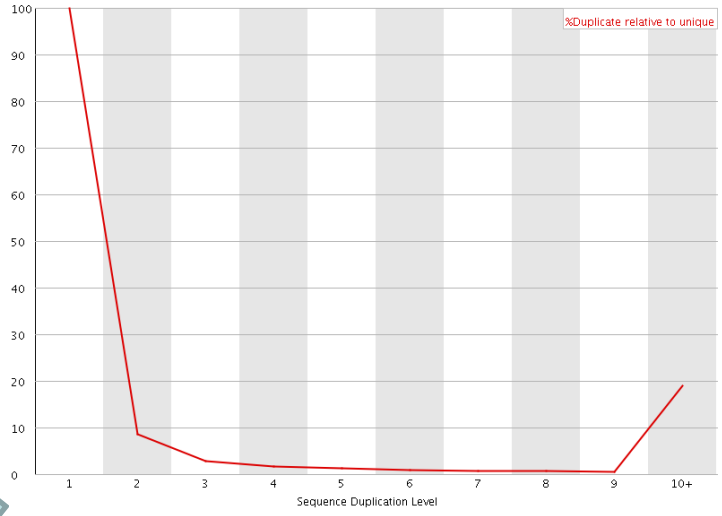
Credits: Jason Miller, JCVI

Run FASTQC first!

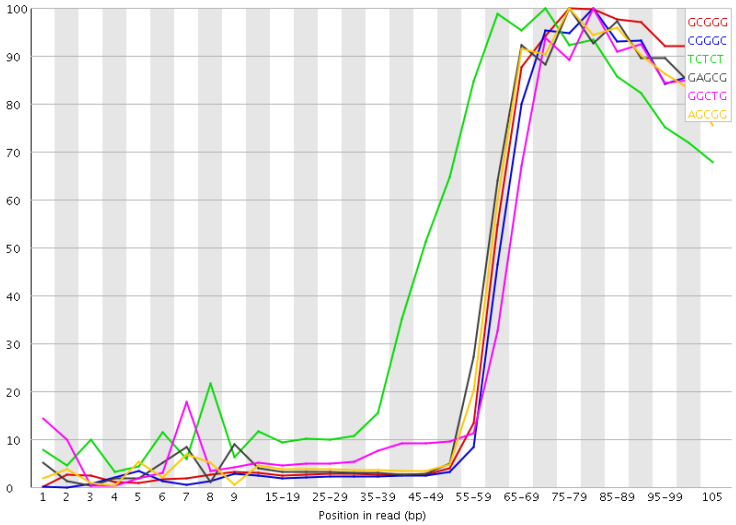
- Quality trimming
  Based on quality scores

- Basic Statistics
- Per base sequence quality
- Per sequence quality scores
- Per base sequence content
- Per base GC content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Kmer Content

Sequence Duplication Level



Relative Enrichment Over Read Length



Read trimming using FASTQ files

```
@SOLEXA2:1:1:2:1561#0/1
TTGACGGTTAATGCTGGTAATNGTGGTTCTTTTCATTTCATTCNTATAGATACATCTTTT
+SOLEXA2:1:1:2:1561#0/1
a``aaU]]aaaa]aa^\]bb[BBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB
@SOLEXA2:1:1:2:1381#0/1
AAGGCGGTTCCTGAATGAATGNGAAGCCTTCAAGAAGGTGATANGCAGGAGAAACATACG
+SOLEXA2:1:1:2:1381#0/1
a_SW`RVS[^^YLV]]QS^\ODU^]]]]X\_ZZRTBBBBBBBBBBBBBBBBBBBBBBBB
@SOLEXA2:1:1:2:391#0/1
CTGTTGATGCTAAAGGTGAGCNGCTTAAAGCTACCAGTTATATNGCTGTTTGTTTCTATT
+SOLEXA2:1:1:2:391#0/1
aaRaaZa`SaaaabUS]UaU^DOaabab`Raaaa`aY_aa_`YBBBBBBBBBBBBBBBB
```

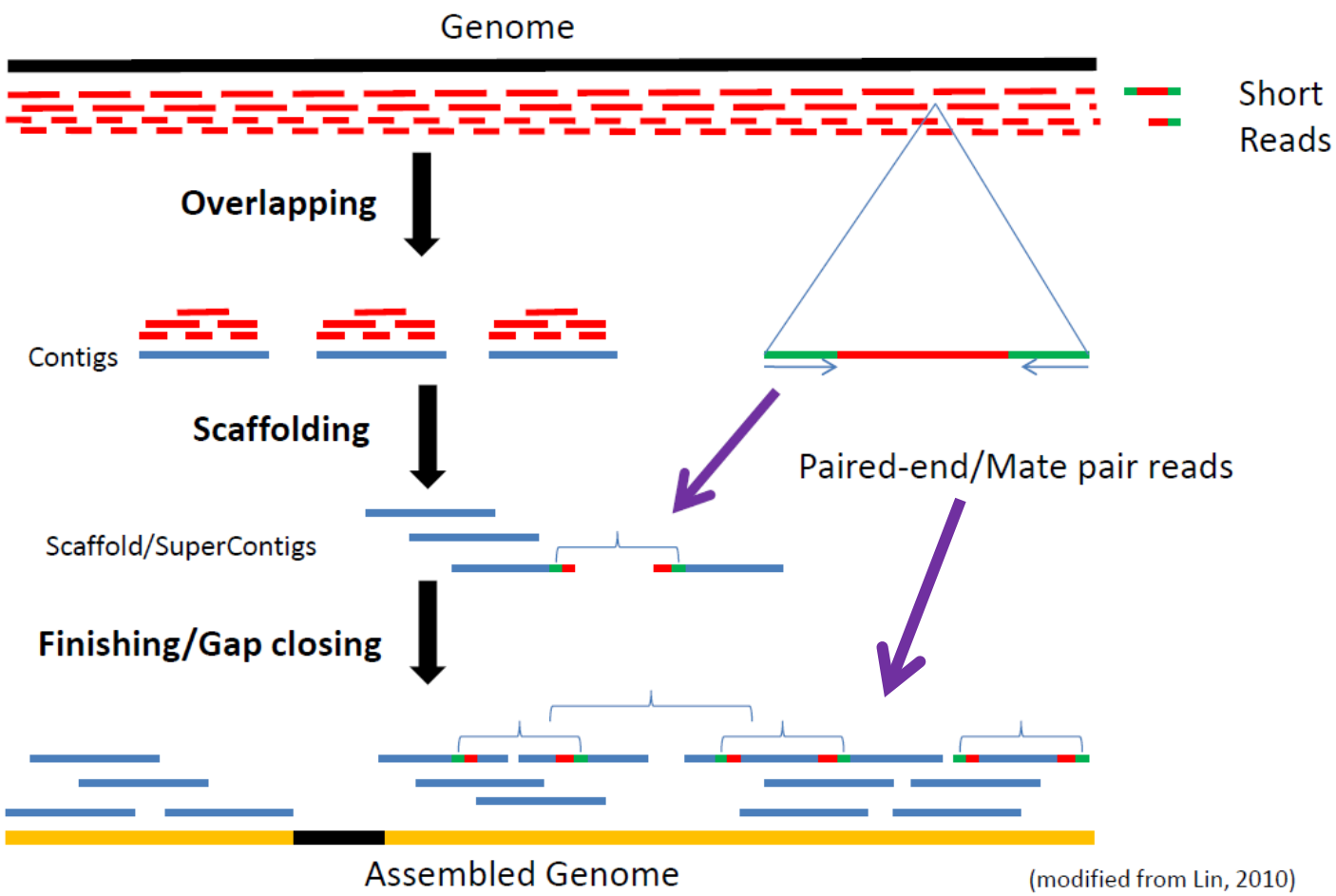
In "Illumina fastq" ...
B = "bad"
(not Phred score of 2)

```
@SOLEXA2:1:1:2:1561#0/1
TTGACGGTTAATGCTGGTAAT
+SOLEXA2:1:1:2:1561#0/1
a``aaU]]aaaa]aa^\]bb[
@SOLEXA2:1:1:2:1381#0/1
AAGGCGGTTCCTGAATGAATGNGAAGCCTTCAAGA
+SOLEXA2:1:1:2:1381#0/1
a_SW`RVS[^^YLV]]QS^\ODU^]]]]X\_ZZRT
@SOLEXA2:1:1:2:391#0/1
CTGTTGATGCTAAAGGTGAGCNGCTTAAAGCTACCAGTTATAT
+SOLEXA2:1:1:2:391#0/1
aaRaaZa`SaaaabUS]UaU^DOaabab`Raaaa`aY_aa_`Y
```

J. Craig Venter INSTITUTE

2. Assembly

Whole genome shotgun sequencing



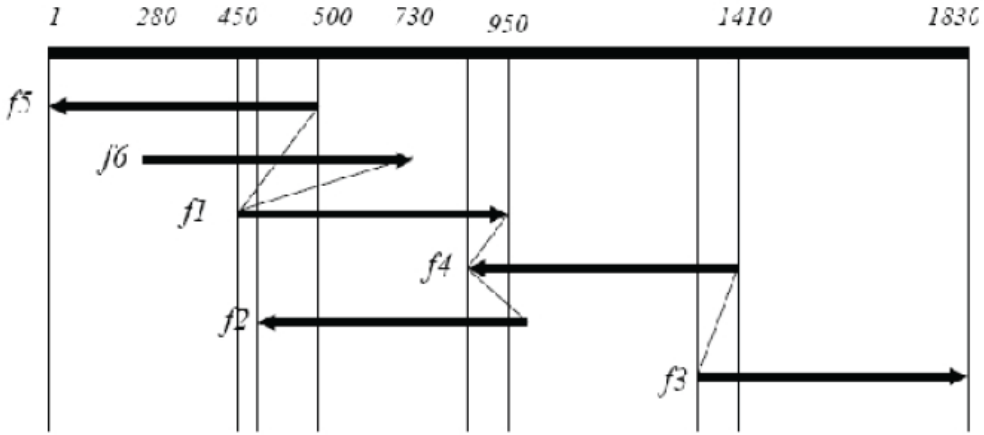
(modified from Lin, 2010)

J. Craig Venter INSTITUTE

Overlap – Layout - Consensus

Overlap:

Layout:



Consensus:

```
R1          ACGCTCCAACCGCTAATACG
R2                    ATCGCTAATCCACGCCCGCCCCGC
R2        AAAC–CTCCAACCG
R3                              TGCGCGCCCGCCCCGAAACCGC
Consensus AAAC–CTCCAACCGCTAATGCGCGCCCGCCCCGAAACCGC
```

de Bruijn graph assemblers: break the reads into K-mers

Read 1: AGTCGAG

AGTC ⇒ GTCG ⇒ TCGA ⇒ CGAG

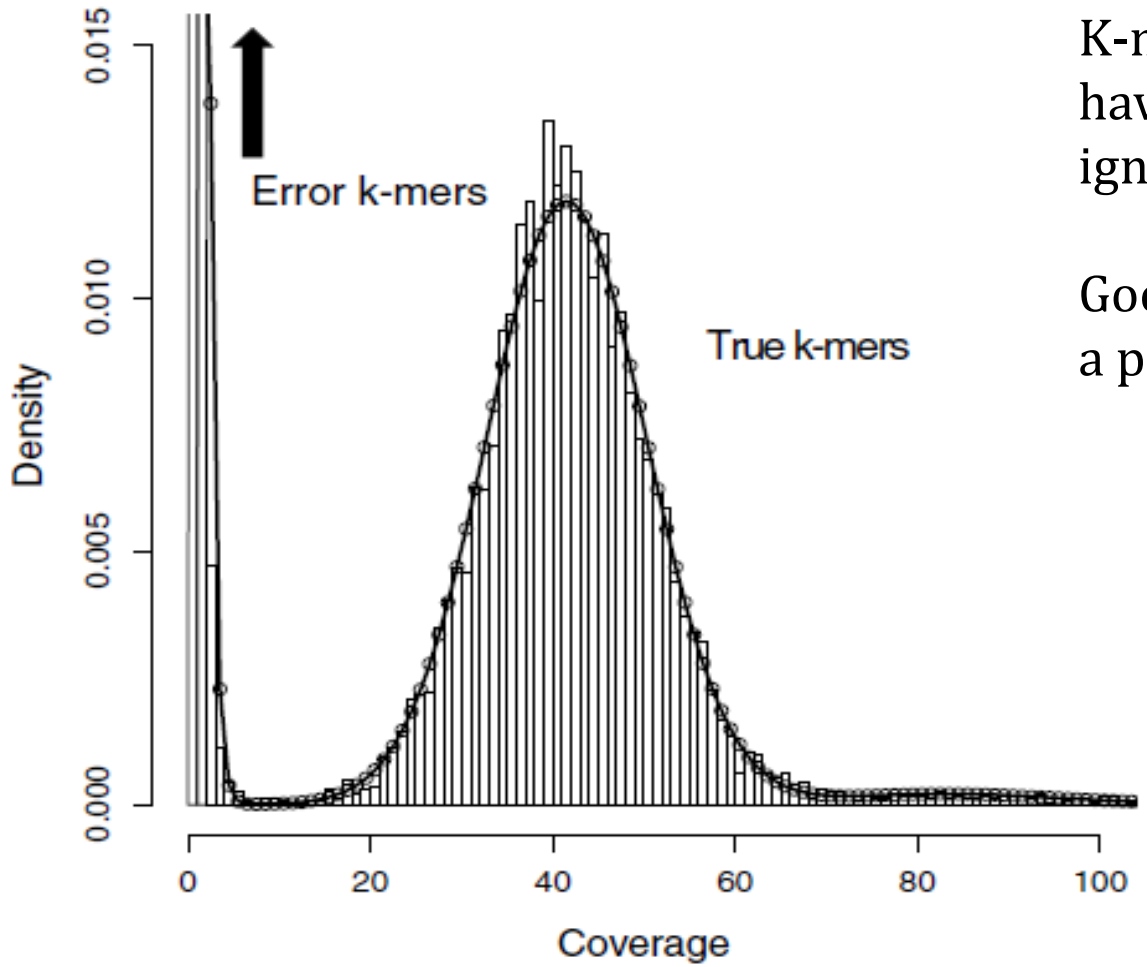
Read 2: TCGAGGC

TCGA ⇒ CGAG ⇒ GAGG ⇒ AGGC

AGTC ⇒ GTCG ⇒ TCGA (2x) ⇒ CGAG (2x) ⇒ GAGG ⇒ AGGC

Contig: AGTCGAGGC

K-mer histogram



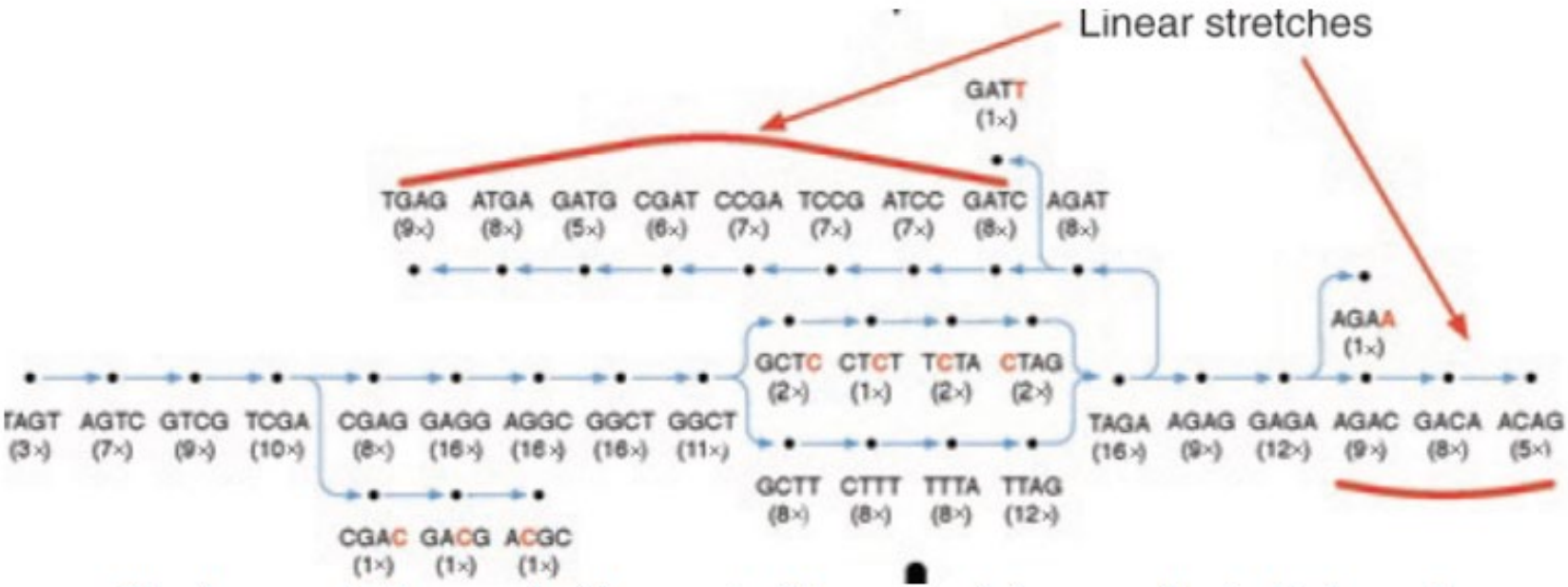
K-mers at low coverage have many errors and are ignored

Good K-mers should show a peak

Kelley et al., 2010

J. Craig Venter INSTITUTE

de-Bruijn graph assembly



The *k*-mers in the reads (4-mers in this example) are collected into nodes and the coverage at each node is recorded (numbers at nodes)

Features

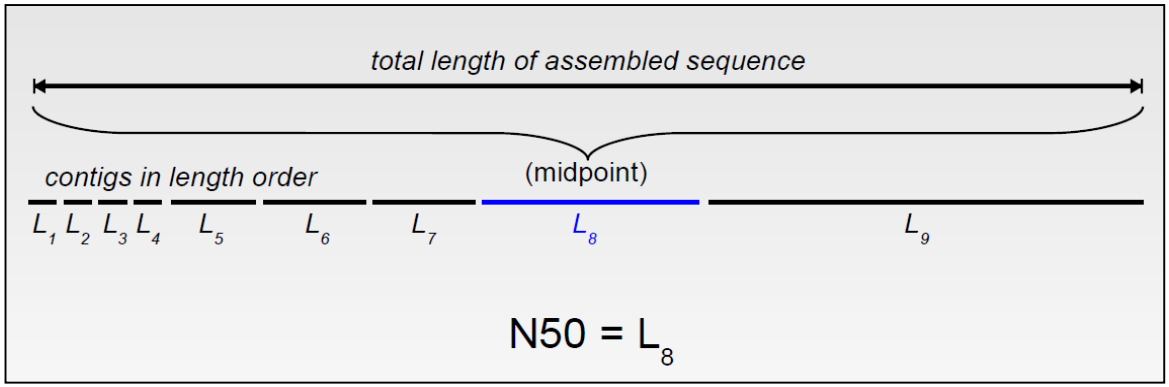
- continuous linear stretches within the graph
- Sequencing errors are low frequency tips in the graph

J. Craig Venter INSTITUTE

3. Assembly QC

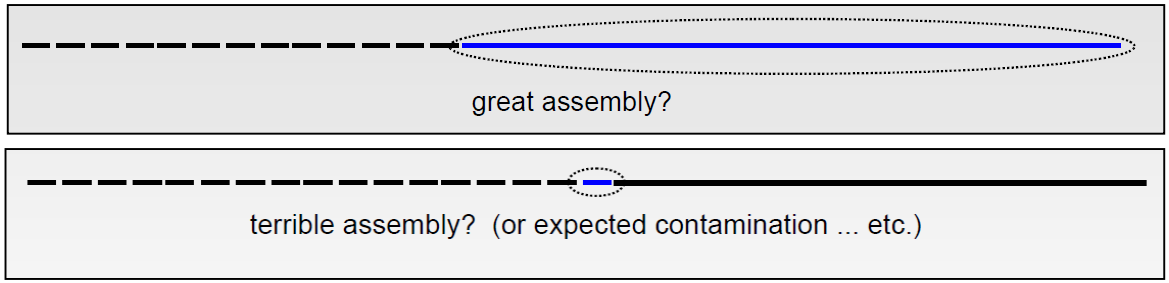
Assembly QC – assess continuity

- **N50** captures how much of the assembly is covered by relatively large contigs
- "When ordered, half of all the sequences in contigs larger than {N50} bp …"
- Others: Average length, min and max length, combined total length (N%)



Bigger is usually better!

- Watch out for:



J. Craig Venter INSTITUTE

Which assembly is better?

	Assembly 1	Assembly 2		Assembly 1	Assembly 2
N50	51kb	42Kb		50Kb	20Kb
Total length	2.7Gb	2.69Gb		1.2Gb	2.7Gb
Avg. length	45Kb	39kb		40Kb	18Kb
Mapping rate	0.82	0.78		0.6	0.85
SNP rate	0.02	0.02		0.02	0.02
Indel rate	0.01	0.01		0.01	0.012
Pairing rate	0.8	0.9		0.9	0.88
Misassemblies	15	5		2	2

Fewest misassemblies with Highest N50

OKLAHOMA STATE UNIVERSITY

J. Craig Venter INSTITUTE

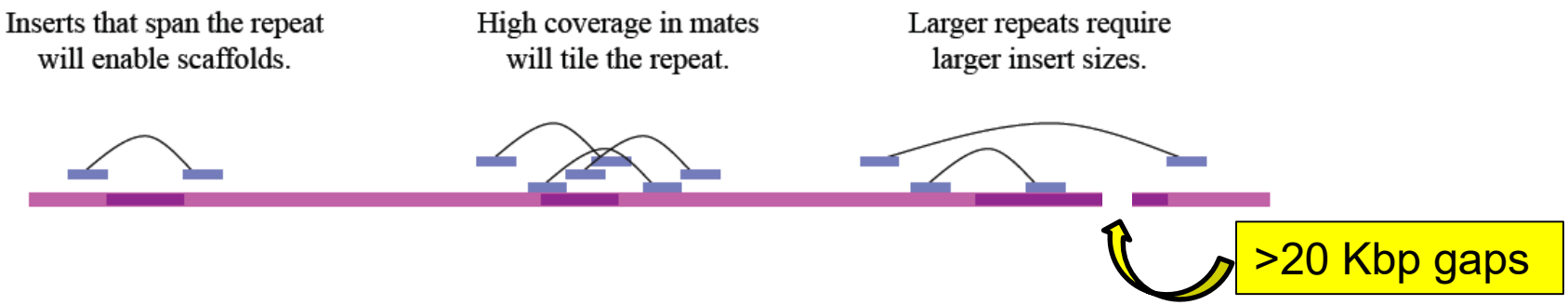
4. Assembly Curation

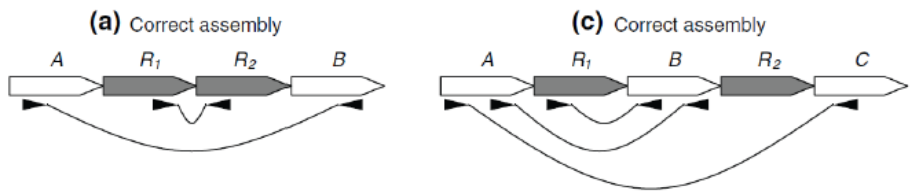
Assembly curation

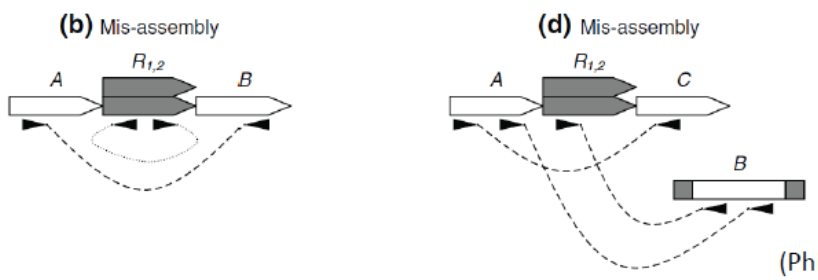
What to do after assembly?

- **Two goals:**
  - Fix chimeric (mis-joined) scaffolds
  - Build larger scaffolds towards chromosomes
- **Methods:**
  - Large insert mate pair library **
  - Optical (restriction) map
  - Genetic mapping (or use fosmids, BACs)
  - Synteny**

Large insert mate-pair libraries: Spanning repeats and closing gaps







(Phillippy, 2008)

Long distance mate pairs allow substantially longer scaffolds to form

You can then use the long mates directly in assembler (e.g. ALLPATHS) or use a standalone scaffolder (e.g. BAMBUS)

J. Craig Venter INSTITUTE

Talk summary

- Good genome assembly is dependent on good preparation of data
- Don't rely on the results of your assembler, check adequately and double-check using any references you can find
- External scaffolding using maps (genetic map, physical map, optical map) allow repair of chimeric scaffolds, and anchor onto chromosomes

THANK-YOU!

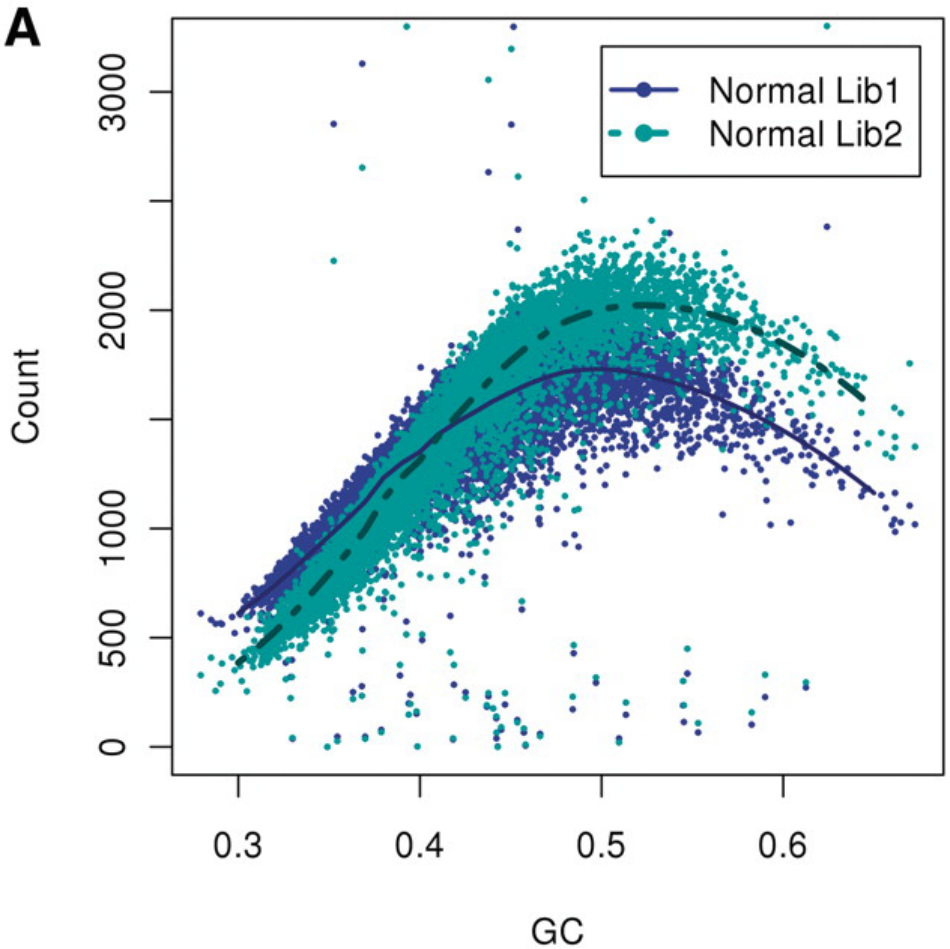
OKLAHOMA STATE UNIVERSITY

J. Craig Venter INSTITUTE

J. Craig Venter™
INSTITUTE

- The following slides were edited out:

Not really random



As GC contents vary along the genome, the depth will be uneven too.

High AT regions, like promoter regions (think TATA box) will often have very low depth and sometimes not assembled

With short reads technology, you typically need at least **20x-40x** coverage

Benjamini and Speed, 2012

TRIMMOMATIC example

```
$ java -cp trimmomatic-0.15.jar org.usadellab.trimmomatic.TrimmomaticPE
s_1_1_sequence.txt.gz s_1_2_sequence.txt.gz lane1_forward_paired.fq.gz
lane1_forward_unpaired.fq.gz lane1_reverse_paired.fq.gz
lane1_reverse_unpaired.fq.gz ILLUMINACLIP:adapters.fasta:2:40:15
LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36
```

- This will perform the following:
- Remove adapters
- Remove leading low quality or N bases (below quality 3)
- Remove trailing low quality or N bases (below quality 3)
- Scan the read with a 4-base wide sliding window, cutting when the average quality per base drops below 15
- Drop reads below the 36 bases long
- Read and write files in gzipped format

Errors / polymorphisms / repeats

"Bubble"

SNP or Sequencing Error

Repeat Sequence

"Frayed rope"

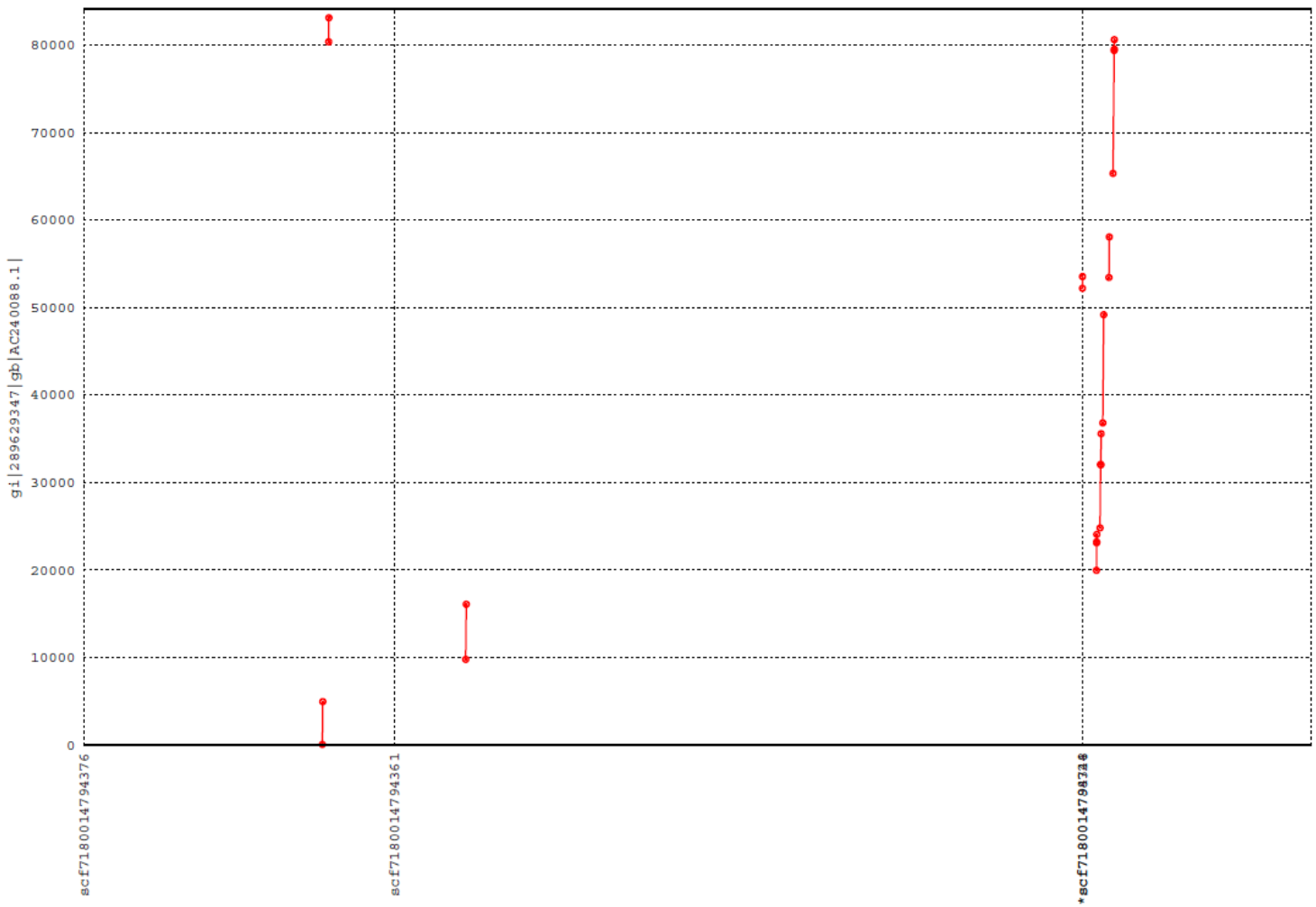
Early *Brassica* assembly

- True story – **Good N50 ≠ Good assembly!**
- One of the earliest large genomes to exploit CABOG to assemble a mixture of 454 and Illumina reads
- CABOG assembly stats:
  ctg: 353 Mb, N50=6.3 Kb
  scf: 488 Mb (27.5% Ns), N50=3.8 Mb
- Awesome scaffold N50 !!!!
- The CABOG assembly had serious flaws due to bad data

OKLAHOMA STATE UNIVERSITY

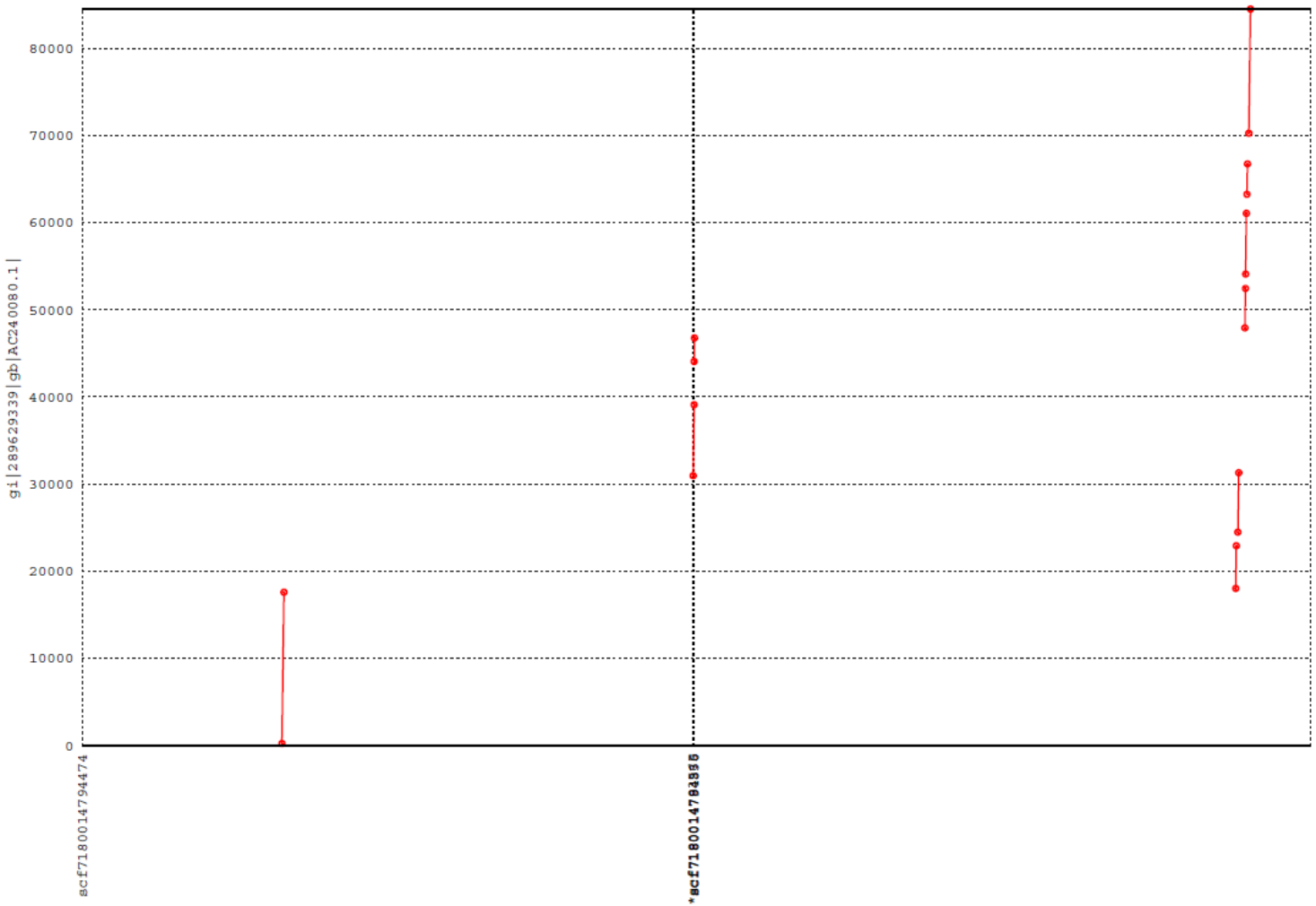
J. Craig Venter INSTITUTE

Brassica v1 assembly against BAC



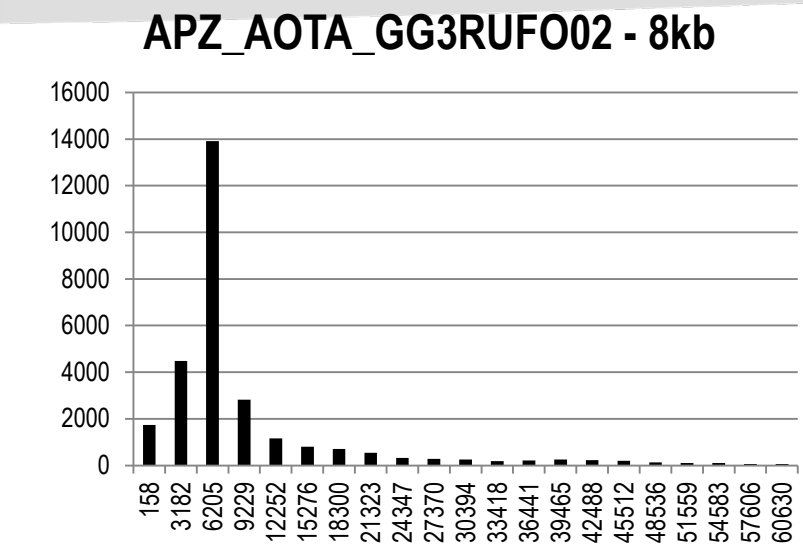
J. Craig Venter INSTITUTE

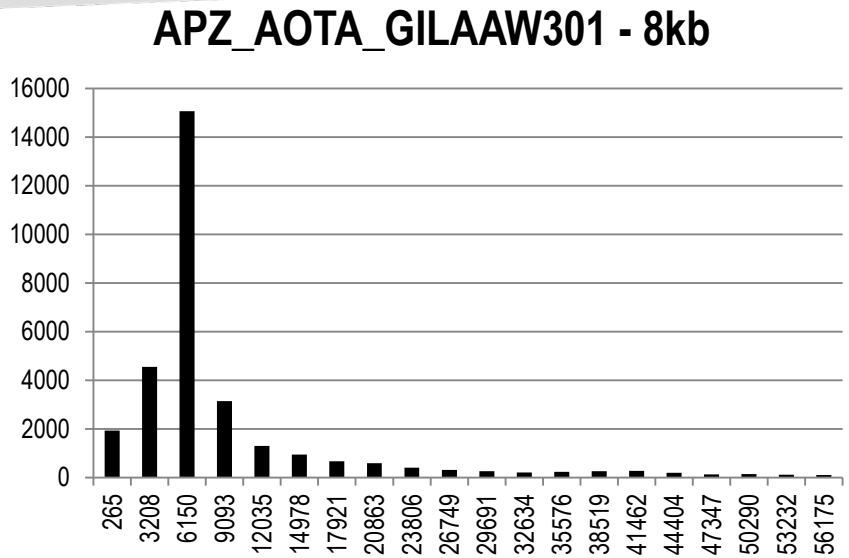
Brassica v1 assembly against BAC

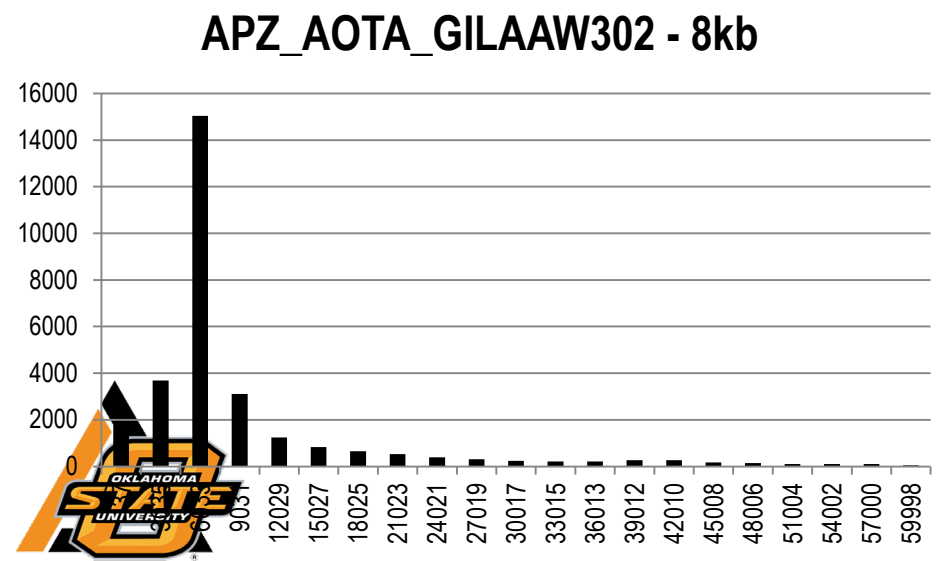


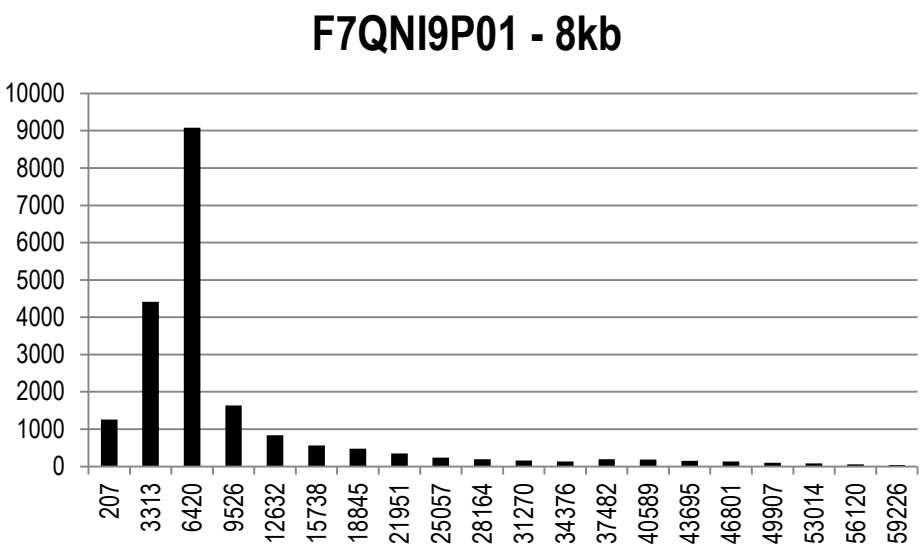
J. Craig Venter INSTITUTE

454 mate libraries – 8kb

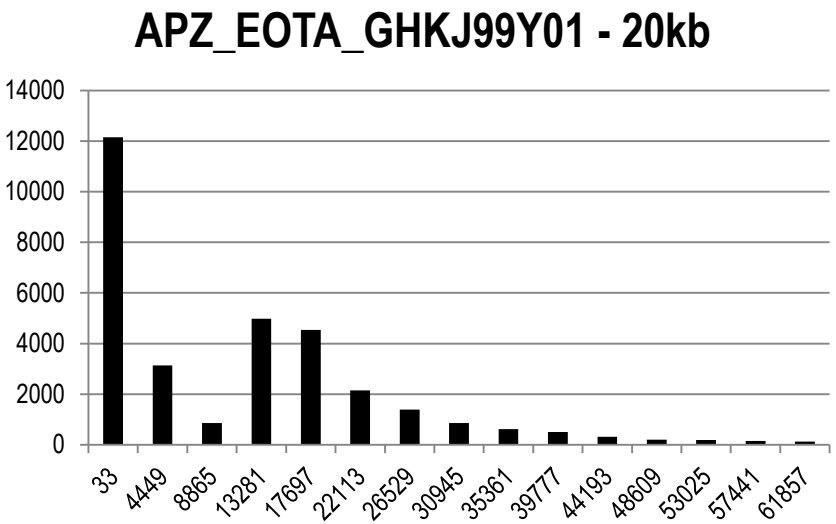


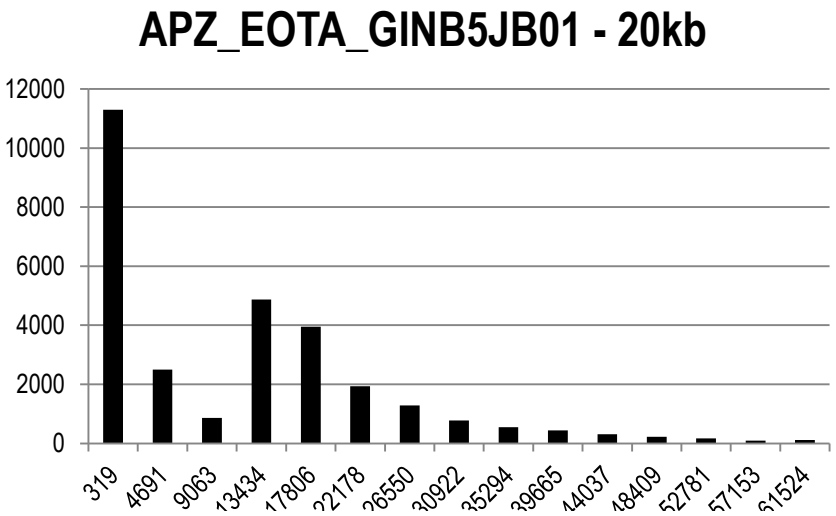


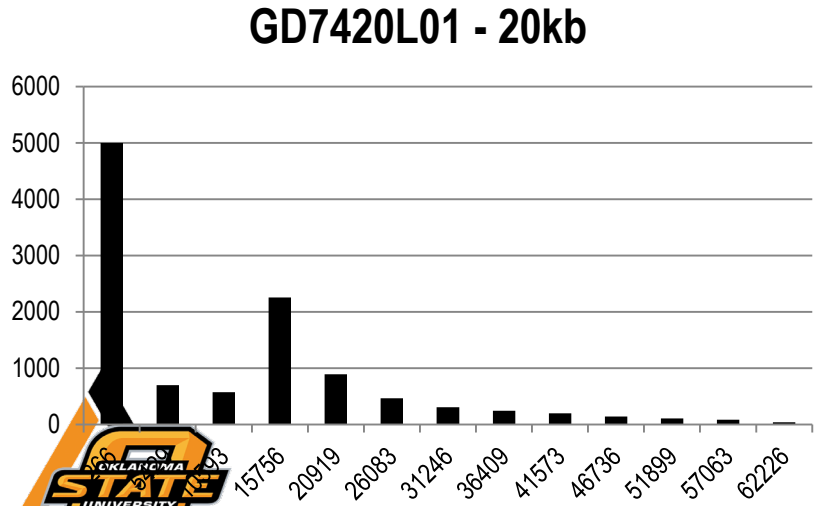


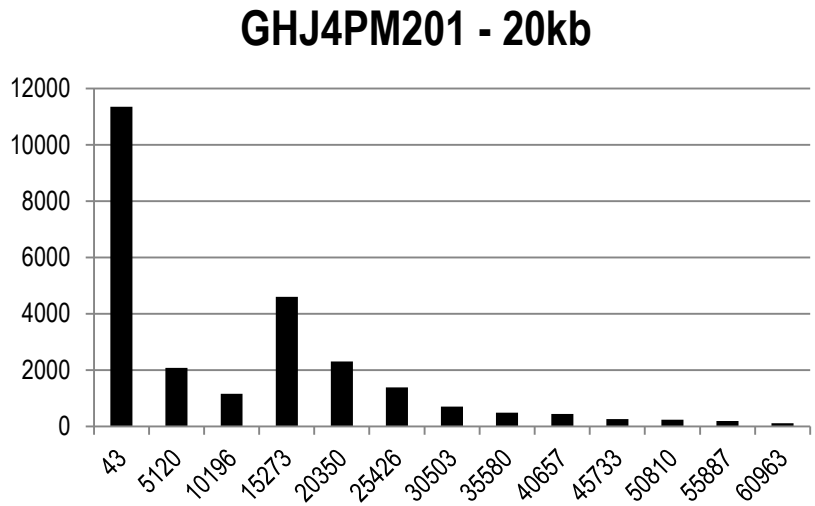


454 20kb libraries

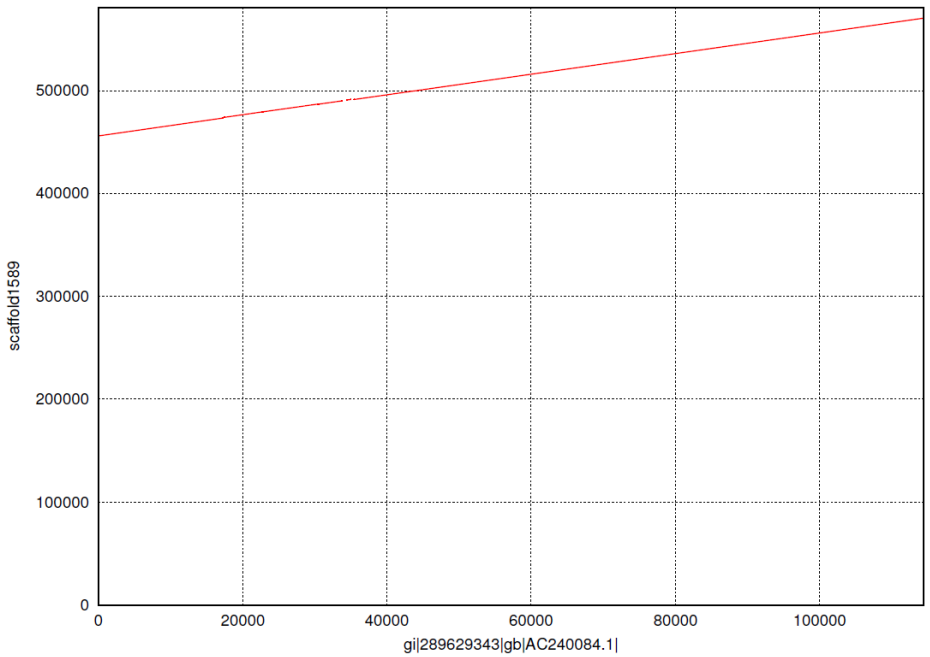


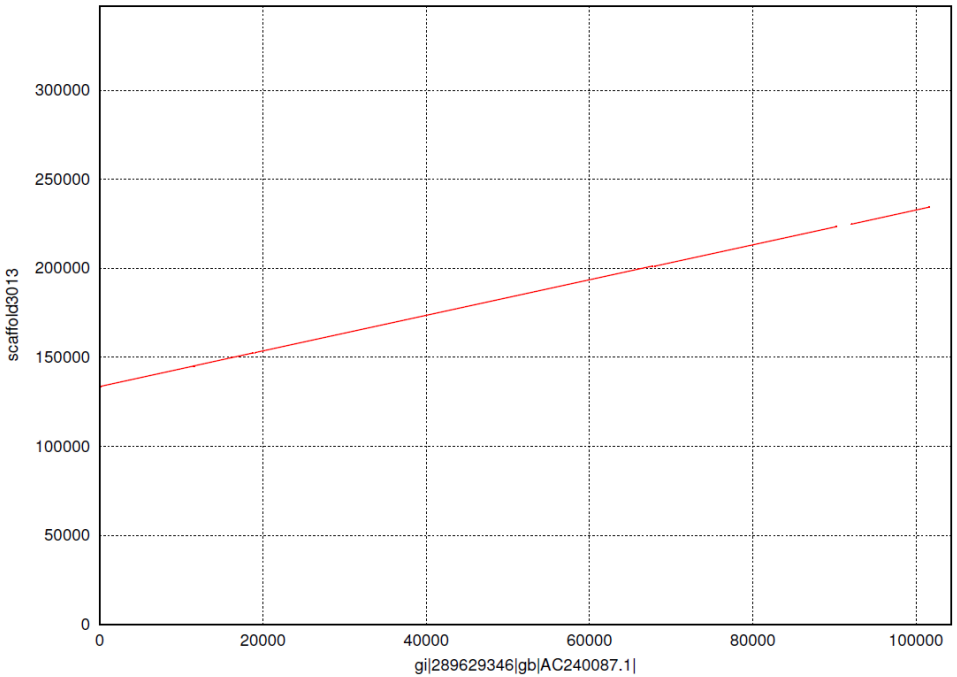






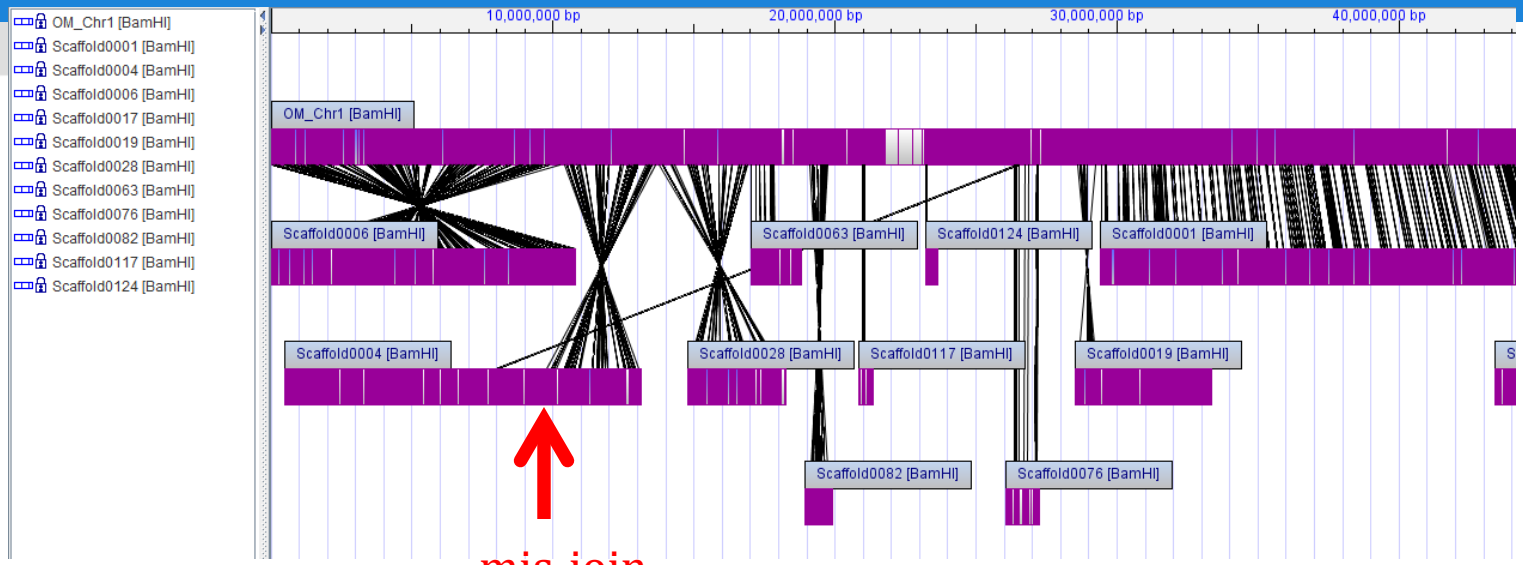
New improved assembly against BACs



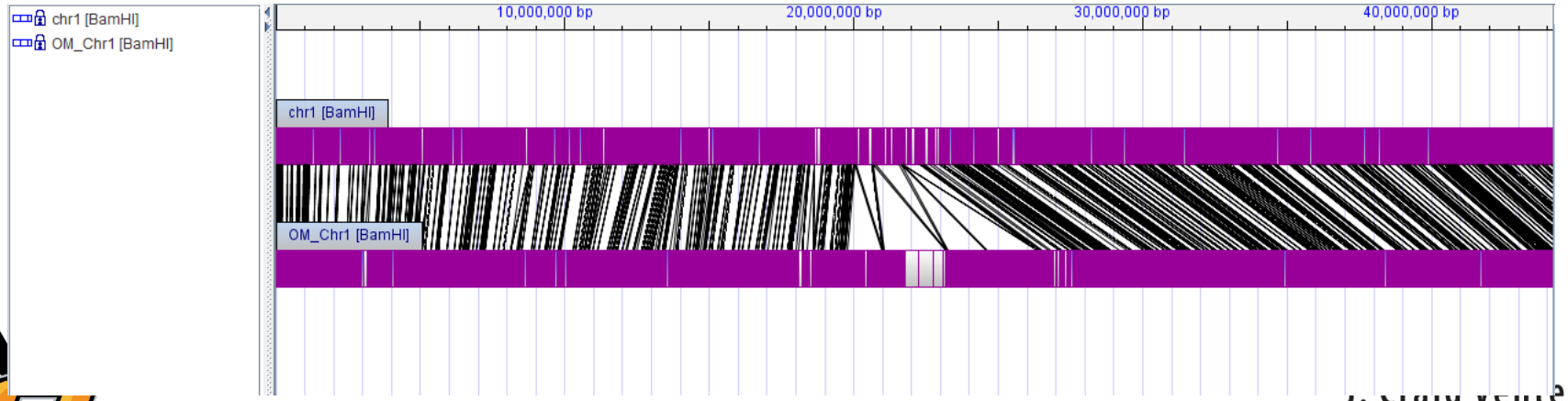


J. Craig Venter INSTITUTE

Optical map



mis-join



OKLAHOMA STATE UNIVERSITY

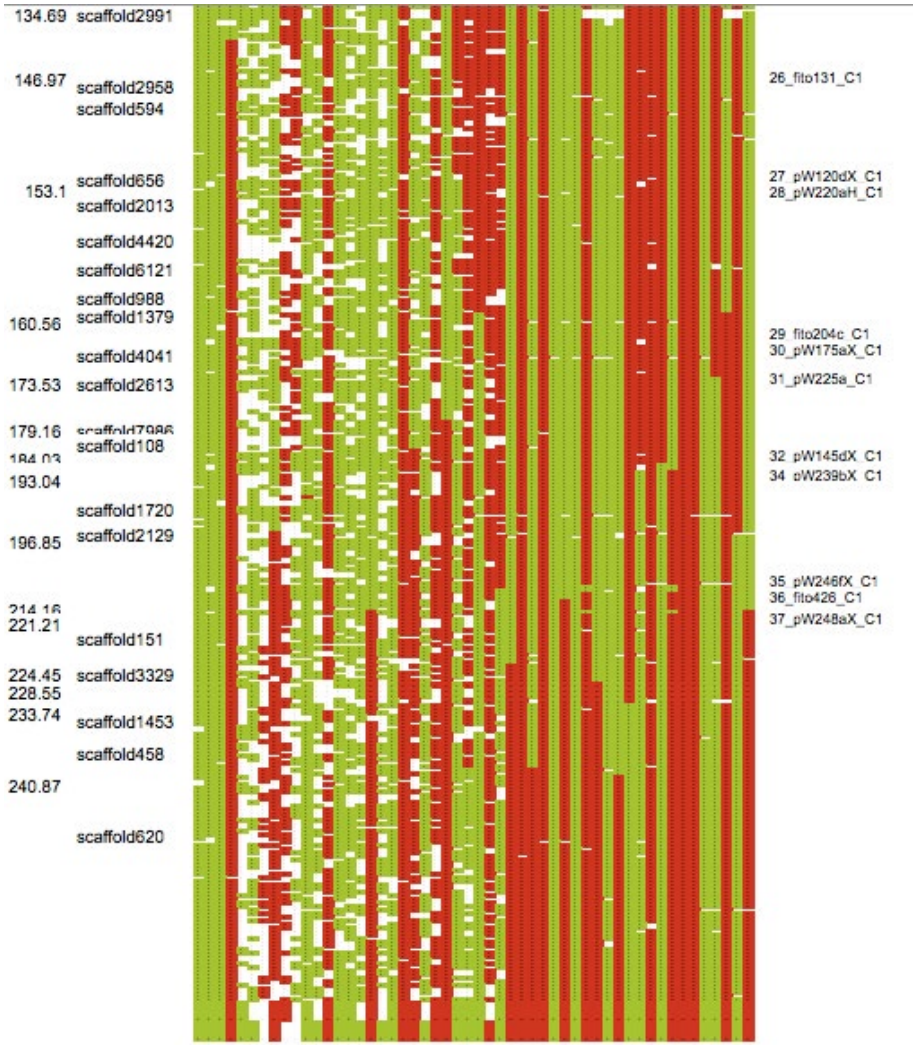
J. Craig Venter INSTITUTE

Mapping the RAD tags to scaffolds

GenomeView (http://genomeview.org)
Credits: Andy Sharpe, CANSEQ

J. Craig Venter INSTITUTE

RAD segregation data for linkage group 1

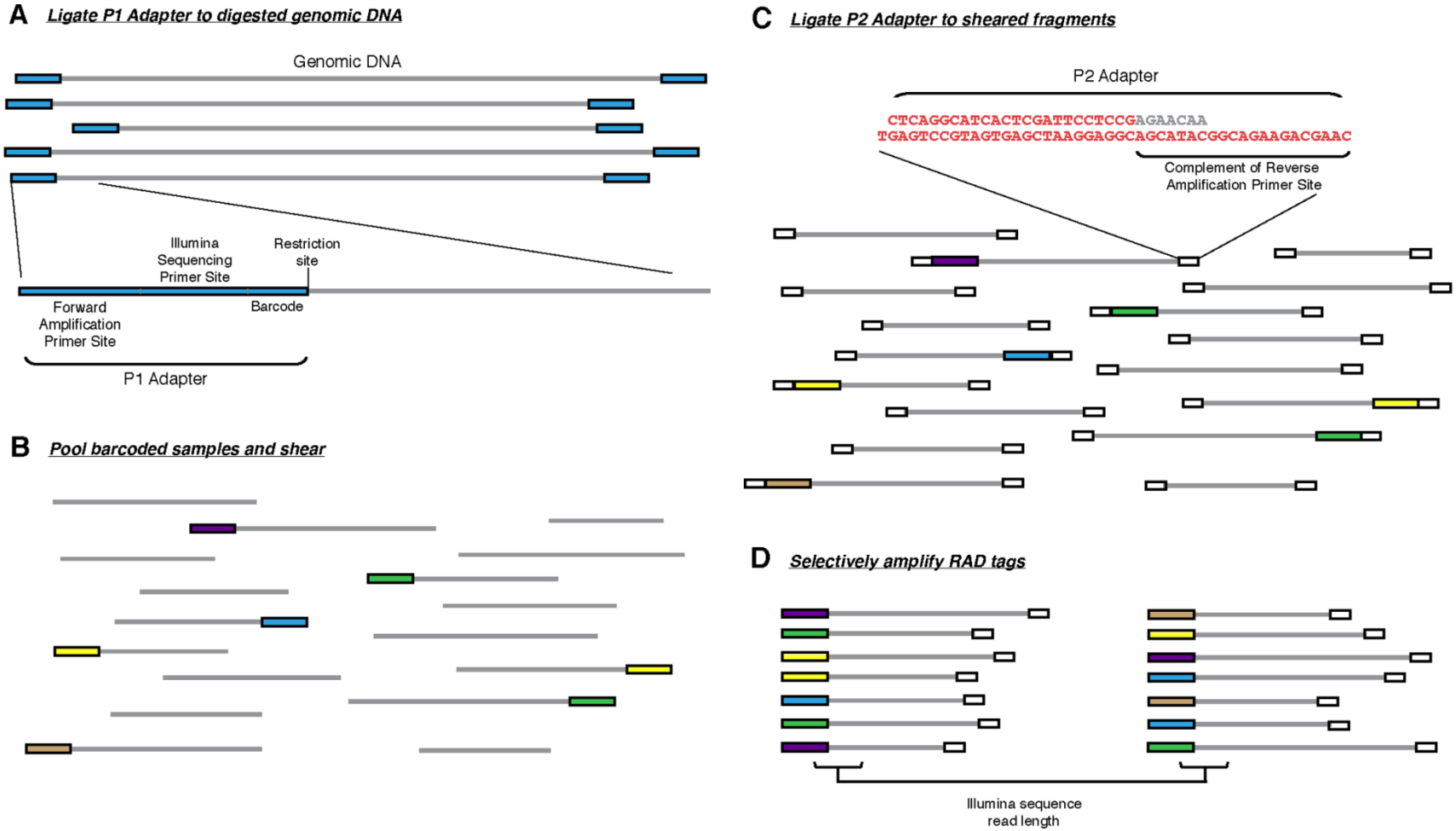


The genetic map is an ordered list of markers

We can anchor the scaffolds onto the genetic map to build chromosomes

J. Craig Venter INSTITUTE

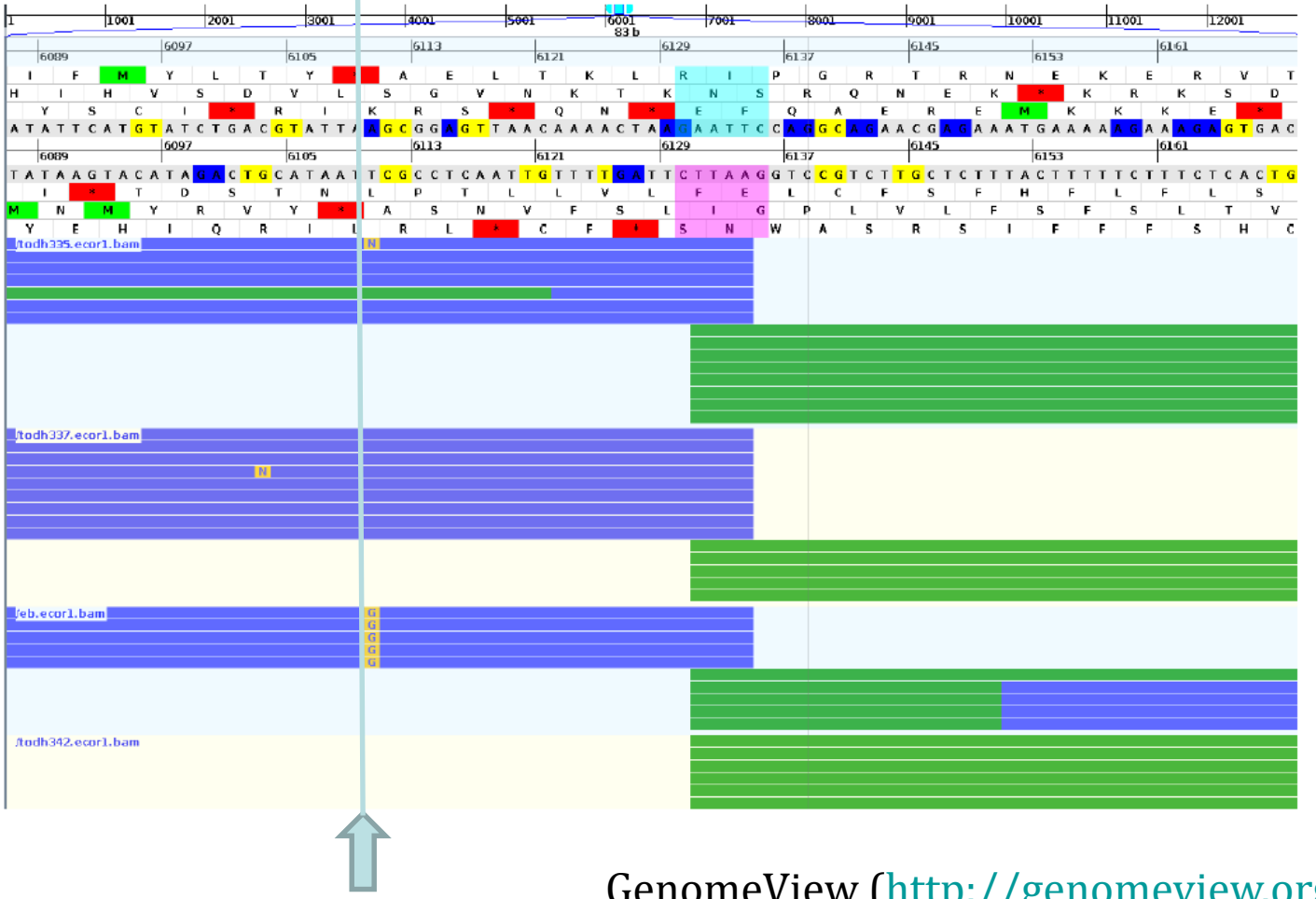
Restriction site associated DNA (RAD-tag)



Baird et al., 2011

J. Craig Venter INSTITUTE

Map Rad-tags to scaffolds & SNP calling



SNP site:
Scaffold126_6110

Individual 1: a

Individual 2: a

Individual 3: b

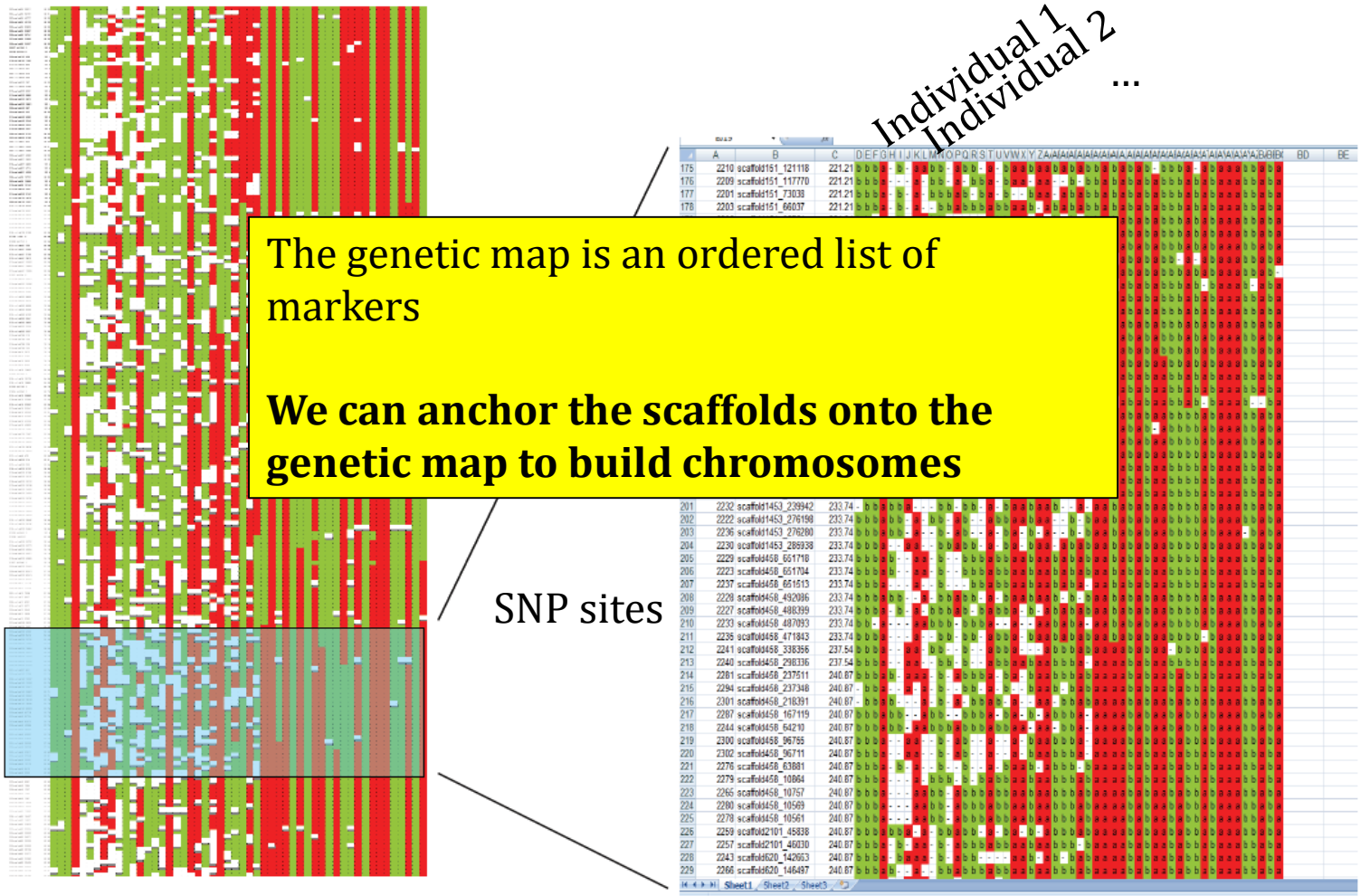
Individual 4: -

GenomeView (http://genomeview.org)
Credits: Andy Sharpe, CANSEQ

J. Craig Venter INSTITUTE

RAD segregation data

a=ref allele, b=alt allele, - = missing



The genetic map is an ordered list of markers

We can anchor the scaffolds onto the genetic map to build chromosomes

MSTmap
(http://alumni.cs.ucr.edu/~yonghui/mstmap.html)

GBS map

locus_name	DZA315.16	Jemalong.J6	LR4-1	LR4-4	LR4-7	LR4-9	LR4-17	LR4-20	LR4-24	LR4-30	LR4-32	LR4-6	LR4-8	LR4-10	LR4-18	LR4-23	LR4-14	LR4-27	LR4-43	LR4-42	LR4-46	LR4-49	LR4-51	LR4-58	LR4-60	LR4-62	LR4-65	LR4-67	LR4-72	LR4-44	LR4-48	LR4-50	LR4-55	LR4-59	LR4-61	LR4-63	LR4-66	LR4-162	LR4-166	LR4-73	LR4-75	LR4-76	LR4-78	LR4-80	LR4-83	LR4-95	LR4-186	LR4-203	LR4-77	LR4-79	LR4-81	LR4-84	LR4-91	LR4-93	LR4-96	LR4-207	LR4-209	LR4-119	LR4-141	LR4-143
Scaffold0004.6045383	B	A	A	B	B	B	B	A	B	A	-	-	A	B	B	-	A	B	A	A	B	A	B	A	A	B	B	A	A	A	-	A	B	B	A	A	B	A	B	B	B	A	B	A	-	B	A	A	B	A	A	A	B	A	A	B	A	A	B	B
Scaffold0004.6045642	B	A	A	B	B	B	B	A	B	A	-	-	A	B	B	-	-	B	A	A	B	-	B	-	-	B	B	A	A	A	A	-	B	B	A	A	B	A	B	B	-	A	B	A	B	B	A	A	B	A	A	A	B	A	A	B	A	A	B	B
Scaffold0004.6057970	B	A	-	B	B	B	B	A	B	A	B	-	A	B	B	B	-	-	A	A	B	-	B	-	A	B	B	A	A	A	-	A	B	-	A	A	B	A	B	B	B	A	B	A	-	B	A	A	B	A	A	A	B	A	A	B	A	A	-	B
Scaffold0004.6089918	B	A	A	-	B	B	B	-	B	A	B	-	A	B	-	B	-	-	A	A	B	-	B	A	-	B	B	A	A	A	A	-	B	B	A	A	B	A	B	B	B	A	B	A	B	B	-	A	B	A	A	A	B	A	A	B	A	A	B	B
Scaffold0004.6214421	B	A	A	B	B	B	B	A	B	-	-	-	A	B	B	B	A	B	A	A	-	-	B	A	A	B	B	A	A	A	-	A	-	B	A	A	B	A	-	B	B	A	B	A	B	B	A	A	B	A	A	A	B	A	A	B	A	A	B	B
Scaffold0004.6238635	B	A	A	B	B	B	B	A	B	A	-	-	A	B	-	-	-	B	A	A	B	-	B	-	-	-	B	A	A	A	-	A	B	B	A	A	B	A	-	B	-	A	B	A	-	B	A	A	B	A	A	A	B	A	A	B	A	A	B	B
Scaffold0004.6344717	B	A	A	B	B	B	B	-	B	-	-	B	A	B	B	-	-	B	-	A	B	A	B	A	A	-	B	A	A	A	-	A	B	B	A	A	B	-	B	B	B	-	B	A	-	B	A	A	B	A	A	A	B	A	A	B	A	A	B	B
Scaffold0004.6351076	B	A	A	-	B	-	B	-	B	A	B	-	-	B	B	-	-	B	A	A	B	-	B	-	-	B	B	A	A	A	-	-	B	-	A	A	B	A	B	B	B	A	B	A	B	B	A	A	B	A	A	A	B	A	A	-	A	A	B	B
Scaffold0004.6393511	B	A	A	B	B	-	B	A	B	A	-	-	A	-	B	B	B	B	A	A	B	A	B	A	-	B	B	A	A	A	-	A	B	B	A	A	B	A	B	B	B	A	B	A	-	B	A	A	B	A	A	A	B	A	A	B	A	A	B	B
Scaffold0004.6578739	B	A	A	B	B	B	B	-	B	A	-	-	A	B	B	-	B	B	A	A	B	-	-	-	-	B	B	A	A	A	-	-	B	-	A	A	B	A	B	B	B	A	B	A	-	B	A	-	B	A	A	A	B	A	A	B	A	A	B	B
Scaffold0004.6578745	B	A	A	B	B	B	B	-	B	A	-	-	A	B	B	-	B	B	A	A	B	-	-	-	-	B	B	A	A	A	-	-	B	-	A	A	B	A	B	B	B	A	B	A	-	B	A	-	B	A	A	A	B	A	A	B	A	A	B	B
Scaffold0004.6706408	B	A	B	A	-	B	-	-	B	B	-	-	B	-	A	B	A	A	B	-	B	-	B	A	B	A	-	B	B	B	A	A	B	B	B	A	B	A	B	B	B	B	A	A	B	-	B	A	B	A	B	B	B	B	B	B	A	A	A	B
Scaffold0004.6706459	B	A	B	A	-	B	-	-	B	B	-	-	B	-	A	B	A	A	B	-	B	-	B	A	B	A	-	B	B	B	A	A	B	B	B	A	B	A	B	B	B	B	A	A	B	-	B	A	B	A	B	B	B	B	B	B	A	A	A	B
Scaffold0004.6706468	B	A	B	A	-	B	-	-	B	B	-	-	B	-	A	B	A	A	B	-	B	-	B	A	-	A	-	B	B	B	A	A	B	B	B	A	B	A	-	B	-	B	A	A	B	-	B	A	B	A	B	B	B	B	B	B	A	A	A	-
Scaffold0004.6896740	B	A	-	A	-	B	A	A	B	B	-	-	B	A	A	B	-	A	B	-	B	A	B	A	B	A	-	-	B	B	-	A	B	B	B	A	B	A	B	B	B	B	A	A	B	-	B	A	B	A	B	B	B	B	B	B	A	A	A	B
Scaffold0004.7095336	B	A	-	-	B	-	-	A	B	B	-	-	B	A	A	B	A	A	B	B	B	A	B	A	-	A	-	-	B	B	A	-	B	B	B	A	B	A	B	B	B	B	A	A	B	-	B	A	B	A	B	B	B	B	B	B	A	A	A	B
Scaffold0004.7095378	B	A	-	-	B	-	-	A	B	B	-	-	B	A	A	B	A	A	B	B	B	A	B	A	-	A	-	-	B	B	A	-	B	B	B	A	B	A	B	B	B	B	A	A	B	-	B	A	B	A	B	B	B	B	B	B	A	A	A	B
Scaffold0004.7095393	B	A	-	-	B	-	-	A	B	B	-	-	B	A	A	B	A	A	B	B	B	A	B	A	-	A	-	-	B	B	A	-	B	B	B	A	B	A	B	B	B	B	A	A	B	-	B	A	B	A	B	B	B	B	B	B	A	A	A	B
Scaffold0004.7133581	B	A	B	A	-	B	A	A	B	B	-	-	B	A	A	B	A	A	B	B	B	-	B	A	B	A	-	-	B	B	A	A	B	B	B	A	B	A	B	B	B	B	A	A	B	-	B	A	-	A	B	B	B	B	B	B	A	A	A	B
Scaffold0004.7167006	B	A	-	A	-	B	A	A	B	B	-	-	B	A	-	B	-	-	B	B	B	A	B	A	B	A	-	B	B	B	A	A	B	B	B	A	B	A	B	B	B	B	A	A	B	A	-	A	B	A	B	B	B	B	B	B	A	A	A	B
Scaffold0004.7167020	B	A	-	A	-	B	A	A	B	B	-	-	B	A	-	B	-	-	B	B	B	A	B	A	B	A	-	B	B	B	A	A	B	B	B	A	B	A	B	B	B	B	A	A	B	A	-	A	B	A	B	B	B	B	B	B	A	A	A	B
Scaffold0004.7289758	B	A	-	-	-	B	A	A	B	B	-	A	-	A	A	-	-	A	B	B	B	A	B	A	-	A	B	-	-	B	-	A	-	-	B	A	B	A	B	B	-	B	A	A	B	A	B	-	B	A	B	B	B	B	B	B	A	A	A	B
Scaffold0004.7389389	B	A	-	A	A	B	A	B	B	B	-	-	B	-	A	B	A	A	B	-	B	A	B	A	-	A	-	B	B	B	A	A	-	A	B	A	B	A	B	B	B	B	A	A	B	A	B	A	B	A	B	B	B	B	B	B	A	A	A	-
Scaffold0004.7430007	B	A	-	A	-	B	-	B	B	B	B	-	B	-	A	-	A	A	B	-	B	A	B	-	B	A	-	-	-	B	-	A	B	A	B	A	B	A	-	B	B	B	A	-	B	-	B	A	B	A	B	B	B	B	B	B	A	A	A	B
Scaffold0004.7430133	B	A	-	A	B	B	-	B	B	B	-	-	-	A	A	B	-	A	B	B	B	A	B	-	-	A	B	-	B	B	A	A	B	-	B	A	B	A	B	B	B	B	A	A	B	-	B	A	B	A	B	B	B	B	B	B	A	A	A	B

mis-join