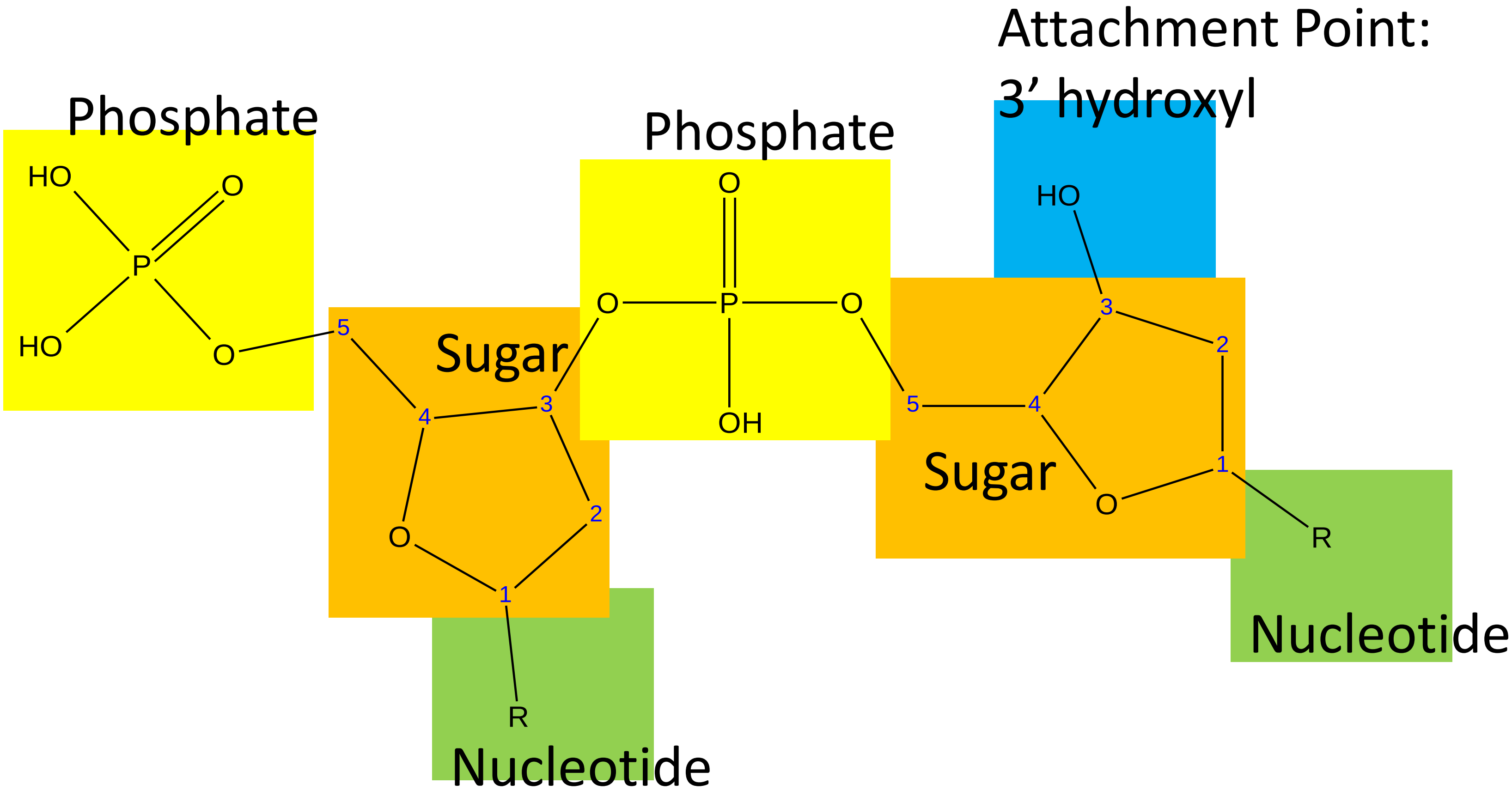
DNA Sequencing, Mapping and Assembly

DAVID L. TABB, PH.D.

Overview

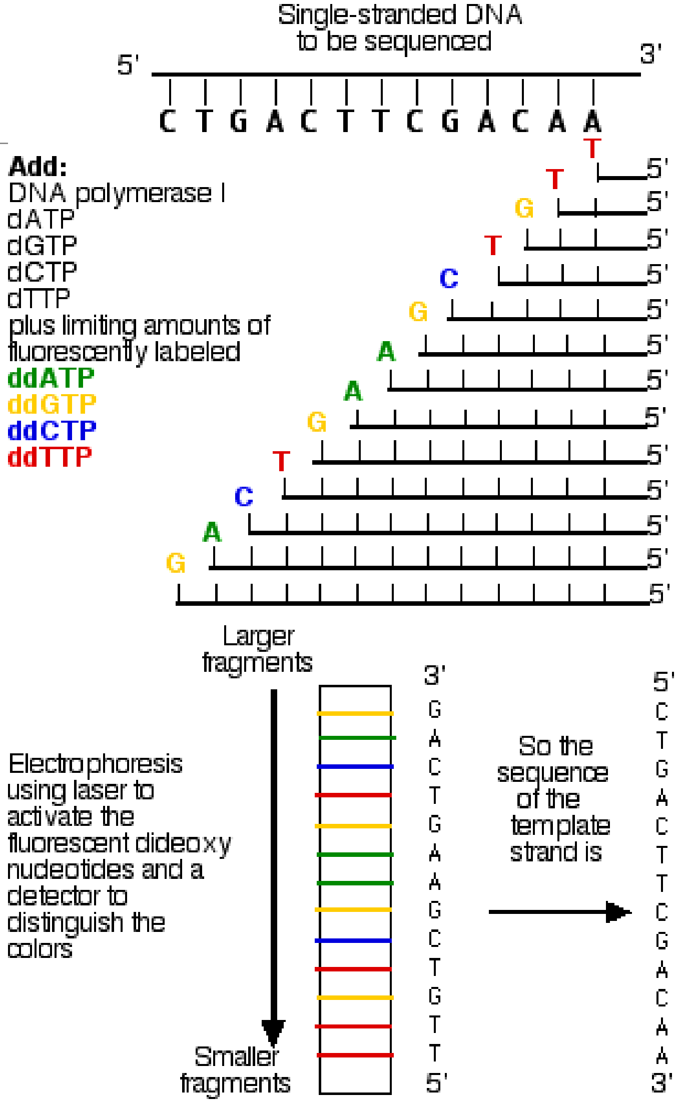
- Sequencing chemistry
- Base-calling errors: the Phred algorithm
- Mapping versus Assembly strategies
- Mapping via Burrows-Wheeler algorithm
- Assembly via *k*-mer graphs
- FASTQ, SAM/BAM, FASTA

DNA backbone structure



Sanger sequencing

- Given a template, generate complementary sequence.
- After dideoxynucleotide is incorporated, no more extension is possible.
- Fragment ladder is separated through electrophoresis.



Single-stranded DNA to be sequenced

Add:
DNA polymerase I
dATP
dGTP
dCTP
dTTP
plus limiting amounts of fluorescently labeled
ddATP
ddGTP
ddCTP
ddTTP

Electrophoresis using laser to activate the fluorescent dideoxy nucleotides and a detector to distinguish the colors

So the sequence of the template strand is

5'
3'
C T G A C T T C G A C A A
T
T
G
T
C
G
A
A
G
T
C
A
G
5'
Larger fragments
Smaller fragments
3'
G A C T G A A G C T G T T
5'
5'
C T G A C T T C G A C A A
3'

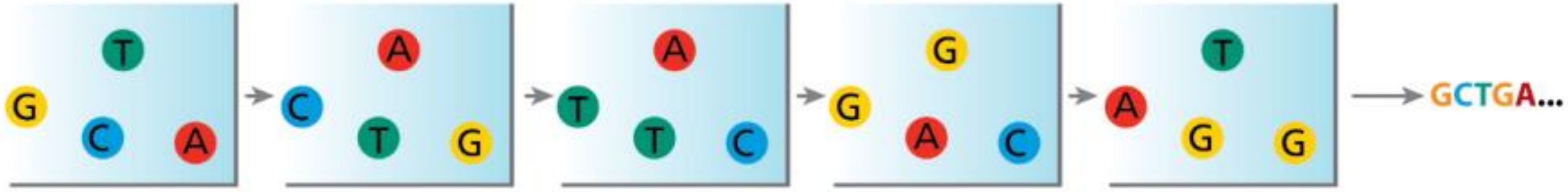
http://www.biology-pages.info/D/DNAsequencing.html

Massively parallel sequencing

- Detects nucleotide incorporation during DNA strand synthesis for millions of templates
- Base detection and strand synthesis typically differentiate competing technologies.

How it works:

- Separate millions of single-stranded DNA templates
- Fix the templates' location on a substrate & amplify
- Detect the incorporation of each base in each location



Sequencing jargon

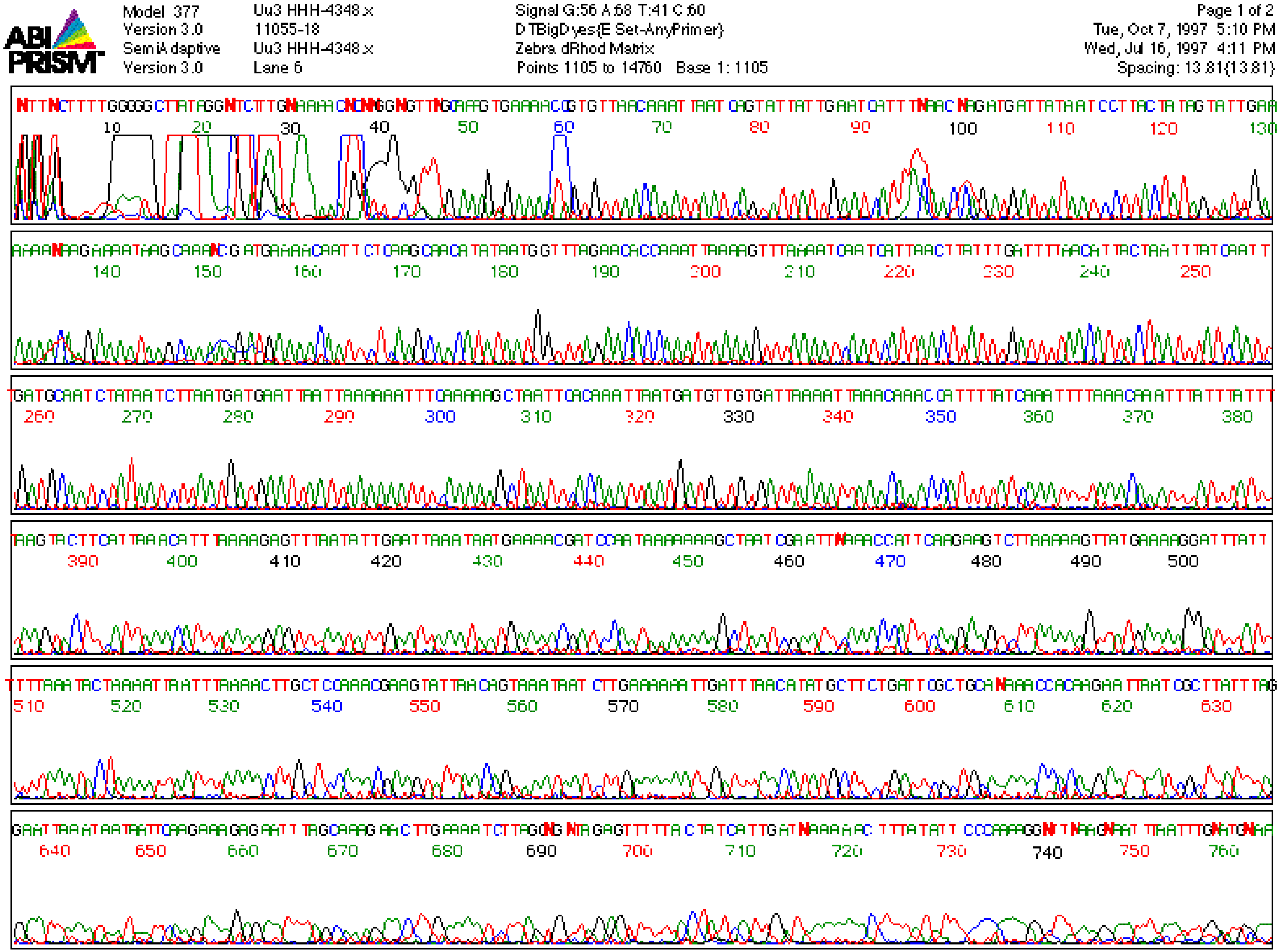
- A “template” is a piece of DNA to be sequenced, often ends of an “insert.”
- A “read” is a sequence corresponding to a single template, output by sequencer.
- “Shotgun sequencing” generates reads at random locations in the target DNA.
- “Fold coverage” divides the sum of read lengths by the target DNA length.

I prefer “massively parallel sequencing” to “next-gen sequencing.”

Old sequencing versus new

- Sanger sequencing typically produces longer reads (600 bp versus 150 bp).
- Sanger sequencing produces more accurate base calls for individual reads, but massively parallel sequencing overlaps many reads for each position.
- Massively parallel sequencing produces sequence from an incomparably larger number of templates in each experiment (millions rather than tens).

Electropherogram output



Model 377
Version 3.0
SemiAdaptive
Version 3.0

Uu3 HHH-4348.x
11055-18
Uu3 HHH-4348.x
Lane 6

Signal G:56 A:68 T:41 C:60
DTBigDyes{E Set-AnyPrimer}
Zebra dRhod Matrix
Points 1105 to 14760 Base 1: 1105

Page 1 of 2
Tue, Oct 7, 1997 5:10 PM
Wed, Jul 16, 1997 4:11 PM
Spacing: 13.81{13.81}

http://www.genome.uab.edu/uu/ASM_Genomes_Poster/genome_seq_recipe.htm

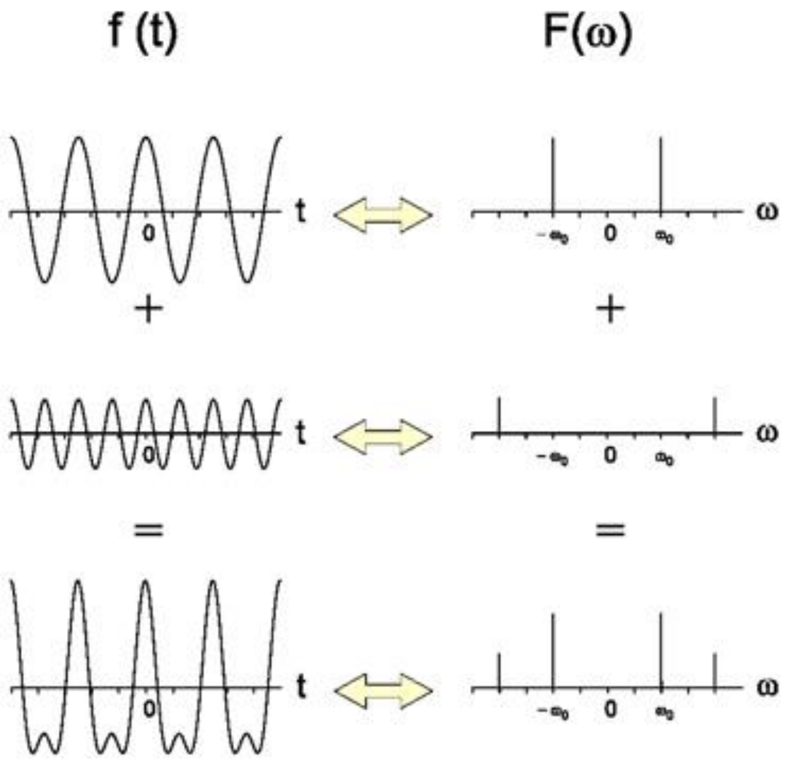
Phred: estimate basecalling errors

- Infer sequence from electropherogram
- Associate each basecall with a probability of error for this letter.
- A good basecall has these properties:
  - Peak matches the “beat” of its neighbors.
  - Only one trace is concave down at this call.
  - Highly-probable errors are far away.

B Ewing et al. *Genome Res.* (1998) 8: 175-185.

B Ewing et al. *Genome Res.* (1998) 8: 186-194.

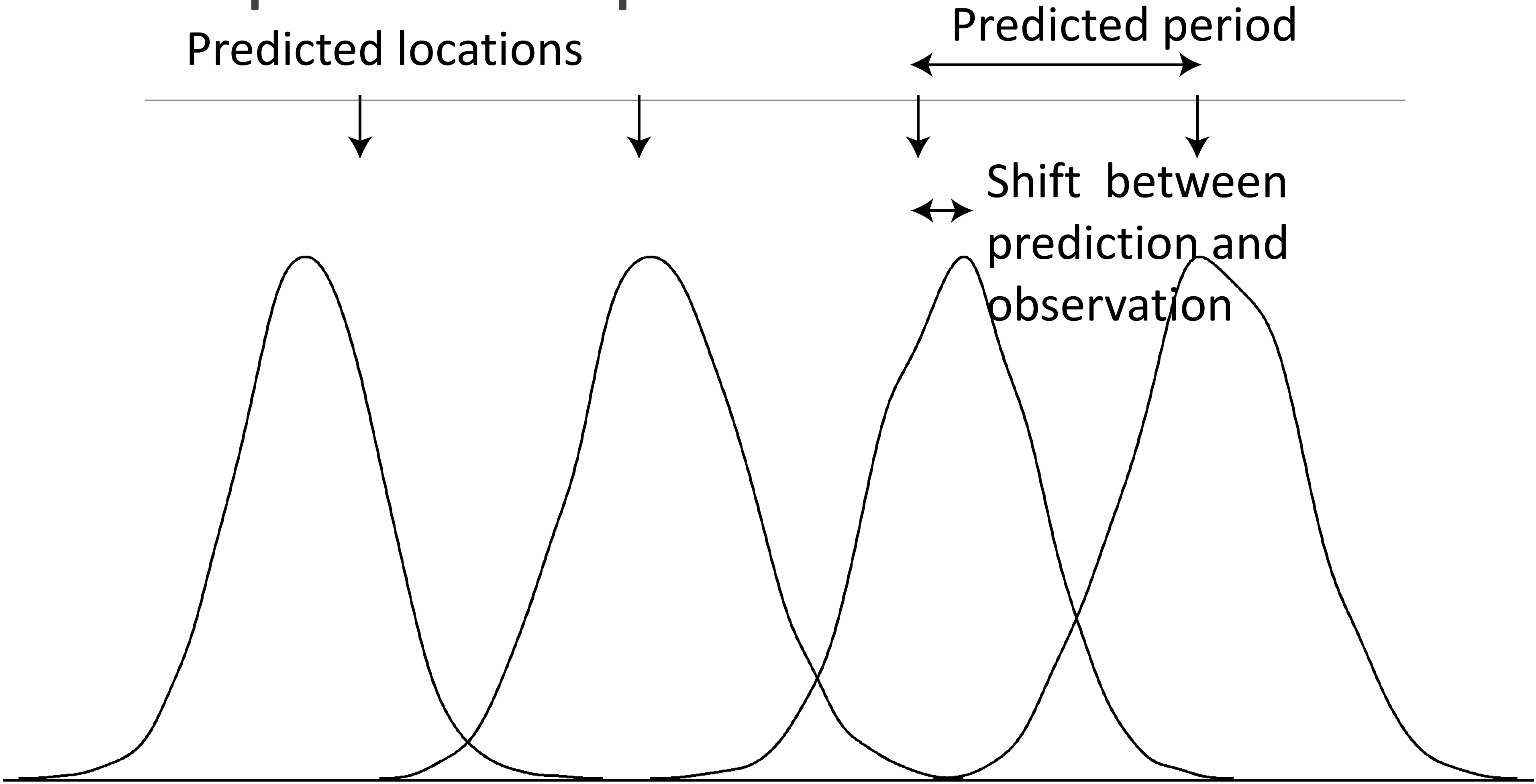
Fourier Transforms shift signal to sums of frequencies



- FTs decompose signals in time to frequencies.
- FTs are ubiquitous for recognizing frequencies.
- MP3 encoding uses FT to compress music.
- Sequencing requires that we recognize frequencies of base calls.

Image originally downloaded from http://www.bruker-biospin.com

Aligning observed and predicted peaks



Predicted locations

Predicted period

Shift between prediction and observation

Observed locations

Negative log scores help differentiate rare events

- P = probability of base call error, ranging from 0 to 1. Proximity to zero matters a lot.
- $Q = -10\log_{10} P$
- If P=.01 (1E-2) (1%) Q=20.
- If P=.001 (1E-3) (0.1%) Q=30.
- If P=.0001 (1E-4) (0.01%) Q=40.
- Q is the “Phred score.”

FASTQ: the output from DNA sequencers

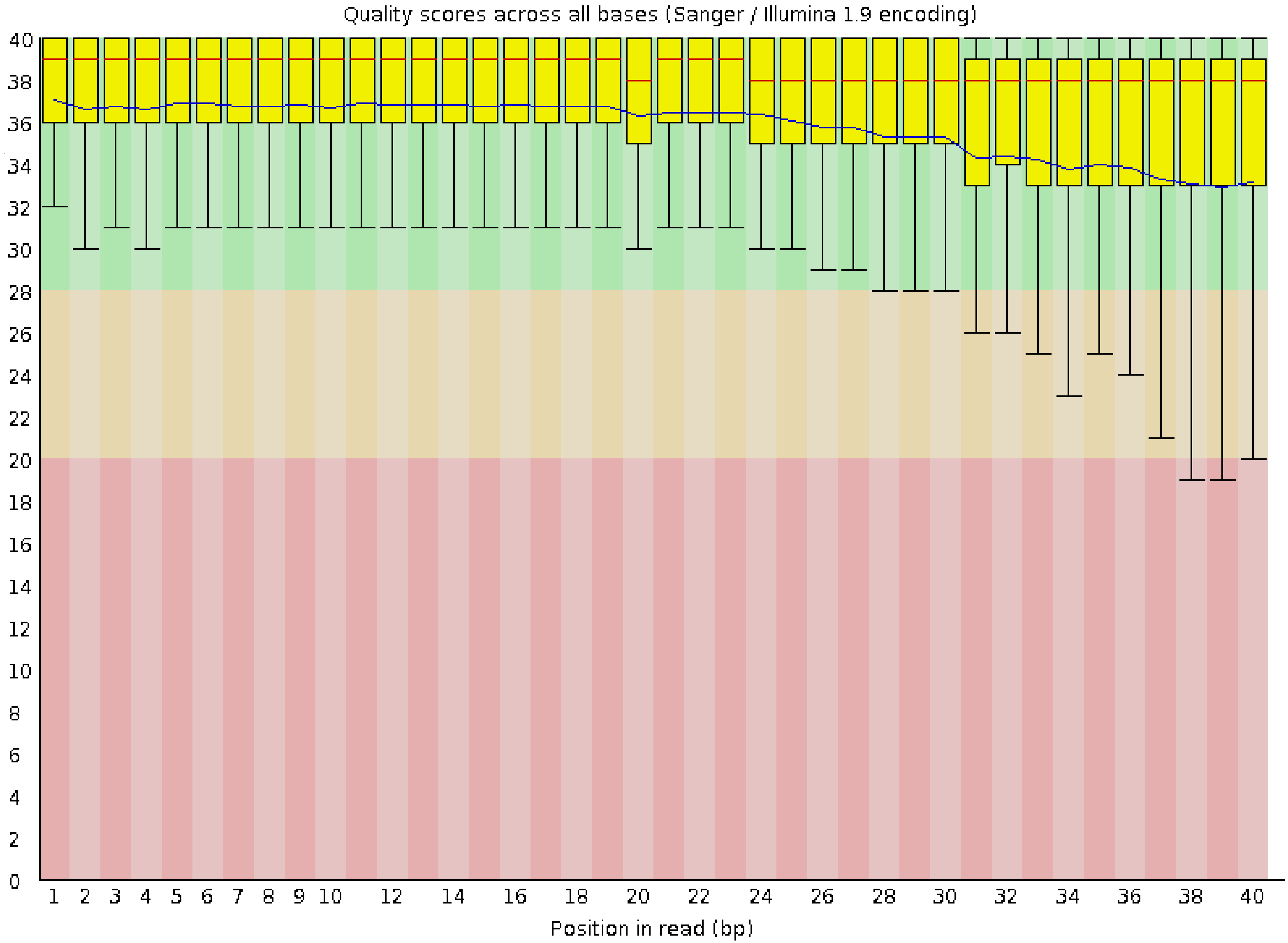
```
@NB501496:55:HMMT2AFXX:1:11101:13769:2030 1:N:0:CGCTCATT+AGGCTATA
CCCGCACTTCACATACCGAAGCCGCCTGTGCCGCTCCTGACCGCCTAATCCCGGAGGGGGGTGAGTGTGTGTT
+
AAAAAEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEAEEEEEEEAEEEEEEEEEEA/
```

- Line1: ‘@’ followed by a record identifier
- Line2: Sequencing read base calls
- Line3: ‘+’ means next line is quality scores
- Line4: Phred scores, from ‘!’ to ‘~’; A=32
- A FASTQ typically contains millions of reads and requires compression (.fastq.gz or .fastq.bz2)

PJA Cock et al. *Nucl. Acids. Res*. (2010) 38: 1767

FASTQC

www.bioinformatics.babraham.ac.uk



Phred scores are not uniform across lengths of reads

Intermission

Two chief uses for reads: *Mapping* and *Assembly*

MAPPING

- Determines where each read matches to the annotation
- Basis for exome / WGS variant calling

ASSEMBLY

- Infers large “contig” regions from overlapping reads
- Necessary for non-model organisms and for unmapped reads

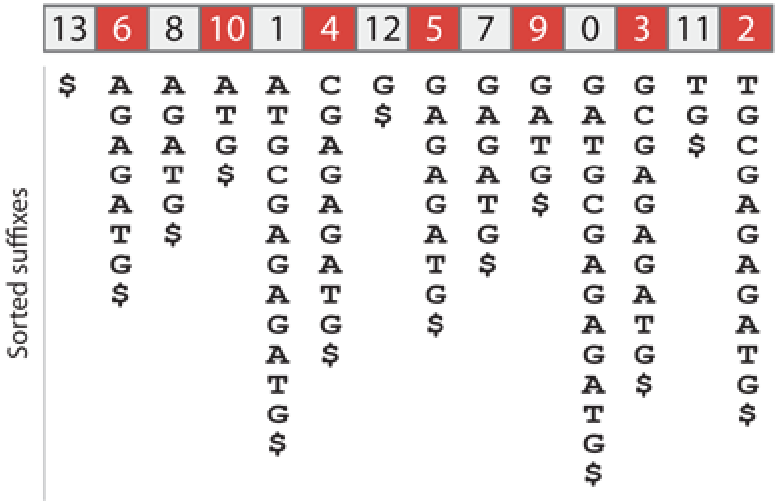
Why emphasize *mapping*: aligning short reads to reference?

- Assembling reads *de novo* is time-consuming and relatively error-prone.
- Recognition of sequence variants is easier when annotation provides typical sequence.
- Short read massively parallel sequencers (e.g. Illumina) produce >100 Gbp per day.
- QC, trimming and alignment are assumed for many downstream tools.

H. Li and N. Homer. *Brief. Bioinform*. (2010) 11: 473-483

Burrows-Wheeler Transform prepares a genome index.

- A suffix array shows all truncations from 5’ of sequence.
- The array of suffixes is sorted.
- BW transform stores the letter preceding truncated sequence and where it appears in original.



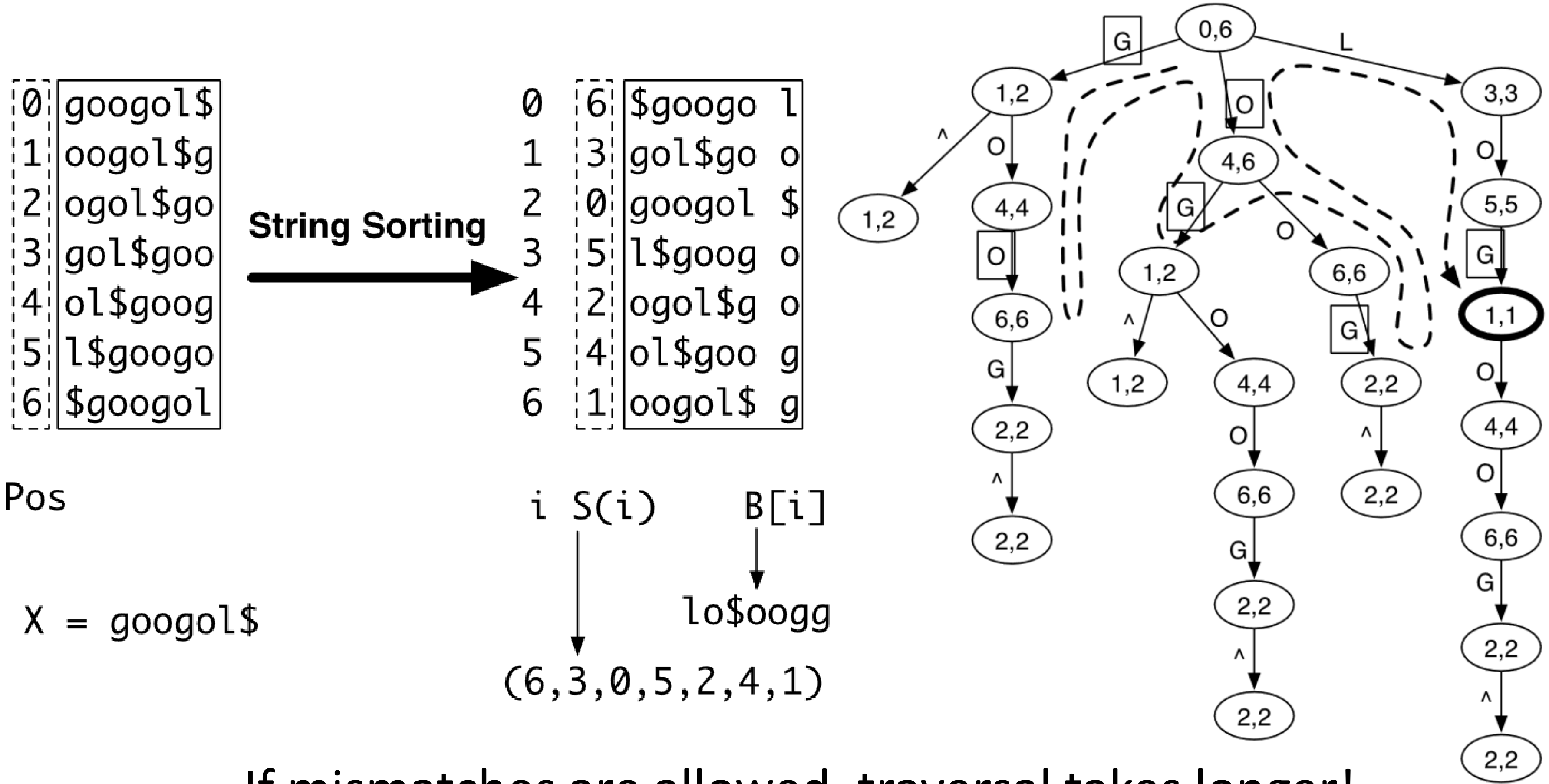
13	6	8	10	1	4	12	5	7	9	0	3	11	2
$	AGAGATG$	AGATG$	ATG$	ATGCGAGAGATG$	CGAGAGATG$	G$	GAGAGATG$	GAGATG$	GATG$	GATGCGAGAGATG$	GCGAGAGATG$	TG$	TGCGAGAGATG$

Suffix array of GATGCGAGAGATG

http://blog.thegrandlocus.com/2016/07/a-tutorial-on-burrows-wheeler-indexing-methods

B. Langmead et al. *Genome Biology* (2009) 10: R25

The index from BWT can be traversed like a tree

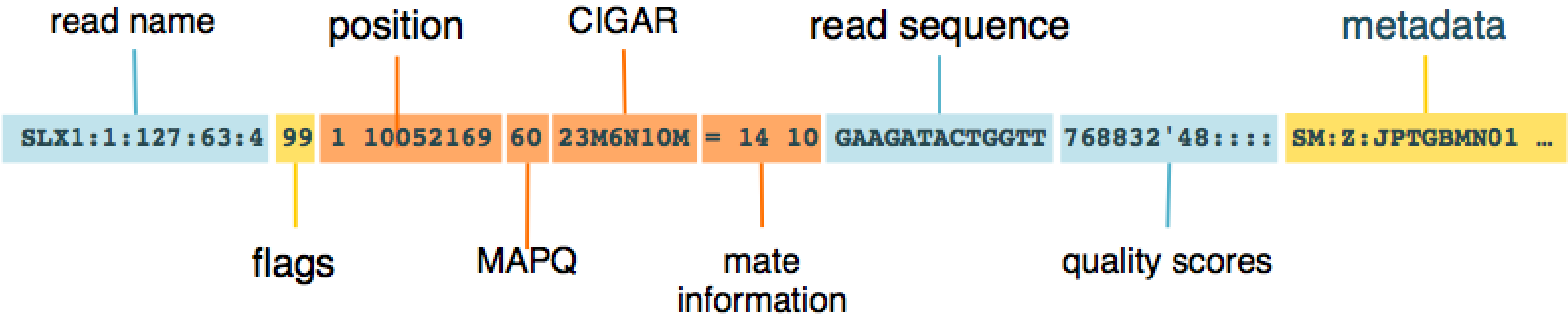


If mismatches are allowed, traversal takes longer!

Heng Li and R. Durban. *Bioinformatics* (2009) 25: 1754-1760.

BAM files: Sequence Alignment Maps, in binary

HEADER containing metadata (sequence dictionary, read group definitions etc)
RECORDS containing structured read information (1 line per read record)



```
SLX1:1:127:63:4 99 1 10052169 60 23M6N10M = 14 10 GAAGATACTGGTT 768832'48:::: SM:Z:JPTGBMN01 …
```

https://gatk.broadinstitute.org/hc/en-us/articles/360035890791-SAM-or-BAM-or-CRAM-Mapped-sequence-data-formats

MAPQ is mapping quality; CIGAR reflects diffs versus annotation. Mate Info relates paired-end sequences. Basecall PHRED scores are given in single-character format (ASCII-33 through 126).

Li et al. *Bioinformatics* (2009) 25: 2078-2079

Sequence Assembly

Why do we need *assembly*?

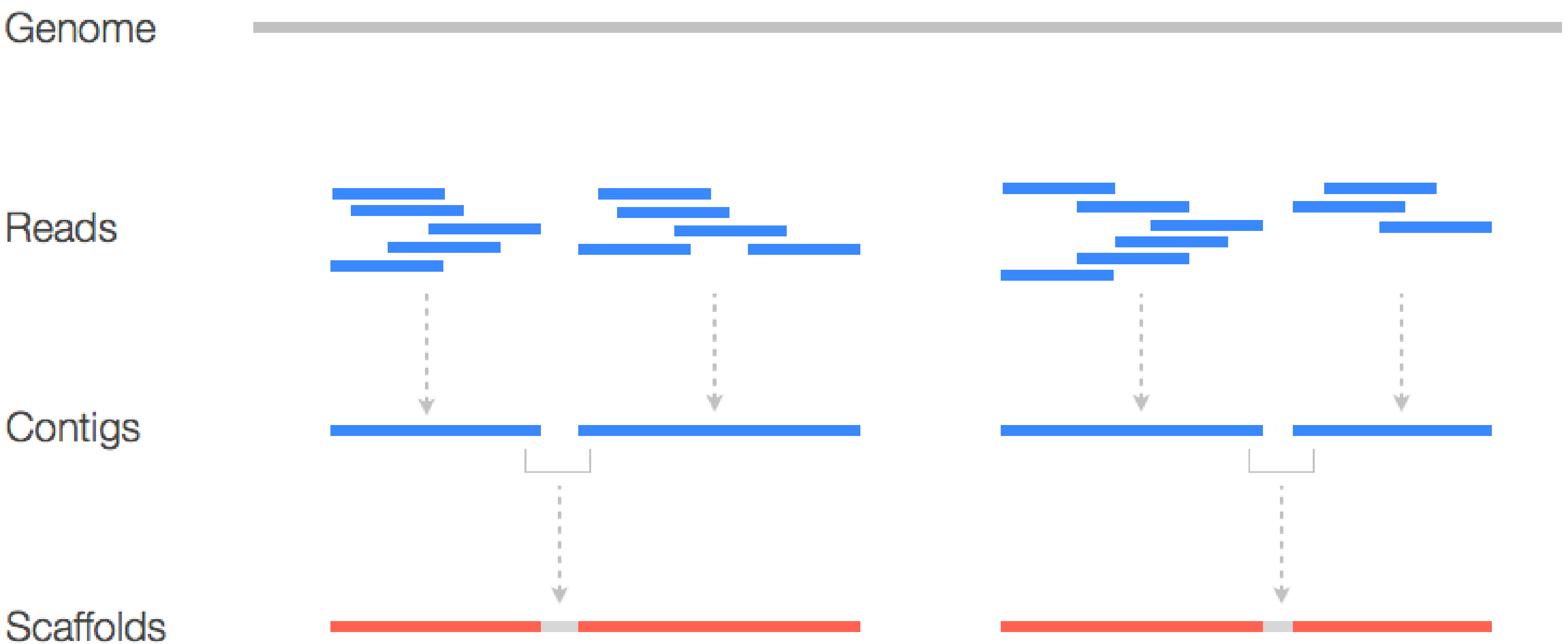
- We have millions of short reads, but we need scaffolds, ideally one per chromosome.
- Even if we map our reads to an annotation, many *unmapped reads* may not align to our selected reference annotation.
- We may work in a non-model organism that lacks a reference genome annotation!

Assembly definitions

- A ***contig*** is a contiguous length of genomic sequence in which the order of bases is known to a high confidence level.
- ***Scaffolds*** are composed of contigs and gaps.
- Within a scaffold, the ordering and orientation of contigs with respect to each other has been established.

https://mycocosm.jgi.doe.gov/help/scaffolds.jsf

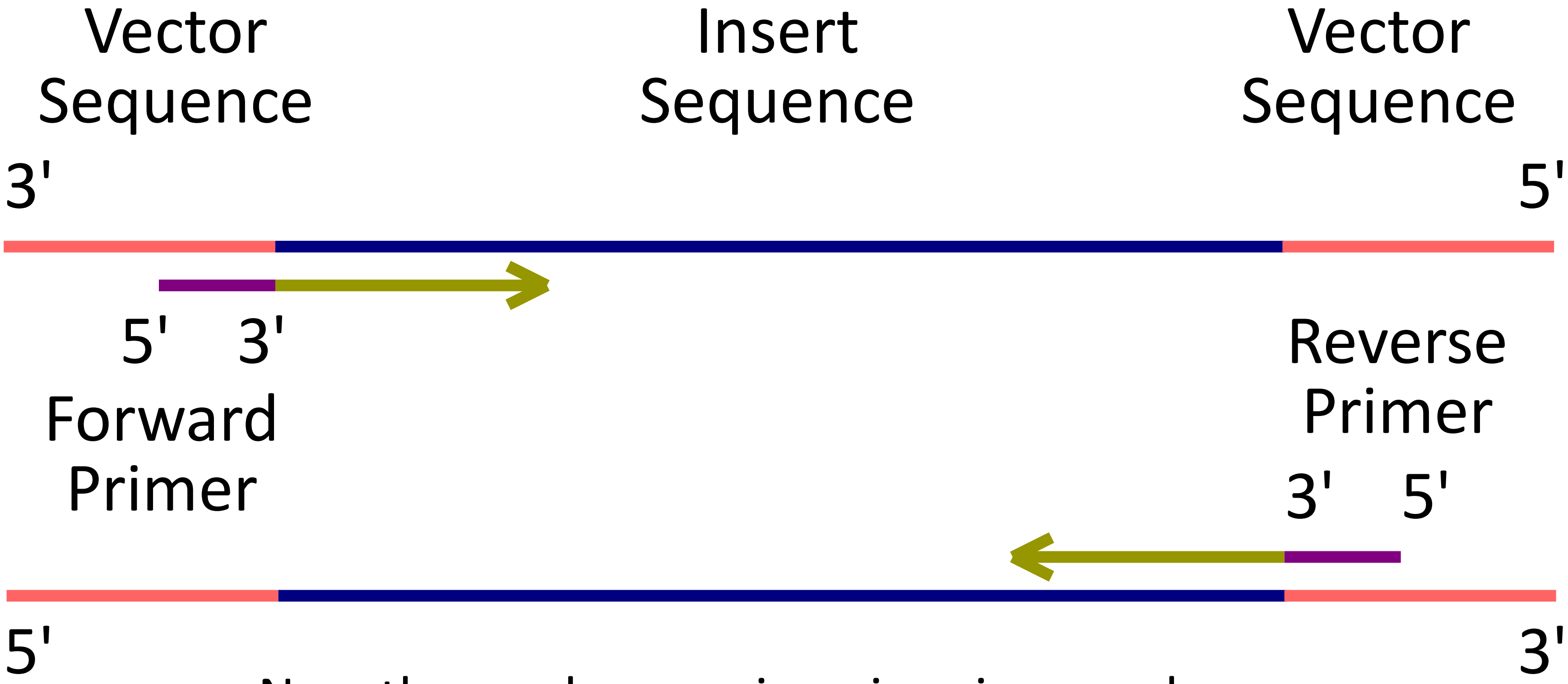
Assembly builds long sequences from many short ones



If reads are independent, how do we judge which contigs are near each other?

Image from Vijay Lakhujani, Biostars

Paired End Sequencing



Vector Sequence

Insert Sequence

Vector Sequence

3'

5'

5' 3'

Forward Primer

Reverse Primer

3' 5'

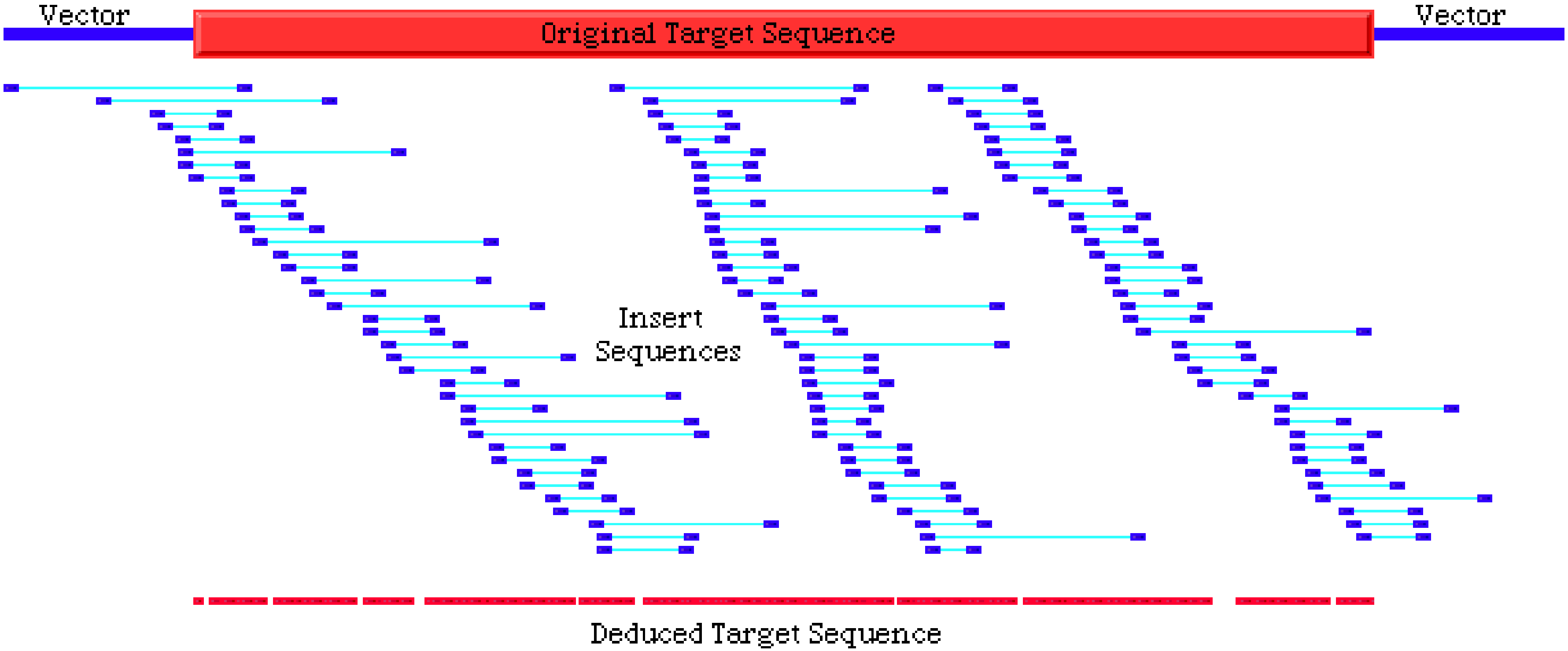
5'

3'

Now the reads come in pairs, since each comes from the opposite end of a single piece of DNA.

A Edwards and CT Caskey, *Methods: a Companion to Methods in Enzymology* (1991) 3: 41-47.

Shotgun sequencing of genomes

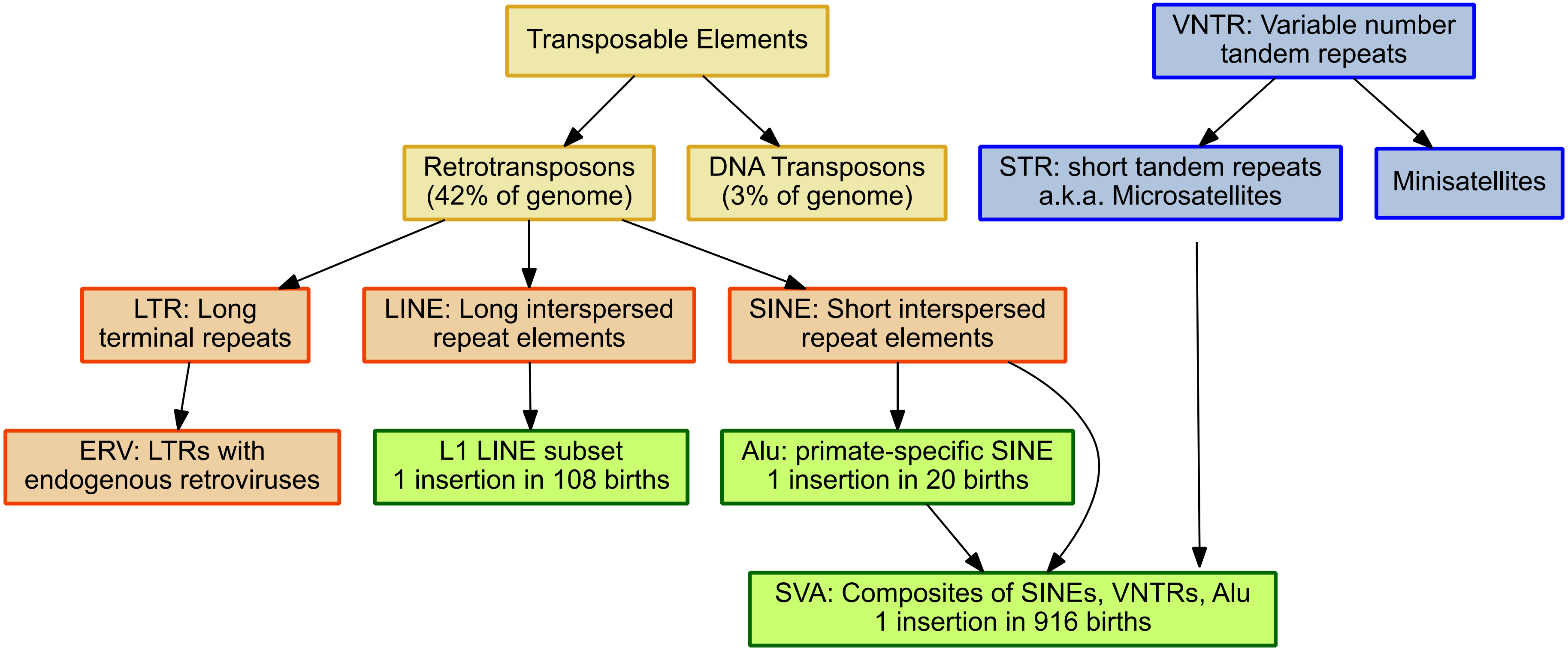


Given millions of reads of 100 nucleotides, assemble contigs of overlapping sequences. Determine which contigs are neighbors.

Figure courtesy of Jared Roach

South Africa operates >14 sequencers in 9 facilities.

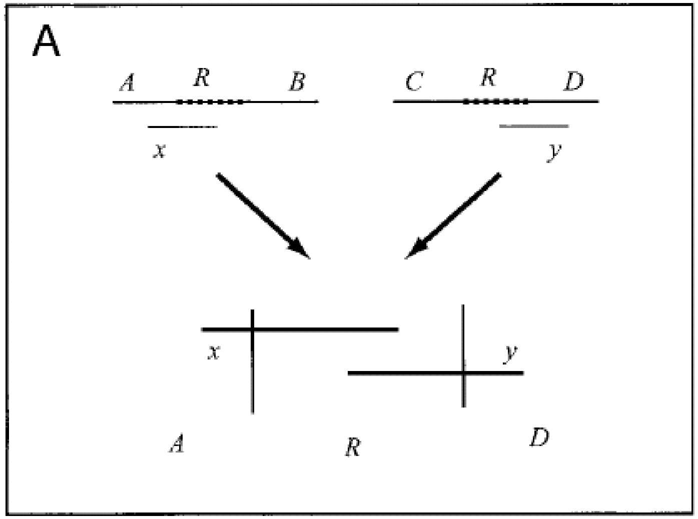
Noncoding DNA

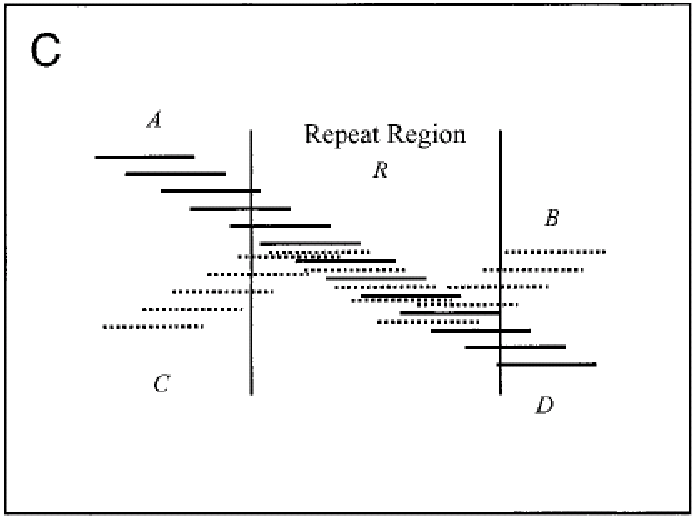

Transposable Elements
VNTR: Variable number tandem repeats
Retrotransposons (42% of genome)
DNA Transposons (3% of genome)
STR: short tandem repeats a.k.a. Microsatellites
Minisatellites
LTR: Long terminal repeats
LINE: Long interspersed repeat elements
SINE: Short interspersed repeat elements
ERV: LTRs with endogenous retroviruses
L1 LINE subset 1 insertion in 108 births
Alu: primate-specific SINE 1 insertion in 20 births
SVA: Composites of SINEs, VNTRs, Alu 1 insertion in 916 births

Cowley and Oakey. *PLOS Genetics*. (2013) 9: e1003234

ARACHNE assembler: The danger of assembling repeats

- The same repeat appears between genes A and B and between C and D.
- If these two repeats are treated as one sequence, neither upstream nor downstream genes will assemble correctly.



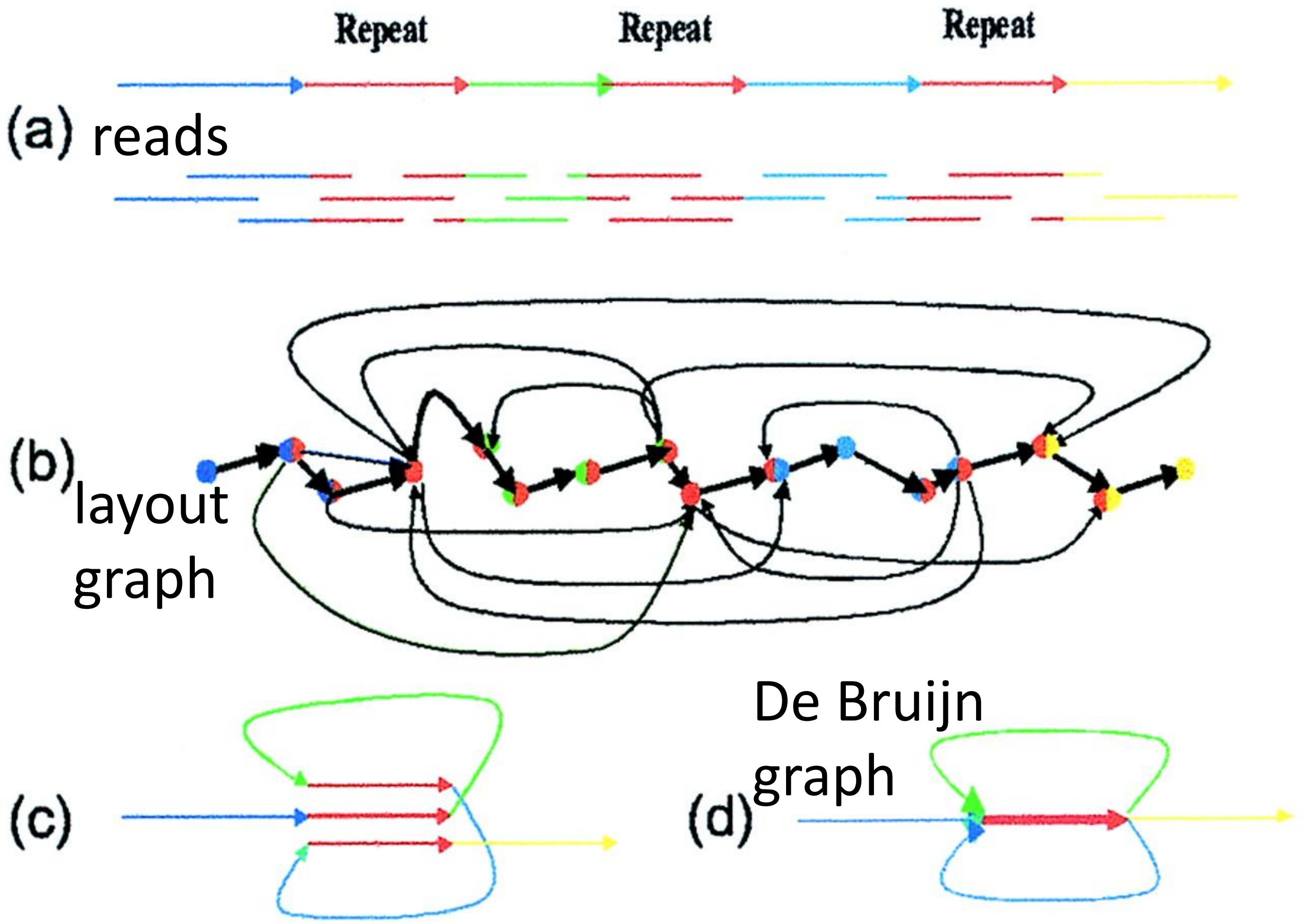


S Batzoglou et al. *Genome Res*. (2002) 12: 177-189.

Initial assemblers: Overlap-Layout-Consensus

Each read (a) is a vertex in overlap graph (b). Edges represent overlap.

Old assemblers sought *Hamiltonian Path*, visiting every vertex once. This is *NP-complete*.



reads

layout graph

De Bruijn graph

PA Pevzner et al. *PNAS* (2001) 98: 9748-9753.

Making a “*k*-mers” catalog

- Create a sorted list of all *k* bp sequences along with their positions within reads.
- Exclude high-count *k*-mers as repeats.
- Overlapping reads will share *k*-mers unless:
  - Overlap is less than *k* in length
  - Base calling errors obfuscate overlap

Q. Zhang et al. *PLOS One* (2014) 9: e101271

k-mer traversal yields contig

TGGTTTTGATTATTTGCTGGTTGCC

GGTTTTGATTATTTGCTGGTTGCCA

GTTTTGATTATTTGCTGGTTGCCAA

TTTTGATTATTTGCTGGTTGCCAAA

TTTGATTATTTGCTGGTTGCCAAAC

TTGATTATTTGCTGGTTGCCAAACA

TGATTATTTGCTGGTTGCCAAACAT

GATTATTTGCTGGTTGCCAAACATC

k-mers of 25bp

TGGTTTTGATTATTTGCTGGTTGCCAAACATC

Contig of 32bp

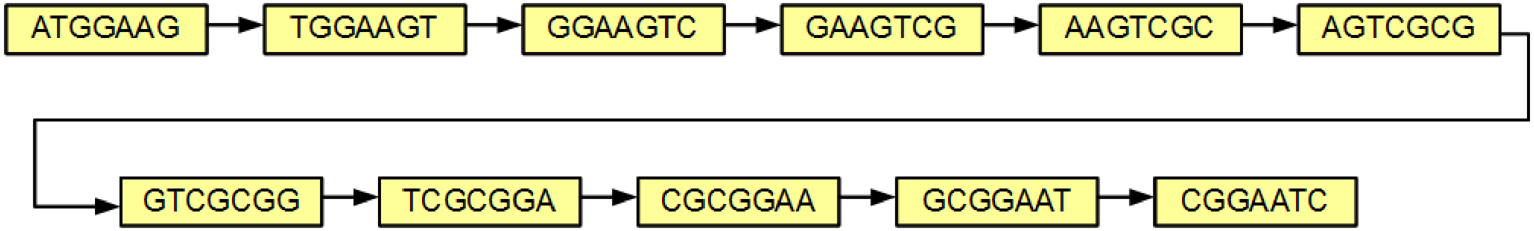
Eulerian path through de Bruijn graph visits each edge once

sequence ATGGAAGTCGCGGAATC

7mers

```
ATGGAAG
 TGGAAGT
  GGAAGTC
   GAAGTCG
    AAGTCGC
     AGTCGCG
      GTCGCGG
       TCGCGGA
        CGCGGAA
         GCGGAAT
          CGGAATC
```

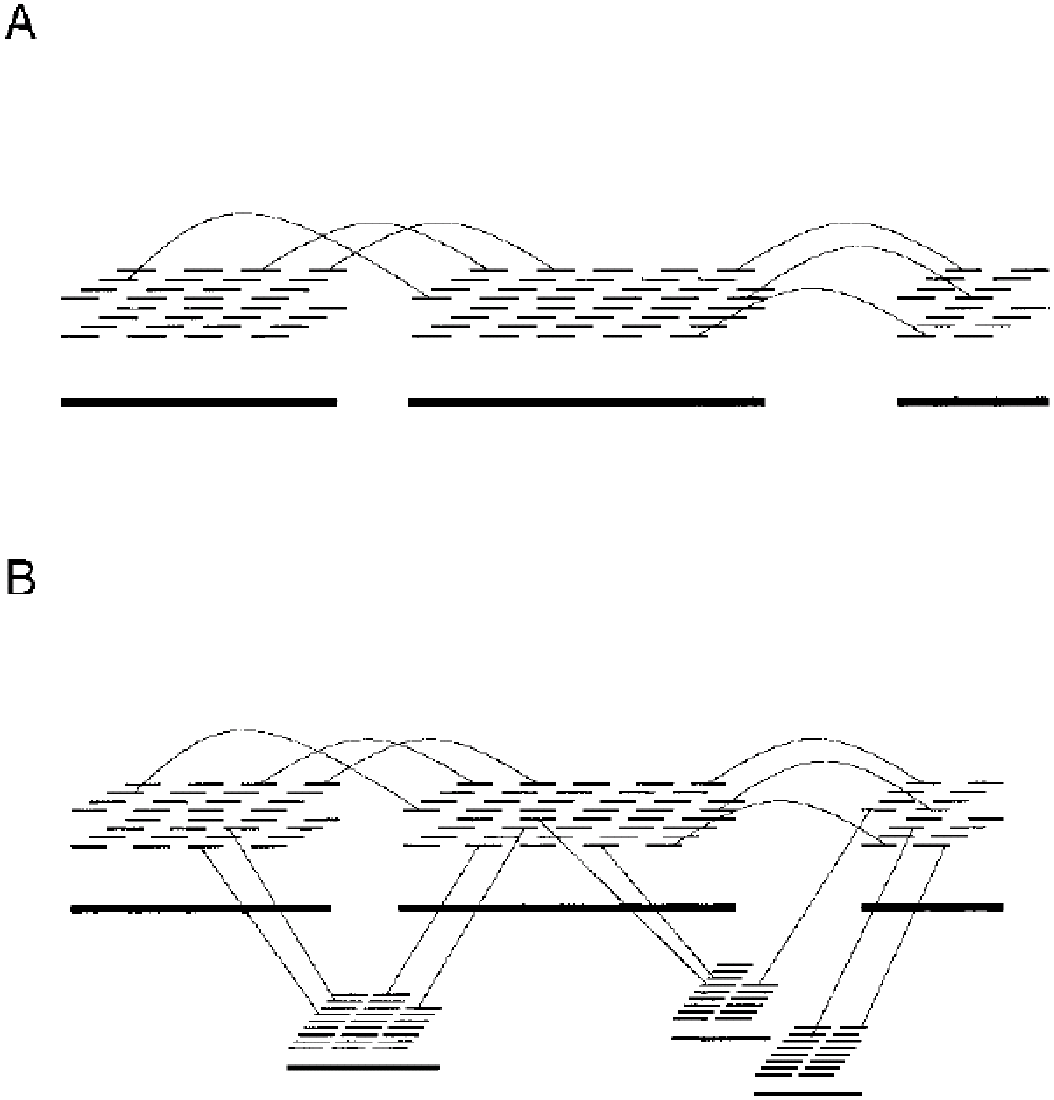
de Bruijn graph



Paired reads enable contig linking

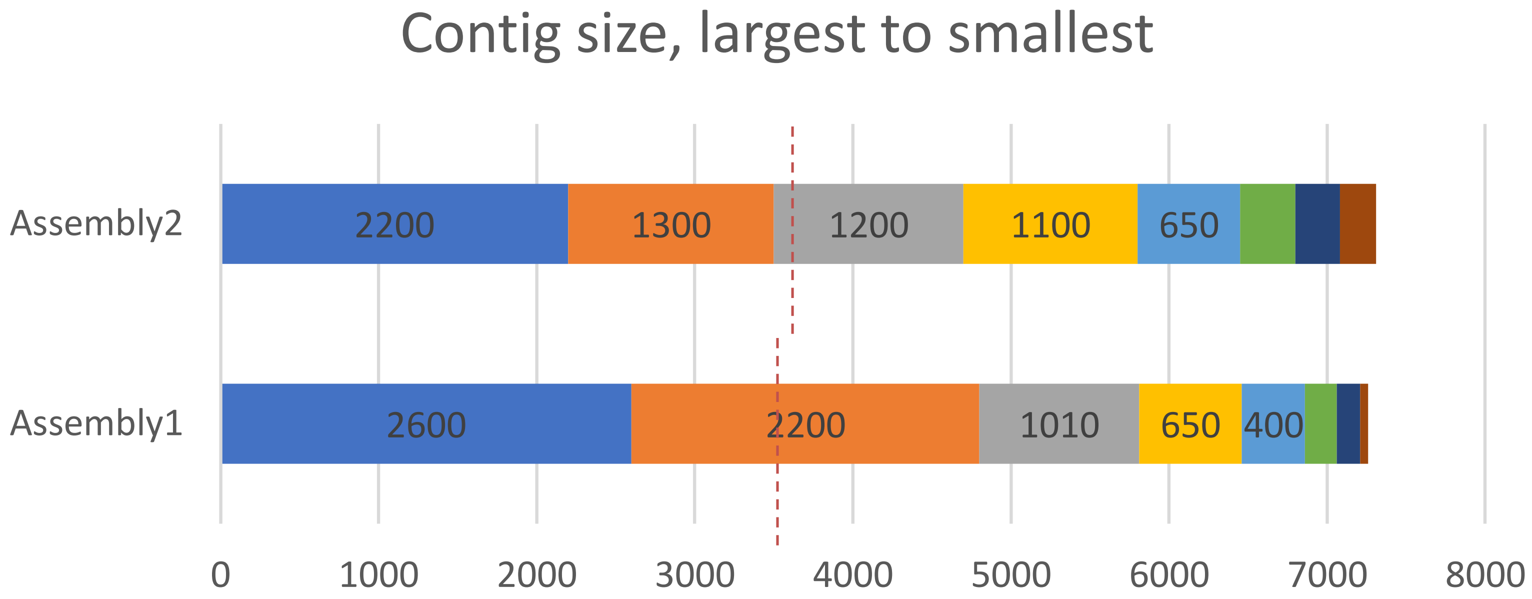
If two separate contigs contain each end of a particular insert, those contigs are near each other.

A set of neighboring contigs is a ***scaffold***.



S Batzoglou et al.

N50: how “chopped up” is my assembly?



- Sort contigs from big to small. Stack them.
- N50 is size of contig at half of length sum.

D Earl et al. *Genome Res*. (2011) 21: 2224-2241.

FASTA “Database” Format

```
>ENSP00000396333.1 pep chromosome:GRCh38:1:154963677:154966490:-1
gene:ENSG00000160691.18 transcript:ENST00000444664.5
gene_symbol:SHC1 description:SHC adaptor protein 1 [Source:HGNC
Symbol;Acc:HGNC:10840]
XDEEEEEPPDHQYYNDFPGKEPPLGGVVDMRLREGAAPGAARPTAPNAQTPSHLGATLPV
GQPVGGDPEVRKQMPPPPPCPGRELFDDPSYVNVQNLDKARQAVGGAGPPNPAINGSAPR
DLFDMKPFEDALRVPPPPQSVSMAEQLRGEPWFHGKLSRREAEALLQLNGDFLVRESTTT
PGQYVLTGLQSGQPKHLLLVDPEGVRWGFAMLPKLFLNSRAQVIRLPRPPRVLGLQARTT
MPSLHIFFCTVYTLLRHANFLQVKKGVYSSQLHSFRADVAFAFSHFTDLSIPTTVSF
```

- Line 1: ‘>’ + accession + whitespace + description
- Following lines: sequence

WR Pearson and DJ Lipman. *PNAS* (1988) 85: 2444-2448.

Closing thoughts

- Phred helps us to evaluate each basecall.
- Mapping aligns reads to existing annotation; assembly builds long sequences from reads.
- Mapping speeds increase dramatically when a BWT-based index is available.
- *De novo* assembly starts with a *k*-mer catalog and seeks a Eulerian path through it.