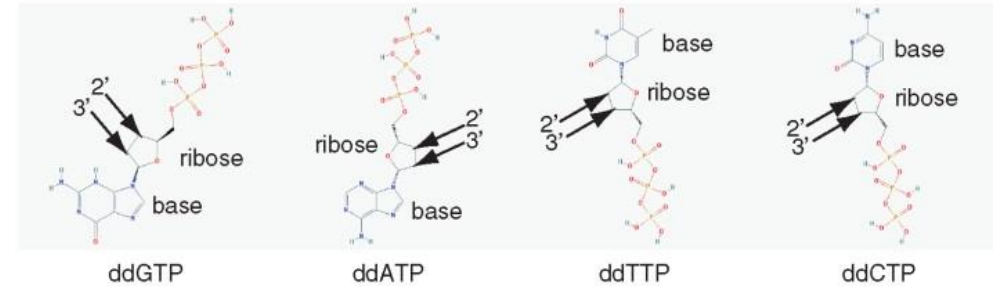


Sequencing and assembly

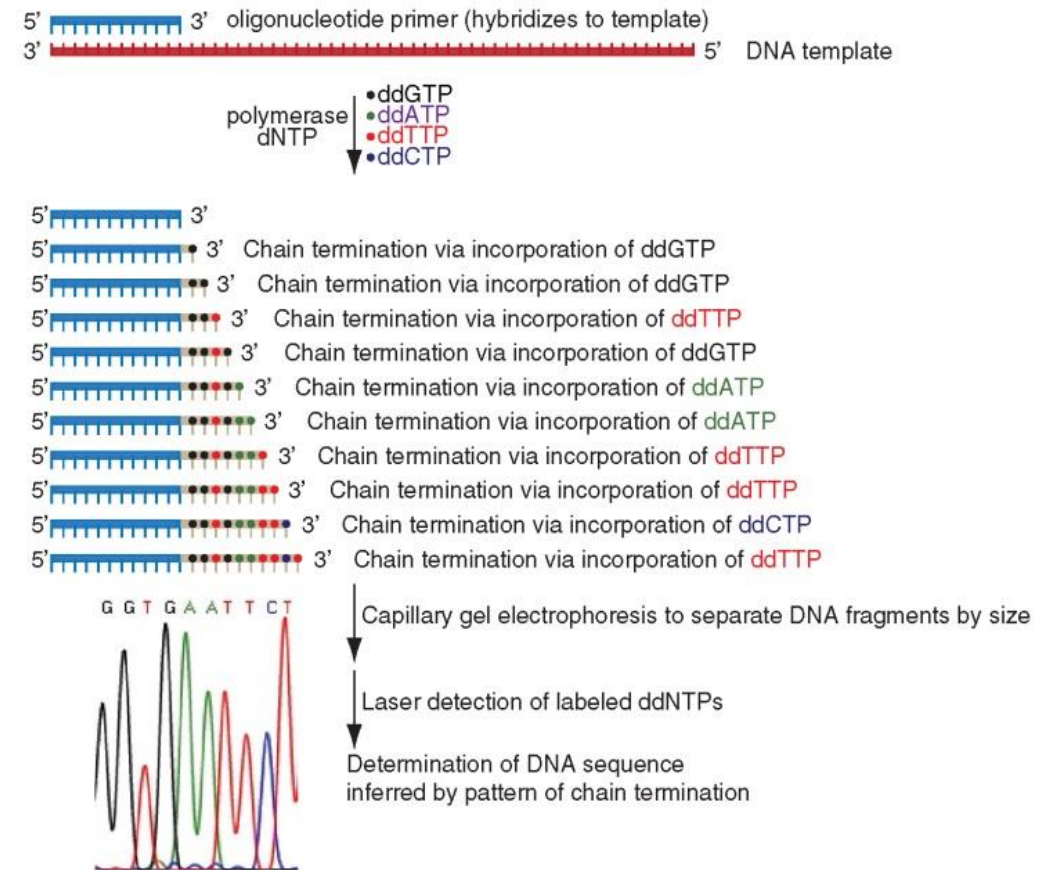
Sanger sequencing

- Invented by Frederick Sanger in 1970
- Starts with a primer template- a short sequence complementary to the sequence of interest
- Specific
- Read length up to 800
- Accurate
- Still used today to sequence clones or verify NGS results

(a) Dideoxynucleotides (ddNTPs) (-OH of dNTP is replaced by -H of ddNTP at the 2' ribose position)

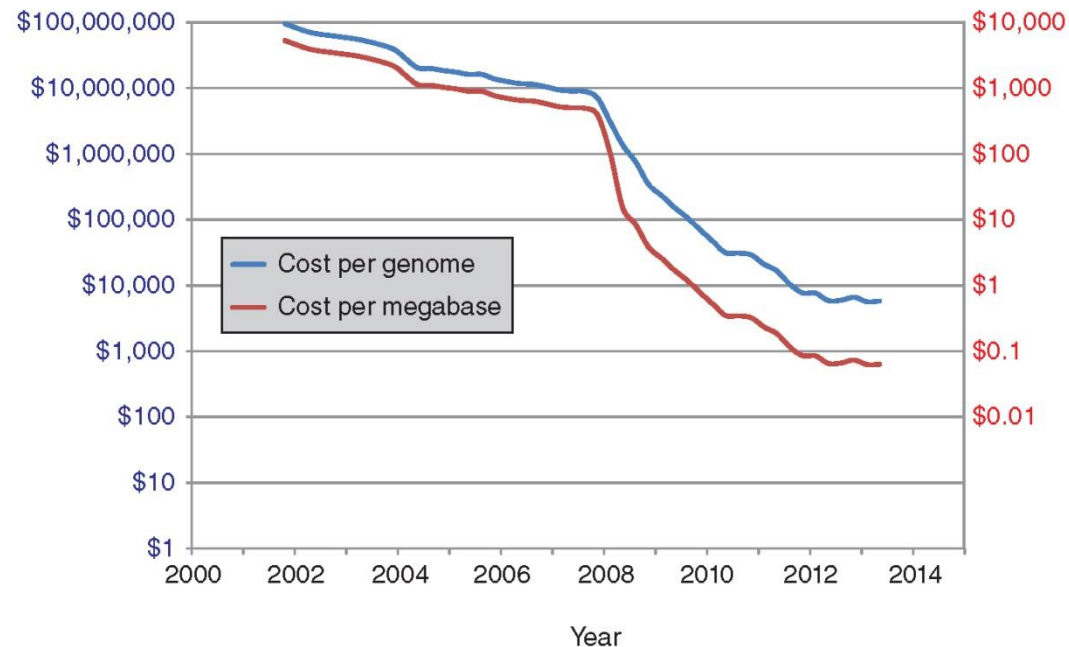


(b) Primer elongation, chain termination upon incorporation of ddNTP, separation, detection



Next Generation Sequencing (NGS)

- A series of different technologies enabling fast automated sequencing
- Most technologies have short read length
- Cost is dramatically reduced
- Each technology has its own error profile



MIT Technology Review

NEWS & ANALYSIS | FEATURES | VIEWS | MULTIMEDIA | DISCUSSIONS | TOPICS

POPULAR: NET NEUTRALITY ABROAD | TESLA'S SMART FIXES | CLOUDLESS

BIOMEDICINE NEWS | 2 COMMENTS

Does Illumina Have the First \$1,000 Genome?

Illumina announces a new high-end sequencer made for “factory-scale” sequencing of human genomes.

By Susan Young on January 14, 2014

The \$1,000 genome has been a catchphrase of the sequencing industry for years, but despite bold promises from different companies, this benchmark hasn't been met. Now, thanks to a new sequencing machine from Illumina, it may finally be within reach.

WHY IT MATTERS

Genome sequencing is still too costly for many medical applications, or for large-scale sequencing projects that could uncover the genetic basis of disease.

At the J.P. Morgan Healthcare Conference on Tuesday, Illumina CEO Jay Flatley announced a new

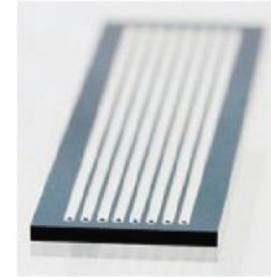
Next | Previous | Highlight all | Match case | Phrase not found

Illumina

- DNA is randomly fragmented followed by size selection
- Adapters are ligated at each end so that the fragment can bind to the flow cell surface
- Each single fragment is amplified in place with “bridge amplification”
- There are 4 **reversible** terminators which are added **at the same time**
- Locations of the added bases is read out by laser scanning
- Most widely used platform

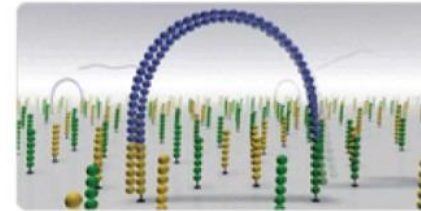
Randomly fragment genomic DNA

Library preparation



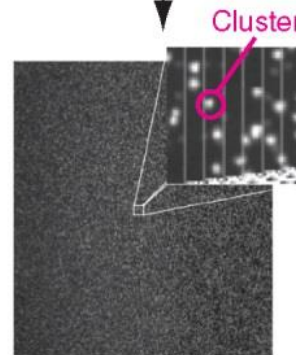
Samples immobilized on surface of a flow cell (8 lanes)

Solid phase amplification



- Bridge amplification (inverted U) generates clusters on surface of flow cell
- ~Ten million single-molecule clusters per square centimeter

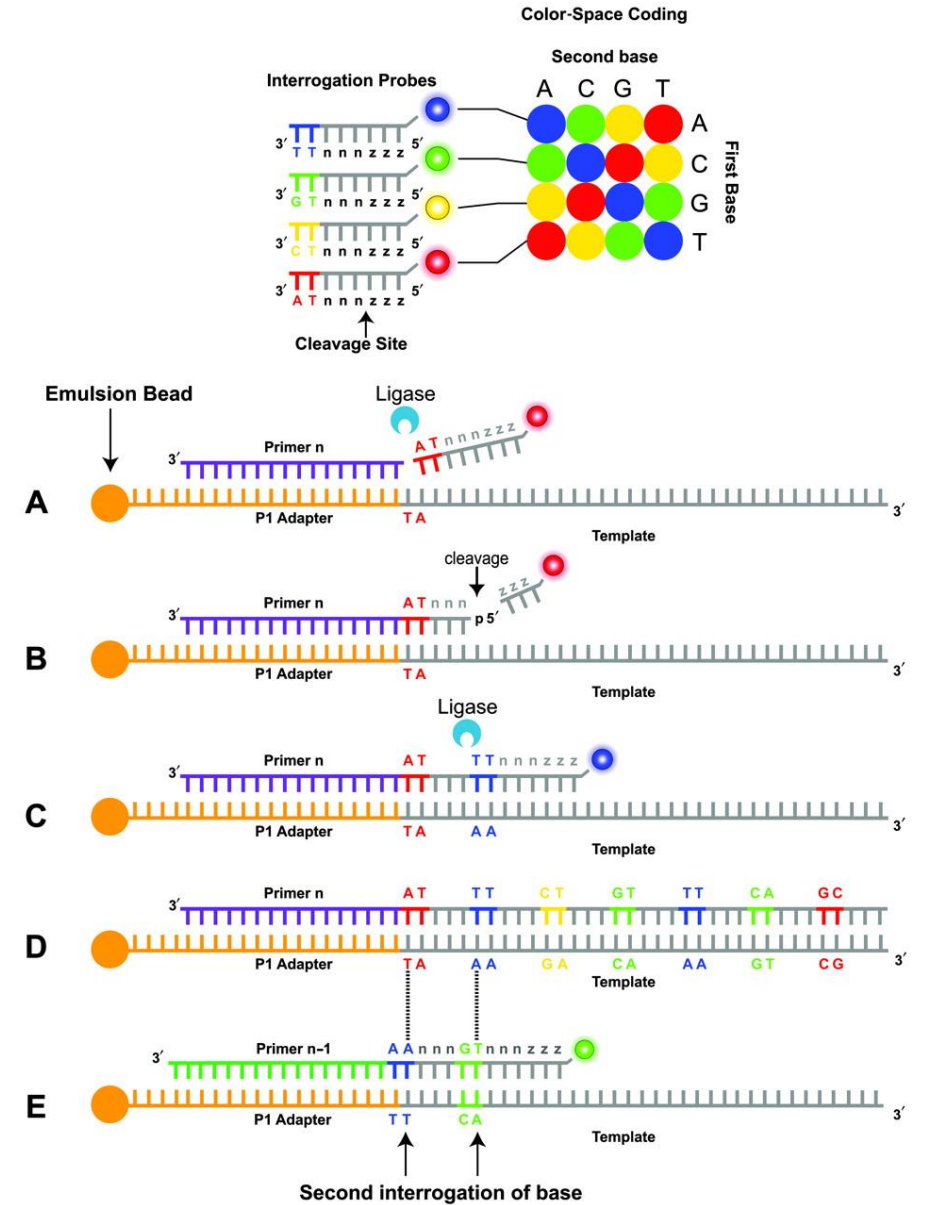
Sequencing by synthesis



- Each cycle: add polymerase, one labeled deoxynucleoside triphosphate (dNTP) at a time (four labeled dNTPs per cycle)
- Image fluorescent dyes
- Call nucleotide
- Enzymatic cleavage to remove

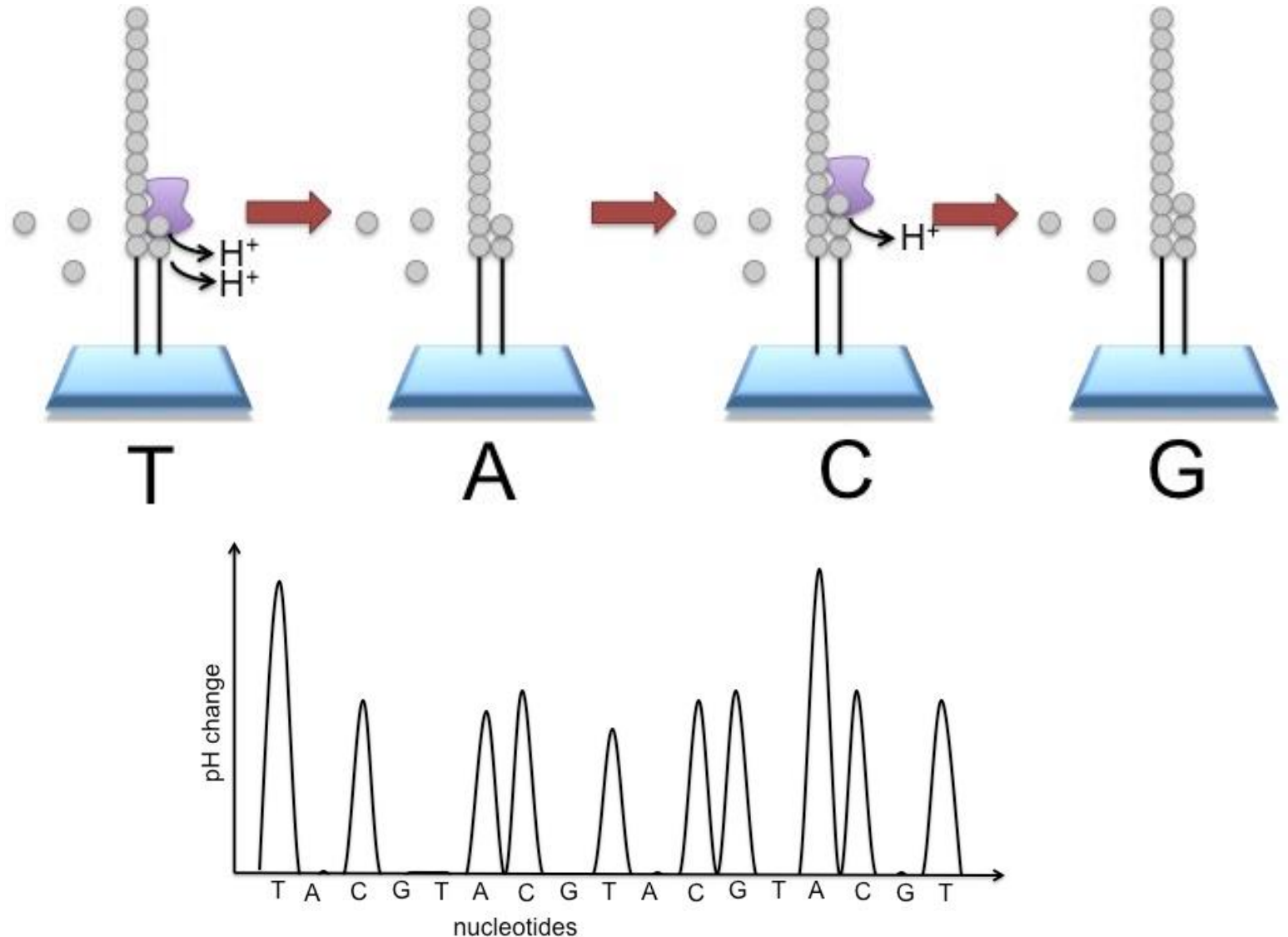
ABI Solid

- Library prep is the same
- Amplification is emulsion based
- DNA sequenced by ligation
- Very low error rate
- Complex error model



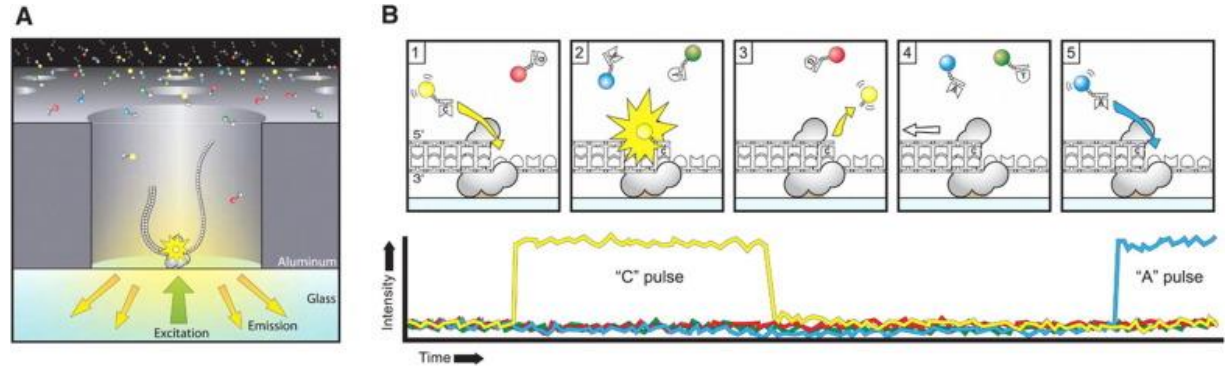
Ion Torrent

- Direct detection of nucleotide incorporation
- Basically a very good pH meter
- Very fast
- Used for clinical sequencing
- High error rate
- Non-trivial error model
- Sequential nucleotide addition --problems counting consecutive identical nucleotides



Pacific Biosciences

- Immobilized polymerase
- Allows for fluorescent pulse detection
- Very long read length
- Can span repetitive elements
- Can be used in hybrid sequencing, combining short and long reads

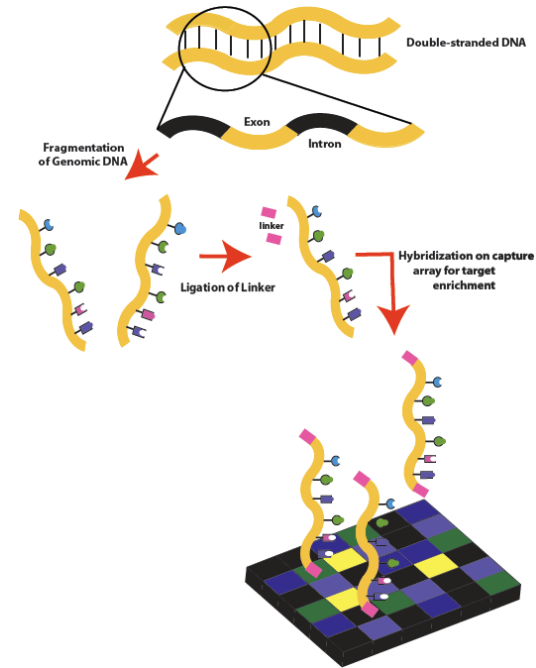


Comparison

Platform	Read length	Reads per run	Run time	Cost per megabsse	Accuracy	Error type
Sanger	400-900		<3hours	2400	99.9	Single nucleotide substitutions
Illumina	50–250	3 billion	1-10 days	~0.10	98	Single nucleotide substitutions
SOLiD	50	~1.4 billion	7–14 days	0.13	99.9	AT bias
Ion torrent	200	<5 million	2 hours	2	99	deletions
PacBio	2900	75,000	<2 hours	2	99	GC deletions

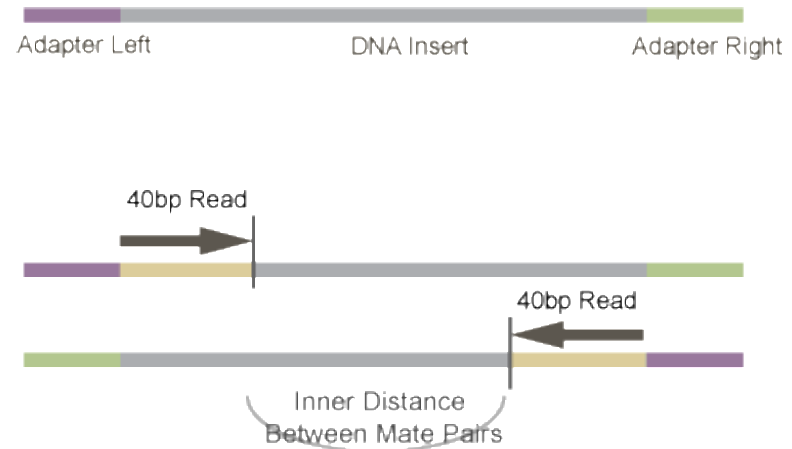
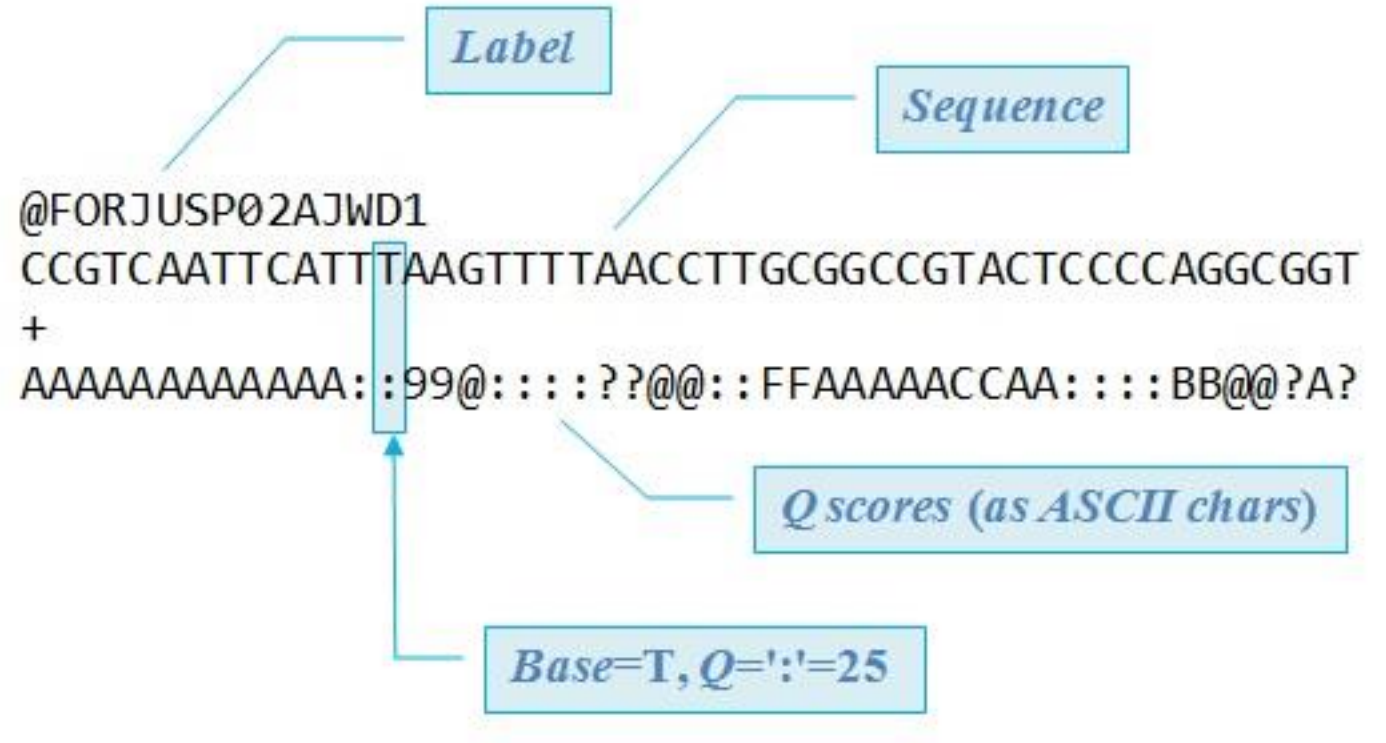
What do we sequence

- Human genome-want to look for variation
 - Cancer genome often looks very different from reference
- Human exome- looking for variation in coding genes
- A new organism's genome
- A new organisms transcriptome
 - Much smaller than genome
 - Produces sequence of mRNA only
- Sequence from functional assays
 - Transcript quantification-RNAseq
 - Chromatin (histones, transcription factors, etc.) capture

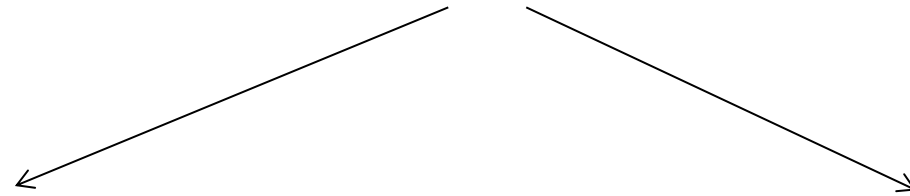
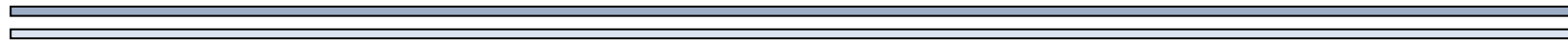


Output:Fastq

- Sequence followed by quality line
- Most platforms can sequence fragments at both ends – two matches fastq files
- Paired sequence information is crucial in assembly task
- Mate pairs—should be in the correct orientation and approximately the right distance apart

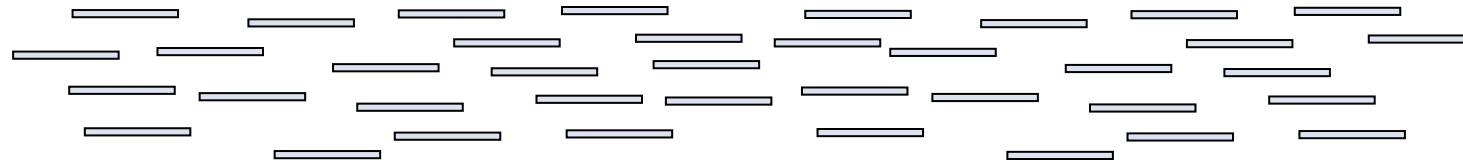


Using a reference genome: Alignment of reads to a reference

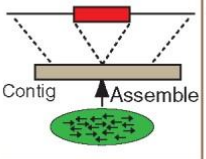
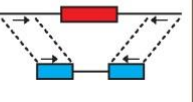

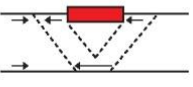
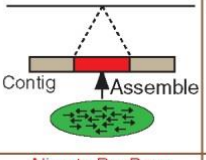
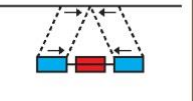
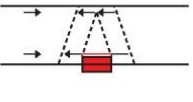
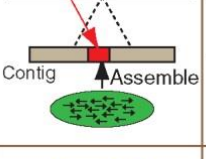
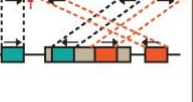
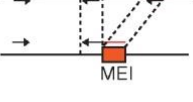
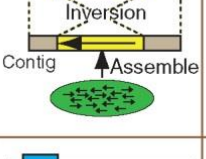
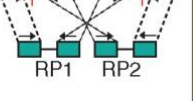
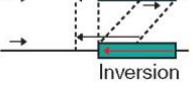
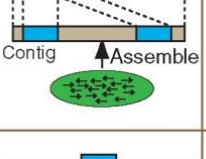
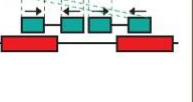

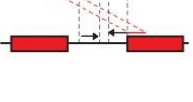
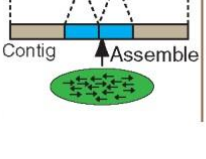
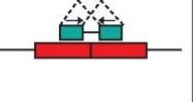

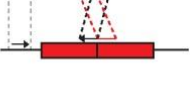


..ACTGGGTCATCGTACGATCGATCGATCGATCGATCGGCTAGCTAGCTA.. Reference

..ACTGGGTCATCGTACGATCGATAGATCGATCGCTAGCTAGCTA.. Sample

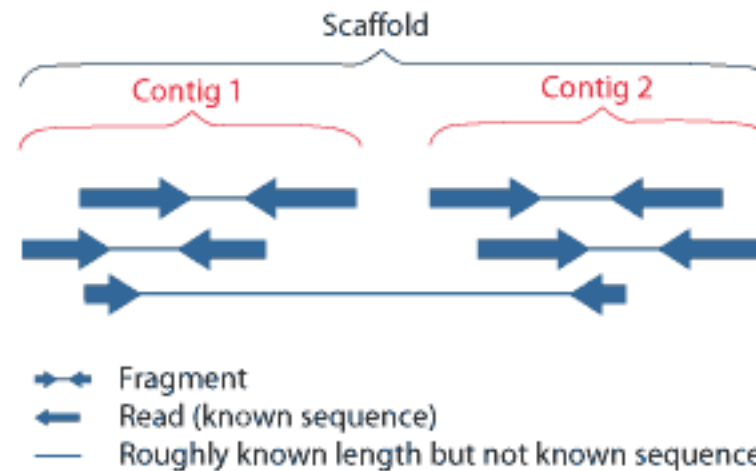


Structural variation

SV class	Assembly	Read pair	Read depth	Split read
Deletion				
Novel sequence insertion			Not applicable	
Mobile-element insertion			Not applicable	
Inversion			Not applicable	
Interspersed duplication				
Tandem duplication				

Sequencing a new genome

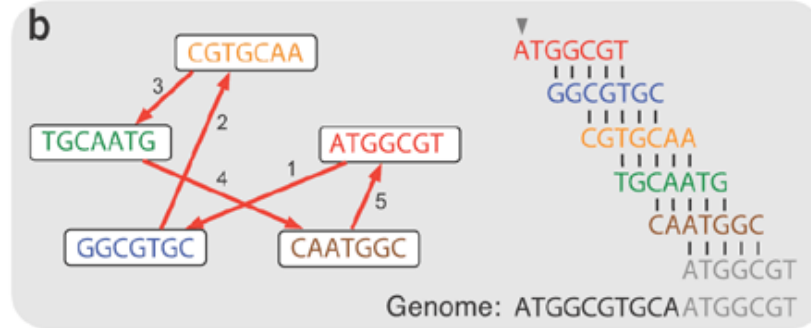
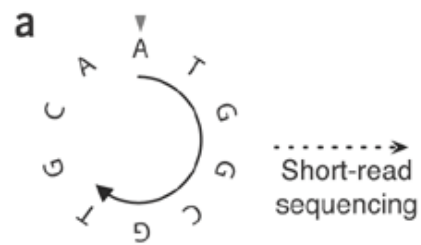
- We want a complete assembly with all nucleotides of each chromosome in the right order
- Human genome project—reads spanning 800bp are put together into 3 billion pairs
- Some definitions
 - Contig: a continuous piece of DNA sequence where base pair identity is known with high confidence
 - Scaffold: a series of contigs assembled in the correct order but possibly with gaps



Assembly strategies

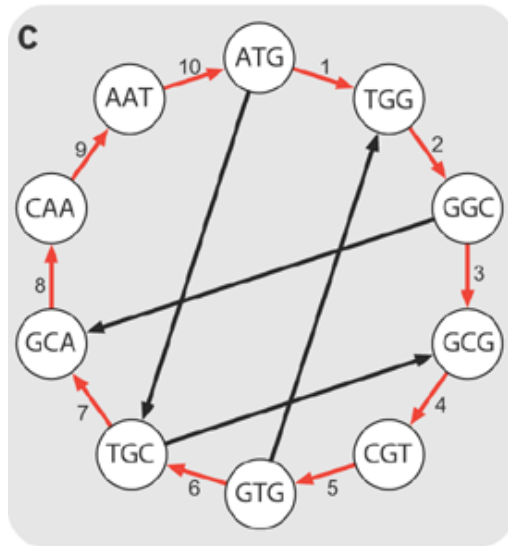
- Overlap layout consensus
 - Greedy
 - Merge largest overlap
 - Proceed until no overlaps remain
 - Uses local information only
 - Graph based
 - Create a graph representing sequence overlaps
 - Reduce/prune
 - Find consensus contigs
 - Can use global information such as mate-pair distance
 - Graph representation
 - Overlap
 - DeBruijn graph

Two graph representations

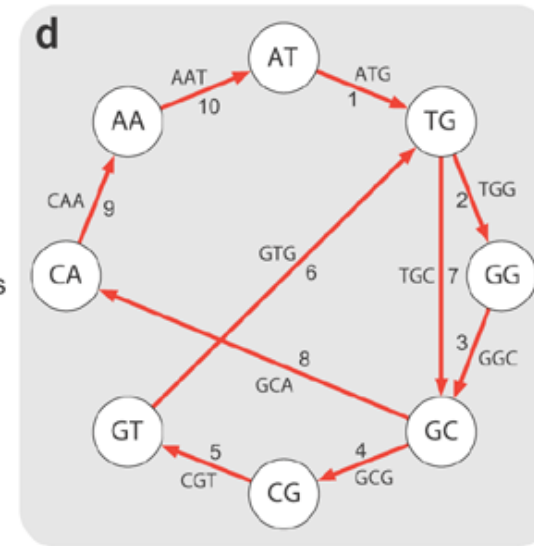


Vertices are k -mers
Edges are pairwise alignments

Vertices are $(k-1)$ -mers
Edges are k -mers



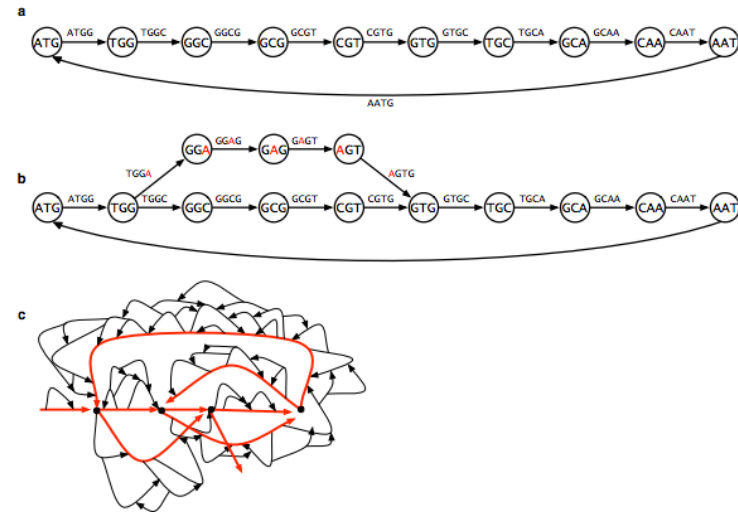
Hamiltonian cycle
Visit each vertex once
(harder to solve)



Eulerian cycle
Visit each edge once
(easier to solve)

Sequencing errors

- Reads contain errors
- We allow for imperfect overlap in overlap graphs
- In DeBruijn graphs errors are bulges specifying alternative paths
- Bulges are pruned based on read number, error model, and other heuristics



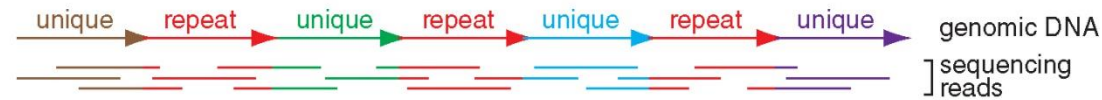
- **Overlap graph**

- Works well with large reads: sanger sequenced genomes
- Human genome project
- Computationally expensive
- Doesn't handle repeats well

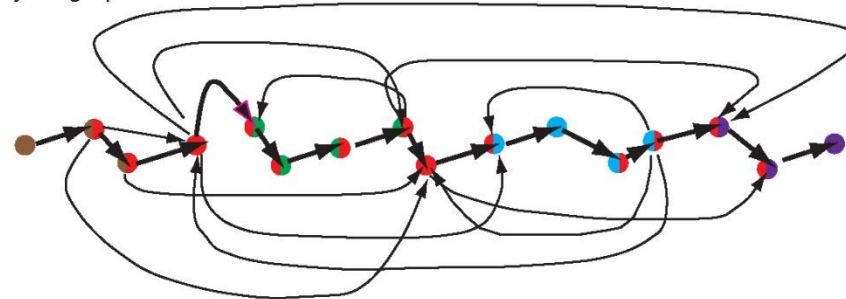
- **DeBruijn graph**

- Preferred for shorter reads
- Path finding is more efficient
- Compactly handles repeats

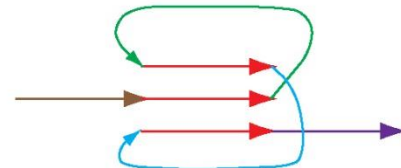
(a) DNA sequence with a triple repeat



(b) layout graph



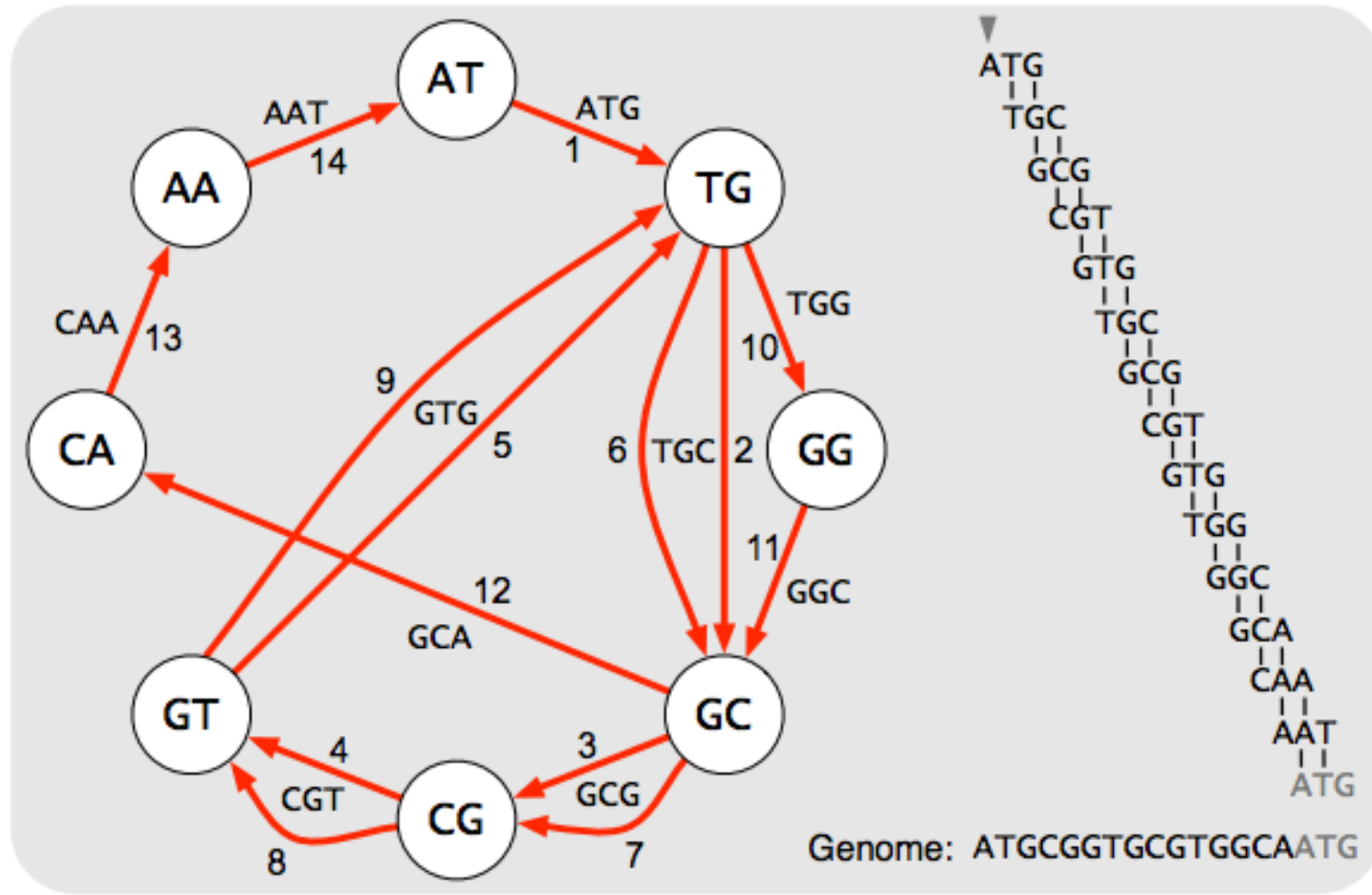
(c) Construction of de Bruijn graph by gluing repeats



(d) de Bruijn graph



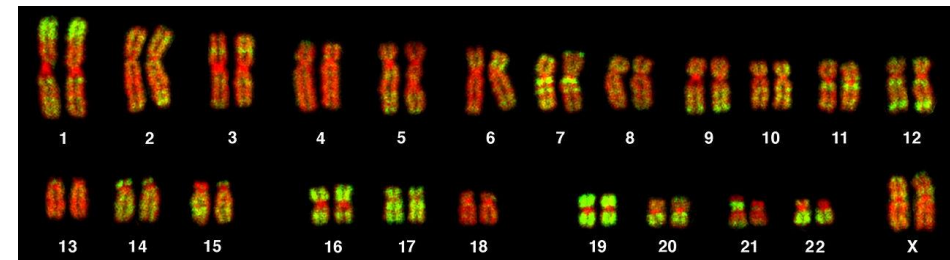
DeBruijn graph with repeats



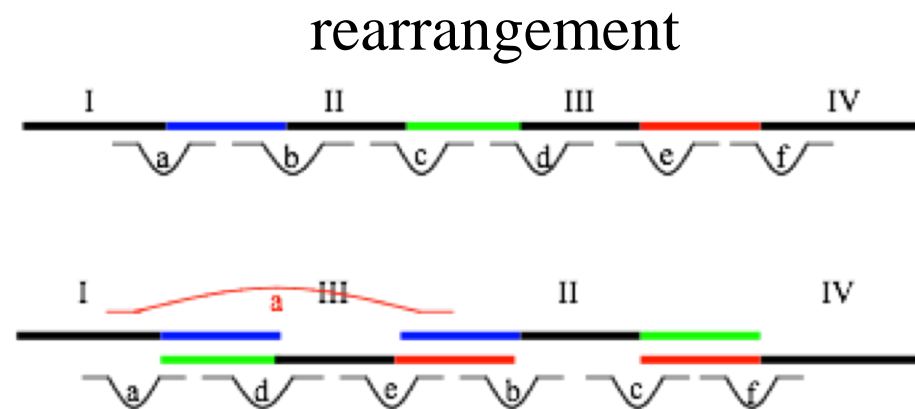
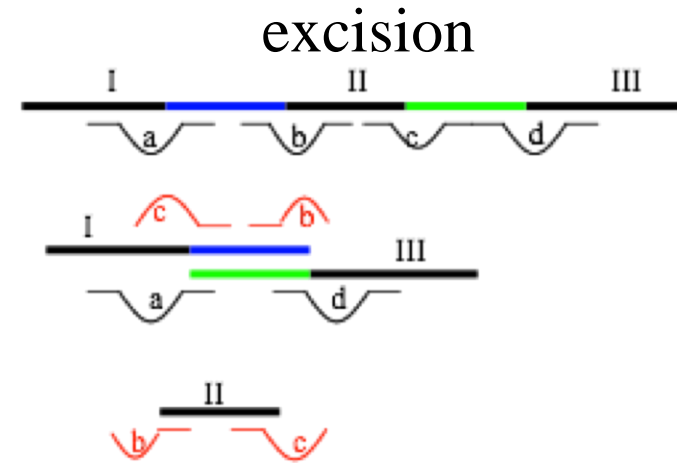
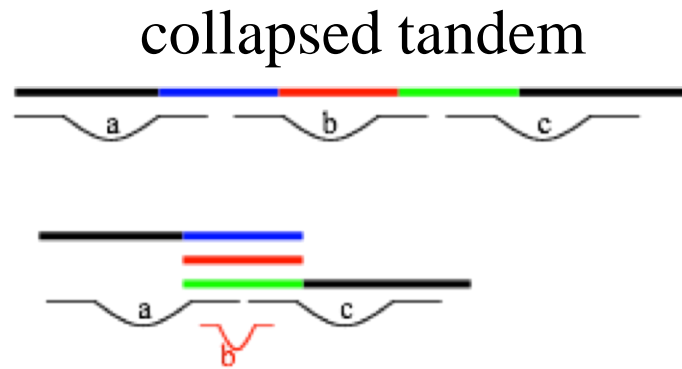
Repeats in Human Genome

- Repetitive DNA makes up a very large part of big eukaryotic genomes.
- Characterized by size and abundance.
- Many of these elements are remnants of virus-like sequences that once hopped around our genome.
- All but the SINEs contain functional sequence encoding genes such as transposase that are responsible for this hopping behavior. Most are inactive.
- The most abundant SINES family are the *Alu* repeats, with over 1 million copies comprising 10% of the genome.
- Also, families of recently duplicated genes

Element	Length (kb)	Human number	Genome Fraction
Retroviruses/ Retroposons	1-11	450,000	8%
LINES (long interspersed elements)	6-8	850,000	17%
SINES (short interspersed elements)	~0.3	1,500,000	15%
Transposons	2-3	300,000	3%



Mis-assembled repeats



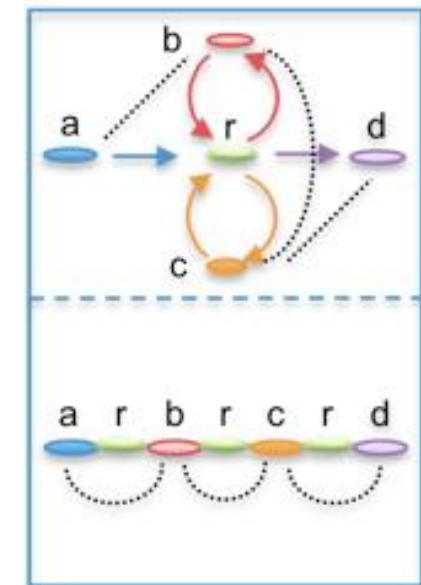
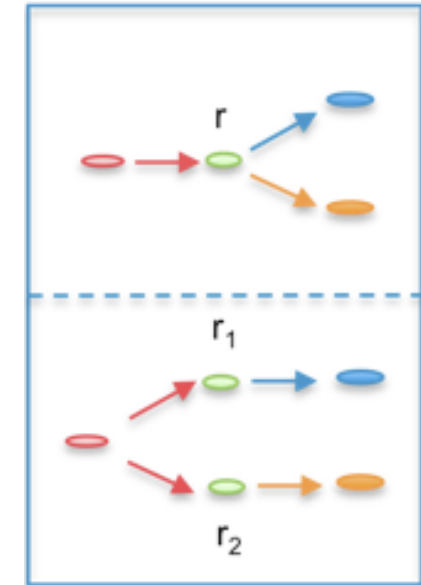
Human Genome assembly is not complete

- *chr1_random chr2_random chr3_random chr4_random chr5_random
chr6_random chr7_random chr8_random chr9_random chr10_random
chr11_random chr13_random chr15_random chr16_random chr17_random
chr18_random chr19_random chr21_random chr22_random chrX_random*

Handling repeats

Repeat detection

- **pre-assembly:** find fragments that belong to repeats
 - statistically (most existing assemblers)
 - repeat database (*RepeatMasker*)
- **during assembly:** detect graph structures indicative of repeats and resolve them
 - Use mate-pair information
- **post-assembly:** find repetitive regions and potential mis-assemblies.
 - *RepeatMasker*
 - "unhappy" mate-pairs (too close, too far, mis-oriented)



Statistical repeat detection

- Significant deviations from average coverage flagged as repeats.
 - frequent k-mers are ignored
 - “arrival” rate of reads in contigs compared with theoretical value

(e.g., 800 bp reads & 8x coverage - reads "arrive" every 100 bp)

Problem 1: assumption of uniform distribution of fragments - leads to false positives

non-random libraries

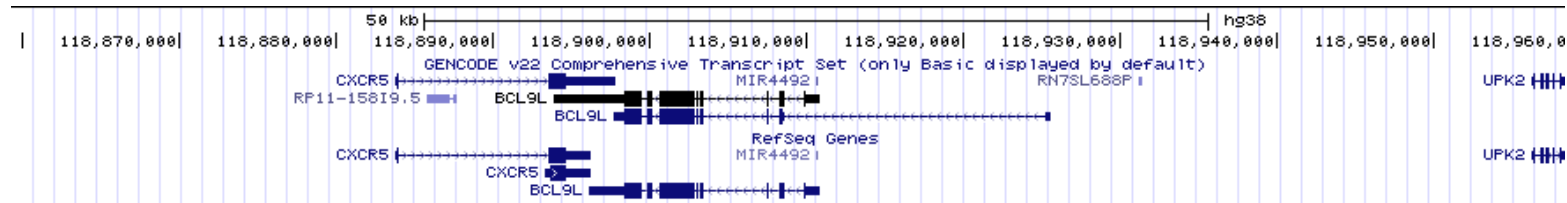
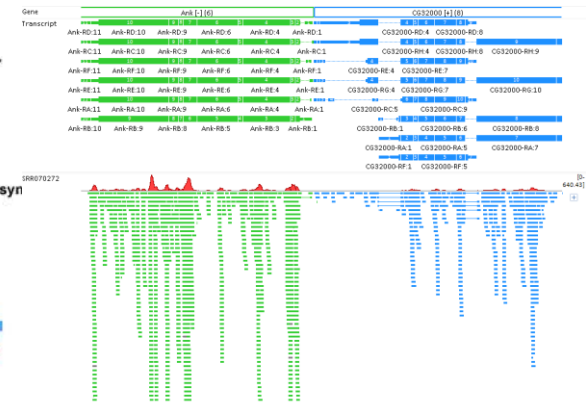
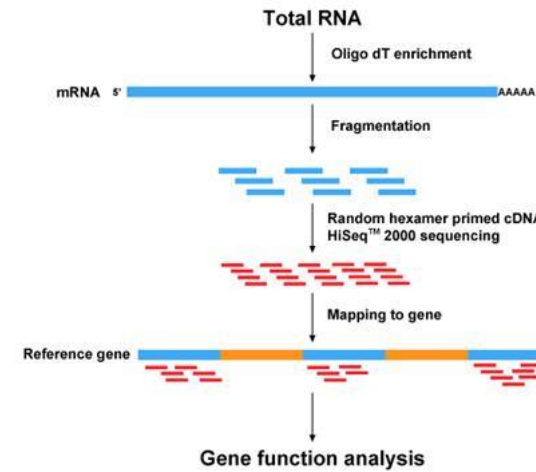
Problem 2: repeats with low copy number are missed - leads to false negatives

Most widely used genome sequence is filtered for repeats

- hg38.fa.gz - "Soft-masked" assembly sequence in one file. Repeats from RepeatMasker and Tandem Repeats Finder (with period of 12 or less) are shown in lower case; non-repeating sequence is shown in upper case.
- hg38.fa.masked.gz - "Hard-masked" assembly sequence in one file. Repeats are masked by capital Ns; non-repeating sequence is shown in upper case.
- RepeatMasker screens DNA sequences for interspersed repeats and low complexity DNA sequences.
- Tandem Repeat Finder looks for short tandem repeats – ATGATGATGATG

Aligning to a reference – not necessarily so simple

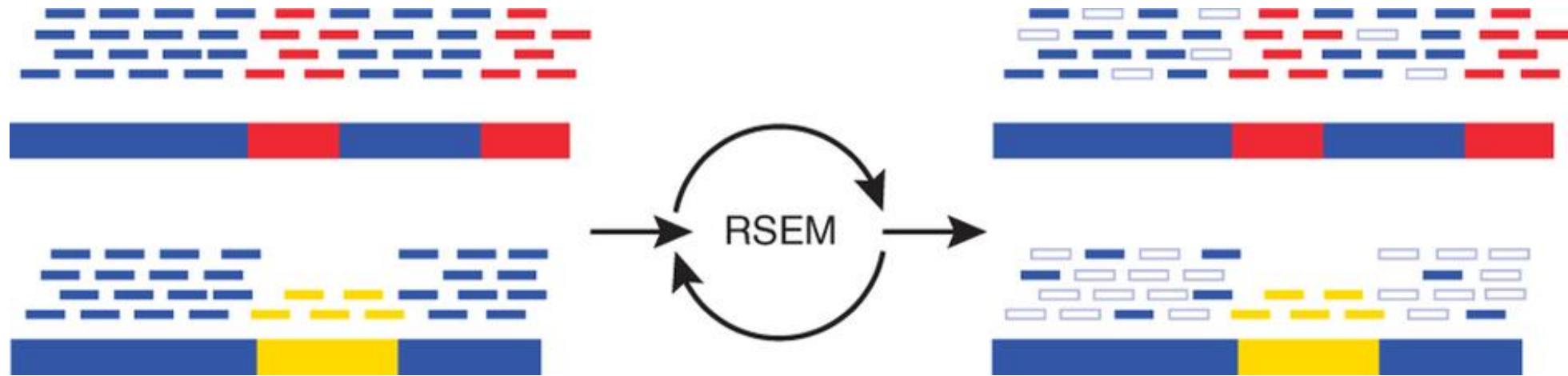
- RNAseq quantification by sequencing
- Genes producing more transcripts will have more reads
- We need to map each read to the right gene
- Problems
 - Genes overlap-often in the UTR



- Homologous genes
- Sequencing errors in high abundance genes can map to a completely different location

One solution—RSEM

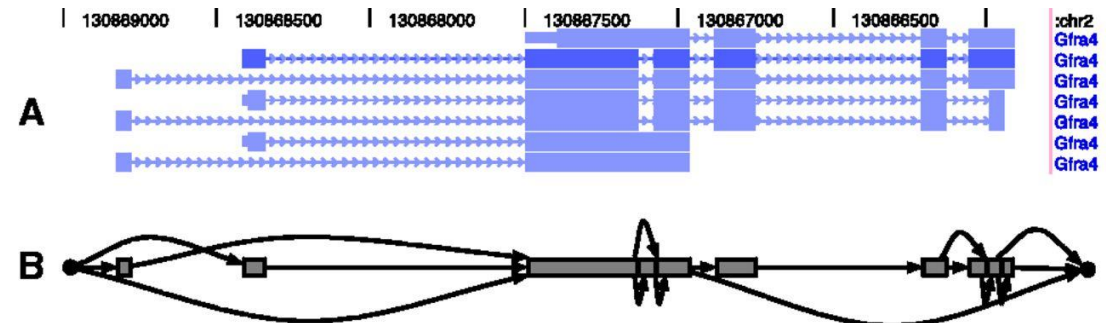
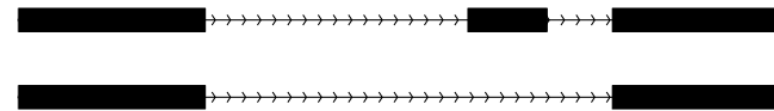
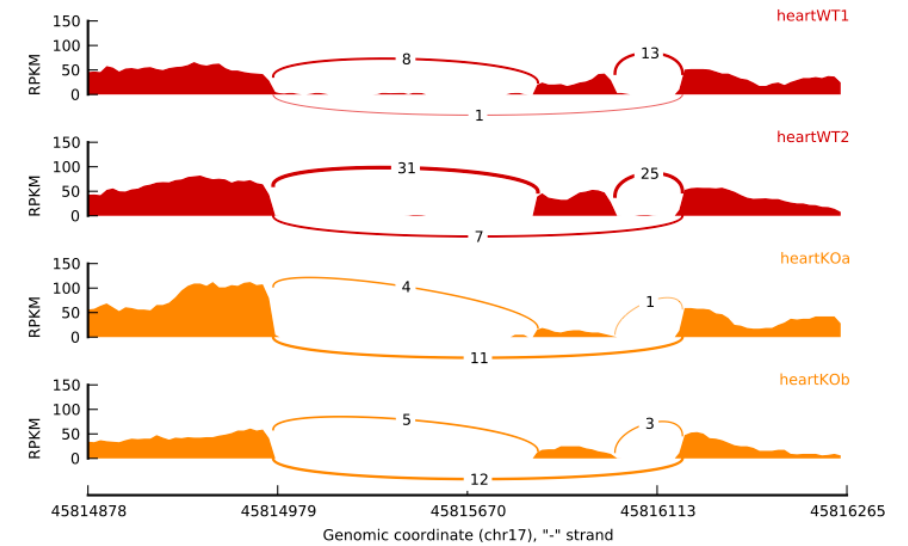
- Assign ambiguous reads in proportion with unambiguous reads



RNAseq– identifying complete transcripts

- Reads contained in an exon don't tell us how the exons are connected
- Junction spanning reads tell us local connectivity
 - Such reads do not align continuously in the genome and present an alignment challenge
- With short reads only local connectivity is known –still don't know the relative proportion of full length transcripts

chr17:45816186:45816265:-@chr17:45815912:45815950:-@chr17:458



Assembling transcripts

- **splice graph**, nodes represent exons or parts of exons, and paths through the graph represent possible splice variants supported by junction reads
- StringTie
 - Find maximum flow through “heaviest” path
 - Assemble transcript
 - Compute expression
 - Remove reads that contributed to the total expression
 - Repeat

