Bioinformatics for High-Throughput Sequencing

Misha Kapushesky

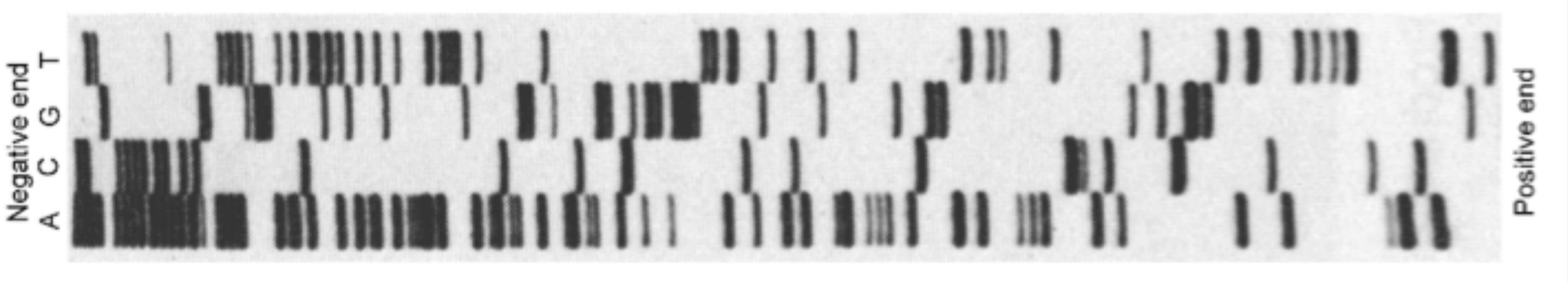
St. Petersburg Russia 2010

EMBL-EBI

Slides: Nicolas Delhomme, Simon Anders, EMBL-EBI

High-throughput Sequencing

- Key differences from Sanger sequencing
  - Library not constructed by cloning
  - Fragments sequenced in parallel in a flow cell
  - Observed by a microscope + CCD camera



Roche 454

- 2005 (first to market)
- Pyrosequencing
- Read length: 250bp
- Paired read separation: 3kb
- 300Mb per day
- $60 per Mb
- Error rate: ~ 5% per bp
- Dominant error: indel, especially in homopolymers

Illumina/Solexa

- Second to market
- Bridge PCR
- Sequencing by synthesis
- Read length: 32…40bp, newest models up to 100bp
- Paired read separation: 200bp
- 400Mb per day (and increasing)
- $2 per Mb
- Error rate: 1% per bp, sometimes as good as 0.1%
- Dominant error: substitutions

ABI SOLiD

- Third to market (2007)
- Emulsion PCR, ligase-based sequencing
- Read length 50bp
- Paired read separation 3kb
- 600Mb per day
- Reads in colour space
- $1 per Mb
- Very low error rate <0.1% per bp (Sanger error 0.001%)
- Dominant error: substitutions

Helicos

- Recent
- No amplification
- Single-molecule polymerase sequencing
- Read length: 25..45bp
- 1200Mb per day
- $1 per Mb
- Error <1% (manufacturer)

Polonator

- Recent
- Emulsion PCR, ligase-based sequencing
- Very short read length: 13bp
- Low-cost instrument ($150K)
- <$1 per Mb

Uses for HTS

- De-novo sequencing, assembly of small genomes
- Transcriptome analysis (RNA-seq)
- Resequencing to identify genetic polymorphisms
    - SNPs, CNVs
- ChIP-seq
- DNA methylation studies
- Metagenomics
- …

Multiplexing

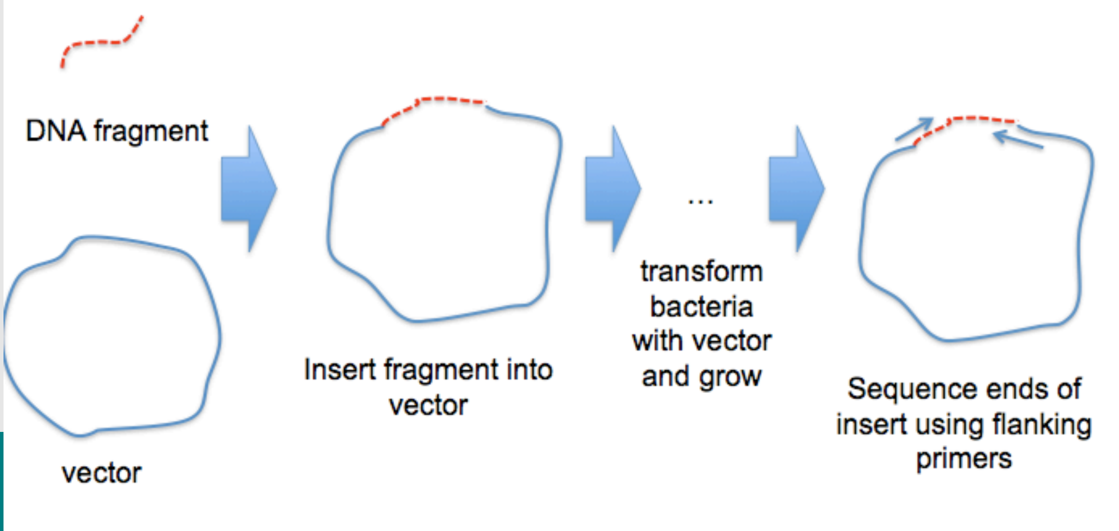
- Solexa: 6-12 mln 36bp reads per lane
- One lane for one sample – wasteful
- Multiplexing: incorporate tags between sequencing primer and sample fragments to distinguish several samples in the same lane

Targeted Sequencing

- Instead of whole genome, sequence only regions of interest but **deep**
- Microarrays can help to select fragments of interest

Paired end sequencing

- The two ends of the fragments get different adapters
- Hence, one can sequence from one end with one primer, then repeat to get the other end with the other primer.
- This yields “pairs” of reads, separated by a known distance (200bp for Illumina).
- For large distances, “circularisation” might be needed and generates “mate pairs”.


DNA fragment
vector
Insert fragment into vector
...
transform bacteria with vector and grow
Sequence ends of insert using flanking primers

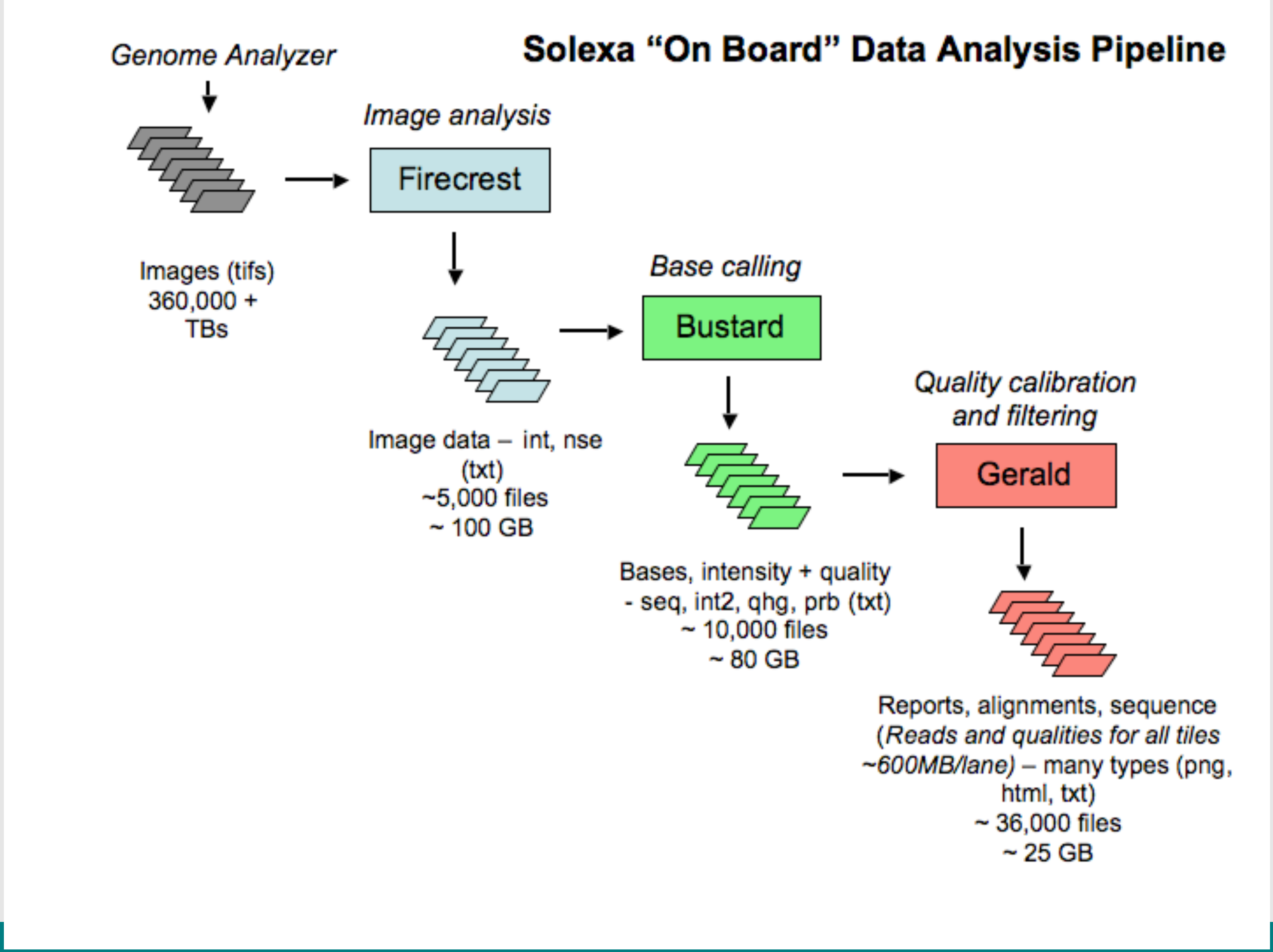
Paired end read uses

- Useful to find:
  - Micro indels
  - Copy-number variations
  - Assembly tasks
  - Splice variants

Bioinformatics Problems

- Assembly
- Alignment
- Statistics
- Visualization

Solexa Pipeline



Solexa "On Board" Data Analysis Pipeline

Genome Analyzer

Images (tifs)
360,000 +
TBs

Image analysis

Firecrest

Image data – int, nse
(txt)
~5,000 files
~ 100 GB

Base calling

Bustard

Bases, intensity + quality
- seq, int2, qhg, prb (txt)
~ 10,000 files
~ 80 GB

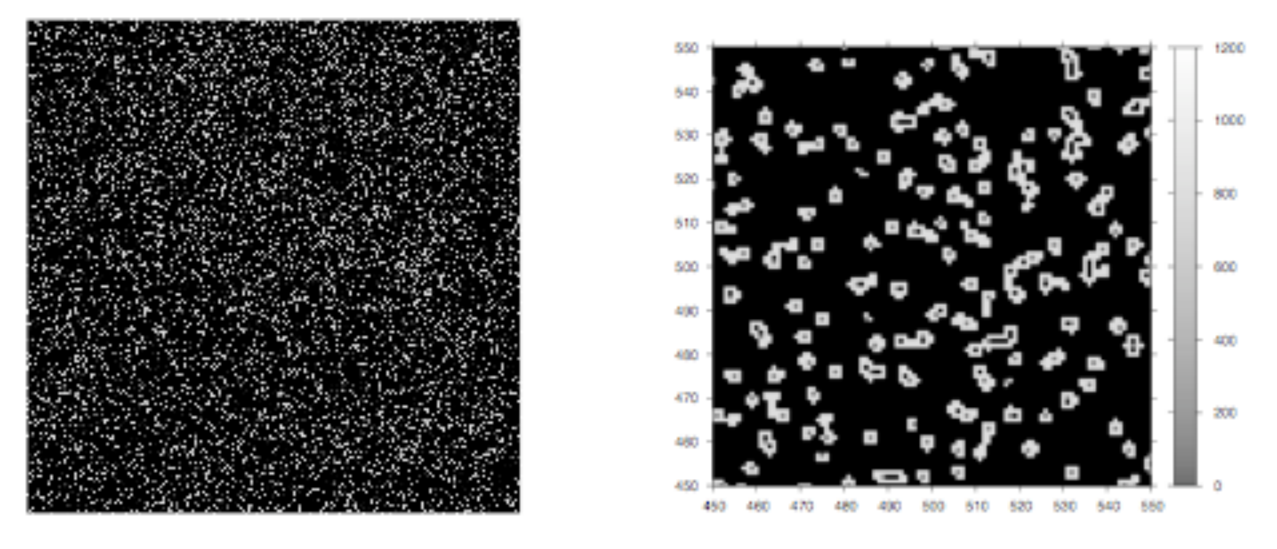
Quality calibration
and filtering

Gerald

Reports, alignments, sequence
(*Reads and qualities for all tiles*
~600MB/lane) – many types (png,
html, txt)
~ 36,000 files
~ 25 GB

Firecrest Output

- Tab-separated text files, one row per cluster
- Lane & tile index
- X,Y coordinates of cluster
- For each cycle, group of four numbers – fluorescence intensities for A, G, C, T



Bustard output

- Two tab-separated text files, one row per cluster
- “seq.txt”
  - Lane and tile index, x and y coordinates
  - Called sequence as string of A, G, C, T
- “prb.txt”
  - Phred-like scores [-40,40]
  - One value per called base

FASTQ format

```
@HWI-EAS225:3:1:2:854#0/1
GGGGGGAAGTCGGCAAAATAGATCCGTAACTTCG
GG +HWI-EAS225:3:1:2:854#0/1
a`abbbbabaabbababb^`[aaa`_N]b^ab^``a  @HWI-
EAS225:3:1:2:1595#0/1
GGGAAGATCTCAAAAACAGAAGTAAAACATCGAAC
G +HWI-EAS225:3:1:2:1595#0/1
a`abbbababbbabbbbbabb`aaababab\aa_`
```

FASTQ Format

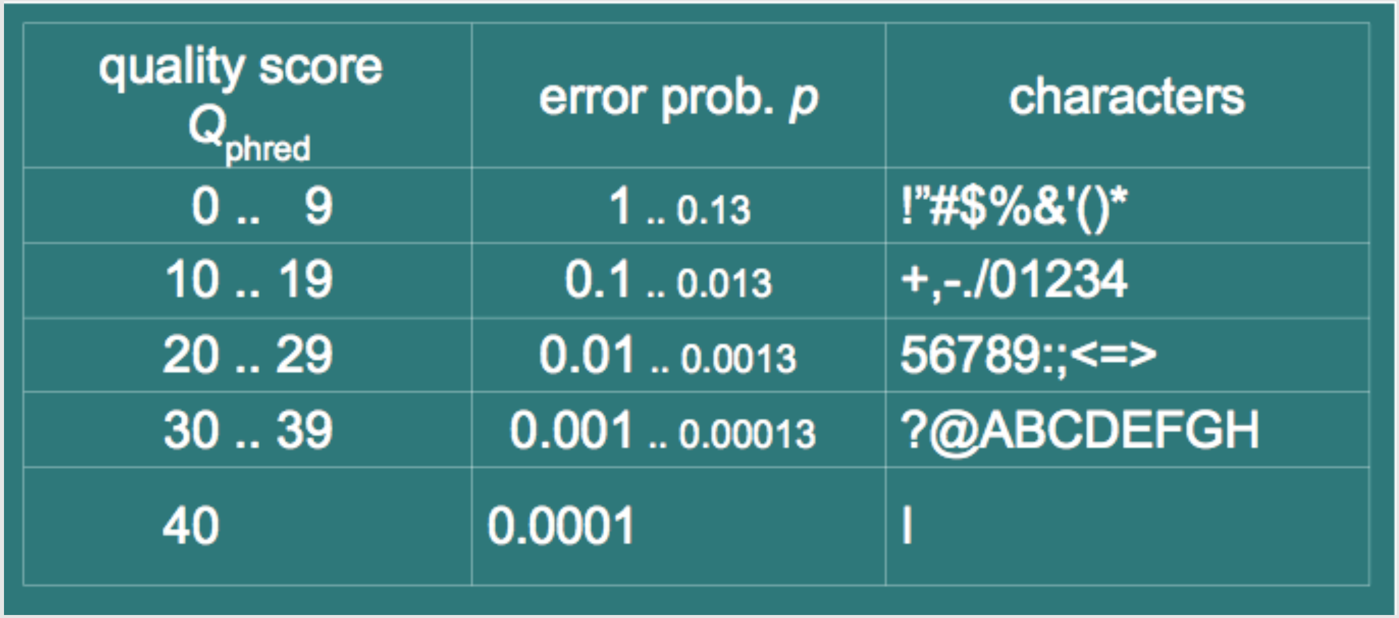
- Each read is represented by four lines
- @ + read ID
- Sequence
- “+”, optionally followed by repeated read ID
- Quality string
  - Same length as sequence
  - Each character coding for base-call quality per 1 base

Base call quality strings

- If *p* is the probability that the base call is wrong, the (standard Sanger) Phred score is:

  $$Q_{\text{Phred}} = -10 \log_{10} p$$

  Score written with character – ascii code Q + 33.
- Solexa slightly different, but changing



quality score Q_{phred}	error prob. *p*	characters
0 .. 9	1 .. 0.13	!”#$%&'()*
10 .. 19	0.1 .. 0.013	+,-./01234
20 .. 29	0.01 .. 0.0013	56789:;<=>
30 .. 39	0.001 .. 0.00013	?@ABCDEFGH
40	0.0001	I

Short Read Alignment

- Read mapping – position within a reference sequence

Challenges of mapping short reads

- Speed: if the genome is large and we have billions of reads?

- Memory: suffix array approach requires 12GB for human genome indexing reads in-memory

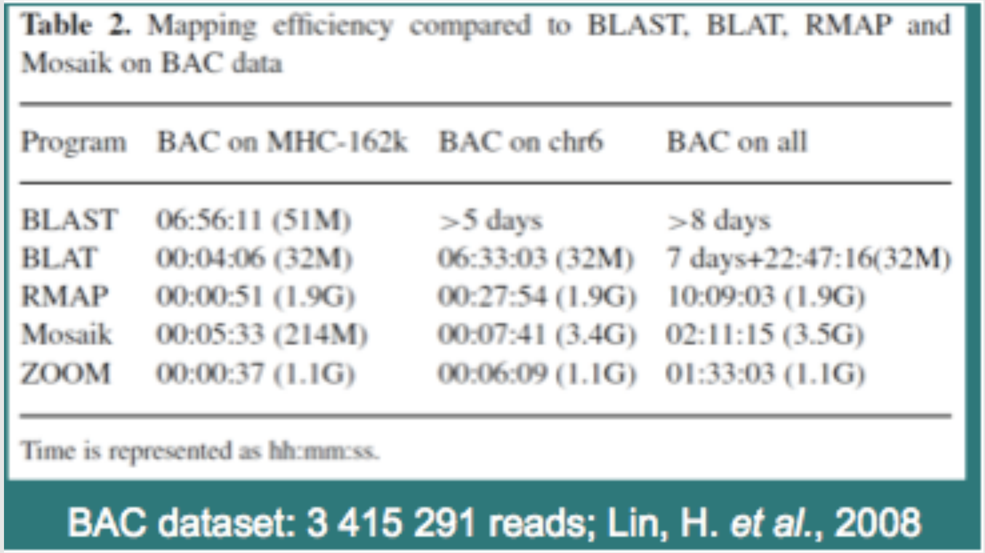


Table 2. Mapping efficiency compared to BLAST, BLAT, RMAP and Mosaik on BAC data

Program	BAC on MHC-162k	BAC on chr6	BAC on all
BLAST	06:56:11 (51M)	>5 days	>8 days
BLAT	00:04:06 (32M)	06:33:03 (32M)	7 days+22:47:16(32M)
RMAP	00:00:51 (1.9G)	00:27:54 (1.9G)	10:09:03 (1.9G)
Mosaik	00:05:33 (214M)	00:07:41 (3.4G)	02:11:15 (3.5G)
ZOOM	00:00:37 (1.1G)	00:06:09 (1.1G)	01:33:03 (1.1G)

Time is represented as hh:mm:ss.

BAC dataset: 3 415 291 reads; Lin, H. *et al.*, 2008

Additional Challenges

- Read errors
  - Dominant cause for mismatches
  - Detection of substitutions?
  - Importance of base-call quality

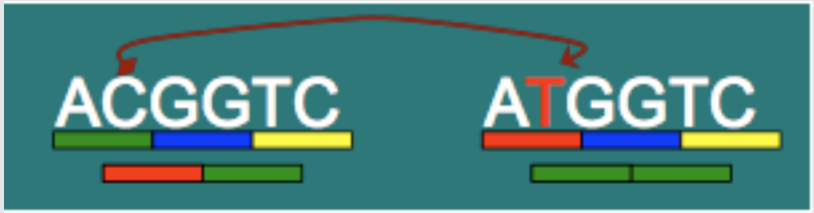
- Unknown reference genome
  - De-novo assembly

- Repetitive regions/accuracy
  - ~20% of human genome is repetitive for 32bp reads
  - Use paired-end information

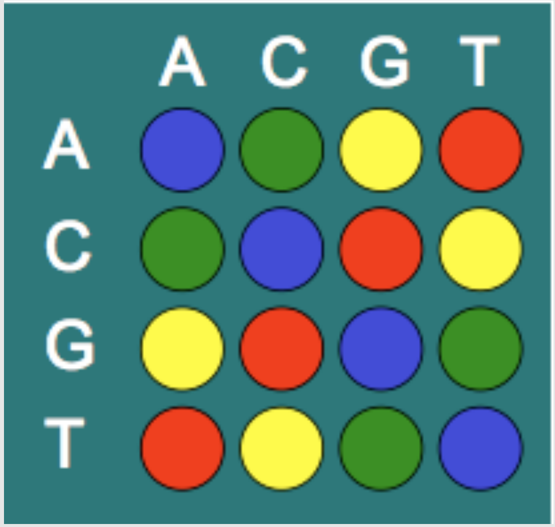
Technical Challenges

- 454 – longer reads may require different tools

- SOLiD
  - Use colour space
  - Sequencing error vs. polymorphism



  - Deletion shifts colors
  - Not easy to convert to bases, needs aligning to color space reference



	A	C	G	T
A	🔵	🟢	🟡	🔴
C	🟢	🔵	🔴	🟡
G	🟡	🔴	🔵	🟢
T	🔴	🟡	🟢	🔵

Alignment Tools

- Many tools have been published
- Eland
- MAQ
- Bowtie
- BWA
- SOAP2
- …

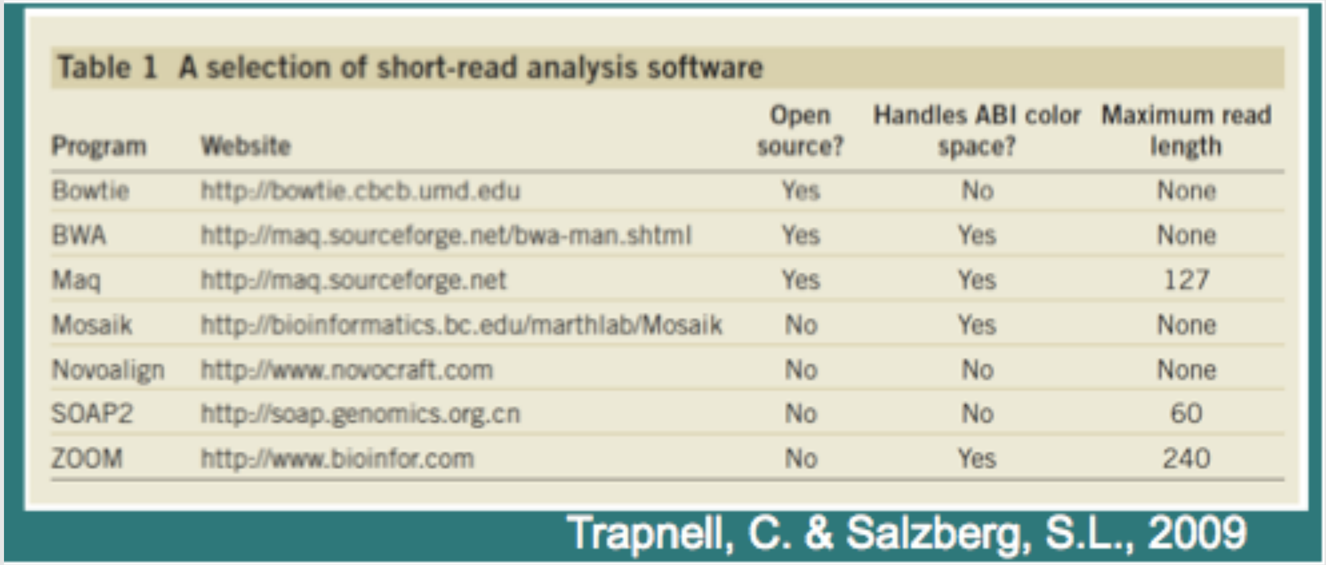


Table 1 A selection of short-read analysis software

Program	Website	Open source?	Handles ABI color space?	Maximum read length
Bowtie	http://bowtie.cbcb.umd.edu	Yes	No	None
BWA	http://maq.sourceforge.net/bwa-man.shtml	Yes	Yes	None
Maq	http://maq.sourceforge.net	Yes	Yes	127
Mosaik	http://bioinformatics.bc.edu/marthlab/Mosaik	No	Yes	None
Novoalign	http://www.novocraft.com	No	No	None
SOAP2	http://soap.genomics.org.cn	No	No	60
ZOOM	http://www.bioinfor.com	No	Yes	240

Trapnell, C. & Salzberg, S.L., 2009

Short read aligners - differences

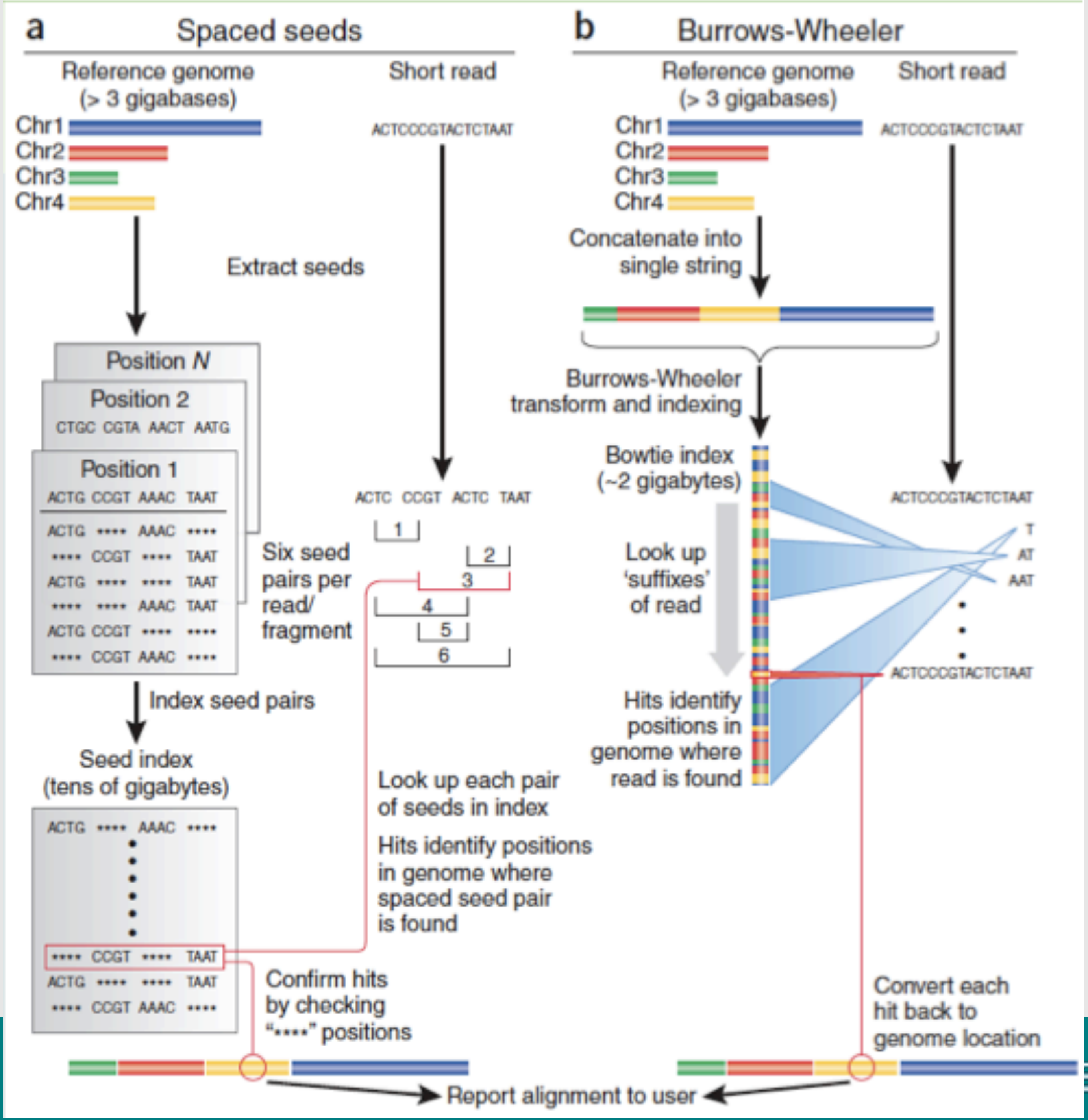
- Speed
- Use on clusters
- Memory requirements
- Accuracy
  - Good match always found?
  - Allowed mismatches
- Downstream analysis tools
  - SNP/indel callers for output format
  - R package to read the output?

Other differences

- Alignment tools also differs in whether they can
  - make use of base-call quality scores
  - estimate alignment quality
  - work with paired-end data
  - report multiple matches
  - work with longer than normal reads
  - match in colour space (for SOLiD systems)
  - deal with splice junctions

Alignment Algorithm Approaches

- Hashing
  (seed-and-extend paradigm, k-mers + Smith-Waterman)
  - The entire genome
    - Straightforward, easy multi-threading, but large memory
  - The read sequences
    - Flexible memory footprint, harder to multi-thread
- Alignment by merge sorting
  - Pros: flexible memory
  - Cons: not easy to adapt for paired-end reads
- Indexing by Burrows-Wheeler Transform
  - Pros: fast and relatively small memory
  - Cons: not easily applicable to longer reads


a
Spaced seeds
Reference genome
(> 3 gigabases)
Short read
Chr1
Chr2
Chr3
Chr4
ACTCCCGTACTCTAAT
Extract seeds
Position N
Position 2
CTGC CGTA AACT AATG
Position 1
ACTG CCGT AAAC TAAT
ACTG **** AAAC ****
**** CCGT **** TAAT
ACTG **** **** TAAT
**** **** AAAC TAAT
ACTG CCGT **** ****
**** CCGT AAAC ****
ACTC CCGT ACTC TAAT
1
2
3
4
5
6
Six seed pairs per read/ fragment
Index seed pairs
Seed index
(tens of gigabytes)
ACTG **** AAAC ****
**** CCGT **** TAAT
ACTG **** **** TAAT
**** CCGT AAAC ****
Look up each pair of seeds in index
Hits identify positions in genome where spaced seed pair is found
Confirm hits by checking "****" positions
b
Burrows-Wheeler
Reference genome
(> 3 gigabases)
Short read
Chr1
Chr2
Chr3
Chr4
ACTCCCGTACTCTAAT
Concatenate into single string
Burrows-Wheeler transform and indexing
Bowtie index
(~2 gigabytes)
ACTCCCGTACTCTAAT
T
AT
AAT
ACTCCCGTACTCTAAT
Look up 'suffixes' of read
Hits identify positions in genome where read is found
Convert each hit back to genome location
Report alignment to user

Burrows-Wheeler Transform

- BWT seems to be a winning idea
  - Very fast
  - Accurate
  - Bowtie, SOAP2, BWA – latest tools

Others

- ELAND
  - Part of Solexa Pipeline
  - Very fast, does not use quality scores
- MAQ (Li et al., Sanger Institute)
  - Widely used hashing-based approach
  - Quality scores used to estimate alignment score
  - Compatible with downstream analysis tools
  - Can deal with SOLiD colour space
  - To be replaced with BWA
- Bowtie (Langmead et al., Maryland U)
  - Burrows-Wheeler Transform based
  - Very fast, good accuracy
  - Downstream tools available

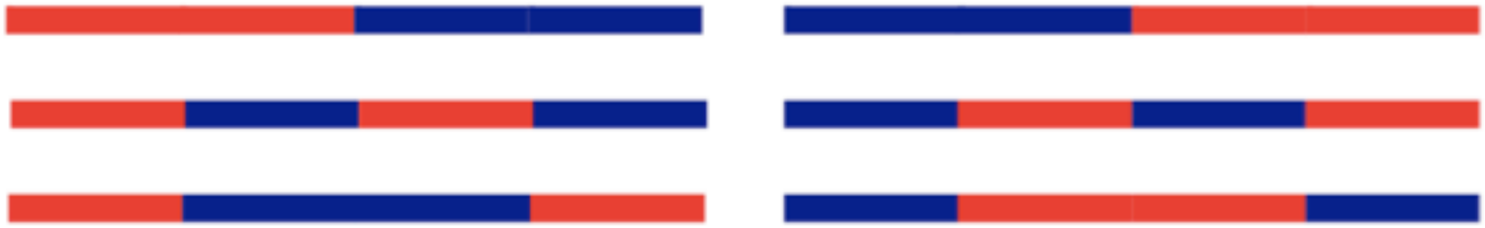
Hashed Read Alignment

- Naïve Algorithm
  - Make a hash table of the first 28mers of each read, so that for each 28mer, we can look up quickly which reads start with it.
  - Then, go through the genome, base for base. For each 28mer, look up in the hash table whether reads start with it, and if so, add a note of the current genome position to these reads.
- Problem: What if there are read errors in the first 28 base pairs?

MAQ: basic algorithm

- Index reads and scan the genome.
  - ✓ Avoid aligning too few reads
- 28bp seed; Eland-like indexing
  - ✓ Able to find more mismatches beyond the seed
- Guarantee to find 2-mismatch seed hits

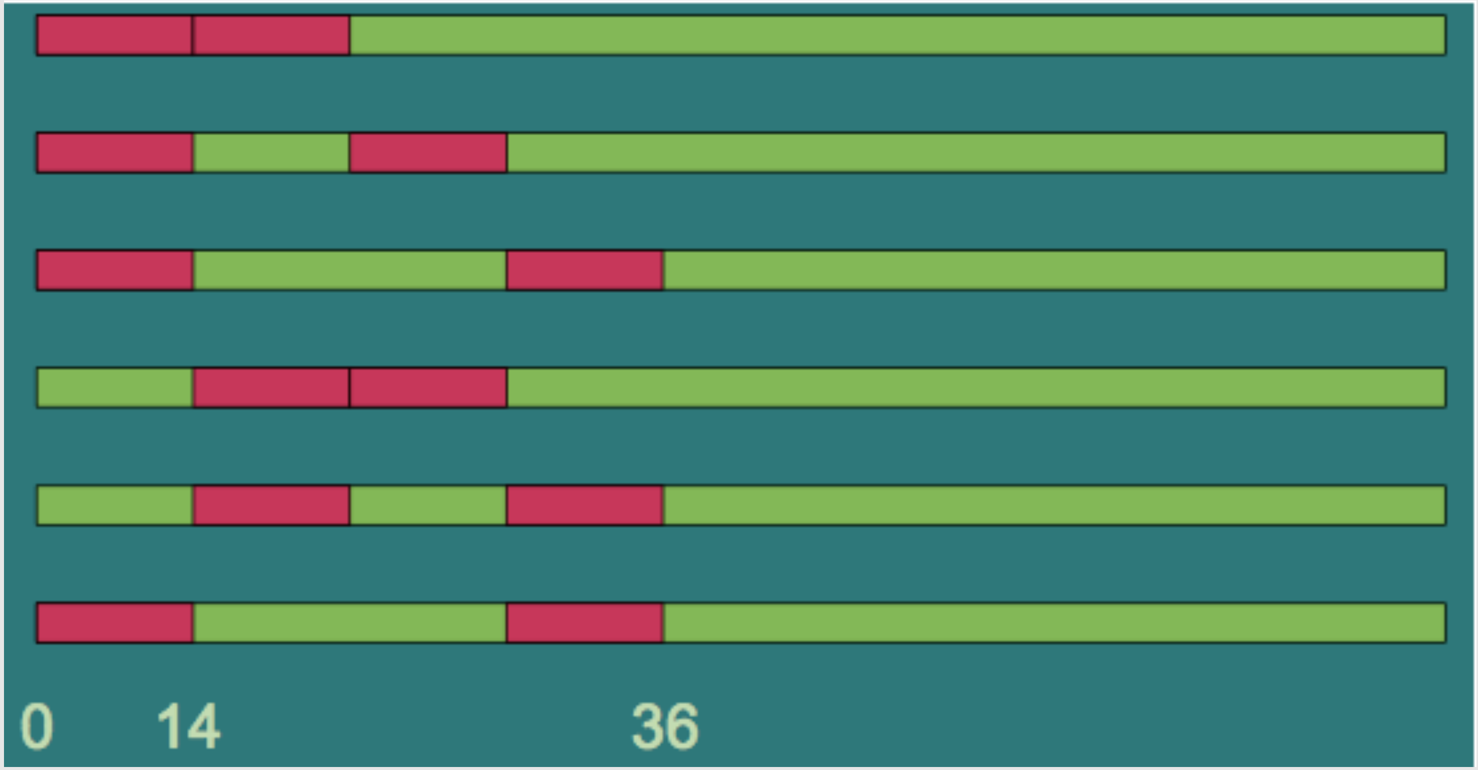
Seed templates:



Courtesy of Heng Li (Welcome Trust Sanger Institute)

Spaced seeds

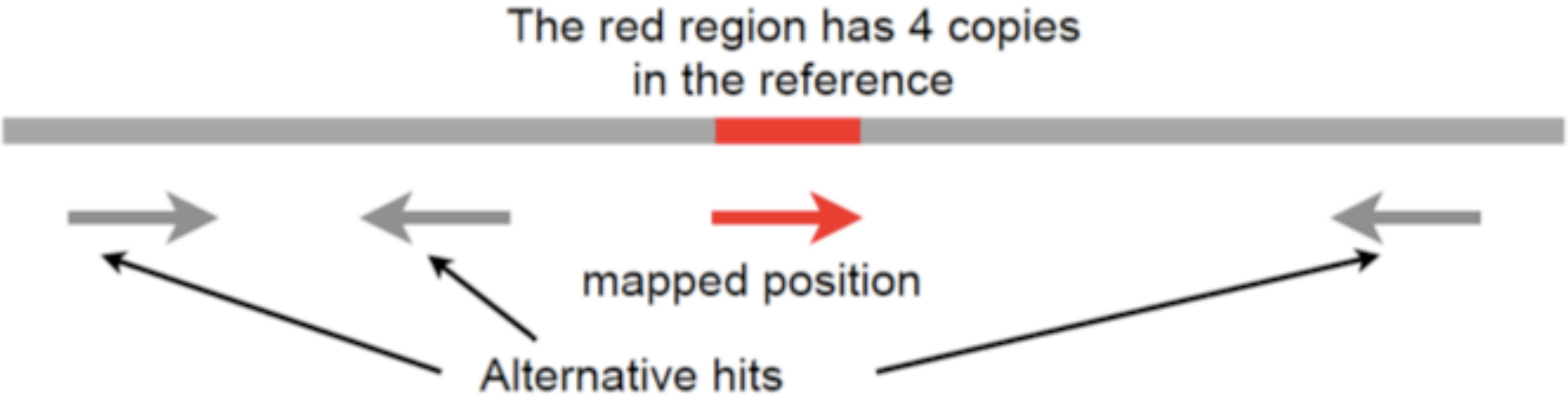
- Maq prepares six hash table, each indexing 28 of the first 36 bases of the reads, selected as follows:



Hence, Maq finds all alignments with at most 2 mismatches in the first 36 bases.

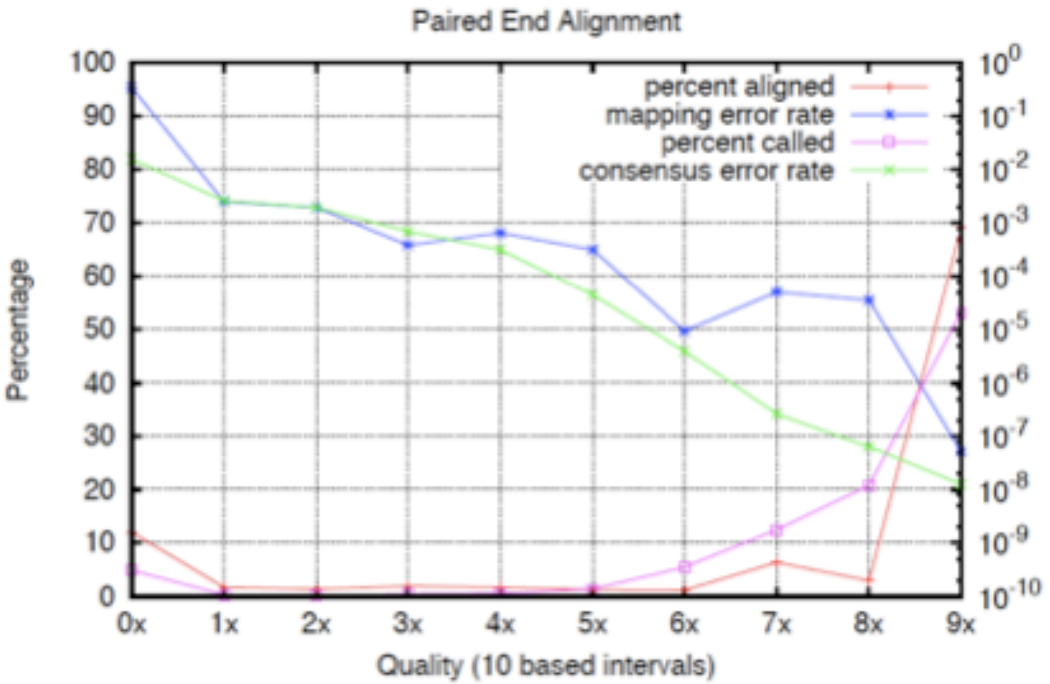
MAQ: random mapping

- Randomly place a read if it has multiple equally best hits
- Advantages:
  - ✓ tell if a read is mapped
  - ✓ tell if a region has reads mapped (avoid holes due to repeats)



MAQ: mapping quality

- Mapping quality is the phred-scaled probability of the alignment being wrong.
- Discriminate good mappings from bad ones, e.g.:
  - ✓ repetitive reads
  - ✓ top hit is perfect but there are 100 1-mismatch hits
  - ✓ top hit is perfect but the second best hit has one Q5 mismatch
- Proved to be effective for SV detections where wrong alignments dominate.

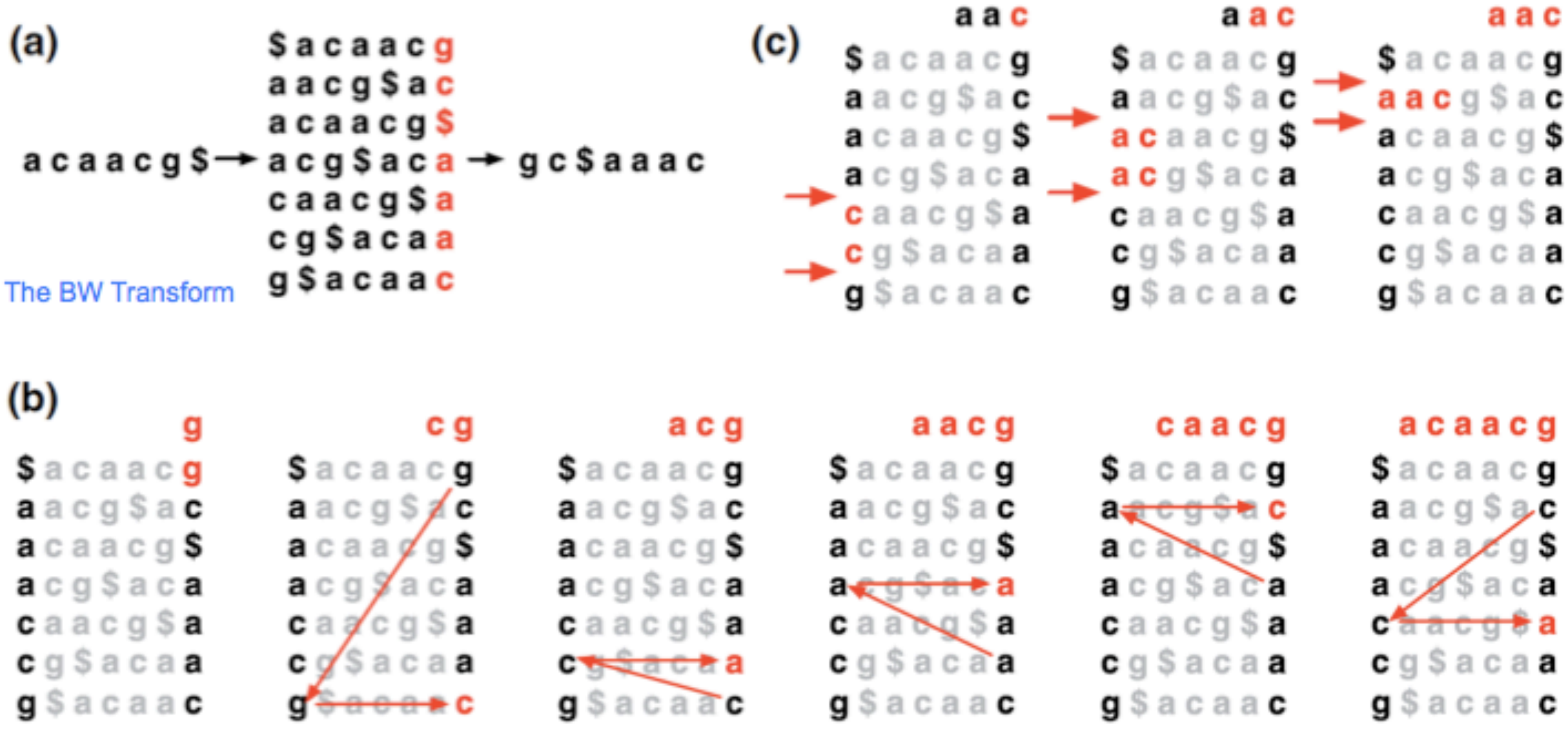


Courtesy of Heng Li (Welcome Trust Sanger Institute)

Burrows-Wheeler Transform

- Burrows & Wheeler (1994, DEC Research)
- Data compression algorithm (e.g. in bzip2)

Bowtie: Burrows-Wheeler Indexing



First Last Mapping

first occurrence of g ➔ second occurrence of c

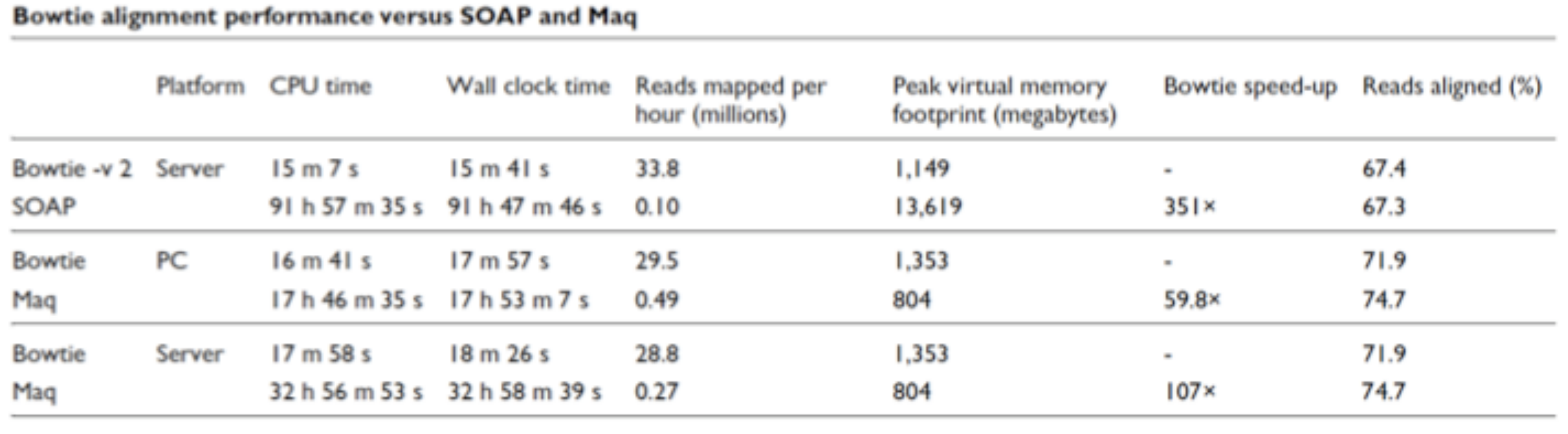
third occurrence of a ➔ first occurrence of a

second occurrence of c ➔ third occurrence of a

and so on…

Bowtie

- Reference genome suffix arrays are BW transformed and indexed
- Model organism genome indexes are available for download from the Bowtie webpage



Bowtie alignment performance versus SOAP and Maq

	Platform	CPU time	Wall clock time	Reads mapped per hour (millions)	Peak virtual memory footprint (megabytes)	Bowtie speed-up	Reads aligned (%)
Bowtie -v 2	Server	15 m 7 s	15 m 41 s	33.8	1,149	-	67.4
SOAP		91 h 57 m 35 s	91 h 47 m 46 s	0.10	13,619	351×	67.3
Bowtie	PC	16 m 41 s	17 m 57 s	29.5	1,353	-	71.9
Maq		17 h 46 m 35 s	17 h 53 m 7 s	0.49	804	59.8×	74.7
Bowtie	Server	17 m 58 s	18 m 26 s	28.8	1,353	-	71.9
Maq		32 h 56 m 53 s	32 h 58 m 39 s	0.27	804	107×	74.7

Bowtie

- Pros
  - small memory footprint (1.3GB for the human genome)
  - fast (8M reads aligned in 8 mins against the Drosophila genome)
  - paired-end able (gapped alignment)
- Cons
  - less accurate than MAQ
  - does not support SOLiD, Helicos
  - no gapped alignment

Other commonly used aligners

- SOAP and SOAP2 (Beijing Genomics Institute)
  - with downstream tools
  - SOAP2 uses BWT
- SSAHA, SSAHA2 (Sanger Institute)
  - one of the first short-read aligners
- Exonerate (EBI)
  - not really designed for short reads but still useful
- Biostrings (Bioconductor)
  - R package under development

References

- Langmead, B. et al., 2009. Ultrafast and memory- efficient alignment of short DNA sequences to the human genome. *Genome Biology, 10(3), R25.*

- Lin, H. et al., 2008. ZOOM! Zillions of oligos mapped. *Bioinformatics, 24(21), 2431-2437.*

- Trapnell, C. & Salzberg, S.L., 2009. How to map billions of short reads onto genomes. *Nat Biotech, 27(5), 455-457.*

HTS Assembly

- NGS offers the possibility to sequence anything and aligning the reads against “reference” genome is straightforward.

- But what if there is no such “reference” genome?
    - “de novo” assembly

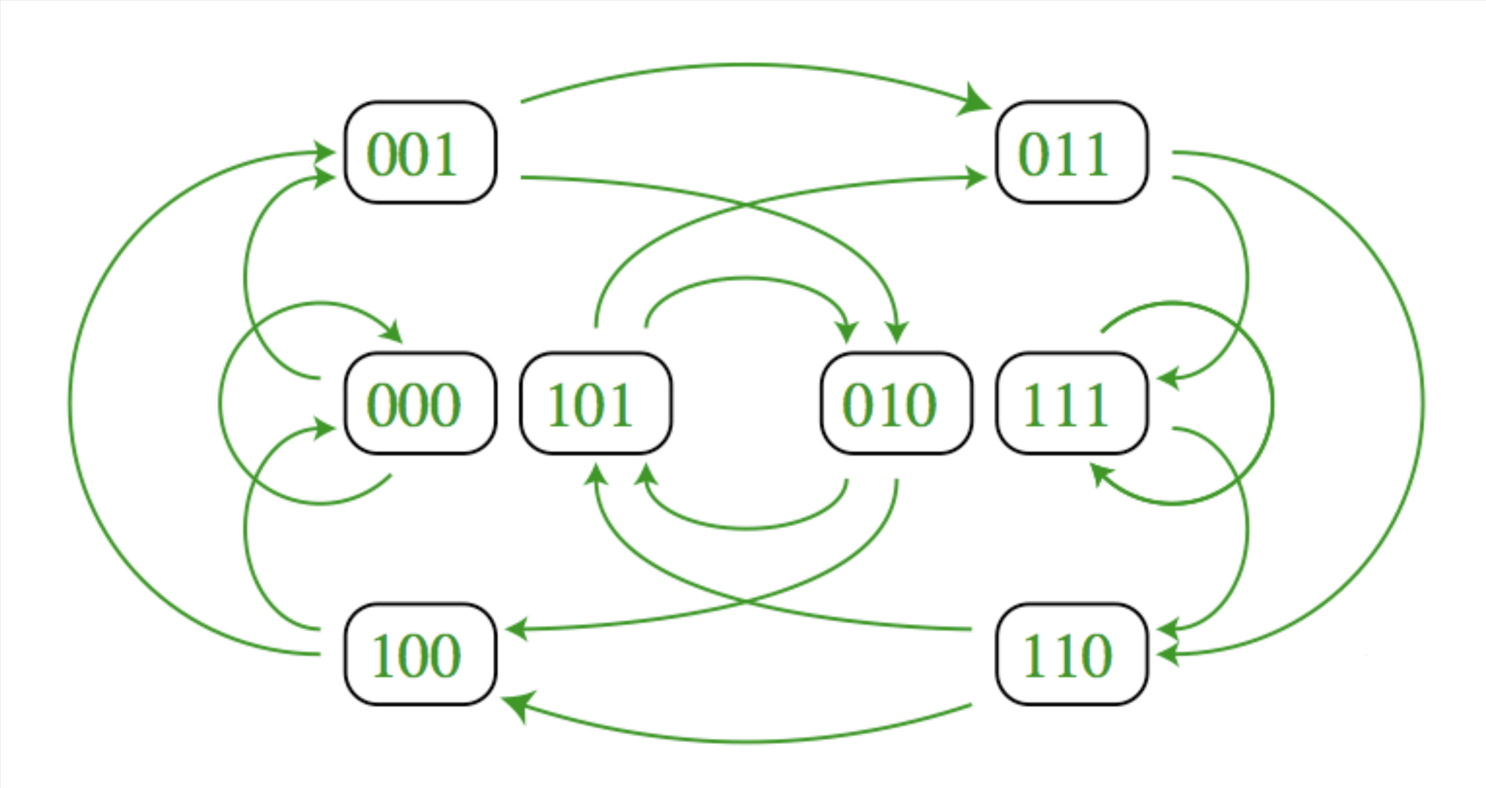
- Aligning the reads is only the first step

Assembly

- Solexa reads are too short for de novo assembly of large genomes.
- However, for prokaryotes and simple eukaryotes, reasonably large contigs can be assembled.
- Using paired-end reads with very large end separation is crucial.
- Assembly requires specialized software, typically based on de Brujin graphs
- Most popular assembly tools:
  - Velvet (Zerbino et al.)
  - ABySS (Simpson et al.)

Velvet Assembler

- De Bruijn Graphs

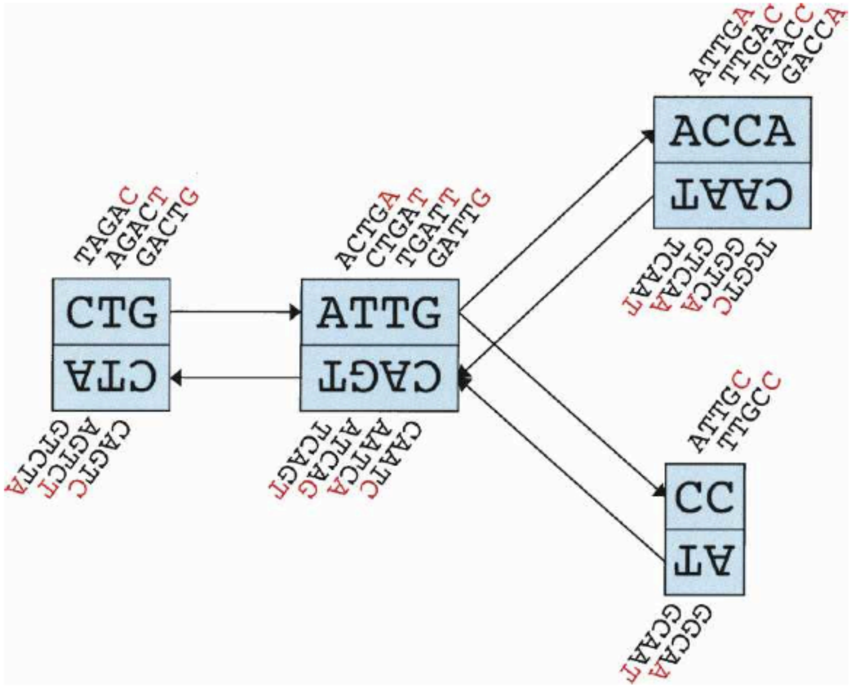


- Nodes are (sub)sequences, edges indicate overlap
- Each sequence is a path through the graph

Graph Construction

- sequence of each read is parsed into k-mers
- typical k=21 for read length of 25
- series of matches(k-1 long) are aligned together called a block
- the information of each block is the last bp of each k-mer in of the block

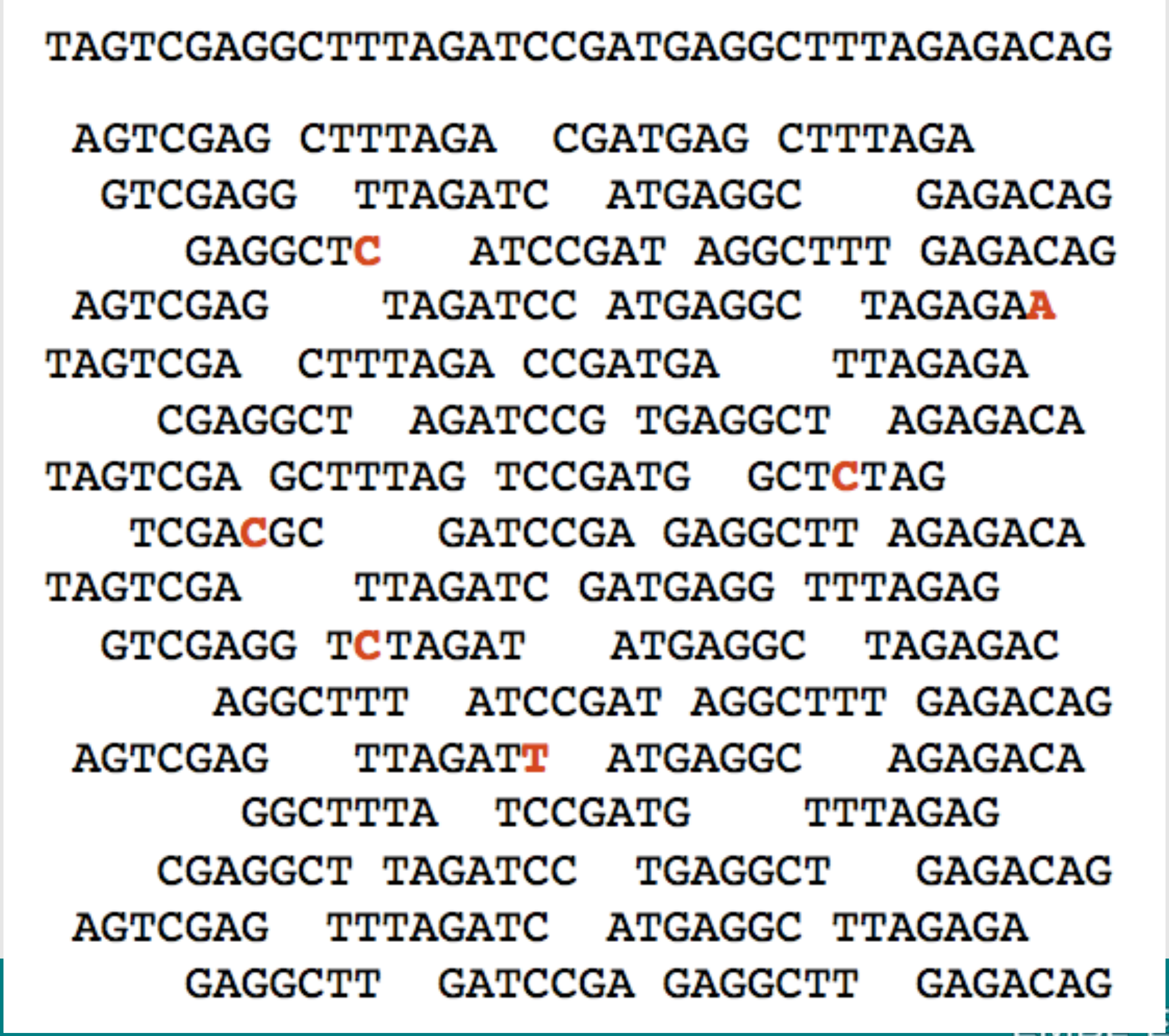
Alignment


TAGAC
AGACT
GACTG
CTG
CTA
GTCTA
AGTCT
CAGTC
ACTGA
CTGAT
TGATT
GATTG
ATTG
CAGT
TCAGT
ATCAG
AATCA
CAATC
ATTGA
TTGAC
TGACC
GACCA
ACCA
CAAT
TCAAT
GTCAA
GGTCA
TGGTC
ATTGC
TTGCC
CC
AT
GCAAT
GGCAA

Links

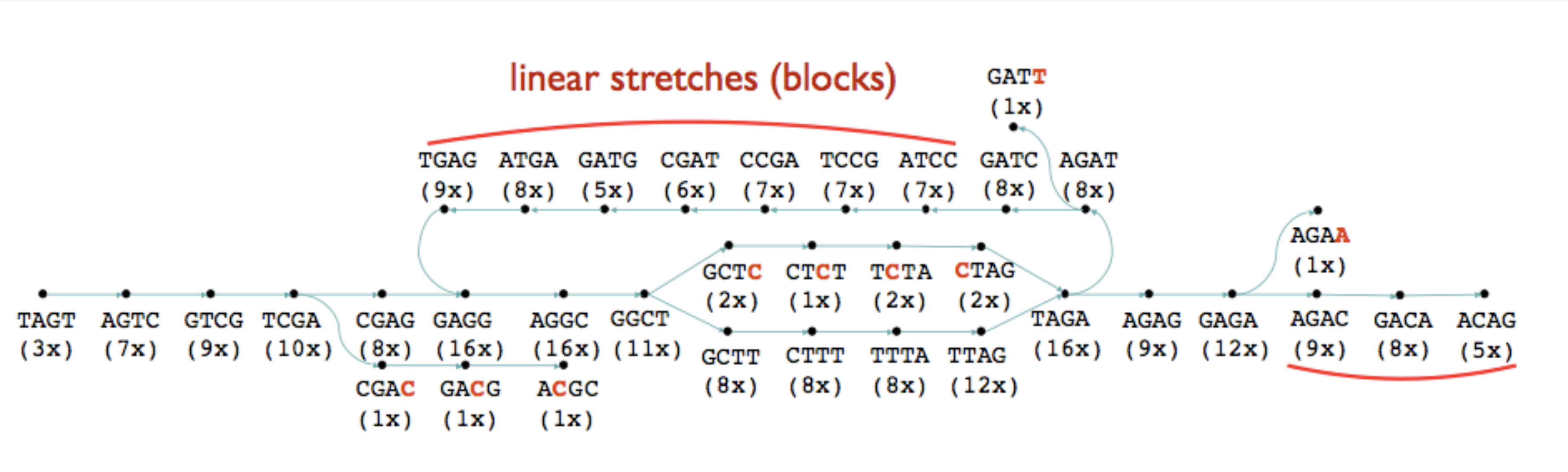
- a directed link is drawn if there exists a (k-1) long match between two blocks
- if everything is perfect, an underlying sequence follows all links in the de Bruijn graph while tracing through every block
- however, due to the noisy measurement and sequence repeats, many more steps are required

Example



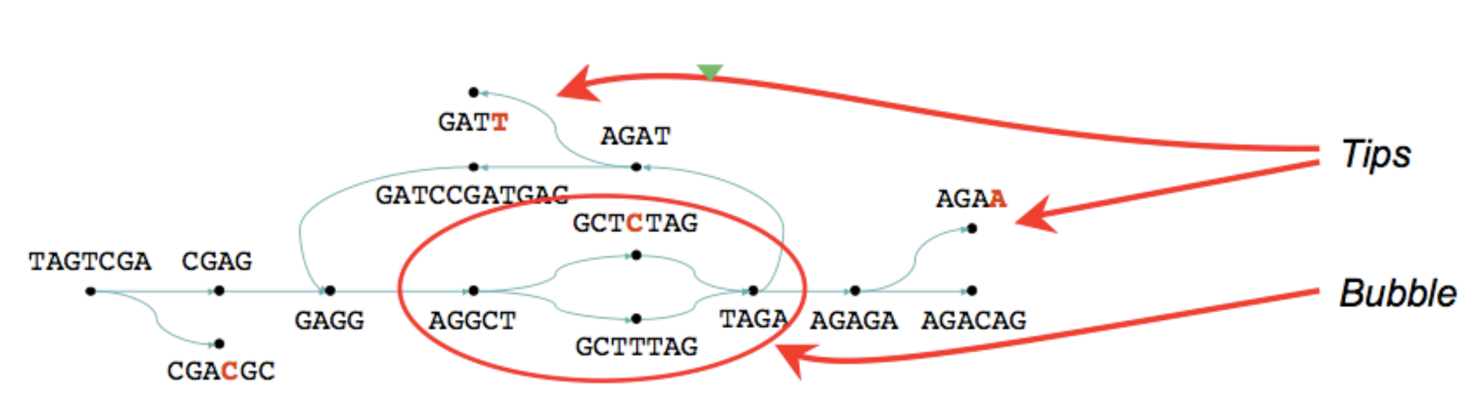
```
TAGTCGAGGCTTTAGATCCGATGAGGCTTTAGAGACAG

 AGTCGAG CTTTAGA  CGATGAG CTTTAGA
  GTCGAGG  TTAGATC  ATGAGGC    GAGACAG
     GAGGCTC   ATCCGAT AGGCTTT GAGACAG
 AGTCGAG    TAGATCC ATGAGGC  TAGAGAA
TAGTCGA  CTTTAGA CCGATGA    TTAGAGA
    CGAGGCT  AGATCCG TGAGGCT  AGAGACA
TAGTCGA GCTTTAG TCCGATG  GCTCTAG
   TCGACGC    GATCCGA GAGGCTT AGAGACA
TAGTCGA    TTAGATC GATGAGG TTTAGAG
  GTCGAGG TCTAGAT   ATGAGGC  TAGAGAC
      AGGCTTT  ATCCGAT AGGCTTT GAGACAG
 AGTCGAG   TTAGATT  ATGAGGC   AGAGACA
       GGCTTTA  TCCGATG    TTTAGAG
    CGAGGCT TAGATCC  TGAGGCT   GAGACAG
 AGTCGAG  TTTAGATC  ATGAGGC TTAGAGA
     GAGGCTT  GATCCGA GAGGCTT  GAGACAG
```

4-mer parsing

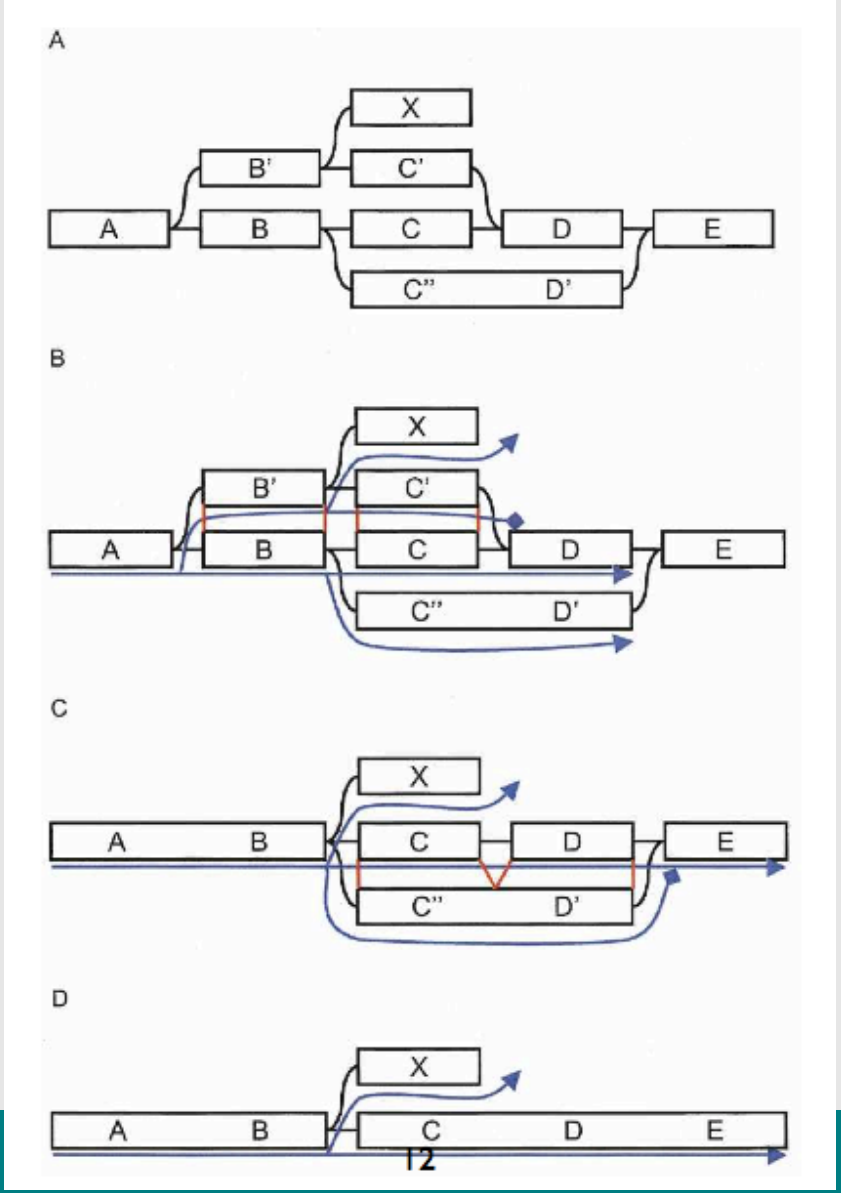


Mistakes

- Hanging tips (blocks that do not connect to anything) are likely due to mistakes, especially low-coverage ones
- Bubbles (cycles in the graph) likely due to errors

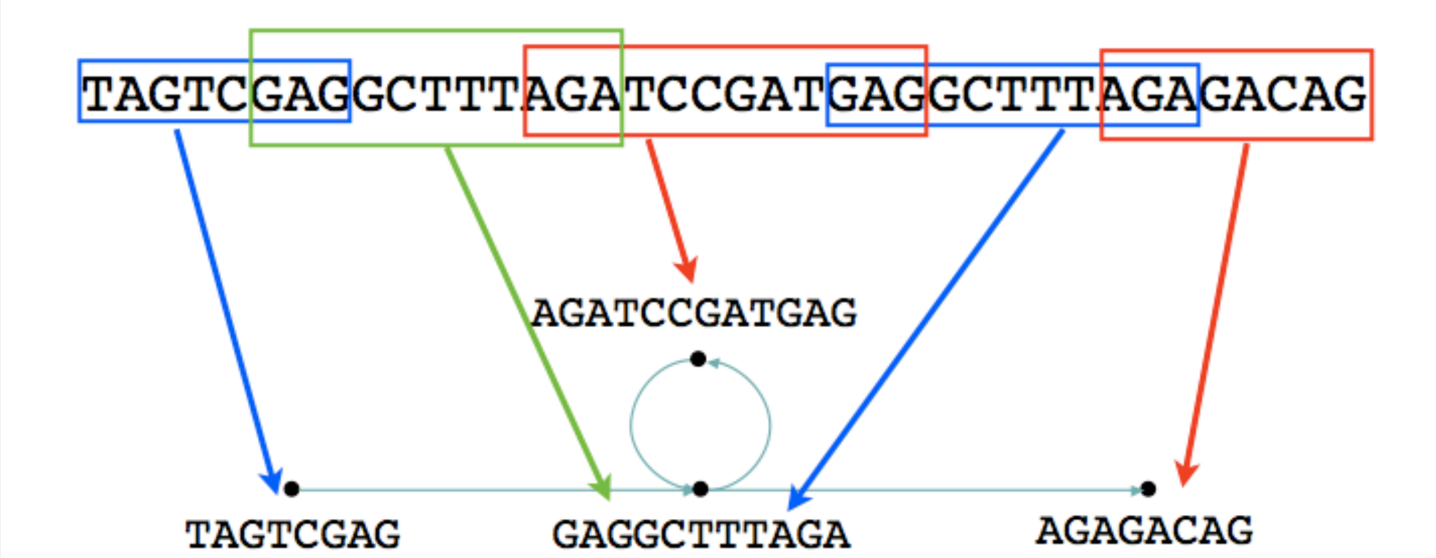


Bubble Removal



Example Construction

- In the example, sequence length=38 bp, read length=7bp , coverage~10X, error rate~ 3%, with one major repeat = 11bp
- k is chosen to be 5 bp
- Velvet is able to resolve this toy example!



On real data - harder

- a 173 kbp human BAC was sequenced by Solexa with a coverage of 970X
- read length are 35 bp
- k set to 31
- an virtual ideal sequencer (error free, gap free) that looks at the reference sequence is compared with Velvet

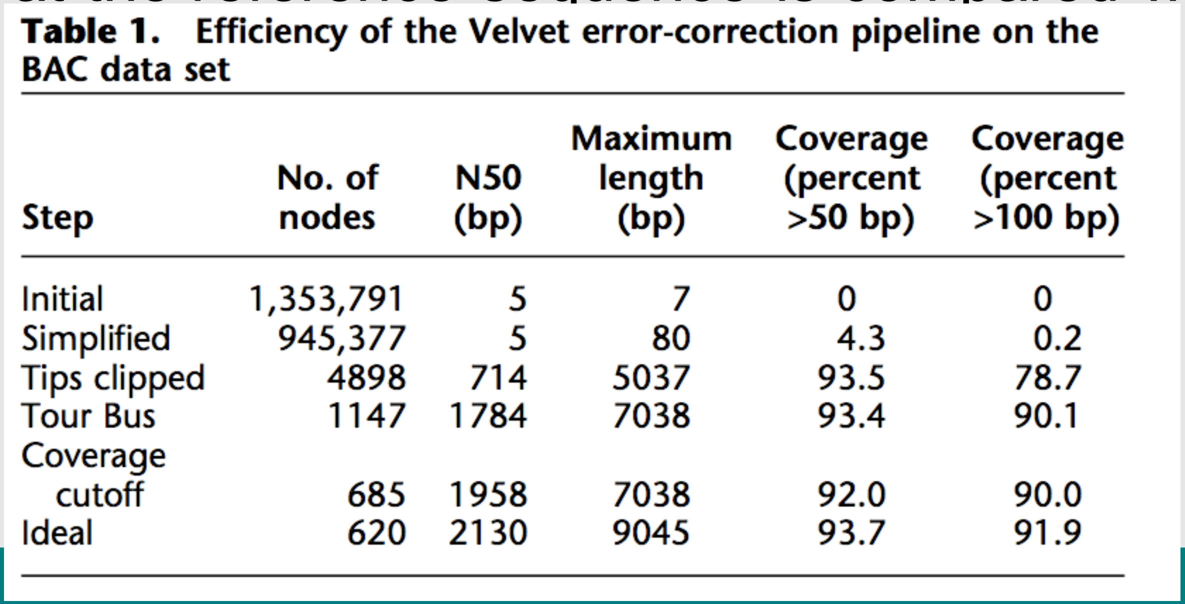


Table 1. **Efficiency of the Velvet error-correction pipeline on the BAC data set**

Step	No. of nodes	N50 (bp)	Maximum length (bp)	Coverage (percent >50 bp)	Coverage (percent >100 bp)
Initial	1,353,791	5	7	0	0
Simplified	945,377	5	80	4.3	0.2
Tips clipped	4898	714	5037	93.5	78.7
Tour Bus	1147	1784	7038	93.4	90.1
Coverage cutoff	685	1958	7038	92.0	90.0
Ideal	620	2130	9045	93.7	91.9

RNA-seq

- RNA-Seq has additional challenges
  - Reads may straddle splice junctions
  - Paralogy between genes prevent unique mappings
  - One may want to incorporate or amend known gene models
- Specialized tools for RNA-Seq alignment are
  - ERANGE
  - TopHat
  - To call differential expression
    - edgeR
    - BayesSeq

Summary

- Alignment & assembly
  - far from solved
- A lot of areas for improvement
  - Performance
  - Accuracy
- Tool pipelines non-existent, everyone writes their own
  - C/C++
  - R/Python

Large numbers of genomes sequenced

- 1000+ Bacterial and Archaeal genomes
- 100s of fungal genomes
- 10s of animal and plant genomes
- 10s of other eukaryotes
- 1000 human genome project - http://1000genomes.org
- 1001 Arabidopsis genomes - http://1001genomes.org
  1000 Drosophila genomes - http://dpgp.org
- 15,000 vertebrate genome project (proposed)
- Metagenomics

More open areas in computational biology

- **Image processing**
- **Protein 3-D structure analysis and prediction**
- **RNA structure prediction**
- **Gene network reconstruction from time series data**
- **Gene identification and annotation**
- **Gene function prediction**

Acknowledgements

- These presentations have been supported by funding from:
  - Sponsors of the CS Club (St. Petersburg)
  - EMBL
  - EC – SYBARIS Project

- We thank again all the authors of the slides used in these presentations.