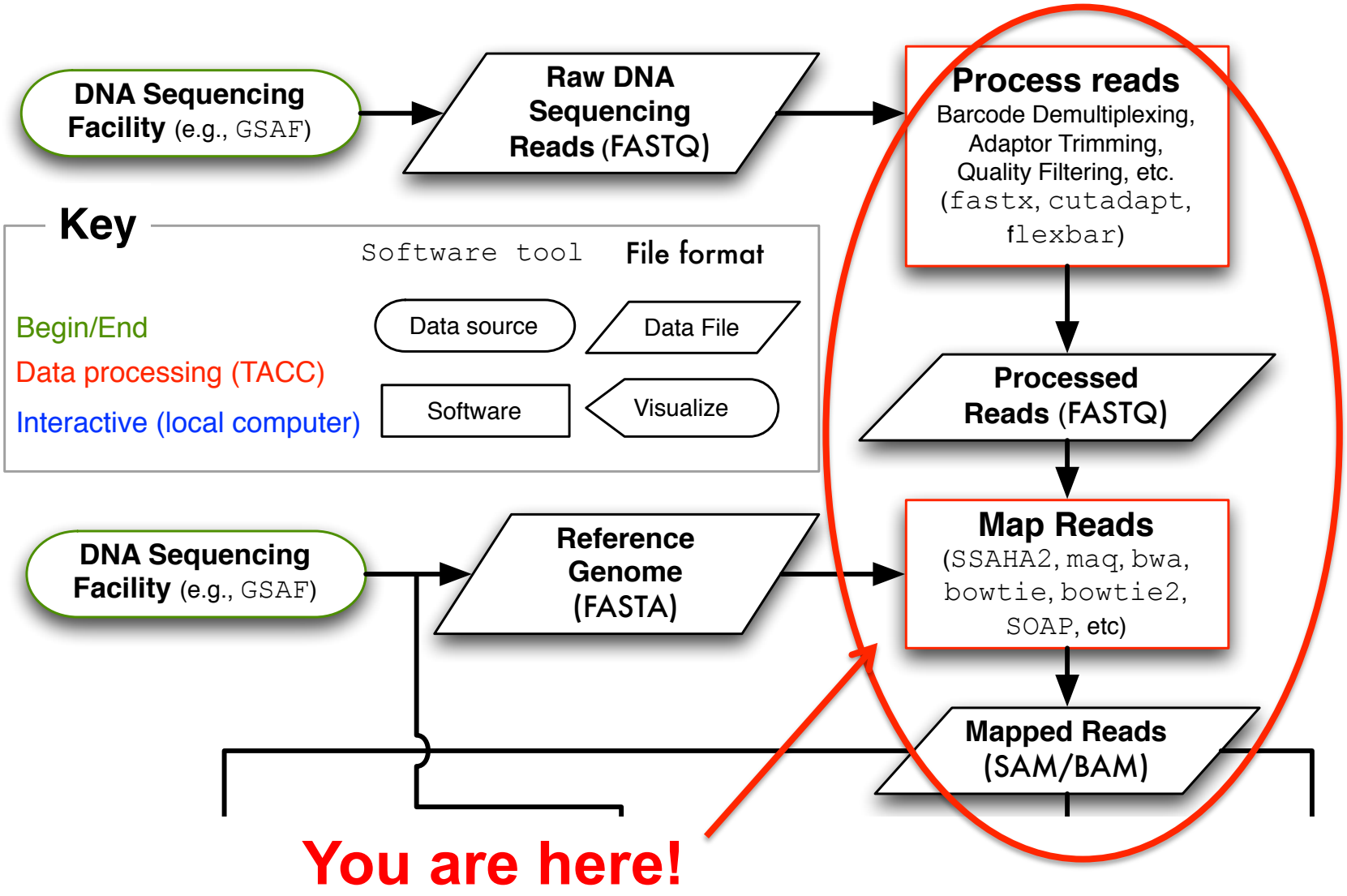
Introduction to Mapping



Basic steps of mapping reads

1. Read file quality control
2. Build reference sequence index
3. Map DNA sequencing reads
   - Exact tool/approach depends on sequencing technology and DNA fragment library type
4. Convert result to SAM/BAM database
5. Application specific analysis...
   - These steps are common to any reference-based (opposed to *de novo*) data analysis.
   - We will look at variant calling first.

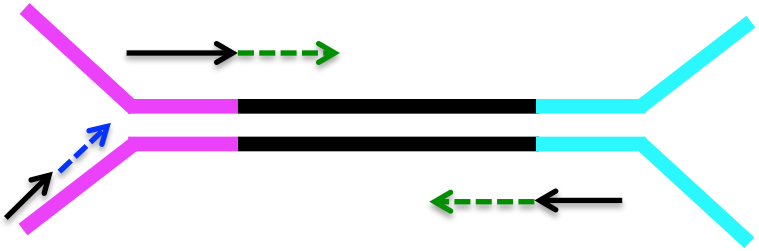
Read sequence terminology

Fragment library (input DNA sample)

Library prep

Sequencing library
Double-stranded or Y- **adaptors** added

DNA sequencing



Primers

Barcode (6–12 bases) – so many samples can be run in one lane. Data is **demultiplexed**.

Reads (36–1000+ bases)

Types of Illumina fragment libraries

single-end

independent reads

paired-end

two inwardly oriented reads separated by ~200 nt

mate-paired

two outwardly oriented reads separated by ~3000 nt

Read sequence quality control

Garbage in = garbage out

- Contaminated with other samples?
- Sample barcodes removed?
- Adaptor sequences trimmed?
  - RNAseq, MiSeq data
- Trim ends of reads with poor quality?
  - *de novo* Assembly
- Know your data
  - Paired reads? Relative orientations?
  - Technology specific concerns?
    - Indels with 454

Read sequences

FASTQ Format

```
@HWI-EAS216_91209:1:2:454:192#0/1
GTTGATGAATTTCTCCAGCGCGAATTTGTGGGCT
+HWI-EAS216_91209:1:2:454:192#0/1
B@BBBBBB@BBBBAAAA>@AABA?BBBAAB??>A?
```

Line 1: @read name

Line 2: called base sequence

Line 3: +read name (optional after +)

Line 4: base quality scores

Deciphering base quality scores

http://www.asciitable.com/

```
Quality character   !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHI
                    |         |         |         |         |
ASCII Value         33        43        53        63        73
Base Quality (Q)    0         10        20        30        40
```

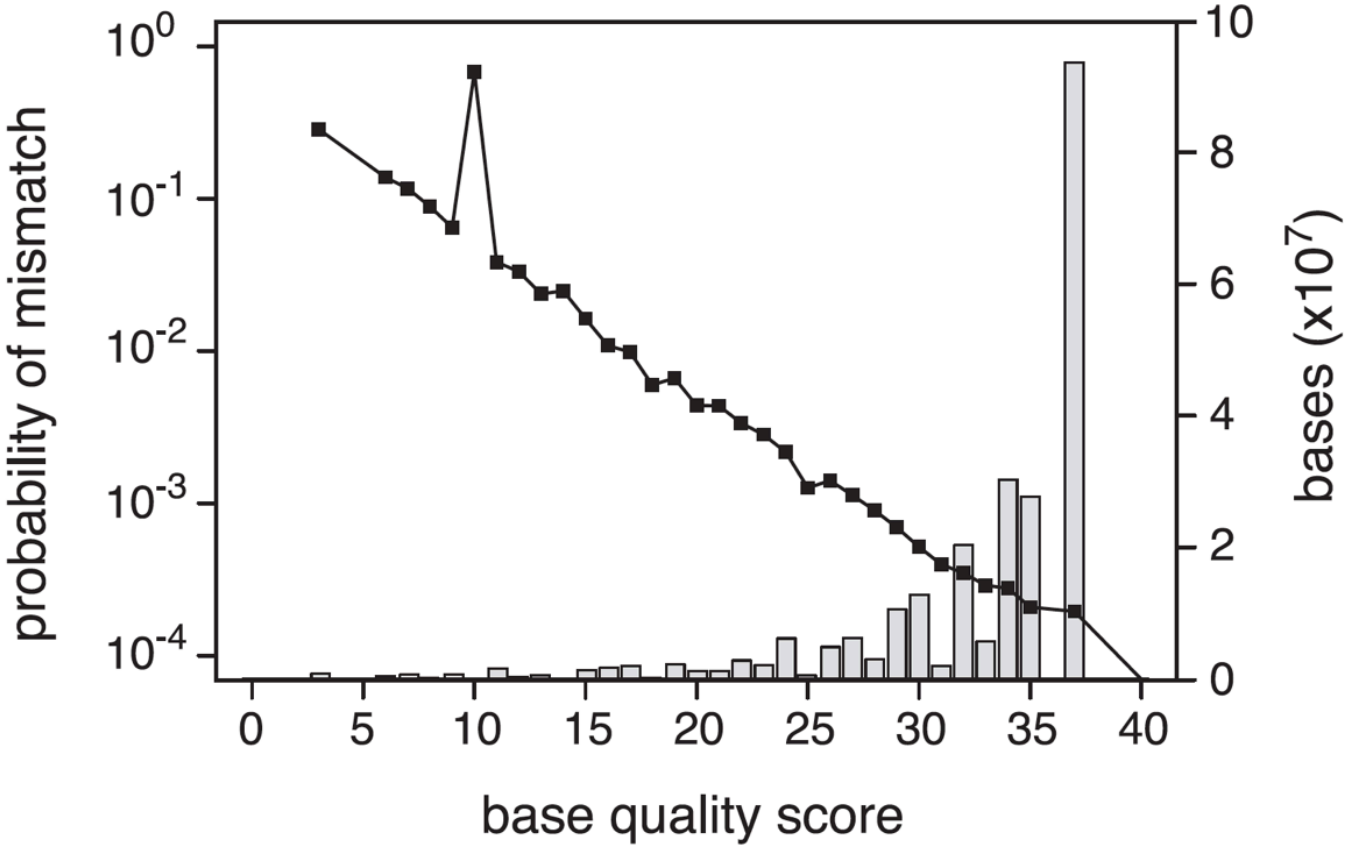
$$\text{Probability of Error} = 10^{-Q/10}$$

(This is a **Phred** score, also used for other types of qualities.)

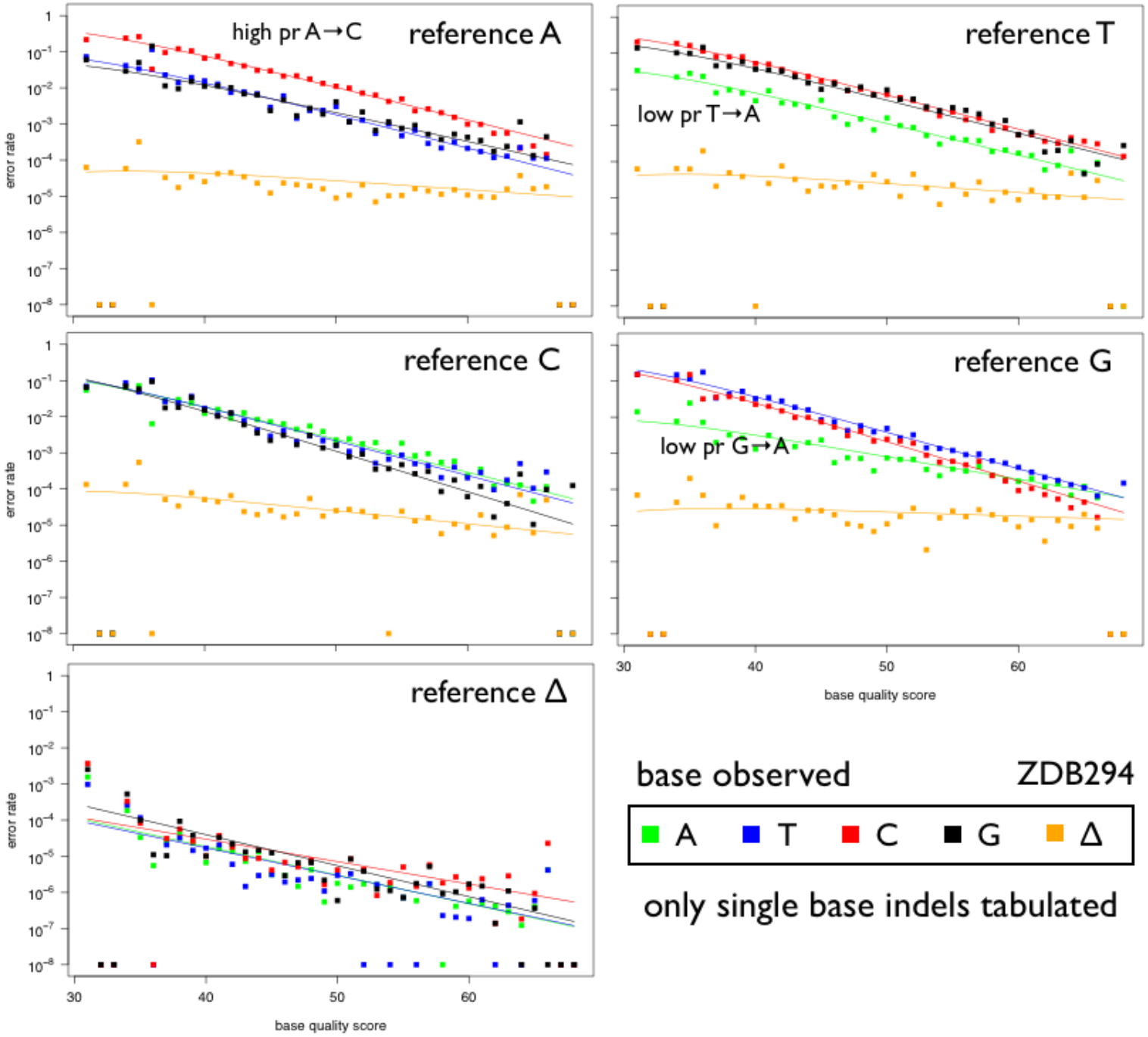
* Very low quality scores can mean something special – Illumina Q ≤ 3 means something like: "I'm lost, you might want to stop believing sequencing cycles from here on out."

* In older FASTQ files, the formula and ASCII offset might differ. Consult: http://en.wikipedia.org/wiki/FASTQ_format

Example of Illumina data



- Most bases have high qualities (Q>30).
- Overall qualities are well calibrated*.


high pr A→C
reference A
reference T
low pr T→A
reference C
reference G
low pr G→A
reference Δ
error rate
base quality score
base observed
ZDB294
A
T
C
G
Δ
only single base indels tabulated

Reference considerations

http://microbialgenomics.energy.gov

- Is it appropriate to your study?
  - Close enough to your species?
  - Complete?
- Which version?
  - Make sure you use an agreed-on standard
- Does it contain repeats? What kinds?
  - Know this up front or you will be confused
- What annotations exist?
  - References lacking feature annotations are much more challenging to use

Reference sequences

FASTA Format

```
>gi|254160123|ref|NC_012967.1| Escherichia coli B str. REL606
agcttttcattctgactgcaacgggcaatatgtctctgtgtggattaaaaaaagagtgtc
tgatagcagcttctgaactggttacctgccgtgagtaaattaaaatttattgacttagg
tcactaaatactttaaccaatataggcatagcgcacagacagataaaaattacagagtac
acaacatccatgaaacgcattagcaccaccattaccaccaccatcaccattaccacaggt
aacggtgcgggctgacgcgtacaggaaacacagaaaaaagcccgcacctgacagtgcggg
ctttttttcgaccaaaggtaacgaggtaacaaccatgcgagtgttgaagttcggcggta
....
```

Using complex reference sequence names is a common problem during analysis. Might rename:

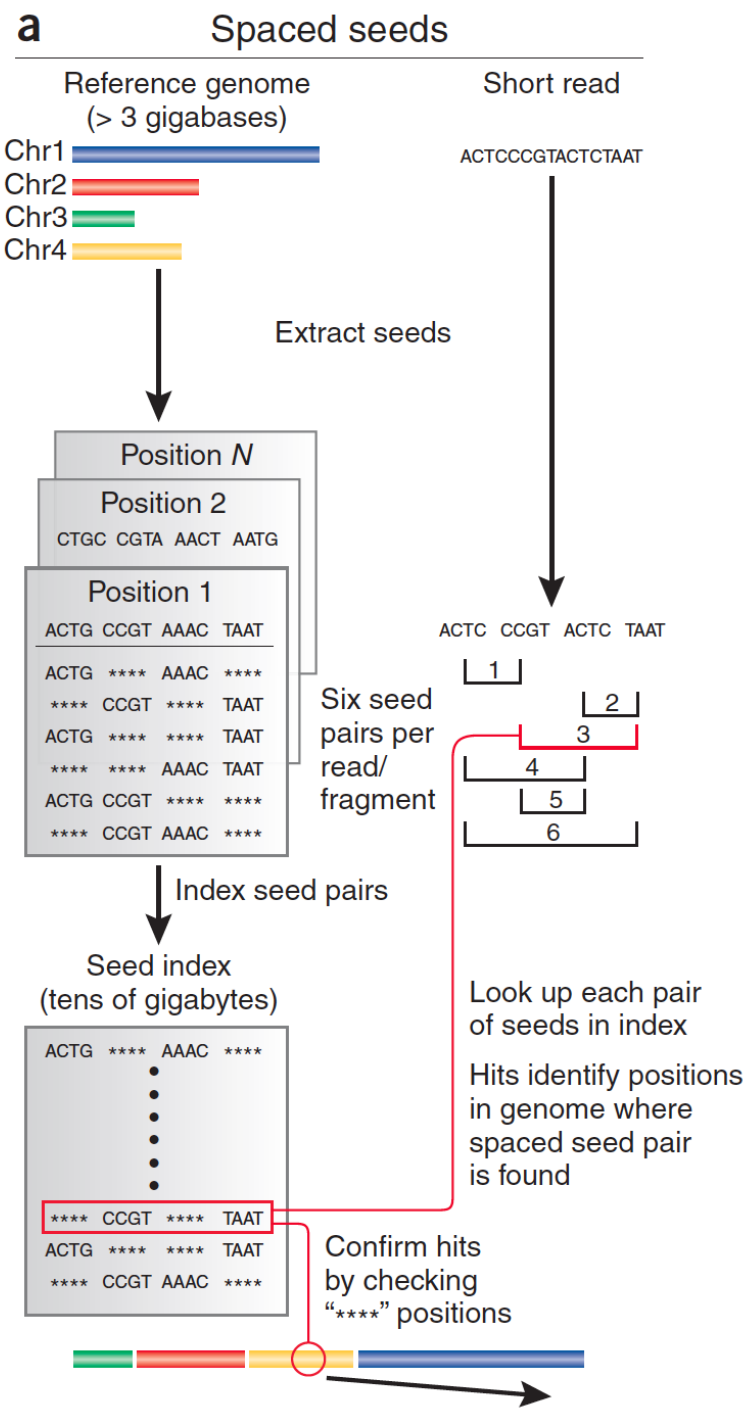
```
>REL606
agcttttcattctgactgcaacgggcaatatgtctctgtgtggattaaaaaaagagtgtc
tgatagcagcttctgaactggttacctgccgtgagtaaattaaaatttattgacttagg

....
```

Finding a Reference

- **Microbes (<10Mb)**: download FASTA containing in sequence and/or GenBank/ EMBL/GFF flat files encapsulating both sequence and features.
- **Macrobes (>100Mb)**: download specific consortium "build" of reference (Ex: hg19), consisting of FASTA, and various files used to construct a database of features.
- **Non-model organisms**: build your own? *See de novo assembly topic later in course.*

Mappers/Aligners

- Algorithms
  - Spaced-seed indexing
  - Burrows-Wheeler transform (BWT)
- Differences
  - Input data (read length, colorspace aware/useful)
  - Speed and scaleability (multithreading, GPUs)
  - Memory requirements (RAM, fat nodes)
  - Sensitivity: esp. indels (gaps)
  - Ease of installation and use. Development phase.
  - Uses base qualities? Outputs mapping scores?
  - Handles of multiple matches, paired end matches
  - Configurability and transparency of options



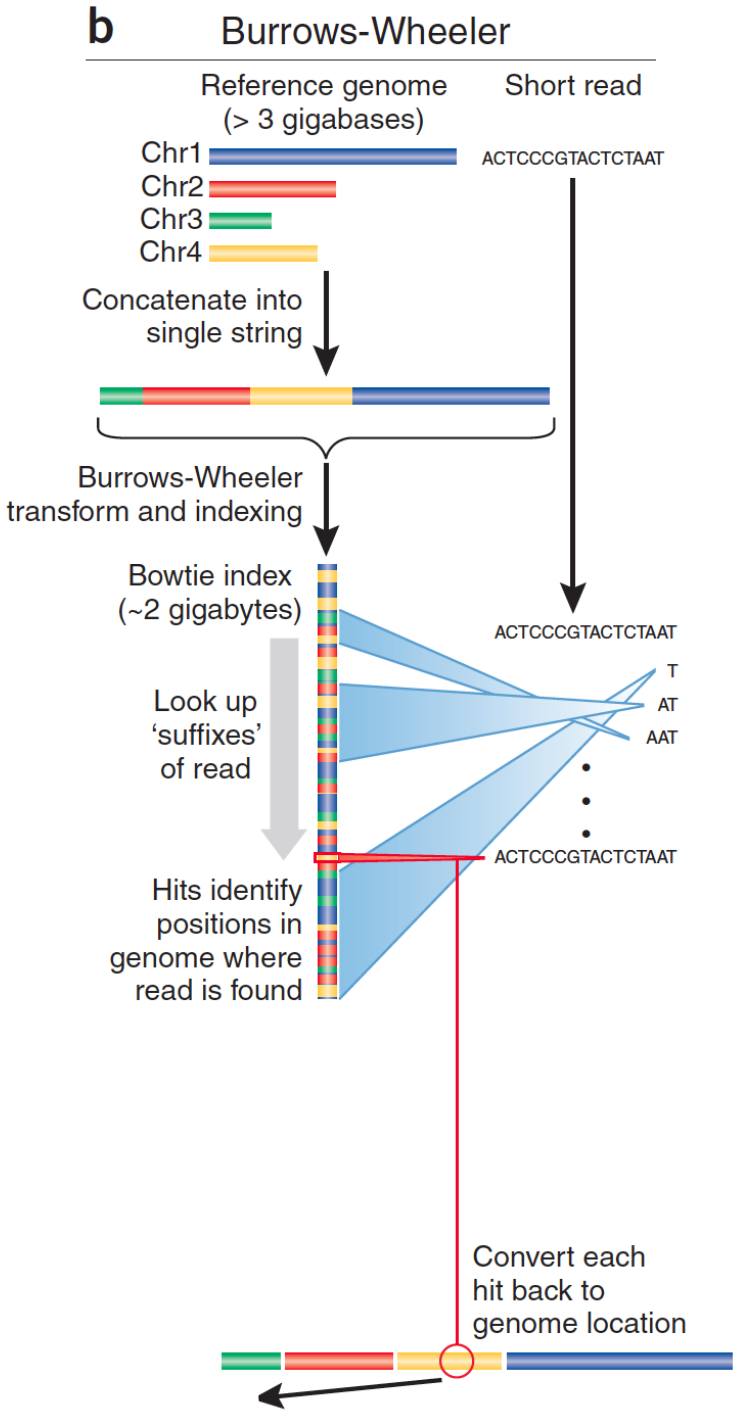
Hash table enables lookup of exact matches.

Subsequence	Reference Positions
ATAGCTAATCCAAA	2341, 2617264
ATAGCTAATCCAAT	
ATAGCTAATCCAAC	134, 13311, 732661,
ATAGCTATCCAAAG	
ATAGCTAATCCATA	
ATAGCTAATCCATT	3452
ATAGCTAATCCATC	
ATAGCTATCCAATG	234456673

Table is sorted and complete so you can jump immediately to matches. (But this can take a lot of memory.)

May include N bases, skip positions, etc.

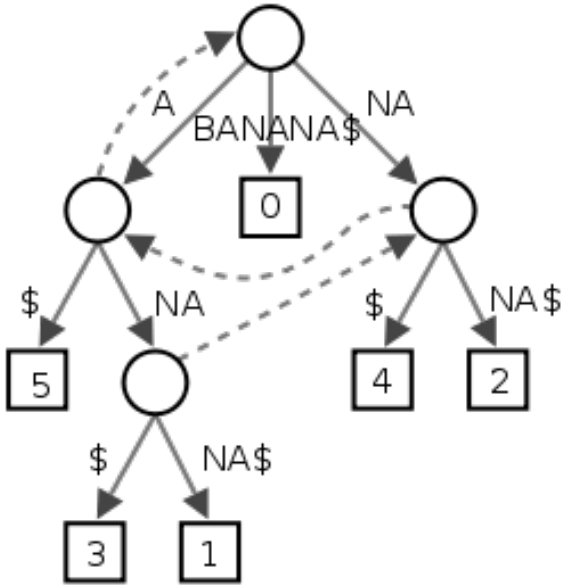
Trapnell, C. & Salzberg, S. L. How to map billions of short reads onto genomes. *Nature Biotech.* **27**, 455–457 (2009).



Burrows-Wheeler transform compresses sequence.

Input	SIX.MIXED.PIXIES.SIFT.SIXTY.PIXIE.DUST.BOXES
Output	TEXYDST.E.IXIXIXXSSMPPS.B..E.S.EUSFXDIIOIIIT

Suffix tree enables fast lookup of subsequences.



http://en.wikipedia.org/wiki/Suffix_tree

Exact matches at all positions below a node.

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

From Mapped Read to Alignment

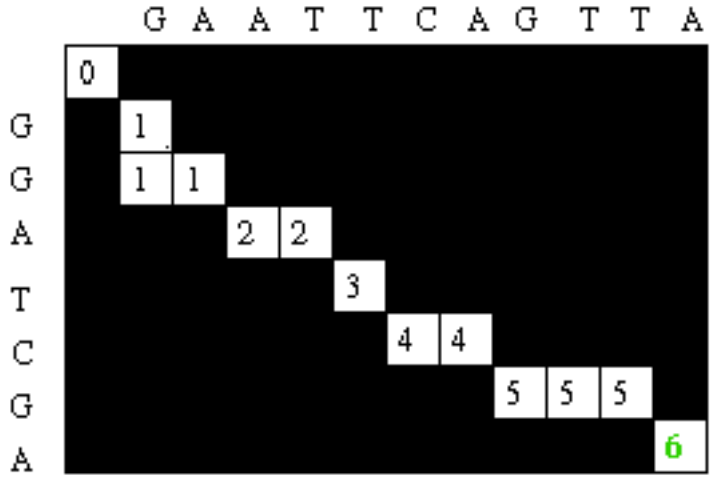
- **Mapping** determines a "seed" position where the read shares a subsequence with the reference. But, is this the best match?
- **Alignment** starts with the seed and determines how the read is best aligned on a base-by-base basis around the seed.

Seed→Alignment score→Mapping quality

Alignment

- Dynamic programming algorithm (Smith-Waterman | Needleman-Wunsch)

		G	A	A	T	T	C	A	G	T	T	A
	0	0	0	0	0	0	0	0	0	0	0	0
G	0	1	1	1	1	1	1	1	1	1	1	1
G	0	1	1	1	1	1	1	1	2	2	2	2
A	0	1	1	2	2	2	2	2	2	2	2	3
T	0	1	2	2	3	3	3	3	3	3	3	3
C	0	1	2	2	3	3	4	4	4	4	4	4
G	0	1	2	2	3	3	4	4	5	5	5	5
A	0	1	2	3	3	3	4	5	5	5	5	6




```
G _ A A T T C A G T T A
|   |   | |   |     |
G G _ A _ T C _ G _ _ A
```

- Various scoring schemes possible... (next slide)

Alignment Score

- Dynamic programming algorithm (Smith-Waterman | Needleman-Wunsch)
- **Alignment score** = Σ
  - match reward
  - base mismatch penalty
  - gap open penalty
  - gap extension penalty
  - rewards and penalties may be adjusted for quality scores of bases involved

Reference sequence

```
ATTTGCGATCGGATGAAGACGAA
||||||||||||||
ATTTGCGATCGGATGTTGACTTT

ATTTGCGATCGGATGAAGACG..AA
||||||||||||||XX|||XXX||
ATTTGCGATCGGATGTTGACTTTAA
```

- **Important: Local** versus **global** alignment

Mapping Quality

Mapping quality– what is the probability that the read is correctly mapped to this location in the reference genome?

High **alignment** score ≠ high **mapping** quality.

Phred score: $P(\text{mismapped}) = 10^{-MQ/10}$

Types of DNA fragment libraries

single-end

independent reads

paired-end

two inwardly oriented reads separated by ~200 nt

mate-paired

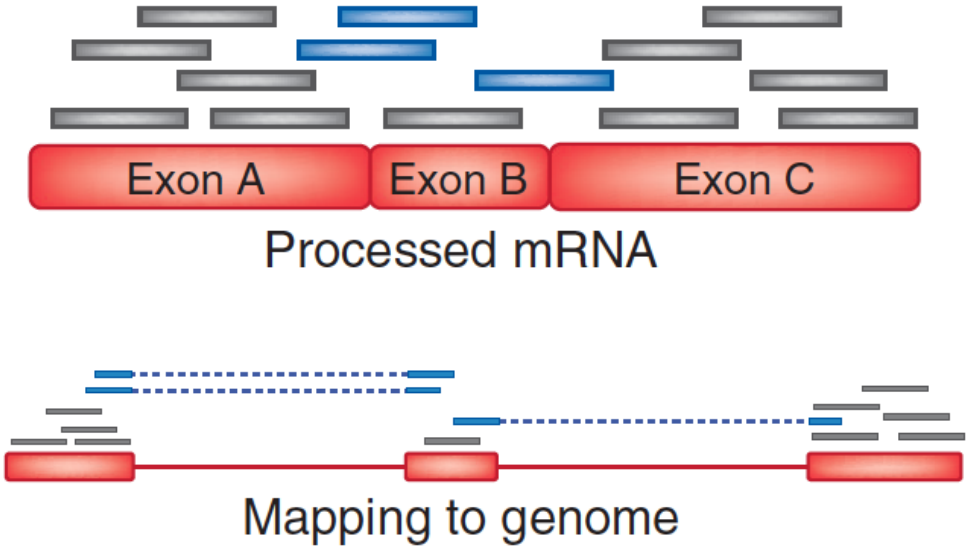
two outwardly oriented reads separated by ~3000 nt

Paired-end mapping (PEM)

- There is an expected insert size distribution based on the DNA fragment library.
- Mapping one read anchors the paired read to a specific location, even if the second read alone maps multiple places equally.
- Only one read in a pair might be mappable. (**singleton/orphan**)
- Both reads can map with an unexpected insert size or orientation (**discordant pair**)

Split-read alignment (SRA)

- Useful for predicting splice variants or structural variants.
- Not many mappers do this directly, usually happens in a post-processing step.



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

List of Mappers/Aligners

	Algorithm	Gapped	Quality-aware	Colorspace aware
BLAST	Hash table	Y	N	N
BLAT/SSHA2	Hash table	N	N	N
MAQ	Spaced seed	N	N	N
RMAP	Spaced seed	N	Y	N
ZOOM	Spaced seed	N	--	N
SOAP	Spaced seed	N	N	N
Eland	Spaced seed	N	N	N
SHRiMP	Q-gram/multi-seed	Y	Y	Y
BFAST	Q-gram/multi-seed	Y	Y	Y
Novoalign	Multi-seed + Vectorized SW	Y	Y	Y
clcBio	Multi-seed + Vectorized SW	Y	Y	Y
MUMmer	Tries	Y	N	N
OASIS	Tries	Y	--	--
VMATCH	Tries	Y	--	--
BWA/BWA-SW	Tries	Y	Y	Y
BOWTIE	Tries	Y	Y	Y
SOAP2	Tries	Y	N	N
Saruman	Exact (GPU)	Y	--	N

trie = tree structure for fast text re**trie**val.

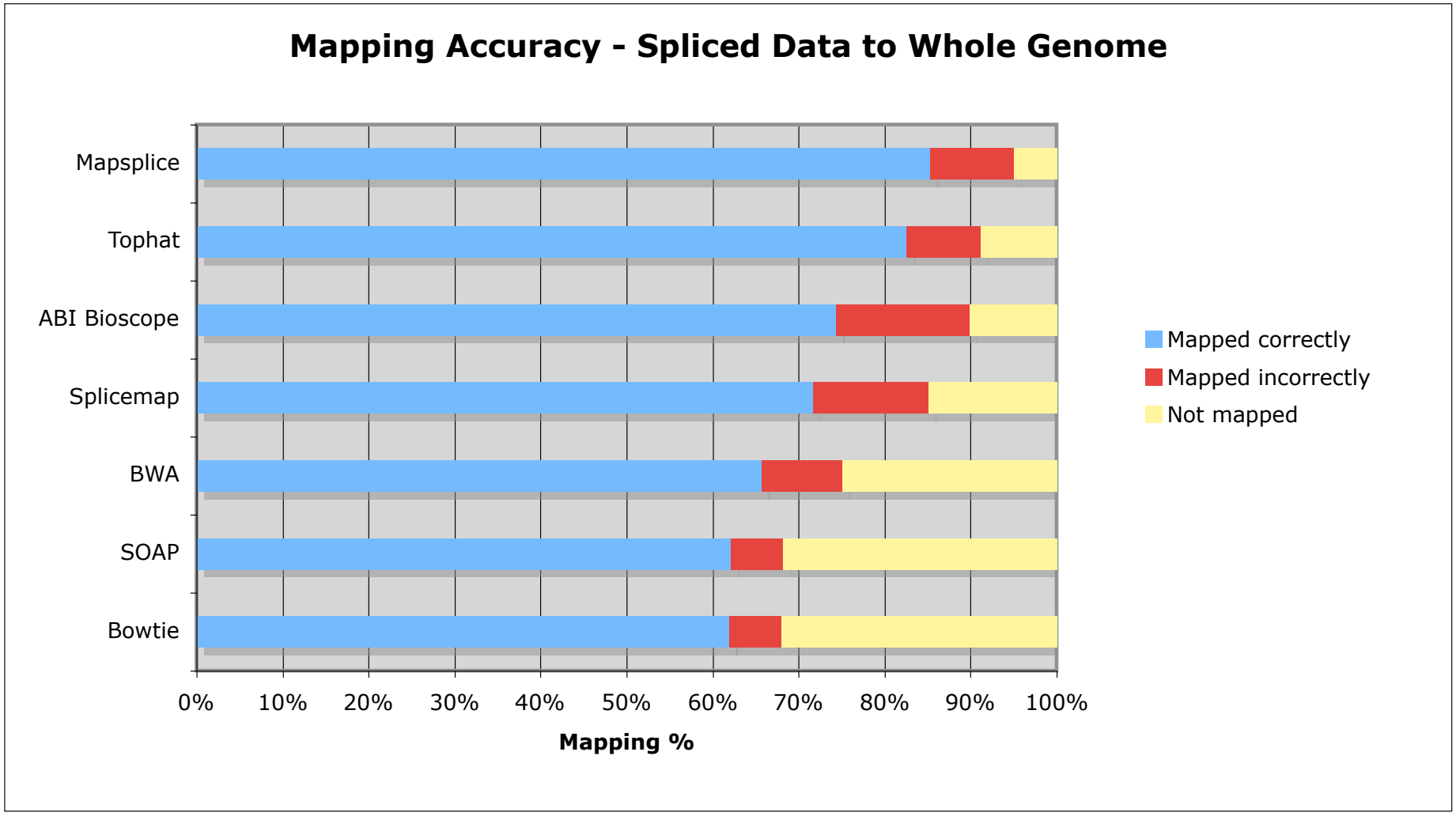
Courtesy Matt Vaughn, TACC

Indexing time

Aligner	Time (mins) to Index 3GB genome
SOAP2	98.06
BWA	110.73
Bowtie	220.82
Bfast	941.10*

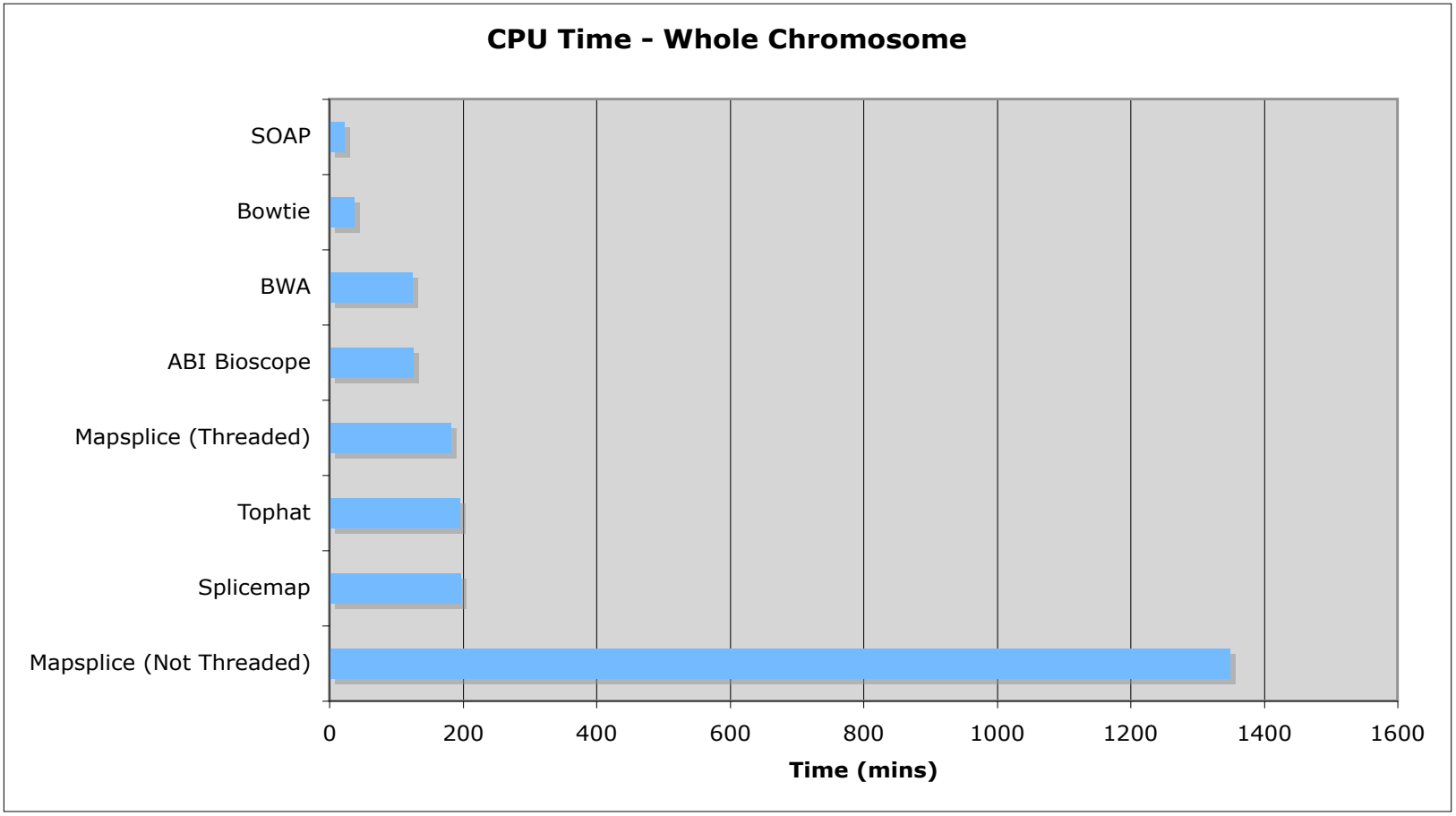
Data courtesy Dhivya Arasappan, GSAF Bioinformatician

Some Comparisons



Data courtesy Dhivya Arasappan, GSAF Bioinformatician

Some Comparisons



Data courtesy Dhivya Arasappan, GSAF Bioinformatician

Final Words

- My personal favorites of the moment...
  - Bowtie2
  - BWA
- Pay close attention to the details...
  - Methylation (bisulfite) analysis
  - RNA editing
  - Hypervariable region analysis
    (Ex: contingency loci / microsatellites)
  - To understand what you might be missing!

SAM File Format

- Community flat file/database format that describes how reads align to a reference (and can also include unmapped reads).
- Can tag reads as being from different instrument runs / technologies / samples.
- Going forward you need the reference file and the SAM, no longer need the FASTQ.
- Tab delimited with fixed columns followed by arbitrary user-extendable key:data values.

SAM File Format

Two example SAM lines:

```
SRR030257.264529      99  NC_012967  1521  29  34M2S  =  1564  79
    CTGGCCATTATCTCGGTGGTAGGACATGGCATGCCC
    AAAAAA;AA;AAAAAA??A%.;?&'3735',()0*,
    XT:A:M  NM:i:3  SM:i:29  AM:i:29  XM:i:3  XO:i:0  XG:i:0  MD:Z:23T0G4T4

SRR030257.2669090     147 NC_012967  1521  60  36M    =  1458  -99
    CTGGCCATTATCTCGGTGGTAGGTGATGGTATGCGC
    <<9:<<AAAAAAAAAAAAAAAAAAAAAAAAAAAA
    XT:A:U  NM:i:0  SM:i:37  AM:i:37  X0:i:1  X1:i:0  XM:i:0  XO:i:0  XG:i:0  MD:Z:36
```

SAM File Format

SAM fixed fields:

http://samtools.sourceforge.net/

Col	Field	Type	Regexp/Range	Brief description
1	QNAME	String	`[!-?A-~]{1,255}`	Query template NAME
2	FLAG	Int	$[0,2^{16}-1]$	bitwise FLAG
3	RNAME	String	`\*|[!-()+-<>-~][!-~]*`	Reference sequence NAME
4	POS	Int	$[0,2^{29}-1]$	1-based leftmost mapping POSition
5	MAPQ	Int	$[0,2^{8}-1]$	MAPping Quality
6	CIGAR	String	`\*|([0-9]+[MIDNSHPX=])+`	CIGAR string
7	RNEXT	String	`\*|=|[!-()+-<>-~][!-~]*`	Ref. name of the mate/next segment
8	PNEXT	Int	$[0,2^{29}-1]$	Position of the mate/next segment
9	TLEN	Int	$[-2^{29}+1,2^{29}-1]$	observed Template LENgth
10	SEQ	String	`\*|[A-Za-z=.]+`	segment SEQuence
11	QUAL	String	`[!-~]+`	ASCII of Phred-scaled base QUALity+33

SRR030257.264529 99 NC_012967 1521 29 34M2S = 1564
79 CTGGCCATTATCTCGGTGGTAGGACATGGCATGCCC
AAAAAA;AA;AAAAAA??A%.;?&'3735',()0*,
XT:A:M NM:i:3 SM:i:29 AM:i:29 XM:i:3 XO:i:0 XG:i:0 MD:Z:23T0G4T4

Sometimes a CIGAR is a just a way of describing how a read is aligned...

```
Ref   CTGGCCATTATCTC--GGTGGTAGGACATGGCATGCCC
Read        aaATGTCGCGGTG.TAGGAggatcc
```

2S5M2I4M1D4M6S

	Op	BAM	Description
	M	0	alignment match (can be a sequence match or mismatch)
	I	1	insertion to the reference
	D	2	deletion from the reference
*	N	3	skipped region from the reference
	S	4	soft clipping (clipped sequences present in SEQ)
*	H	5	hard clipping (clipped sequences NOT present in SEQ)
*	P	6	padding (silent deletion from padded reference)
*	=	7	sequence match
*	X	8	sequence mismatch

*Rarer / newer

CIGAR = "Concise Idiosyncratic Gapped Alignment Report"

BAM format

- "Human readable" text (SAM) and GZIP compressed binary (BAM) versions.
- BAM files can be **sorted** and **indexed**, so that all reads mapped to a given window of the reference genome can be retrieved rapidly (for display or processing).