File Formats and Additional Read Mapping Details

FILE FORMATS

fasta format

>sequence name

ACTGACTGACTG… (sequence)

- Charactersistics:
  - Alternating 2 line structure.
  - Single fasta file can have infinite number of sequences in it.
  - Can be used for reference as well as data.

fastq format

```
@HWI-ST1097:104:D13TNACXX:4:1101:1715:2142 1:N:0:CGATGT
GCGTTGGTGGCATAGTGGTGAGCATAGCTGCCTTCCAAGCAGTTATGGGAG
+
=<@BDDD=A;+2C9F<CB?;CGGA<<ACEE*1?C:D>DE=FC*0BAG?DB6
```

- Charactersistics:
  - 4line structure: read information, sequence, “+”, quality
  - Standard output for NGS data
  - https://en.wikipedia.org/wiki/FASTQ_format

SAM file

- 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 use the reference file and the SAM/BAM, no longer need the FASTQ.
- Tab delimited with fixed columns followed by arbitrary user-extendable key:data values.

2 SAM lines format example

- 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:3XO: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:0XO:i:0 XG:i:0 MD:Z:36
- NOTE: white space is TABS not space

SAM file format continued

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
```

SAM format 'CIGAR' (Concise Idiosyncratic Gapped Alignment Report) strings

```
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

Note: indels relative to reference

*Rarer / newer

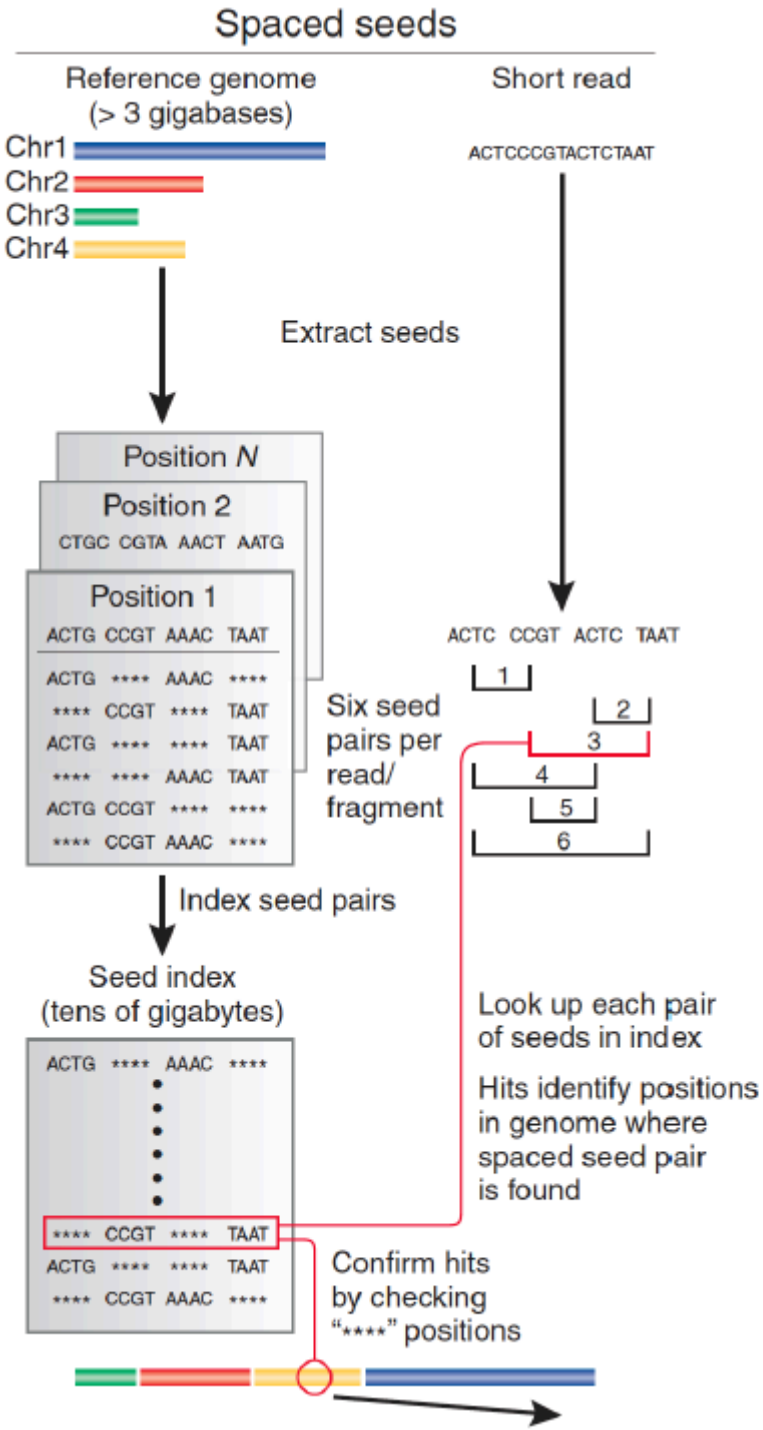
BAM files

- "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).
- SAMtools package can calculate stats and perform basic genome variant calling.

BAM file format

- As much as cigar strings may evoke memories of looking at the matrix code, BAM formats are actually written in binary and as such are converted from SAM files not written by mere humans.

MAPPING DETAILS



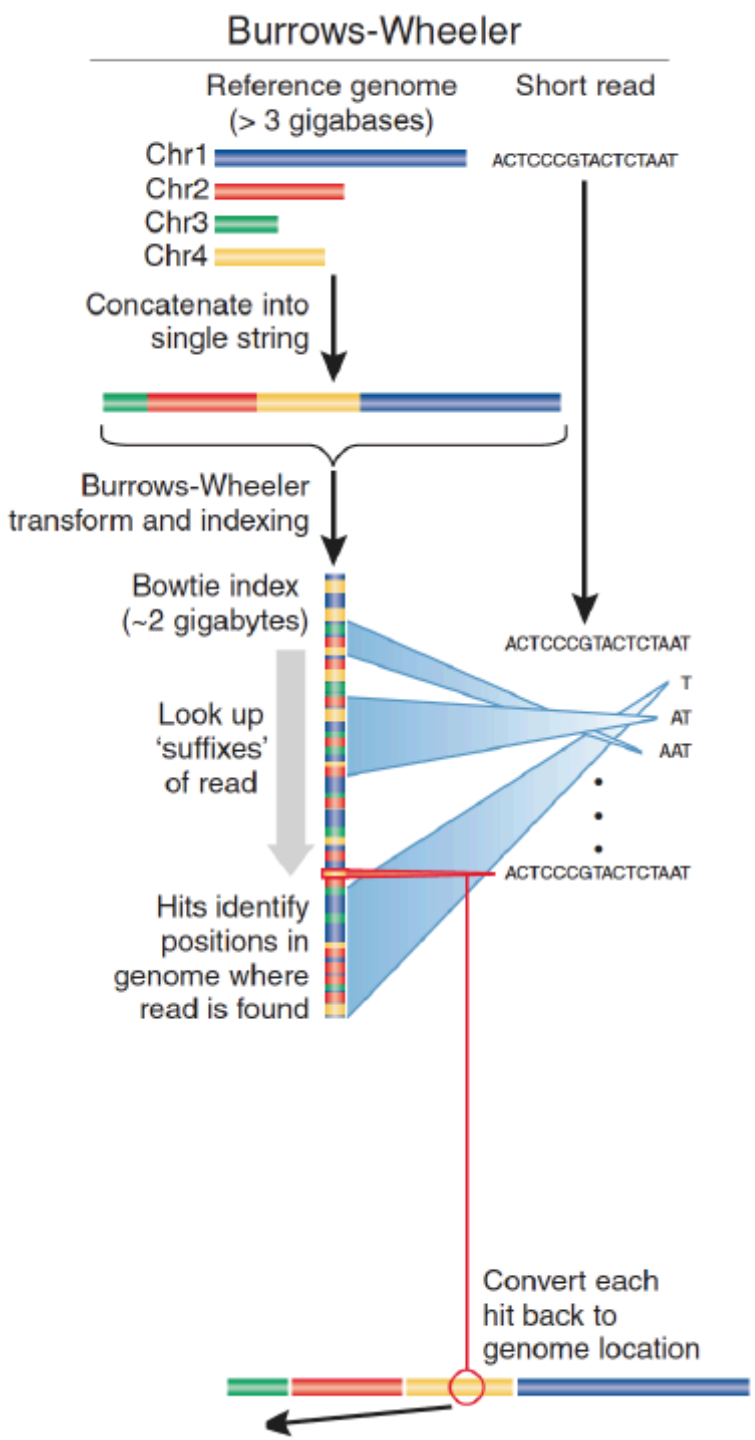
Hash table enables lookup of exact matches.

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.

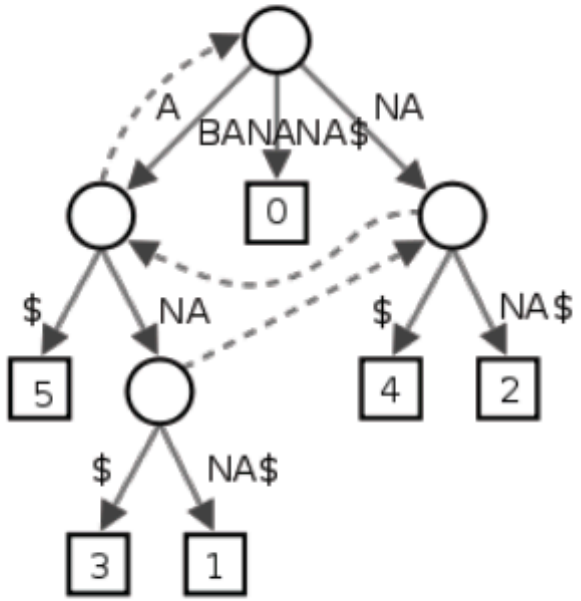
Trapnell, C. & Salzberg, S. L. How to map billions of short reads onto genomes. *Nature Biotech.* **27**, 455–457 (2009).



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

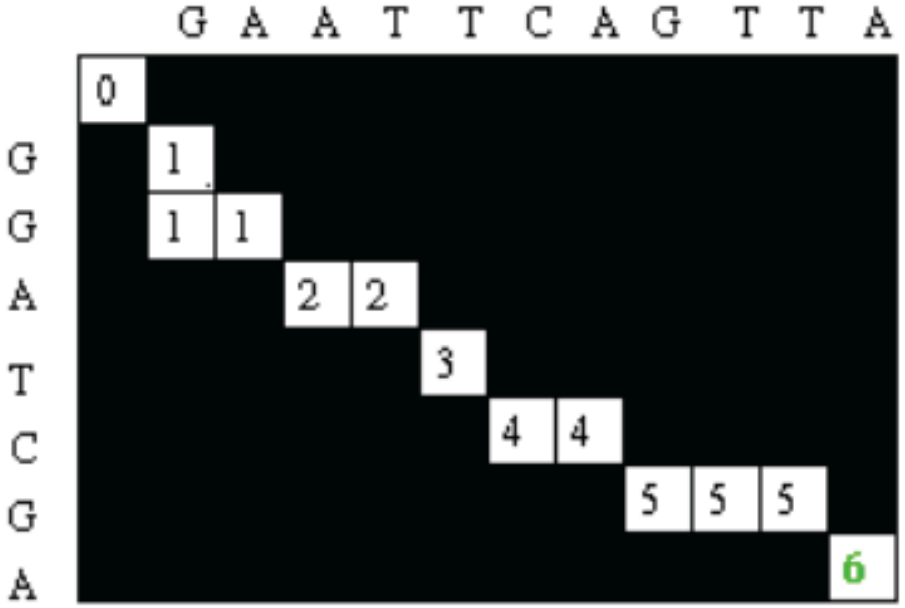
Exact matches at all positions below a node.

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

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
```