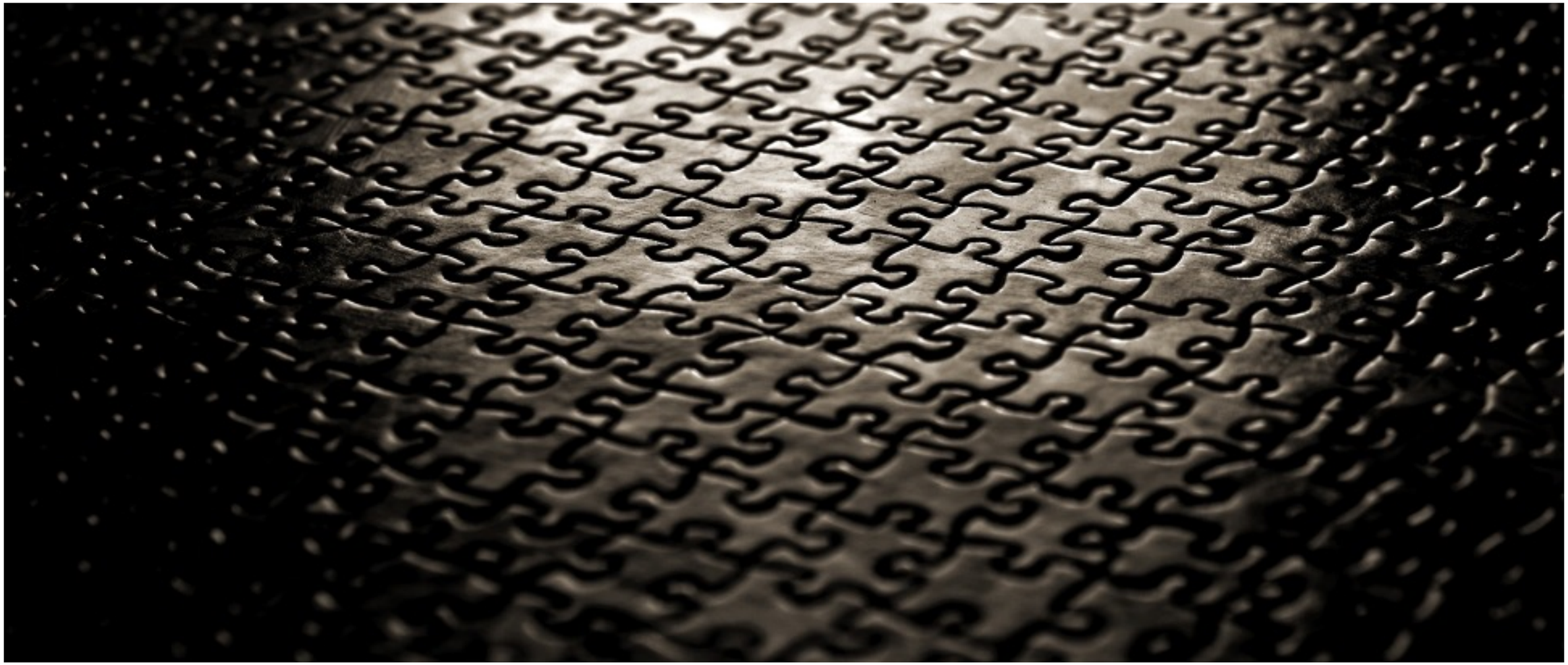


Aligning reads: tools and theory

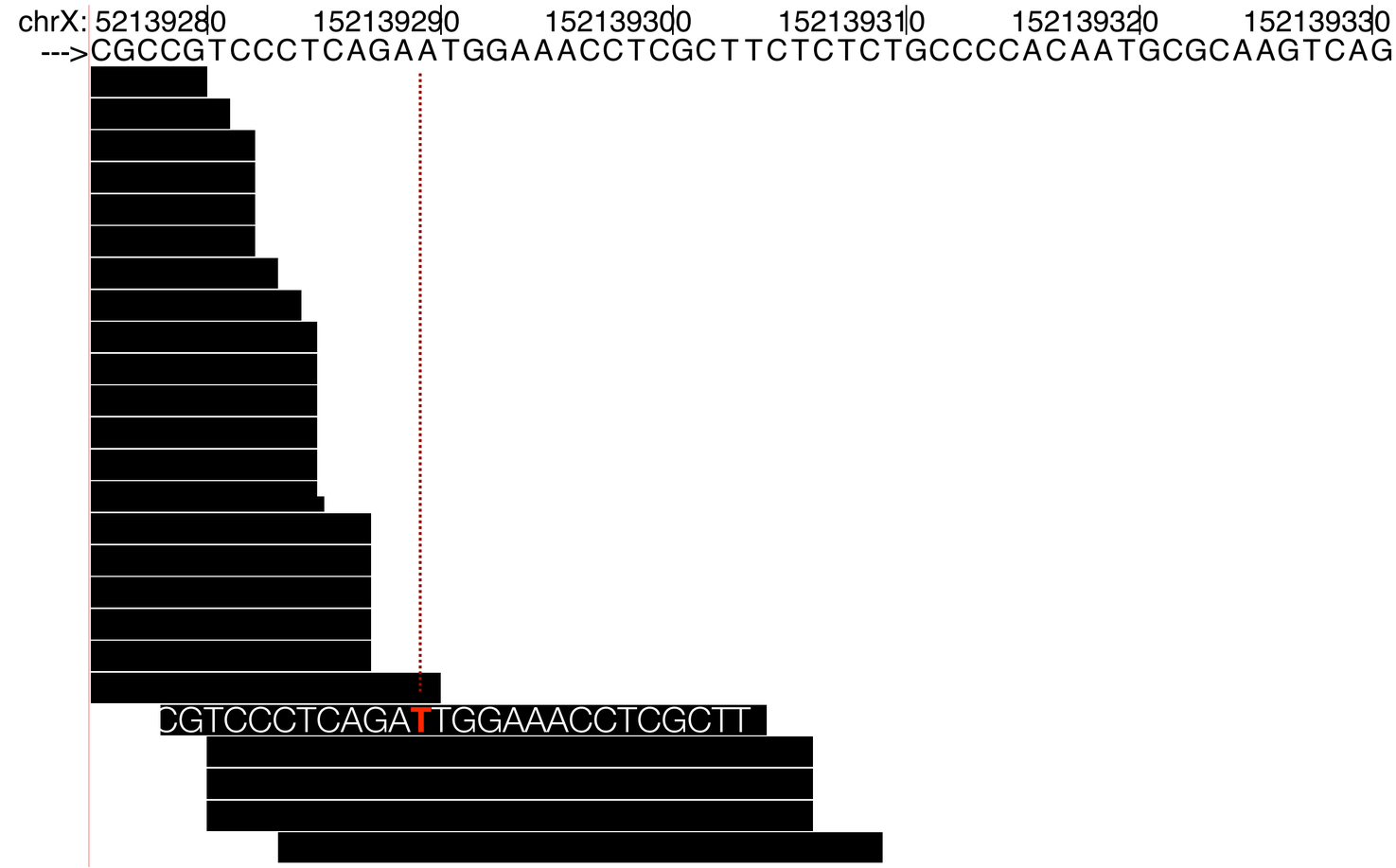
Genome

Sequence read

A simple case of string matching

Genome

Sequence reads



Difficult in practice

- Volume of data: ~3 Gbp
- ~50% of genome is **repeat regions** that cannot be covered by reads
  - Simple repeats, tandem, interspersed
  - Transposons
  - Segmental duplications where mapping is unclear
- **Gaps or unfinished regions**
  - peri-centromere, sub-telomere
  - ~5Mb unique to ethnic groups (e.g., African, Asian)

Challenges:
Human genome is large and complex

- Short reads: 50-150 bp (versus a very long reference)
  - Non-unique alignment
  - Sensitive to sequencing errors
- Massive amount of short reads: one lane produces ≥ 150 million 100 nucleotide reads
- Small insert size: 200-500 bp libraries

Challenges: short read NGS data

Reference	`ATCTCCATAGGACTAGAAGTAG`
Substitution	`ATCTCCATAGCACTAGAAGTAG`
Deletion	`ATCTCCATAGGAC-AGAAGTAG`
Insertion	`ATCTCCATAGGACTAGAAGTTAG`
3bp deletion	`ATCTC--AGGACTAGAAGTAG`

Challenges: non-exact matching

Local alignment vs Global alignment

- **Local alignment** matches the query with a *substring* (k-mer) of the reference
  - Tailored towards finding *regions of highly similar sequence* and aligning around those by working outwards to align the rest

Local Alignment

```
5' ACTACTAGATTACTTACGGATCAGGTACTTTAGAGGCTTGCAACCA 3'
           |||| |||||| ||||||||||||||||
      5' TACTCACGGATGAGGTACTTTAGAGGC 3'
```

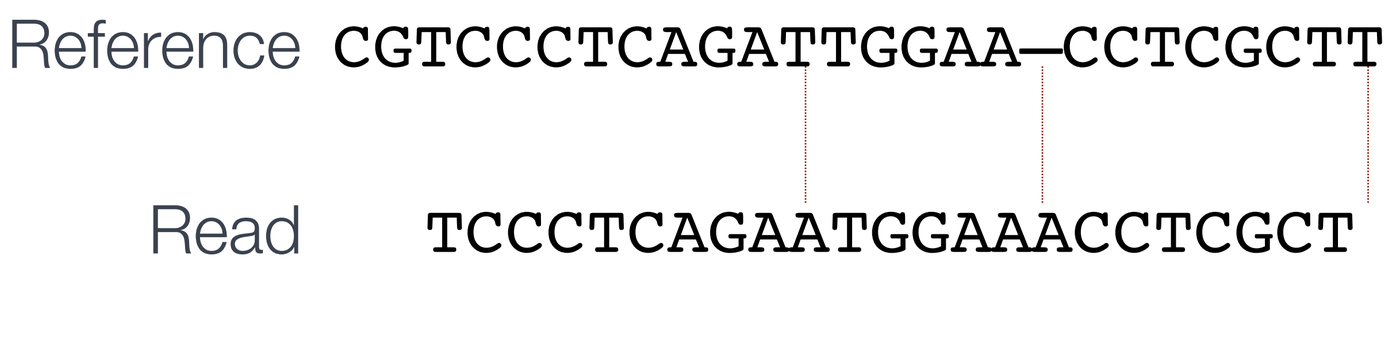
Global Alignment

```
5' ACTACTAGATTACTTACGGATCAGGTACTTTAGAGGCTTGCAACCA 3'
   ||||||||||    |||||||  |||||||||||||| |||||||
5' ACTACTAGATT----ACGGATC--GTACTTTAGAGGCTAGCAACCA 3'
```

- A **global alignment** performs end-to-end alignment between the query and the reference

Reference	`ATCTCCATAGGACTAGAAGTAG`
Substitution	`ATCTCCATAGCACTAGAAGTAG`
Deletion	`ATCTCCATAGGAC-AGAAGTAG`
Insertion	`ATCTCCATAGGACTAGAAGTTAG`
3bp deletion	`ATCTC--AGGACTAGAAGTAG`

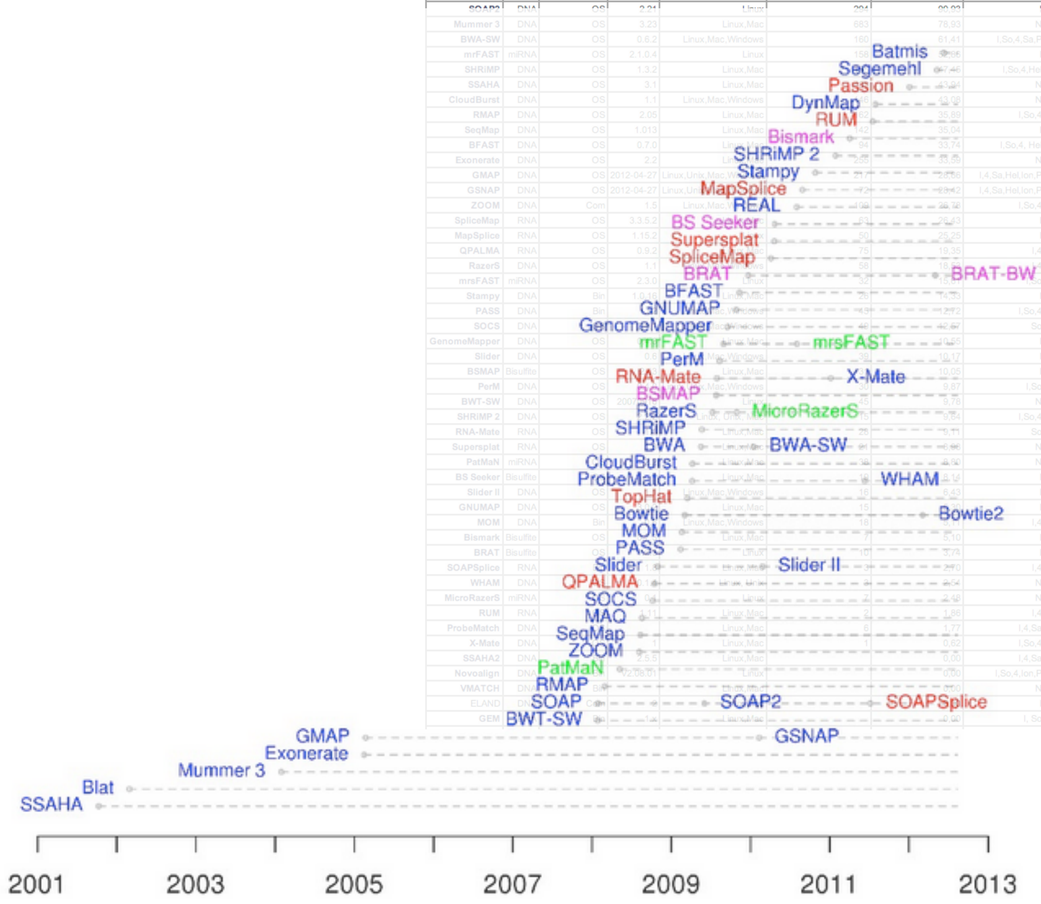
General concepts: edit distance



Edit distance =3

General concepts: edit distance

Mapper	Data	Availability	Version	O.S.	Number Citations	Citations/Years	Seq.Plat.	Input	Output	Min. RL	Max. RL	Mismatches	Indels	Gaps	Align. Reported	Alignment	Parallel	QA	PE	Splicing
Bowtie	DNA	OS	0.12.7	Linux,Mac,Windows	1168	335,04	I,So,4,Sa,P	(C)FAST(A/Q)	SAM TSV	4	1K	Score	Score	N	A,B,R,S	G L	SM	Y	Y	N
Blat	DNA	OS	34	Linux,Mac	2844	268,37	N	FASTA	TSV BLAST	11	5000K	Score	Score	Y	B	L	N	N	N	De novo
MAQ	DNA	OS	0.7.1	Linux,Mac	957	237,27	I,So	(C)FAST(A/Q)	TSV	8	63	Y	Y	N			N	Y	Y	N
BWA	DNA	OS	0.6.2	Linux,Mac,Windows	738	225,15	I,So,4,Sa,P	FASTA/Q	SAM	4	200	Y	8	Y	R,S	G	SM	Y	Y	N
TopHat	RNA	OS	1.4.1	Linux,Mac	389	112,66	I	FASTA/Q, GFF	BAM	-	-	2	0	N	B,S	-	SM	Y	Y	De novo
SOAP	DNA	OS	1.11	Linux,Mac	451	98,04	I	FASTA/Q	TSV	7	60	5	3	N	B,R S		SM	N	Y	N
SOAP3	DNA	OS	2.21	Linux	204	90,93	I	FASTA/Q	SAM TSV	27	1K	2	0	Y	A,B,R	L	SM	N	Y	N
Mummer 3	DNA	OS	3.23	Linux,Mac	683	78,93	N	FASTA	TSV	10	*	Y	Y	Y	A,B	G	N	N	N	N
BWA-SW	DNA	OS	0.6.2	Linux,Mac,Windows	160	61,41	I,So,4,Sa,P	FASTA/Q	SAM	4	1000K	0.1	0.1	Y	R,S	L	SM	Y	N	N
mrFAST	mRNA	OS	2.1.0.4	Linux	156	52,00	I	FASTA/Q	SAM	25	300	Score	6	N	A,B	G	N	N	Y	N
SHRiMP	DNA	OS	1.3.2	Linux,Mac	140	46,67	I,So,4,Hel	(C)FAST(A/Q)	TSV	14	1K	Score	Score	Y	B,S	G	SM	N	Y	N
SSAHA	DNA	OS	3.1	Linux,Mac	134	44,67	N	FASTA/Q	TSV	15	*	Y	Y	Y	B,S	G L	N	N	N	N
CloudBurst	DNA	OS	1.1	Linux,Mac,Windows	132	44,00	N	FASTA	TSV		1K	Y	Y	Y	A,B	G	Cloud	N	N	N
RMAP	DNA	OS	2.05	Linux,Mac	107	35,89	I,So,4	(C)FAST(A/Q)	BED	11	10K	Y	0	N	B,S		N	Y	Y	N
SeqMap	DNA	OS	1.013	Linux,Mac	102	35,04	I	FASTA	ELAND	15	500	5	3	N	A		SM	N	N	N
BFAST	DNA	OS	0.7.0	Linux,Mac,Windows	94	33,74	I,So,4, Hel	(C)FAST(A/Q)	SAM TSV		*	Y	Y	Y	B,R,U	G	SM	N	Y	N
Exonerate	DNA	OS	2.2	Linux,Mac,Windows	94	33,74	N	FASTA	TSV	20	*	Score	Score	Y	B,S	G L	N	N	N	De novo
GMAP	DNA	OS	2012-04-27	Linux,Unix,Mac,Windows	94	33,74	I,4,Sa,Hel,Ion,P	FASTA/Q	SAM, GFF	8	*	Y	Y	Y	B	G L	SM	N	N	De novo
GSNAP	DNA	OS	2012-04-27	Linux,Unix,Mac,Windows	94	33,74	I,4,Sa,Hel,Ion,P	FASTA/Q	SAM	8	250	Y	Y	Y	A,B,U,S	G L	SM	N	Y	Lib and de novo
ZOOM	DNA	Com	1.5	Linux,Mac	92	33,04	I,So,4	(C)FAST(A/Q)	SAM BED GFF	12	240	Y	Y	N	B,U,S	G	SM/DM	Y	Y	N
SpliceMap	RNA	OS	3.3.5.2	Linux,Mac	87	31,23	I	FASTA/Q	SAM BED	-	-	0.1		Y	A	-	SM	N	Y	Lib and/or de novo
MapSplice	RNA	OS	1.15.2	Linux,Mac	84	30,15	I	FASTA/Q	SAM BED	-	-	3		Y	B	-	SM	N	Y	De novo
QPALMA	RNA	OS	0.9.2	Linux,Mac	82	29,44	I,4	Specific	TSV	-	-	Y	Y	Y	B	L	N	Y	N	Lib and de novo
RazerS	DNA	OS	1.1	Linux,Mac,Windows	80	28,72	I	FASTQ	TSV ELAND	11	*	Score	Score	Y	A,B,S	G	N	N	Y	N
mrsFAST	mRNA	OS	2.3.0	Linux	79	28,36	I	FASTA/Q	SAM	25	200	Y	0	N	A	G	N	N	Y	N
Stampy	DNA	Bin	1.0.18	Linux,Mac,Windows	78	28,00	I	FASTA/Q	SAM TSV	4	4K	0.15	30	N	B,R,S	G	N	Y	Y	N
PASS	DNA			Linux,Mac,Windows	74	26,57	I,So,4	(C)FAST(A/Q)	SAM GFF3 BLAST	23	1K	Y	Y	Y	A,B	G	SM	Y	Y	De novo
SOCS	DNA			Linux,Mac,Windows	73	26,21	So	(C)FAST(A/Q)	TSV		64	Y	0	N	A,B		SM	Y	N	N
GenomeMapper	DNA	OS	0.4.0	Linux,Mac,Windows	68	24,42	I	FASTA/Q	BED TSV	12	2K	10	10	Y	A,B,R	G	SM	N	N	N
Slider	DNA	OS	0.9	Linux,Mac,Windows	64	22,97	I	Illumina	TSV		62	3	0	N	B,S		N	Y	Y	N
BSMAP	Bisulfite	OS	2.10	Linux,Mac	61	10,17	I	FASTA/Q	SAM TSV	8	144	15	0	N	B,U,S		SM	N	Y	N
PerM	DNA	OS	0.3.4	Linux,Mac,Windows	59	10,05	I	(C)FAST(A/Q)	SAM TSV	20	128	9	0	Y	A,U	G	DM	Y	Y	N
BWT-SW	DNA	OS	20100601	Linux,Mac	58	9,87	I,So	FASTA	TSV		1K	Score	Score	Y	A		N	N	N	N
SHRiMP 2	DNA	OS	2.0.1	Linux,Mac	57	9,78	N	FASTA	SAM	30	1K	Y	Score	N	B,U,S	G	SM	Y	Y	N
RNA-Mate	RNA	OS	1.0.4	Linux,Mac	57	9,78	I,So,4	FASTA/Q	BED Counts	-	-	Y	0	N	S	-	DM	Y	N	Lib
Supersplat	RNA	OS	1.0.1	Linux,Mac	57	9,78	So	CFASTA	TSV			0	0	N	A,U	G	N	N	N	N
PatMaN	mRNA	OS	1.0.1	Linux,Mac	56	9,61	N	FASTA	TSV	1	*	Y	Y	N	A	G	N	N	N	De novo
BS Seeker	Bisulfite	OS	2.0.4	Linux,Mac	54	9,27	N	FASTA/Q	SAM	-	-	3	0	N	U	-	SM	Y	N	N
Slider II	DNA	OS	0.9.43	Linux,Mac,Windows	54	9,27	I	Illumina	TSV		93	Y		N	B,S		N	N	Y	N
GNUMAP	DNA	OS	2.4.1	Linux,Mac	51	8,76	I	FASTA/Q Illumina	SAM TSV	16	1K	Score	Score	Y	B	G	SM/DM	Y	N	N
MOM	DNA	Bin	2.0	Linux,Mac,Windows	49	8,41	I,4	FASTA	TSV			Y	0	N	A	L	SM	N	Y	N
Bismark	Bisulfite	OS	0.5.0	Linux,Mac	48	8,24	I	FASTA/Q	SAM	16	10K	Score	Score	N	U	-	SM	Y	Y	N
BRAT	Bisulfite	OS	1.2.1	Linux	48	8,24	I	FASTA/Q	TSV			Y	0	N			N	N	Y	N
SOAPSplice	RNA	OS	1.9	Linux,Mac	46	7,90	I,4	FASTA/Q	TSV	13	3K	5	2	Y	U	-	SM	Y	Y	De novo
WHAM	DNA	OS	1.0.4	Linux,Mac	44	7,56	N	FASTQ	SAM	5	128	5	3	N	A,B,R,U,S	G	N	Y	Y	De novo
MicroRazerS	mRNA	OS	1.0.1	Linux,Mac,Windows	43	7,39	N	FASTA	SAM TSV	10	*	Score	0	N	S	G	N	N	N	N
RUM	RNA	OS	1.08	Linux,Mac	34	5,84	I,4	FASTA/Q	SAM TSV BED	-	-	Y	Y	Y	B	-	SM	N	Y	De novo
ProbeMatch	DNA	OS	1.1.1	Linux,Mac	31	5,33	I,4,Sa	FASTA	ELAND	36	50	3	Y	N	A,B		N	N	N	N
X-Mate	DNA	OS	1.0	Linux,Mac	31	5,33	I,So,4	(C)FAST(A/Q)	SAM BED Counts	-	-	Y	0	N	S	-	DM	Y	N	Lib
SSAHA2	DNA	OS	2.5.3	Linux,Mac	31	5,33	I,4,Sa	FASTA/Q	SAM	15	48K	Score	Score	N	B,S	L	N	N	Y	N
Novoalign	DNA	Com	V2.08.01	Linux,Mac,Windows	30	5,16	I,So,4,Ion,P	(C)FAST(A/Q) Illumina	SAM TSV	30	300	8	2	N	A,B,R,U,S	G	SM/DM/Cloud	Y	Y	Lib
VMATCH	DNA	OS	2.2.5.1	Linux,Mac,Windows	29	4,99	N	FASTA	TSV			Score	Score	Y	A,B,S	G L	N	N	N	N
ELAND	DNA	Com		Linux,Mac,Windows	28	4,82	I	FASTA	TSV		32	2	0	N	B		N	N	N	N
GEM	DNA	OS	1.3	Linux,Mac	28	4,82	I, So	FASTA/Q	SAM, Counts		*	Y	Y	Y	A,S	G	SM	Y	Y	Lib and de novo



Short-read aligners: choices

http://wwwdev.ebi.ac.uk/fg/hts_mappers/

Building an index

- Having an index of the reference genome provides an efficient way to search
- Once index is built, it can be queried any number of times
- Indexes are genome and tool-specific
- Different types of indices (i.e hash-tables, suffix arrays, Burrows-Wheeler Transform)

- Bowtie2: indexes with an FM Index to keep low memory footprint. Supports gapped, local and paired-end alignment
- BWA: indexes with the Burrows-Wheeler Transform (BWT). Has three algorithms for varying read lengths.
- SOAP2: uses a 2-way BWT for indexing. Fast and accurate alignment of Illumina sequencing reads. Not open source.
- MAQ: first aligns reads to reference sequences and then calls the consensus. Designed for Illumina reads.

Commonly used aligners for ChIP-seq

Alignment considerations for ChIP-seq

- Percentage of uniquely mapped reads
    - 70% or higher is good. 50% or lower is concerning.
    - Percentages vary between organisms
- Duplicates
    - Usually due to over-amplification or short read length
- Multi-mappers
    - Reduce by setting allowable mismatches according to sequencing platform
    - For situations when protein binds repetitive DNA, use paired-end sequencing

Keep only uniquely mapped reads for peak calling!