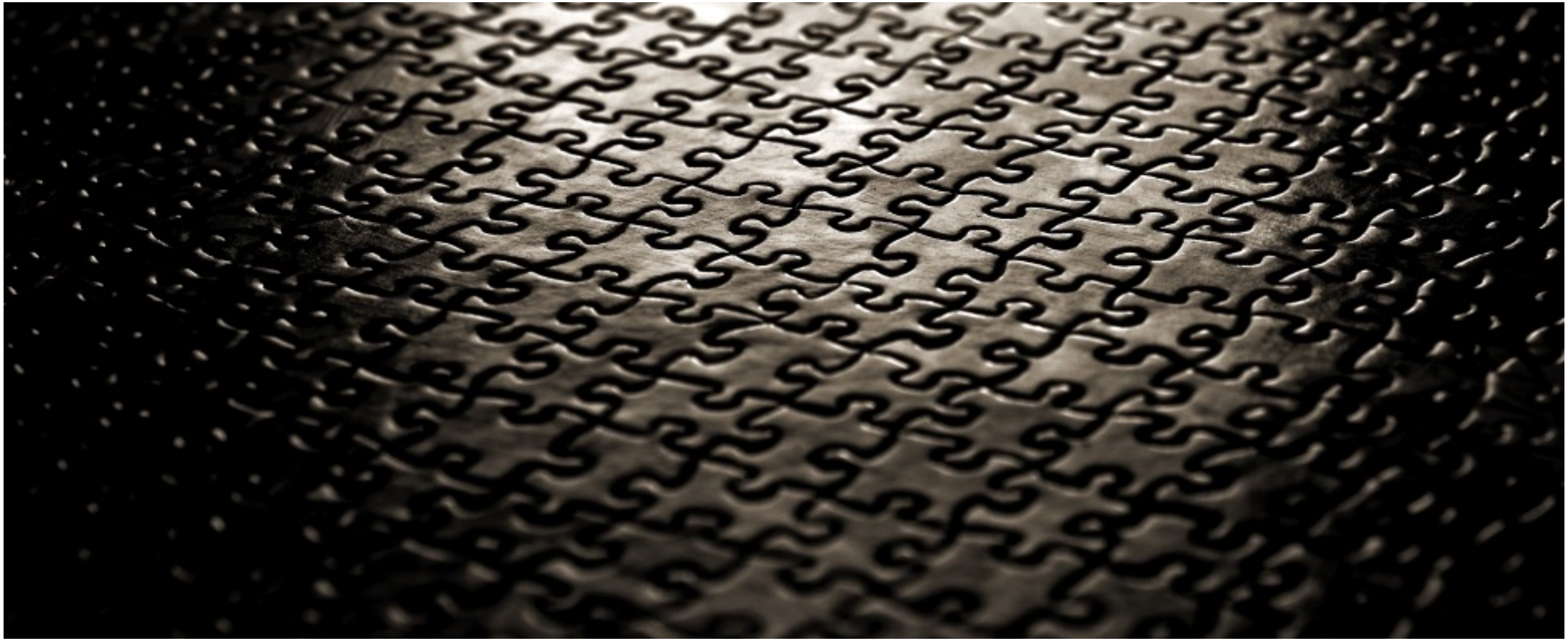
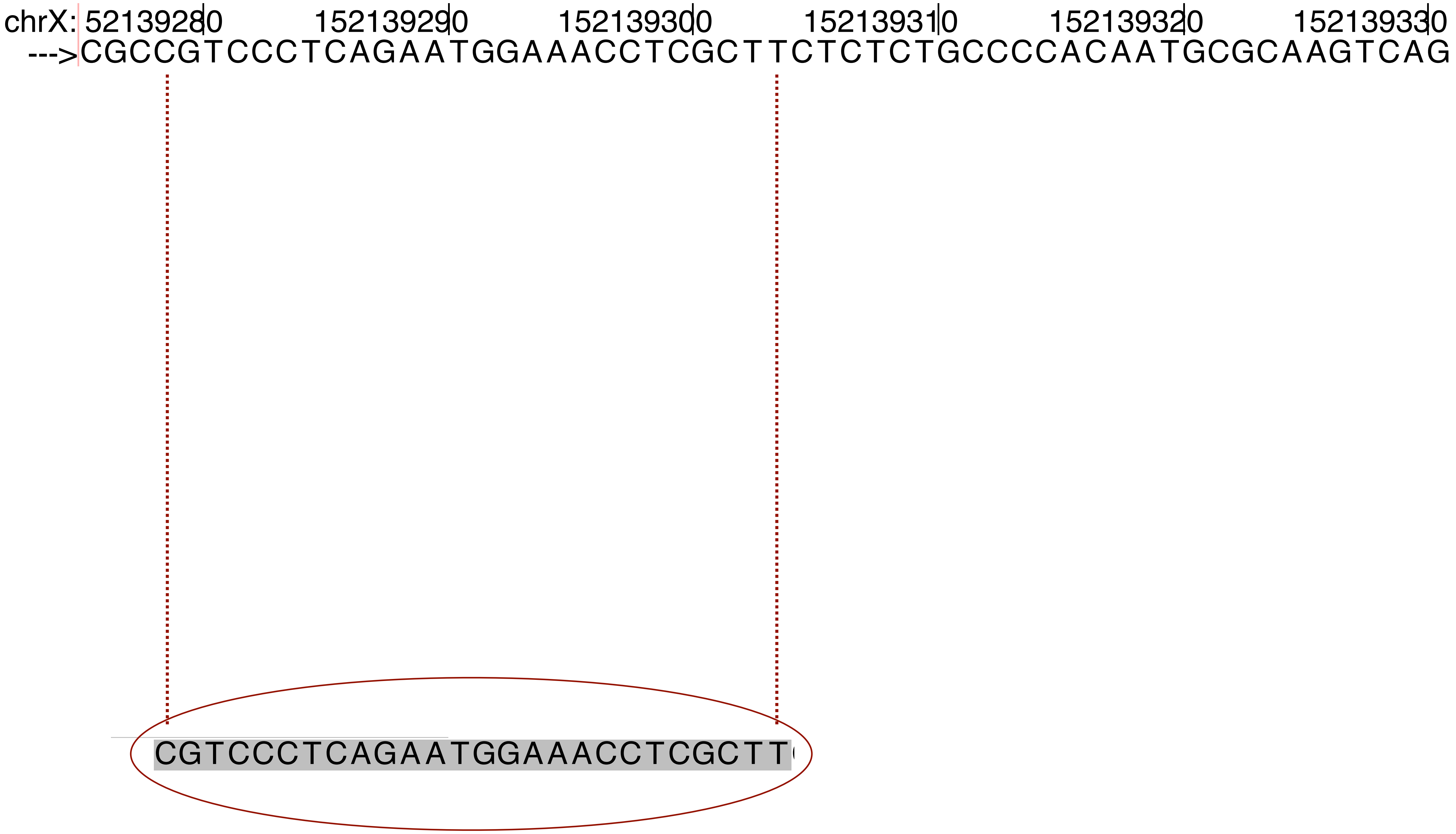


Aligning reads: tools and theory

Genome



chrX: 52139280 152139290 152139300 152139310 152139320 152139330

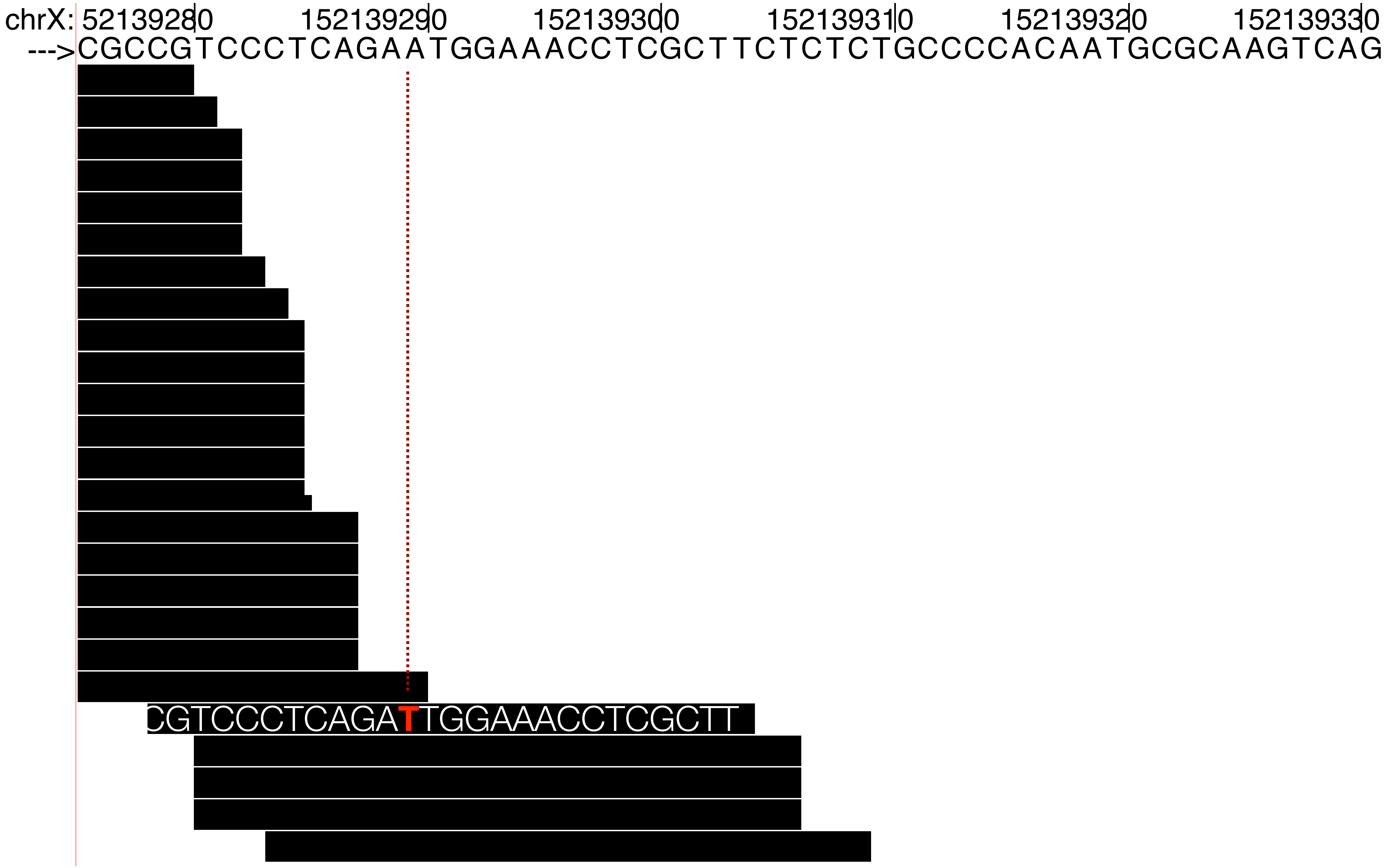
--->CGCCGTCCCTCAGAATGGAAACCTCGCTTCTCTCTGCCCCACAATGCGCAAGTCAG

Sequence read

CGTCCCTCAGAATGGAAACCTCGCTT

A simple case of string matching

Genome


chrX: 52139280
152139290
152139300
152139310
152139320
152139330
--->CGCCGTCCCTCAGAATGGAAACCTCGCTTCTCTCTGCCCCACAATGCGCAAGTCAG
CGTCCCTCAGATTGGAAACCTCGCTT

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

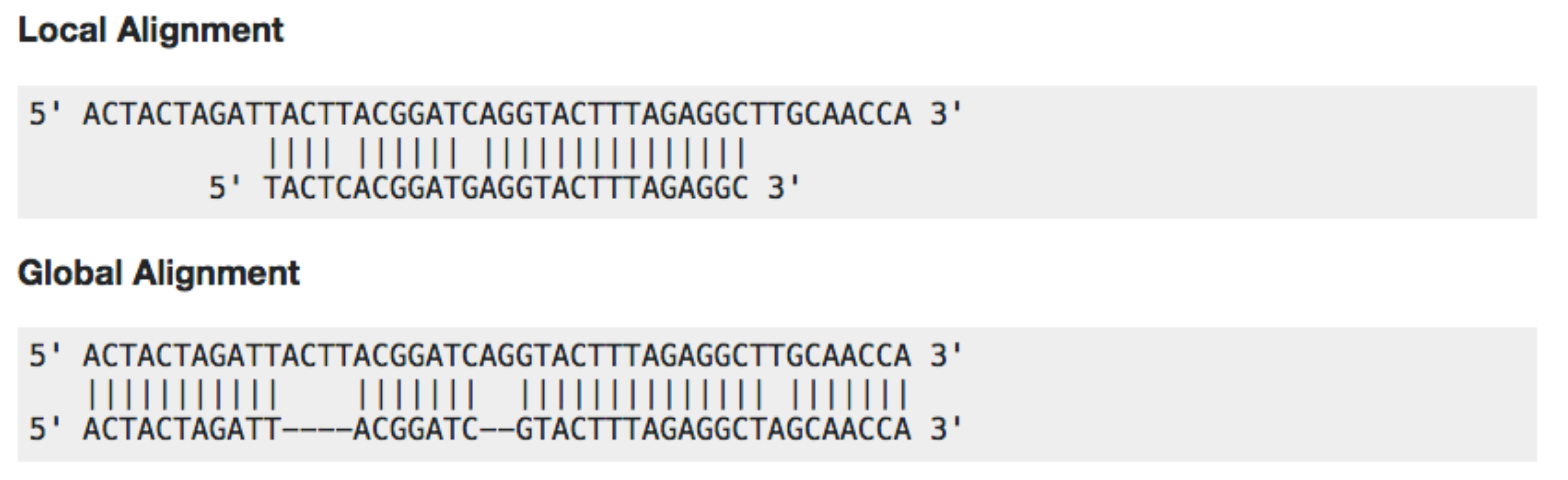
Challenges: short read NGS data

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

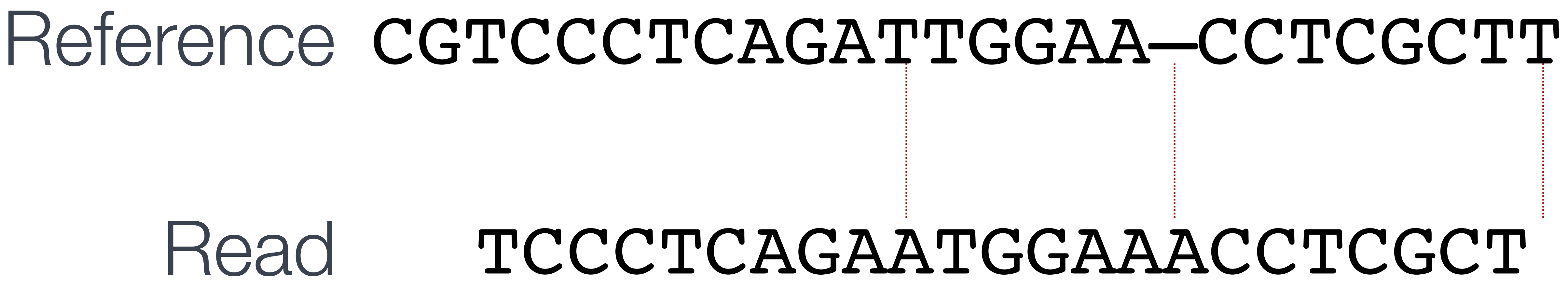
```
5' ACTACTAGATTACTTACGGATCAGGTACTTTAGAGGCTTGCAACCA 3'
   |||||||||||    |||||||  |||||||||||||| |||||||
5' ACTACTAGATT----ACGGATC--GTACTTTAGAGGCTAGCAACCA 3'
```

- 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

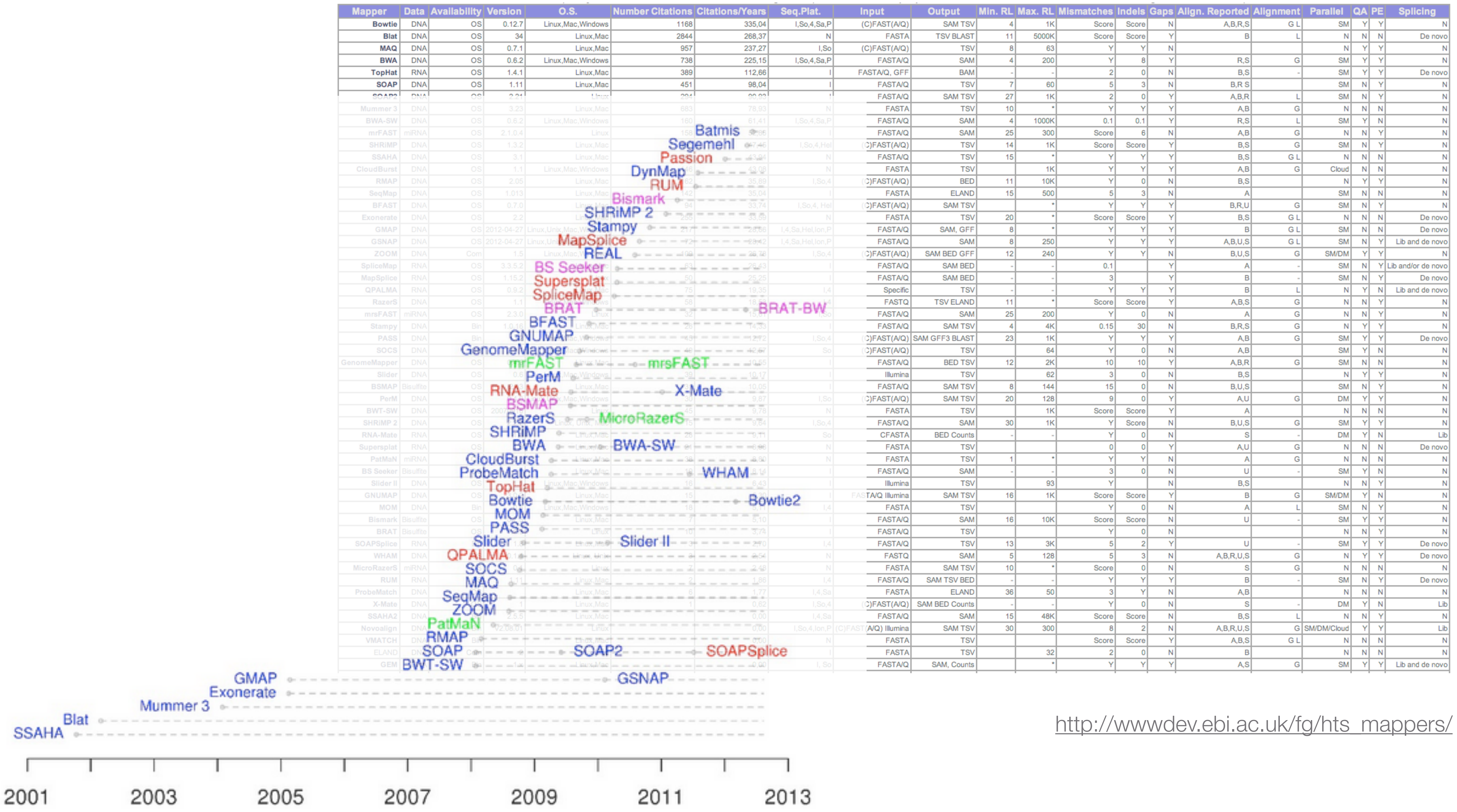
General concepts: edit distance



```
Reference  CGTCCCTCAGATTGGAA—CCTCGCTT

Read         TCCCTCAGAATGGAAACCTCGCT
```

Edit distance =3



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
SOAP2	DNA	OS	2.21	Linux	294	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	miRNA	OS	2.1.0.4	Linux	158	52,86	I	FASTA/Q	SAM	25	300	Score	6	N	A,B	G	N	N	Y	N
SHRiMP	DNA	OS	1.3.2	Linux,Mac	[illegible]	47,46	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	[illegible]	43,94	N	FASTA/Q	TSV	15	*	Y	Y	Y	B,S	G L	N	N	N	N
CloudBurst	DNA	OS	1.1	Linux,Mac,Windows	[illegible]	43,08	N	FASTA	TSV		1K	Y	Y	Y	A,B	G	Cloud	N	N	N
RMAP	DNA	OS	2.05	Linux,Mac	[illegible]	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	142	35,04	I	FASTA	ELAND	15	500	5	3	N	A		SM	N	N	N
BFAST	DNA	OS	0.7.0	Linux,Mac	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	255	33,59	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	217	28,66	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	72	28,42	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,Windows	109	26,78	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	[illegible]	63	26,43	I	FASTA/Q	SAM BED	-	-	0.1		Y	A	-	SM	N	Y	Lib and/or de novo
MapSplice	RNA	OS	1.15.2	Linux	50	25,25	I	FASTA/Q	SAM BED	-	-	3		Y	B	-	SM	N	Y	De novo
QPALMA	RNA	OS	0.9.2	Linux,Mac	75	19,35	I,4	Specific	TSV	-	-	Y	Y	Y	B	L	N	Y	N	Lib and de novo
RazerS	DNA	OS	1.1	Linux,Mac,Windows	58	18,53	I,4	FASTQ	TSV ELAND	11	*	Score	Score	Y	A,B,S	G	N	N	Y	N
mrsFAST	miRNA	OS	2.3.0	Linux	32	15,61	I,So	FASTA/Q	SAM	25	200	Y	0	N	A	G	N	N	Y	N
Stampy	DNA	Bin	1.0.16	Linux,Mac	26	14,33	I	FASTA/Q	SAM TSV	4	4K	0.15	30	N	B,R,S	G	N	Y	Y	N
PASS	DNA	Bin	[illegible]	Linux,Mac,Windows	43	12,72	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	OS	[illegible]	[illegible]	40	12,67	So	(C)FAST(A/Q)	TSV		64	Y	0	N	A,B		SM	Y	N	N
GenomeMapper	DNA	OS	[illegible]	Linux,Mac	[illegible]	10,55	I	FASTA/Q	BED TSV	12	2K	10	10	Y	A,B,R	G	SM	N	N	N
Slider	DNA	OS	0.6	Linux,Mac,Windows	39	10,17	I	Illumina	TSV		62	3	0	N	B,S		N	Y	Y	N
BSMAP	Bisulfite	OS	[illegible]	Linux,Mac	[illegible]	10,05	I	FASTA/Q	SAM TSV	8	144	15	0	N	B,U,S		SM	N	Y	N
PerM	DNA	OS	[illegible]	Linux,Unix,Mac,Windows	30	9,87	I,So	(C)FAST(A/Q)	SAM TSV	20	128	9	0	Y	A,U	G	DM	Y	Y	N
BWT-SW	DNA	OS	[illegible]	Linux	45	9,78	N	FASTA	TSV		1K	Score	Score	Y	A		N	N	N	N
SHRiMP 2	DNA	OS	[illegible]	Linux, Unix, Mac	15	9,64	I,So,4	FASTA/Q	SAM	30	1K	Y	Score	N	B,U,S	G	SM	Y	Y	N
RNA-Mate	RNA	OS	[illegible]	Linux,Mac	28	9,11	So	CFASTA	BED Counts	-	-	Y	0	N	S	-	DM	Y	N	Lib
Supersplat	RNA	OS	[illegible]	Linux,Mac	21	8,98	N	FASTA	TSV			0	0	Y	A,U	G	N	N	N	De novo
PatMaN	miRNA	[illegible]	[illegible]	Linux,Mac	38	8,80	N	FASTA	TSV	1	*	Y	Y	N	A	G	N	N	N	N
BS Seeker	Bisulfite	[illegible]	[illegible]	Linux,Mac	[illegible]	8,14	I	FASTA/Q	SAM	-	-	3	0	N	U	-	SM	Y	N	N
Slider II	DNA	OS	[illegible]	Linux,Mac,Windows	16	6,43	I	Illumina	TSV		93	Y		N	B,S		N	N	Y	N
GNUMAP	DNA	OS	[illegible]	Linux,Mac	15	6,30	I	FASTA/Q Illumina	SAM TSV	16	1K	Score	Score	Y	B	G	SM/DM	Y	N	N
MOM	DNA	Bin	[illegible]	Linux,Mac,Windows	18	5,11	I,4	FASTA	TSV			Y	0	N	A	L	SM	N	Y	N
Bismark	Bisulfite	OS	[illegible]	Linux,Mac	7	5,10	I	FASTA/Q	SAM	16	10K	Score	Score	N	U	-	SM	Y	Y	N
BRAT	Bisulfite	OS	[illegible]	Linux	10	3,74	I	FASTA/Q	TSV			Y	0	N			N	N	Y	N
SOAPSplice	RNA	[illegible]	1.8	Linux,Mac	3	2,70	I,4	FASTA/Q	TSV	13	3K	5	2	Y	U	-	SM	Y	Y	De novo
WHAM	DNA	[illegible]	[illegible]	Linux, Unix	3	2,54	N	FASTQ	SAM	5	128	5	3	N	A,B,R,U,S	G	N	Y	Y	De novo
MicroRazerS	miRNA	[illegible]	[illegible]	Linux	7	2,48	N	FASTA	SAM TSV	10	*	Score	0	N	S	G	N	N	N	N
RUM	RNA	[illegible]	1.11	Linux,Mac	2	1,86	I,4	FASTA/Q	SAM TSV BED	-	-	Y	Y	Y	B	-	SM	N	Y	De novo
ProbeMatch	DNA	[illegible]	[illegible]	Linux,Mac	6	1,77	I,4,Sa	FASTA	ELAND	36	50	3	Y	N	A,B		N	N	N	N
X-Mate	DNA	[illegible]	1	Linux,Mac	1	0,62	I,So,4	(C)FAST(A/Q)	SAM BED Counts	-	-	Y	0	N	S	-	DM	Y	N	Lib
SSAHA2	DNA	[illegible]	2.5.5	Linux,Mac		0,00	I,4,Sa	FASTA/Q	SAM	15	48K	Score	Score	N	B,S	L	N	N	Y	N
Novoalign	DNA	Bin	V2.08.01	Linux		0,00	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	Bin	[illegible]	Linux,Mac		0,00	N	FASTA	TSV			Score	Score	Y	A,B,S	G L	N	N	N	N
ELAND	DNA	Com	2				I	FASTA	TSV		32	2	0	N	B		N	N	N	N
GEM	DNA	Bin	1.x	Linux,Mac		0,00	I, So	FASTA/Q	SAM, Counts		*	Y	Y	Y	A,S	G	SM	Y	Y	Lib and de novo

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

Short-read aligners: choices

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)

Commonly used aligners for ChIP-seq

- 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.

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!

These materials have been developed by members of the teaching team at the Harvard Chan Bioinformatics Core (HBC). These are open access materials distributed under the terms of the Creative Commons Attribution license (CC BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.