AMC Graduate School
Bioinformatics Sequence Analysis
Alignment of nextgen sequences

Barbera van Schaik

Bioinformatics and Biomedical Computing
Epidemiology and Data Science
Amsterdam UMC

b.d.vanschaik@amsterdamumc.nl

March 09, 2021

NGS aligners

This presentation is based on the following review

BIOINFORMATICS REVIEW

Vol. 28 no. 24 2012, pages 3169–3177
doi:10.1093/bioinformatics/bts605

Sequence analysis

Advance Access publication October 11, 2012

Tools for mapping high-throughput sequencing data

Nuno A. Fonseca*, Johan Rung, Alvis Brazma and John C. Marioni

EMBL Outstation, European Bioinformatics Institute (EBI), Hinxton, Cambridge CB10 ISD, UK

Associate Editor: Jonathan Wren

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

The problem

- 100 million - 1 billion reads in 7-15 days
- Sequence reads are relatively short
- Previous alignment methods too slow
- Exploit technological developments (different type of sequencers)
- Support different protocols (single-end, paired-end)

Goal of a sequence aligner

Formal description

- given a set of sequences Q
- a set of reference sequences R
- a set of constraints and a distance threshold K
- find all substrings m of R that follow the constraints and are within a distance k to a sequence q in Q
- $d(q,m)$ ¡= k, where $d()$ is a distance function
- Occurrences of m in R are called *matches*

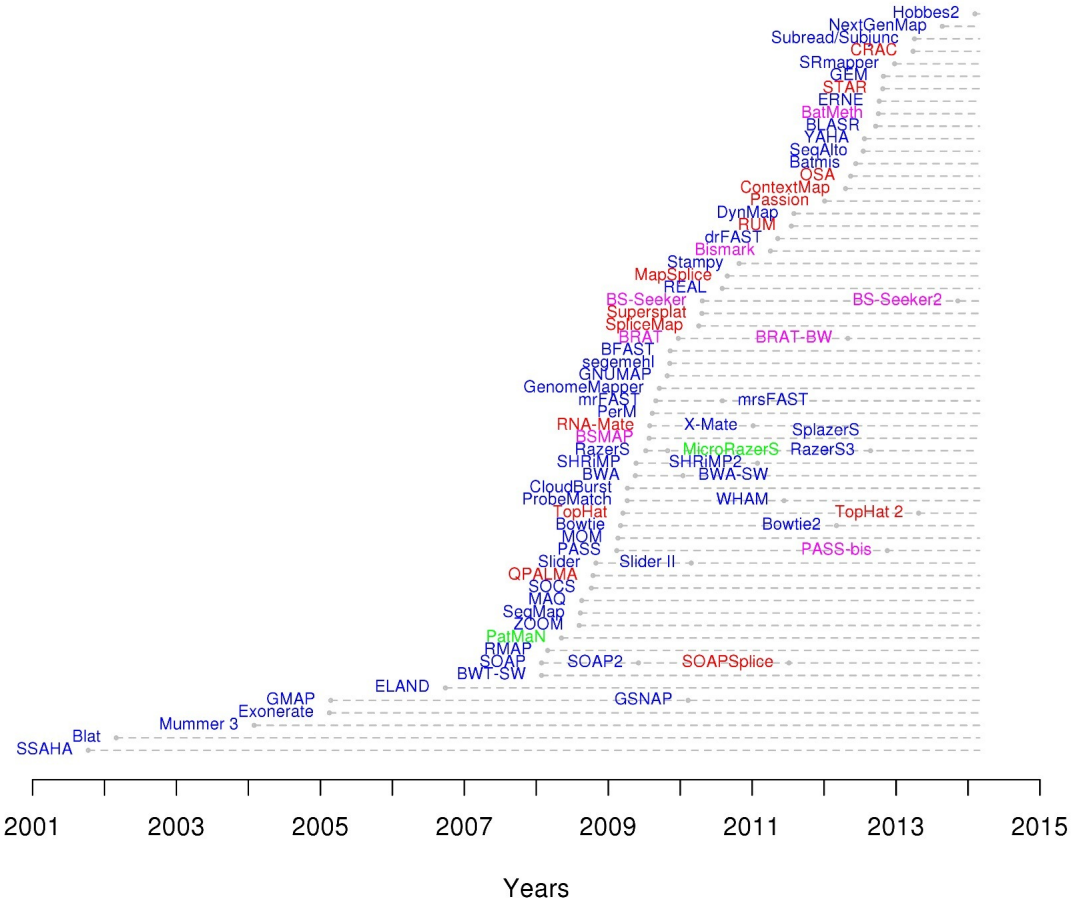
Goal of a sequence aligner

In normal English

- The reads need to be aligned to a reference dataset
- Find the true location of a read in the reference
- Allow for errors and structural variation (in-exact matching)

NGS sequence aligners

Time line



More than 60 mappers are available, most of them developed after 2008 Fonseca et al (2012)

Which one to use for your project?

Starting point

- What is the NGS application? (DNA, RNA, miRNA, bi-sulfite)
- What is the data type? (single, paired-end)
- Which sequencer brand was used?
- How long should the analysis take?
- How accurate does the alignment need to be?
- Can the output be used in follow-up analysis steps?

Overview 1/2

Data

Input format

Open source / commercial

Reference

Sequencing platform

Output format

Times cited

Table 1. List of mappers

Mapper	Data	Seq.Plat.	Input	Output	Avail.	Version	Cit.	Citations/Years	Reference
BFAST	DNA	I,So,4, Hel	(C)FAST(A/Q)	SAM TSV	OS	0.7.0	94	37.11	Homer *et al.* (2009)
Bismark	Bisulphite	I	FASTA/Q	SAM	OS	0.7.3	7	6.21	Krueger and Andrews (2011)
BLAT	DNA	N	FASTA	TSV BLAST	OS	34	2844	275.67	Kent (2002)
Bowtie	DNA	I,So,4,Sa,P	(C)FAST(A/Q)	SAM TSV	OS	0.12.7	1168	363.42	Langmead *et al.* (2009)
Bowtie2	DNA	I,4,Ion	FASTA/Q	SAM TSV	OS	2.0beta5		0.00	Langmead and Salzberg (2012)
BS Seeker	Bisulphite	I	FASTA/Q	SAM	OS		19	9.26	Chen *et al.* (2010)
BSMAP	Bisulphite	I	FASTA/Q	SAM TSV	OS	2.43	31	11.06	Xi and Li (2009)
BWA	DNA	I,So,4,Sa,P	FASTA/Q	SAM	OS	0.6.2	738	224.20	Li and Durbin (2009)
BWA-SW	DNA	I,So,4,Sa,P	FASTA/Q	SAM	OS	0.6.2	160	67.69	Li and Durbin (2010)
BWT-SW	DNA	N	FASTA	TSV	OS	20070916	45	10.42	Lam *et al.* (2008)
CloudBurst	DNA	N	FASTA	TSV	OS	1.1	146	46.97	Schatz (2009)
DynMap	DNA	N	FASTA	TSV	OS	0.0.20		0.00	Flouri *et al.* (2011)
ELAND	DNA	I	FASTA	TSV	Com	2	7	1.09	Unpublished[a]
Exonerate	DNA	N	FASTA	TSV	OS	2.2	255	34.69	Slater and Birney (2005)
GEM	DNA	I, So	FASTA/Q	SAM, counts	Bin	1.x	4	1.35	Unpublished[b]
GenomeMapper	DNA	I	FASTA/Q	BED TSV	OS	0.4.3	31	11.66	Schneeberger *et al.* (2009)
GMAP	DNA	I,4,Sa,Hel,Ion,P	FASTA/Q	SAM, GFF	OS	2012-04-27	217	29.52	Wu and Watanabe (2005)
GNUMAP	DNA	I	FASTA/Q Illumina	SAM TSV	OS	3.0.2	15	5.73	Clement *et al.* (2010)
GSNAP	DNA	I,4,Sa,Hel,Ion,P	FASTA/Q	SAM	OS	2012-04-27	72	31.61	Wu and Nacu (2010)
MapReads	DNA	So	FASTA/Q	TSV	OS	2.4.1		0.00	Unpublished[c]
MapSplice	RNA	I	FASTA/Q	SAM BED	OS	1.15.2	50	28.17	Wang *et al.* (2010)
MAQ	DNA	I,So	(C)FAST(A/Q)	TSV	OS	0.7.1	957	251.66	Li *et al.* (2008a)
MicroRazerS	miRNA	N	FASTA	SAM TSV	OS	0.1	7	2.75	Emde *et al.* (2010)
MOM	DNA	I,4	FASTA	TSV	Bin	0.6	18	5.55	Eaves and Gao (2009)
MOSAIK	DNA	I,So,4,Sa,Hel,Ion,P	(C)FAST(A/Q)	BAM	OS	2.1	4	1.18	Unpublished[d]
mrFAST	miRNA	I	FASTA/Q	SAM	OS	2.1.0.4	158	58.34	Alkan *et al.* (2009)
mrsFAST	miRNA	I,So	FASTA/Q	SAM	OS	2.3.0	32	18.03	Hach *et al.* (2010)
Mummer 3	DNA	N	FASTA	TSV	OS	3.23	683	81.58	Kurtz *et al.* (2004)
Novoalign	DNA	I,So,4,Ion,P	(C)FAST(A/Q) Illumina	SAM TSV	Bin	V2.08.01	137	34.49	Unpublished[e]
PASS	DNA	I,So,4	(C)FAST(A/Q)	SAM GFF3 BLAST	Bin	1.62	45	13.67	Campagna *et al.* (2009)
Passion	RNA	I,4,Sa,P	FASTA/Q	BED	OS	1.2.0		0.00	Zhang *et al.* (2012)
PatMaN	miRNA	N	FASTA	TSV	OS	1.2.2	38	9.36	Prüfer *et al.* (2008)
PerM	DNA	I,So	(C)FAST(A/Q)	SAM TSV	OS	0.4.0	30	10.88	Chen *et al.* (2009)
ProbeMatch	DNA	I,4,Sa	FASTA	ELAND	OS		6	1.92	Kim *et al.* (2009)

(continued)

Overview 2/2

Table 1. Continued

Mapper	Data	Seq.Plat.	Input	Output	Avail.	Version	Cit.	$\frac{Citations}{Years}$	Reference
QPALMA	RNA	I,4	Specific	TSV	OS	0.9.2	75	21.11	De Bona *et al.* (2008)
RazerS	DNA	I,4	FASTQ	TSV ELAND	OS	1.1	58	20.17	Weese *et al.* (2009)
REAL	DNA	I	FASTA/Q	TSV	OS	0.0.28		0.00	Frousios *et al.* (2010)
RMAP	DNA	I,So,4	(C)FAST(A/Q)	BED	OS	2.05	162	38.27	Smith *et al.* (2008)
RNA-Mate	RNA	So	CFASTA	BED Counts	OS	1.1	28	10.04	Cloonan *et al.* (2009)
RUM	RNA	I,4	FASTA/Q	SAM TSV BED	OS	1.11	2	2.36	Grant *et al.* (2011)
SeqMap	DNA	I	FASTA	ELAND	OS	1.013	142	37.34	Jiang and Wong (2008)
SHRiMP	DNA	I,So,4,Hel	(C)FAST(A/Q)	TSV	OS	1.3.2	155	50.91	Rumble *et al.* (2009)
SHRiMP 2	DNA	I,So,4	FASTA/Q	SAM	OS	2.2.2	15	11.76	David *et al.* (2011)
Slider	DNA	I	Illumina	TSV	OS	0.6	39	10.98	Malhis *et al.* (2009)
Slider II	DNA	I	Illumina	TSV	OS	1.1	16	7.25	Malhis and Jones (2010)
Smalt	DNA	I,4,Sa,Ion,P	FASTA/Q	SAM	OS	0.6.1		0.00	Unpublished[f]
SOAP	DNA	I	FASTA/Q	TSV	OS	1.11	451	104.41	Li *et al.* (2008b)
SOAP2	DNA	I	FASTA/Q	SAM TSV	OS	2.21	294	99.38	Li *et al.* (2009b)
SOAPSplice	RNA	I,4	FASTA/Q	TSV	Bin	1.8	3	3.54	Huang *et al.* (2011a)
SOCS	DNA	So	(C)FAST(A/Q)	TSV	OS	2.1.1	49	14.15	Ondov *et al.* (2008)
SpliceMap	RNA	I	FASTA/Q	SAM BED	OS	3.3.5.2	63	29.80	Au *et al.* (2010)
SSAHA	DNA	N	FASTA/Q	TSV	OS	3.1	483	42.29	Ning *et al.* (2001)
SSAHA2	DNA	I,4,Sa	FASTA/Q	SAM	Bin	2.5.5	483	44.99	Ning *et al.* (2001)
Stampy	DNA	I	FASTA/Q	SAM TSV	Bin	1.0.16	26	16.19	Lunter and Goodson (2011)
Supersplat	RNA	N	FASTA	TSV	OS	1.0	21	9.93	Bryant Jr *et al.* (2010)
TopHat	RNA	I	FASTA/Q, GFF	BAM	OS	1.4.1	389	121.04	Trapnell *et al.* (2009)
VMATCH	DNA	N	FASTA	TSV	Bin		26	2.75	Unpublished[g]
WHAM	DNA	N	FASTQ	SAM	OS	0.1.4	3	3.33	Li *et al.* (2011)
X-Mate	DNA	I,So,4	(C)FAST(A/Q)	SAM BED Counts	OS	1	1	0.74	Wood *et al.* (2011)
ZOOM	DNA	I,So,4	(C)FAST(A/Q)	SAM BED GFF	Com	1.5	109	28.66	Lin *et al.* (2008)

The Data column indicates whether the mapper is tailored for DNA, RNA, miRNA or bisulphite sequences. The Seq.Plat. column indicates whether the mapper natively supports reads from a specific sequencing platform (Illumina, ABI Solid, Roche 454, ABI Sanger, Helicos, Ion torrent and PacBio) or not (N). The mappers are available as open source (OS), in binary form (Bin) or commercially (Com). The Input and Output columns indicate, respectively, the file formats accepted and produced by the mappers. Input formats: FASTA, FASTQ, CFASTA, CFASTQ and Illumina sequence and probability files' format. Output formats: SAM (Li *et al.*, 2009a), Tab-separated-values (TSV), BED file format, different versions of General Feature Format (GFF), number of reads mapped to genes/exons (Counts). The Version column indicates the version of the mapper considered in this study. The table also includes the number of citations per year of the associated publication. The number of citations (Cit.) was obtained from Google Scholar on April 14, 2012.

[a]ELAND: Efficient local alignment of nucleotide data.
[b]MapReads: SOLiD System Color Space Mapping Tool.
[c]The Vmatch large scale sequence analysis software.
[d]Mosaik 1.0 documentation.
[e]www.novocraft.com.
[f]SMALT Manual.[g]GEM-GEnomic Multi-tool.

Choices based on sequencing platform

General mappers

- BLAST, BLAT, SSAHA, Exonerate, Mummer
- Align any sequence (DNA, RNA, protein)
- Note that BLAST and BLAT are too slow for large sequence experiments

Platform support

- Not every mapper supports SOLiD, these do: SOCS, RNA-Mate, MapReads, BWA, BFAST
- Illumina: Supported by most aligners, some take Illumina specific errors into account, e.g. SOAP, Bowtie, Novoalign trim low quality bases at the end of the reads

Features

Data-centric

- Read length limits
- Read pairing information
- Parallel processing

Alignment sensitivity and reporting

- Errors allowed
- Support for gaps
- Alignments reported
- Type of alignment
- Usage of read quality information

Read length limits

Short reads

e.g. miRNAs (16-30 bp)

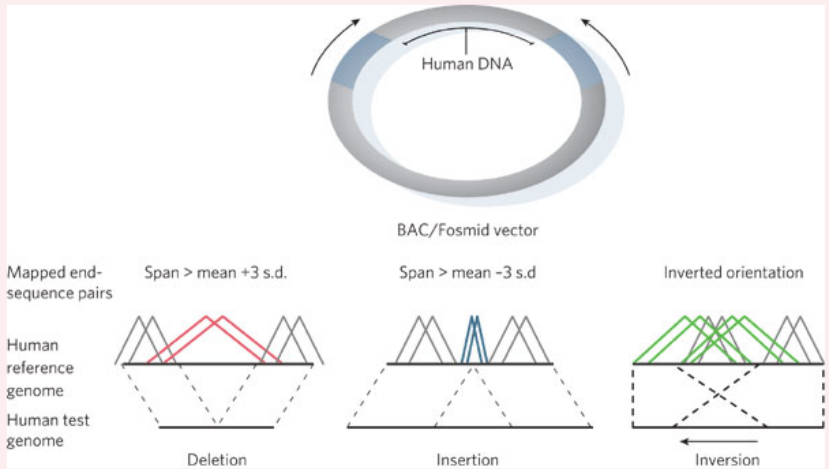
Supported by the miRNA specific aligners, and also: Bowtie, BWA, GNUMAP, MapReads, Maq, Novoalign, SHRiMP, Stampy, SOAP

Long reads

e.g. Roche, PacBio sequencing:

BLAT, RazerS, BWA-SW, SOAP2, RUM, RMAP, SOAPSplice, Bowtie2

Read pairing information



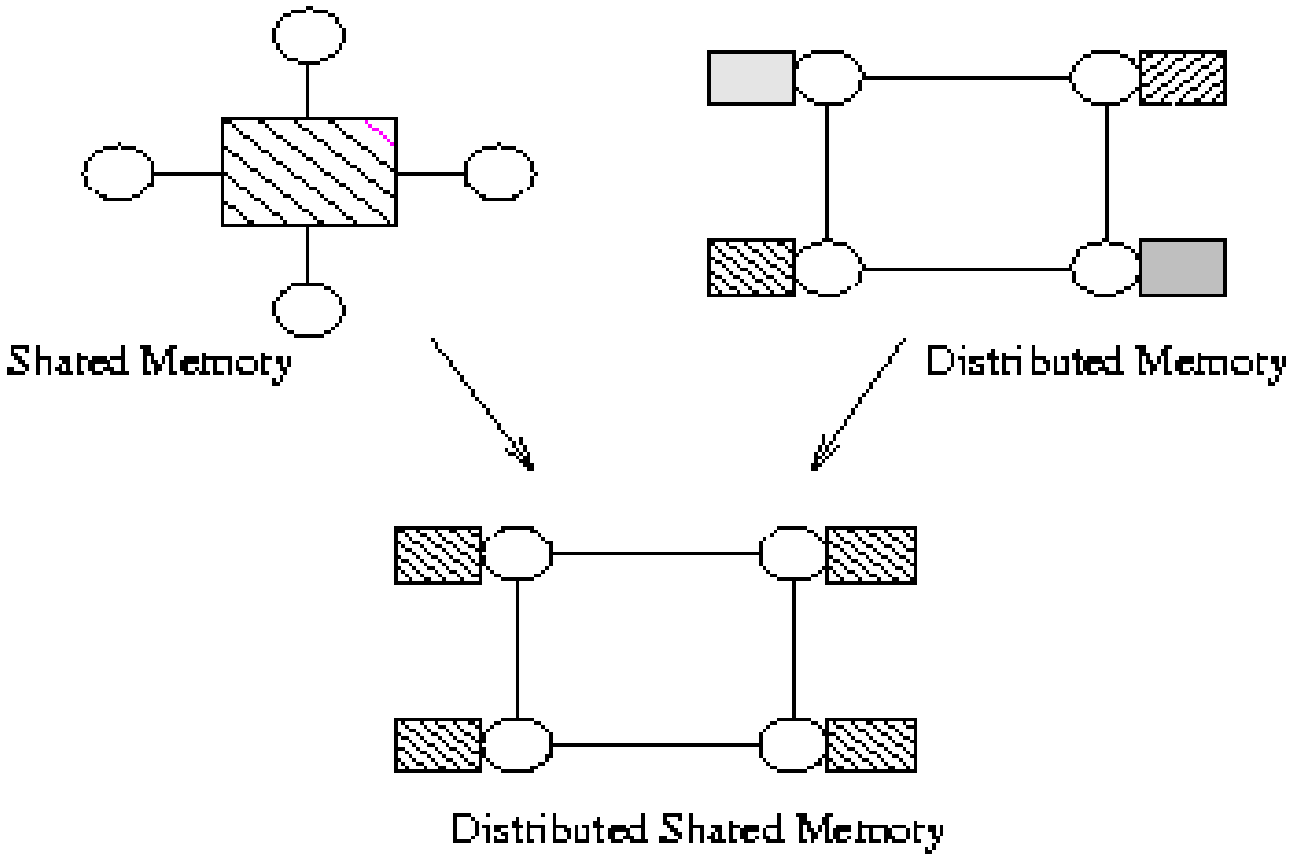
Genomic libraries are constructed from fragmented DNA and subcloned into circular vectors such as BACs or fosmids. The ends of these fragment inserts are directly sequenced from universal vector primers near the subcloning site (arrows) and are termed end-sequence pairs or paired-end sequences. End-sequence pairs are mapped to their best location in the human reference genome sequence assembly. End-sequence pairs that are discordant in terms of length (> 3 s.d. from the mean insert length) and/or orientation when mapped against the reference genome assembly may be indicative of deletions, insertions or inversion, as indicated (red, blue and green, respectively). End-sequence pairs consistent in terms of length and orientation are shown as grey.

The Human Genome Structural Variation Working Group, Nature 2007

Aligning read pairs

- Often both ends are aligned individually
- Distance between the mapped ends can identify insertions and deletions
- Supported by more than half of the sequence aligners

Parallel processing



Data parallelism and using multiple CPU's at the same time.

Using base quality scores

- Using quality scores can reduce alignment errors
- Lower penalties are given to mismatches with bases of low quality
- Used by several aligners, e.g. Bowtie, BWA, GEM-Mapper, PASS, SHRiMP2, ZOOM SOCS, RMAP, GNUMAP

RMAP

No penalty for quality scores below a certain cut-off

Novoalign

Calculate base penalties for Needleman-Wunsch algorithm

GNUMAP

Construct position weigth matrix for each read
Modified Needleman-Wunsch uses these matrices in alignment

Errors allowed

Depends on use case

- Small number of errors for detecting genome variation
- Allow more errors when comparing different species or longer reads

5 mismatches in read of 36 bases (14%)
5 mismatches in read of 150 bases (3%)

Mismatches, indels, gaps

Mutations, insertions and deletions usually supported
Longer indels (gaps) not always supported **Challenge to distinguish true variations and sequencing errors**

Constraints on mismatches and indels

Introducing mismatches and indels is computationally expensive

- **ELAND**: max 2 mismatches
- **VMATCH, WHAM**: up to 5 mismatches
- **BSMAP**: up to 15 mismatches
- **SOAP, SOAP2**: up to 3 and 2 indels
- **MrFast**: max 6 indels
- **BWA**: max 8 indels
- **Bowtie, Bowtie2, GNUMAP, Mosaik, RazerS, SSAHA2, VMATCH, SHRiMP, SHRiMP2**: no constraints

Constraints on gaps

Introducing gaps important for RNA-seq experiments (splicing)

- **SOAPSplice, SpliceMap, WHAM**: allow a single gap, with different gap size constraints
- **SOAP2, QPALMA**: one gap without size constraint
- **BLAT**: multiple gaps with maximum size of 23 kb

RNA sequencing

Difficult due to splicing events

- Align to transcriptome (miss novel transcripts)
- Alternatively align to the genome (large gaps)
- Aligners: MapSplice, TopHat, Supersplat, SOAPSplice, SpliceMap, RNA-mate, RUM, PASS, QPALMA, MapSplice

Some use a two step approach

1. map reads to the genome using unspliced read aligners
2. split unmapped reads and align the parts independently

Other aligners use seed-and-extend search (TopHat)

Alignments reported

Exact alignments

Global: end-to-end of a read

Local: faster. Begin and end of read can often be discarded (MIDs, low quality ends)

Only report locations

E.g. if you are only interested in counts.

Multi-map reads

Reads that map to multiple locations with similar scores:

- Repetitive regions
- Short read lengths

Reporting:

- Randomly pick one and give it mapping score 0 (BWA)
- Return up to N possible locations, discard the rest
- Often configurable, allowing more alignments takes longer
- Report all (mrFast, mrsFast, PatMaN)

Computer requirements

- Varies per aligner
- Some need much memory
- Others a lot of disk space
- Depending on algorithm: longer run times

List of requirements per aligner in Table 3 of the paper

Discussion

Specific aligners for different applications (DNA, miRNA, RNA, ChIP, bisulphite)
What is the best mapper??

- Assess quality with simulated datasets
- RGASP (RNAseq) - `http://www.gencodegenes.org/rgasp/`
- Alignathon (DNA-seq) - `http://compbio.soe.ucsc.edu/alignathon/`

No results available at this moment for Alignathon

Interoperability

- Input formats: FASTA, FASTQ, CFASTQ
- Output formats: tabular, SAM/BAM, CRAM
- Parameter settings also not standardized

What do we use?

- DNA: BWA, STAR2 and GATK package
- RNA: HISAT2
- Long reads: BLAT, BLAST or BWA-MEM
- Fast alignment: Kallisto (pseudo-aligner)