# High-throughout sequencing and using short-read aligners

Simon Anders

FIMM
Institute for Molecular Medicine Finland
Nordic EMBL Partnership for Molecular Medicine

# High-throughput sequencing (HTS)

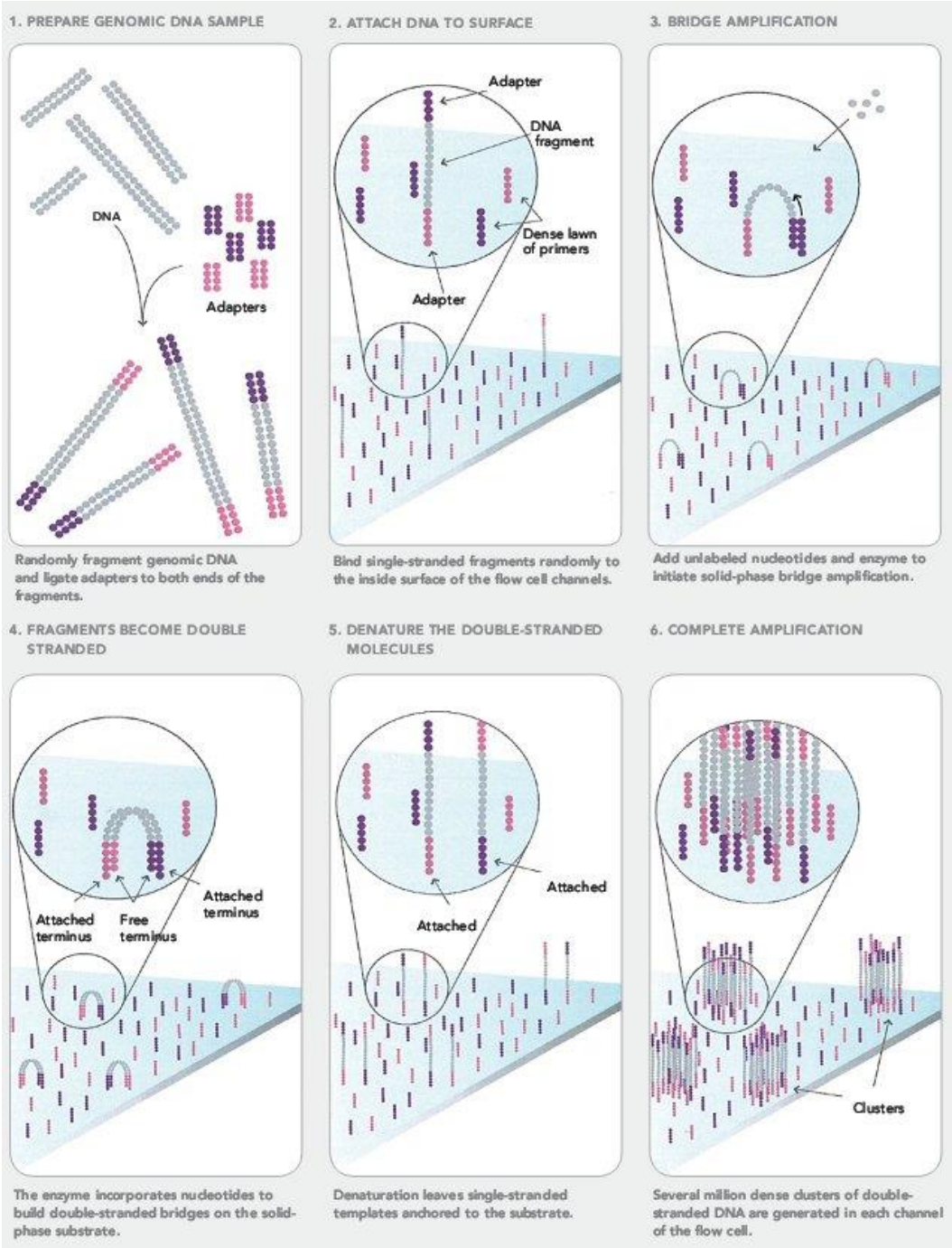
Sequencing millions of short DNA fragments in parallel.

a.k.a.: next-generation sequencing (NGS)
massively-parallel sequencing (MPS)

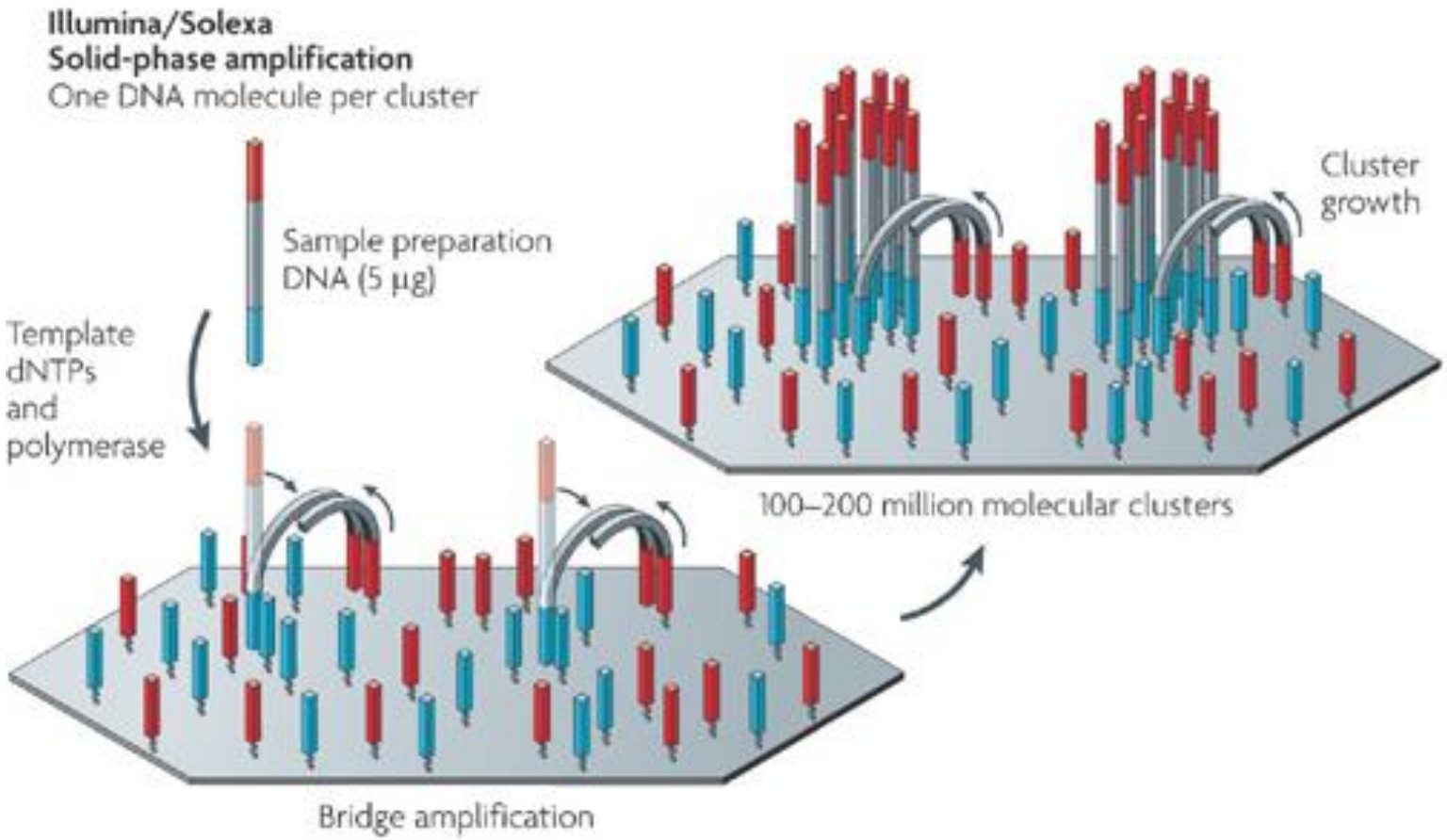
Market leader: Illumina ("HiSeq" instruments)

# Illumina's sequencing technology

- Bridge PCR
- Sequencing-by-synthesis



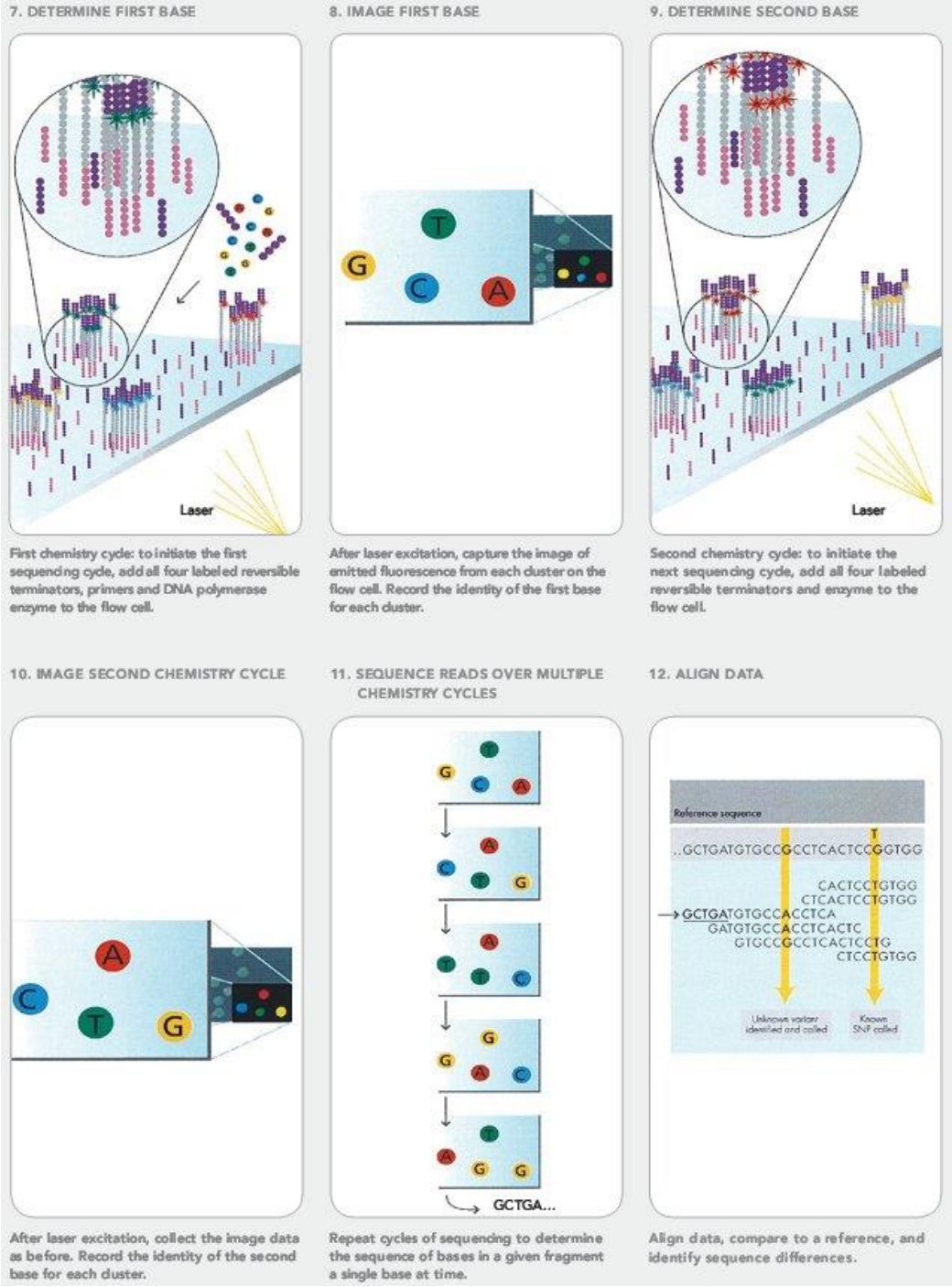
# Bridge PCR




ML Metzker, Nat Rev Genetics (2010)

# Illumina's sequencing technology

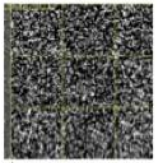
- Bridge PCR
- Sequencing-by-synthesis



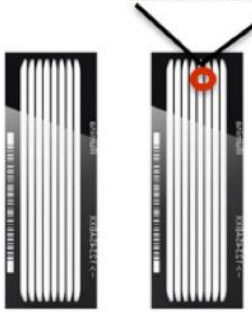
# Illunia HiSeq instrument

**Illumina HiSeq2000**


**c-Bot**
(automated cluster generation)



Cluster density
750-850/mm$^2$



**HiSeq Flow Cells**

- 2 flow cells (can be run independently)
- Up to 320Gb mapped sequence per FC
- 64Gb sequence per day (2 flow cells)

# Illumina HiSeq -- typical numbers

- 1 or 2 flow cells, each with 8 lanes
- up to ~ 200 M clusters per lane
- max 2x 150 bp read length
- 350 GB per flow cell and run
- 3 d per run (23 h in “rapid mode”)

- price per instrument: ~ €500k
- reagent costs per run: ~ €10k

# HTS: Application

- de-novo sequencing
- resequencing, variant calling
  (whole genome or targeted)
- RNA-Seq
- ChIP-Seq and other enrichment assays
- barcodes (tagged mutation collections)
- and many more …

# Short reads and longer reads

Illumina:

maximum fragment length: 1 .. 2 kb

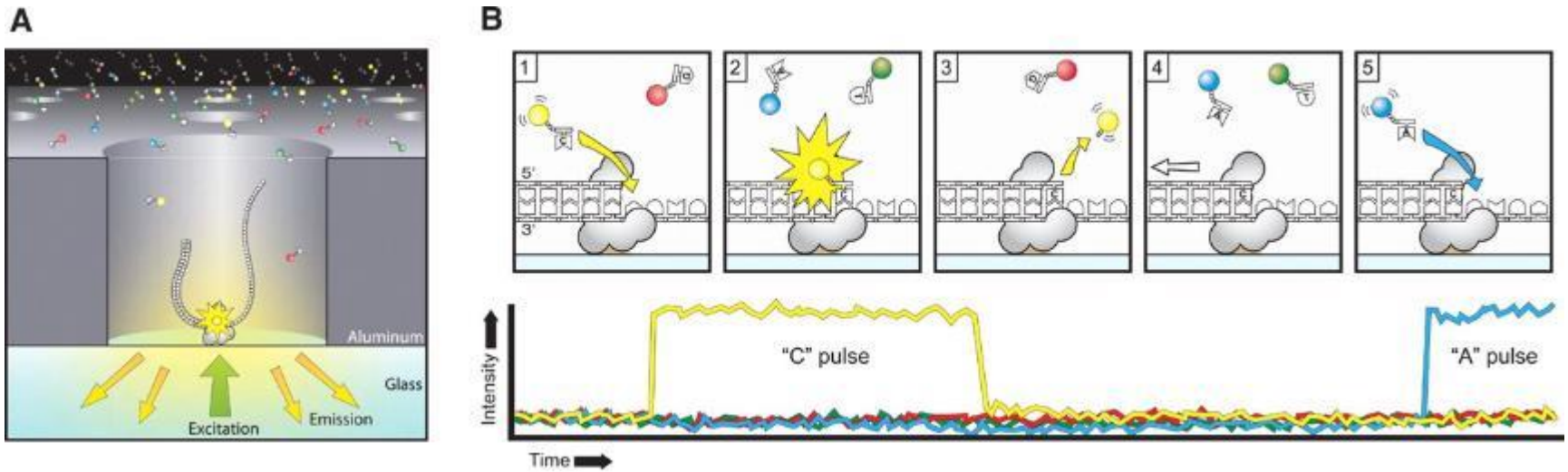
maximum read length: 150 .. 200 bp (x2)

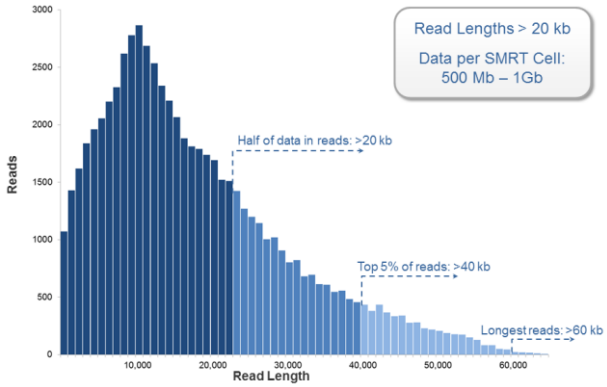
Not really enough for

- calling structural variants
- assembling genomes
- phasing SNPs

# Short reads and longer reads

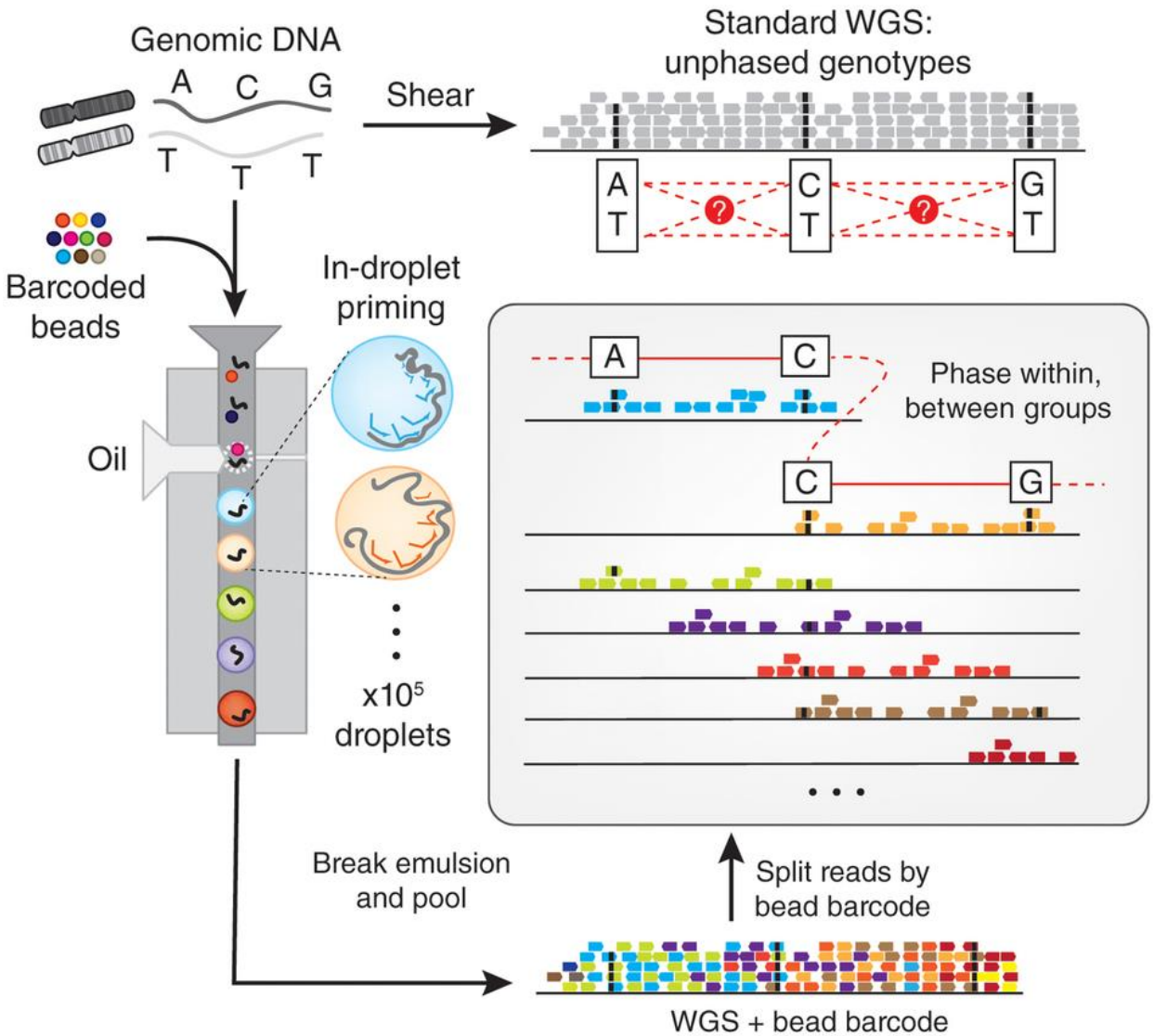
## Pacific Biosystems: SMRT sequencing



# Droplet-based sequencing

10X Genomics



Next hot topic:

# Single-cell sequencing

Each read gets barcoded to indicate which cell it came from, by clever use of microfluidics.

Practical part:

# Getting started with raw sequencing data

# Raw sequencing data: FASTQ files

```
CTCCTTTGTGTGCTTGGAATGTGAAGAATTTAGNAGACAATCCTAGACTGTTGCAGTGTTGGGGGATCCTTGTTTGAATTGGGGACACAATGCAGCCGGCT
+
CC                                                                                  DEEEEEDDDEDDBDDDDDDDDDCDDCDBDD>
@H
TG                                                                                  GCGACATGAAGACACTATGCACGAAGCCTTA
+
BCCFDEFFHHHHHHIIJJJJJJJJJJJJJHIJJJJJJJJJJJJIJJJJIJJIJJJJJIJJJHHHHFFFFFDDDDBBDDDDDDDDDDDDDEDBDBDDDDD>
@HWI-ST1177:156:C38WNACXX:6:1101:8677:2194 1:N:0:GAGTGG
TGTGGTTGTTCCTGGGGCCTCCAGCTAGCACAGCTGGCTCAGCCTGGGCAGTTCTCAGCAGGAGCGTGAGCTGCAGAGTTGCGCAGGTGAAGAGCGAGGAG
+
CCBFFFFFHHHHHJJJJIJJJJJJJJJJJJJJJJJJJIJJJJJJIJJJJJHJJJJGIJJJHHHHFDDDDDDDDDACD@DDDDDDDD?CDDDDDDDDDDDD
@HWI-ST1177:156:C38WNACXX:6:1101:8552:2194 1:N:0:GAGTGG
CGCTCTGCTGGACCTGCTGGGCCATCATCCTGCTGATCCCCATGAGGCTCTCAGTGATGGTACTGGATGTCTGGGCCAGGCTCTCTTTGGTGGTTTTCCTT
+
CCCFFFFFHHHHHJJJJJJJJIJIJJJJJJJJJJJJIJJJJIJJJJJJJJJJJIIJJJJIIJJJJGIJJHGEFHFFFFDDECDDDDDCCBDD88ABDDDD
@HWI-ST1177:156:C38WNACXX:6:1101:8746:2209 1:N:0:GAGTGG
CTTTGGTCAGTGAACAGACACGGGTGAATGCTGCTAAAAACAAGACTGGGGCTGCTCCCATCATTGATGTGGTGCGATCGGGCTACTATAAAGTTCTGGGA
+
CCCFFFFFHHHHHJJJJJJJJHJIFHGJJJJJIJJJJJJIIJIJJJJJIIJJIJJJJJHGHHF@BDDFFDEECBE>@DDDDDDDDDDCECDEDACDDCDDD
@HWI-ST1177:156:C38WNACXX:6:1101:8685:2212 1:N:0:GAGTGG
CAGCAGCTTCCTCTTCCTCTGTCTCTTTTCCAGGTCGATATCCCCAGCCCCTGAAGTGGGCCCTCTGGTCCGGGTGGGAGGGATGTGGGTCTGACTCAGCT
+
CCCFFFFFHHHHHJJJJIJIJHHIJJJJJIJJIJGGHIJJJIJJIJJIIJJJJHJIHGHHIEHGHHHHCEFFDD39?B@BDD@DDCDDD<CDDDDDDDDCD
@HWI-ST1177:156:C38WNACXX:6:1101:8630:2227 1:N:0:GAGTGG
GCTTTATCTTGCTAAAGACAATTTTTCAAGCAATCCTTTAGTTTTAGTTTTCTGGAATAGCTAGTATTGGGTTTTCTAGTTTTTTCACCTTTTAGTTTTTA
+
CCCFFFFFHHHHHJJJJJJJJJJIJJJJJJJJJIJJJJJJJIJIJJJIJHIJJJIJJGIJJJJJJJCGHHIIJ@FHIEHIJ?HHHHFEAEAEEDDDD@ADDD?
@HWI-ST1177:156:C38WNACXX:6:1101:8792:2186 1:N:0:GAGTGG
GTCTGGCATTCCATGTTTCTGCTCCTGTGGCCTCCACGGTGCAACAAGCTAGCGGTTTACTTGGACCTCTGCCTCATCTTTCTTCTTTTGCGCTTCAGCCT
+
CCCFFFFFHHHHHJJJIJGIIIJJJJJHJIIJJJJJJJJHGJJJJJJJGGIJJJJ@GHGIJJHECHHGFDFFFEEEEEEDDDD35;CD@::/'80&(+>3A
@HWI-ST1177:156:C38WNACXX:6:1101:8932:2202 1:N:0:GAGTGG
CCCGCATGTTCTCCTCACACAGTGCGCCCTCCTTGGTGGCCCACTGGAAGCCGGCCACCACACTGTCCTTGATCTCGTTGAGGTACTGCACACCCTTGGTG
+
CCCFFFFFHHHHHJJJJJJJIJIJJJJJJJJJJJJJDHIIJJJJGIJIJJGIJHHFFAEADDDDDDDDDDDDCDECDDDDDDCCDDDDDDDDDBDDCDCCB
@HWI-ST1177:156:C38WNACXX:6:1101:8805:2221 1:N:0:GAGTGG
GCCTCAAGCAGCGTGGTTCCACTGGCATTGCCATCCTTACGGGTGACTTTCCATCCCTTGAACCAAGGCATGTTAGCACTTGGCTCCAGCATGTTGTCACC
+
CCCFFFFFHHHHHJJJIJJJIJJJJIJJIIIJJIJJIEHIIIIGHEHGIJGIIIJDIAAGECEEF=DEFBEEEEEEDDDDDDDDCDDDDDDCDACCDCDD:
@HWI-ST1177:156:C38WNACXX:6:1101:9005:2191 1:N:0:GAGTGG
ATCAATATTTACTCTTGGCTCAAGGTCAATTTCATGATCTTAAGCAACTTTTCATGTCTGCAAATAATAATTTCACTCCCTCCAACAATTCCTCTTCAGAA
```

CTCCTTTGTGTGCTTGGAATGTGAAGAATTTAGNAGACAATCCTAGACTGTTGCAGTGTTGGGGGATCCTTGTTTGAATTGGGGACACAATGCAGCCGGCT
+
CCCFFFFFHHHHHIJJJHIGIHIJJJJJJJJIJ#1?FHIJIJJJJJJJJJIJJJJJIIIGGIIJIHFFFFDEEEEDDDEDDBDDDDDDDDCDDCDBDD>
@HWI-ST1177:156:C38WNACXX:6:1101:8567:2179 1:N:0:GAGTGG
**TGGGTAAGAAAAGTCAAAACACTAATGAGTTGTCCATGAAGCCAACTGCTAAGAACGCGCTCAACTATACGCGACATGAAGACACTATGCACGAAGCCTTA**
+
BCCFDEFFHHHHHHIIJJJJJJJJJJJJJHIJJJJJJJJJJJJIJJJJIJJIJJJJJIJJJJHHHHFFFFFDDDDBBDDDDDDDDDDDDDEDBDBDDDDD>
@HWI-ST1177:156:C38WNACXX:6:1101:8677:2194 1:N:0:GAGTGG
**TGTGGTTGTTCCTGGGGCCTCCAGCTAGCACAGCTGGCTCAGCCTGGGCAGTTCTCAGCAGGAGCGTGAGCTGCAGAGTTGCGCAGGTGAAGAGCGAGGAG**
+
CCBFFFFFHHHHHJJJJIJJJJJJJJJJJJJJJJJJJJIJJJJJJIJJJJJHJJJJGIJJJHHHHFDDDDDDDDDACD@DDDDDDDD?CDDDDDDDDDDDD
@HWI-ST1177:156:C38WNACXX:6:1101:8552:2194 1:N:0:GAGTGG
**CGCTCTGCTGGACCTGCTGGGCCATCATCCTGCTGATCCCCATGAGGCTCTCAGTGATGGTACTGGATGTCTGGGCCAGGCTCTCTTTGGTGGTTTTCCTT**
+
CCCFFFFFHHHHHJJJJJJJJIJIJJJJJJJJJJJJIJJJJIJJJJJJJJJJJJIIJJJJIIJJJJGIJJHGEFHFFFFDDECDDDDDCCBDD88ABDDDD
@HWI-ST1177:156:C38WNACXX:6:1101:8746:2209 1:N:0:GAGTGG
**CTTTGGTCAGTGAACAGACACGGGTGAATGCTGCTAAAAACAAGACTGGGGCTGCTCCCATCATTGATGTGGTGCGATCGGGCTACTATAAAGTTCTGGGA**
+
CCCFFFFFHHHHHJJJJJJJJHJIFHGJJJJJIJJJJJJIIJIJJJJJIIJJIJJJJJHGHHF@BDDFFDEECBE>@DDDDDDDDDDCECDEDACDDCDDD
@HWI-ST1177:156:C38WNACXX:6:1101:8685:2212 1:N:0:GAGTGG
**CAGCAGCTTCCTCTTCCTCTGTCTCTTTTCCAGGTCGATATCCCCAGCCCCTGAAGTGGGCCCTCTGGTCCGGGTGGGAGGGATGTGGGTCTGACTCAGCT**
+
CCCFFFFFHHHHHJJJJIJIJHHIJJJJJIJJIJGGHIJJJIJJIJJIIJJJJHJIHGHHIEHGHHHHCEFFDD39?B@BDD@DDCDDD<CDDDDDDDDCD
@HWI-ST1177:156:C38WNACXX:6:1101:8630:2227 1:N:0:GAGTGG
**GCTTTATCTTGCTAAAGACAATTTTTCAAGCAATCCTTTAGTTTTAGTTTTCTGGAATAGCTAGTATTGGGTTTTCTAGTTTTTTCACCTTTTAGTTTTTA**
+
CCCFFFFFHHHHHJJJJJJJJJIJJJJJJJJJIJJJJJJJIJIJJJIJHIJJJIJJGIJJJJJJJCGHHIIJ@FHIEHIJ?HHHHFEAEAEEDDDD@ADDD?
@HWI-ST1177:156:C38WNACXX:6:1101:8792:2186 1:N:0:GAGTGG
**GTCTGGCATTCCATGTTTCTGCTCCTGTGGCCTCCACGGTGCAACAAGCTAGCGGTTTACTTGGACCTCTGCCTCATCTTTCTTCTTTTGCGCTTCAGCCT**
+
CCCFFFFFHHHHHJJJIJGIIIJJJJJHJIIJJJJJJJJHGJJJJJJGGIJJJJ@GHGIJJHECHHGFDFFFEEEEEDDDD35;CD@::/'80&(+>3A
@HWI-ST1177:156:C38WNACXX:6:1101:8932:2202 1:N:0:GAGTGG
**CCCGCATGTTCTCCTCACACAGTGCGCCCTCCTTGGTGGCCCACTGGAAGCCGGCCACCACACTGTCCTTGATCTCGTTGAGGTACTGCACACCCTTGGTG**
+
CCCFFFFFHHHHHJJJJJJJIJIJJJJJJJJJJJJJDHIIJJJJGIJIJJGIJHHFFAEADDDDDDDDDDDCDECDDDDDDCCDDDDDDDDDBDDCDCCB
@HWI-ST1177:156:C38WNACXX:6:1101:8805:2221 1:N:0:GAGTGG
**GCCTCAAGCAGCGTGGTTCCACTGGCATTGCCATCCTTACGGGTGACTTTCCATCCCTTGAACCAAGGCATGTTAGCACTTGGCTCCAGCATGTTGTCACC**
+
CCCFFFFFHHHHHJJJIJJJIJJJJIJJIIIJJIJJIEHIIIIGHEHGIJGIIIJDIAAGECEEF=DEFBEEEEEDDDDDDDDCDDDDDDCDACCDCDD:
@HWI-ST1177:156:C38WNACXX:6:1101:9005:2191 1:N:0:GAGTGG
**ATCAATATTTACTCTTGGCTCAAGGTCAATTTCATGATCTTAAGCAACTTTTCATGTCTGCAAATAATAATTTCACTCCCTCCAACAATTCCTCTTCAGAA**

# FASTQ format

Each read is represented by four lines:

- '@', followed by read ID
- sequence
- '+', optionally followed by repeated read ID
- quality string:
  - same length as sequence
  - each character encodes the base-call quality of one base

# FASTQ quality scores

| quality score *Q* | error prob. *p* | characters |
|---|---|---|
| 0 .. 9 | 1 .. 0.13 | !”#$%&'()* |
| 10 .. 19 | 0.1 .. 0.013 | +,-./01234 |
| 20 .. 29 | 0.01 .. 0.0013 | 56789:;<=> |
| 30 .. 39 | 0.001 .. 0.00013 | ?@ABCDEFGH |
| 40 | 0.0001 | I |

$Q = -10 \log_{10} p$

# FASTQ files for paired end

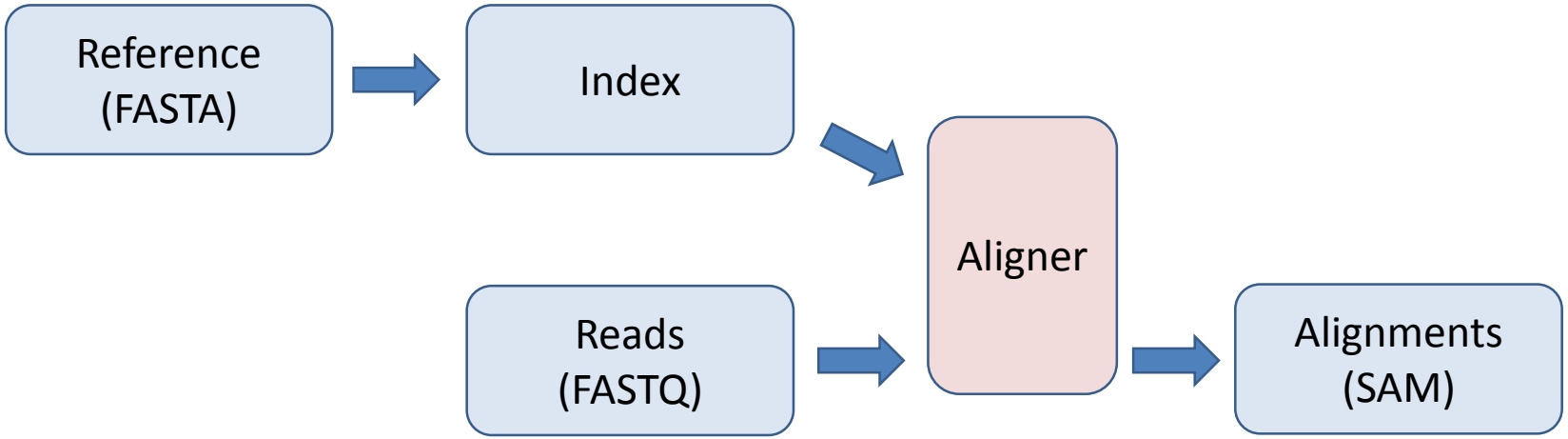
Simple: two FASTQ files,

- one with all first reads, one with all second reads
- Read names and ordering must match.

# Alignment

Given a reference genome, *where* did each read come from?

A *short-read aligner* finds for each read all possible matches in the reference genome.



# Popular aligners

For genomic DNA

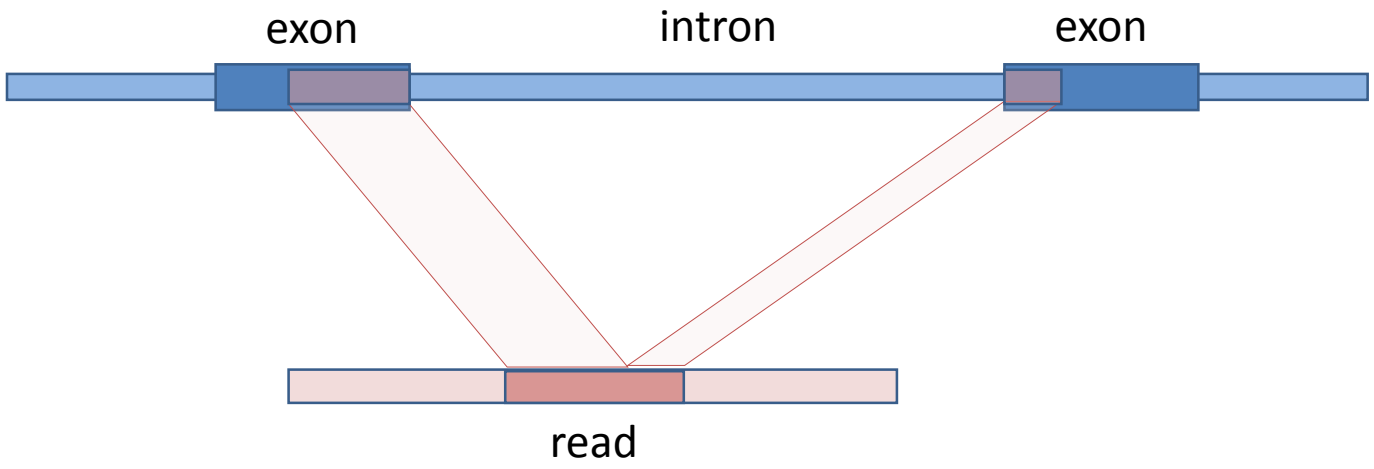
- Bowtie2
- BWA
- …

For RNA

- Tophat2
- GSNAP
- STAR

…

# Spliced alignment




For RNA-Seq data, we need a **splice-aware** aligner.

# Criteria to choose an aligner

- Reputation
- Publication year
- Still developed and maintained?

- Speed and memory requirements
- Sensitivity

- Special functions

  (e.g., bisulfite mapping, fusion gene detection, multiple references, etc.)

# Example:
# Aligning RNA-Seq data with STAR

BIOINFORMATICS **ORIGINAL PAPER** Vol. 29 no. 1 2013, pages 15–21
doi:10.1093/bioinformatics/bts635

*Sequence analysis* Advance Access publication October 25, 2012

## STAR: ultrafast universal RNA-seq aligner

Alexander Dobin[1,*], Carrie A. Davis[1], Felix Schlesinger[1], Jorg Drenkow[1], Chris Zaleski[1], Sonali Jha[1], Philippe Batut[1], Mark Chaisson[2] and Thomas R. Gingeras[1]

[1]Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, USA and [2]Pacific Biosciences, Menlo Park, CA, USA

Associate Editor: Inanc Birol

**ABSTRACT**

**Motivation:** Accurate alignment of high-throughput RNA-seq data is a challenging and yet unsolved problem because of the non-contiguous transcript structure, relatively short read lengths and constantly increasing throughput of the sequencing technologies. Currently available RNA-seq aligners suffer from high mapping error rates, low mapping speed, read length limitation and mapping biases.

**Results:** To align our large (>80 billon reads) ENCODE Transcriptome RNA-seq dataset, we developed the Spliced Transcripts Alignment to a Reference (STAR) software based on a previously undescribed RNA-seq alignment algorithm that uses sequential maximum mappable seed search in uncompressed suffix arrays followed by seed clustering and stitching procedure. STAR outperforms other aligners by a factor of >50 in mapping speed, aligning to the human genome 550 million 2 × 76 bp paired-end reads per hour on a modest 12-core server, while at the same time improving alignment sensitivity and precision. In addition to unbiased *de novo* detection of canonical junctions, STAR can discover non-canonical splices and chimeric (fusion) transcripts, and is also capable of mapping full-length RNA se-

unique challenges to detection and characterization of spliced transcripts. Two key tasks make these analyses computationally intensive. The first task is an accurate alignment of reads that contain mismatches, insertions and deletions caused by genomic variations and sequencing errors. The second task involves mapping sequences derived from non-contiguous genomic regions comprising spliced sequence modules that are joined together to form spliced RNAs. Although the first task is shared with DNA resequencing efforts, the second task is specific and crucial to the RNA-seq, as it provides the connectivity information needed to reconstruct the full extent of spliced RNA molecules. These alignment challenges are further compounded by the presence of multiple copies of identical or related genomic sequences that are themselves transcribed, making precise mapping difficult.

Various sequence alignment algorithms have been recently developed to tackle these challenges (Au *et al.*, 2010; De Bona, *et al.*, 2008; Grant *et al.*, 2011; Han *et al.*, 2011; Trapnell *et al.*, 2009; Wang *et al.*, 2010; Wu and Nacu, 2010; Zhang *et al.*, 2012). However, application of these algorithms invokes compromises

# Using STAR: Creating the index

We need

- a FASTA file with the human reference genome sequence
  - e.g., from Ensembl
- a GTF file with gene models
  - e.g., from Ensembl
- the STAR software:
  - from github.com/alexdobin/STAR

Downloads - Ensem

www.ensembl.org/downloads.html

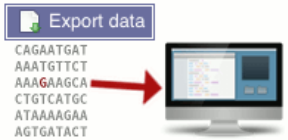
Apps For quick access, place your bookmarks here on the bookmarks bar. Import bookmarks now...

Login/Register

Ensembl BLAST/BLAT | BioMart | Tools | Downloads | Help & Documentation | Blog | Mirrors

Search all species...

## Download a sequence or region




Click on the 'Export data' button in the lefthand menu of most pages to export:

- FASTA sequence
- GTF or GFF features

...and more!

## Customise your download


Custom datasets can be retrieved using the BioMart data-mining tool.

You may find exploring this web-based query tool easier than extracting information direct from our databases.

## Fetch data programmatically

Write your own Perl scripts to retrieve small-to-medium datasets. All our data, as well as added functionality, is available through the Ensembl Perl API.

Use the API to retrieve gene and transcript sets, fetch alignments between sequences, compare allele frequencies and much more!

You can also use our REST API to retrieve data to process in the programming language of your choice.

## Download databases & software


All of our data and software, including pipelines and web code, is available free.

- Download data via FTP
- Ensembl pipeline on GitHub
- Set up your own Ensembl website

Ensembl release 84 - March 2016 © WTSI / EMBL-EBI

Permanent link

About Us | Get help | Our sister sites | Follow us

# Ensembl FTP download server

| ★ | Species | DNA (FASTA) | cDNA (FASTA) | CDS (FASTA) | ncRNA (FASTA) | Protein sequence (FASTA) | Annotated sequence (EMBL) | Annotated sequence (GenBank) | Gene sets | Whole databases | Variation (GVF) | Variation (VCF) | Vari (V |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Y | **Human** *Homo sapiens* | FASTA | FASTA | FASTA | FASTA | FASTA | EMBL | GenBank | GTF GFF3 | MySQL | GVF | VCF | V |
| Y | **Mouse** *Mus musculus* | FASTA | FASTA | FASTA | FASTA | FASTA | EMBL | GenBank | GTF GFF3 | MySQL | GVF | VCF | V |
| Y | **Zebrafish** *Danio rerio* | FASTA | FASTA | FASTA | FASTA | FASTA | EMBL | GenBank | GTF GFF3 | MySQL | GVF | VCF | V |
| | **Alpaca** *Vicugna pacos* | FASTA | FASTA | FASTA | FASTA | FASTA | EMBL | GenBank | GTF GFF3 | MySQL | - | - | V |
| | **Amazon molly** *Poecilia formosa* | FASTA | FASTA | FASTA | FASTA | FASTA | EMBL | GenBank | GTF GFF3 | MySQL | - | - | V |
| | **Anole lizard** *Anolis carolinensis* | FASTA | FASTA | FASTA | FASTA | FASTA | EMBL | GenBank | GTF GFF3 | MySQL | - | - | V |
| | **Armadillo** *Dasypus novemcinctus* | FASTA | FASTA | FASTA | FASTA | FASTA | EMBL | GenBank | GTF GFF3 | MySQL | - | - | V |
| | **Bushbaby** *Otolemur garnettii* | FASTA | FASTA | FASTA | FASTA | FASTA | EMBL | GenBank | GTF GFF3 | MySQL | - | - | V |
| | **C.intestinalis** *Ciona intestinalis* | FASTA | FASTA | FASTA | FASTA | FASTA | EMBL | GenBank | GTF GFF3 | MySQL | - | - | V |
| | **C.savignyi** *Ciona* | FASTA | FASTA | FASTA | FASTA | FASTA | EMBL | GenBank | GTF GFF3 | MySQL | - | - | V |

# Ensembl genome reference sequence

| | | |
|---|---|---|
| Homo_sapiens.GRCh38.dna.chromosome.21.fa.gz | 11.2 MB | 25/02/2016, 04:27:00 |
| Homo_sapiens.GRCh38.dna.chromosome.22.fa.gz | 10.9 MB | 25/02/2016, 04:27:00 |
| Homo_sapiens.GRCh38.dna.chromosome.2.fa.gz | 69.4 MB | 25/02/2016, 04:27:00 |
| Homo_sapiens.GRCh38.dna.chromosome.3.fa.gz | 57.0 MB | 25/02/2016, 04:27:00 |
| Homo_sapiens.GRCh38.dna.chromosome.4.fa.gz | 54.7 MB | 25/02/2016, 04:27:00 |
| Homo_sapiens.GRCh38.dna.chromosome.5.fa.gz | 52.0 MB | 25/02/2016, 04:27:00 |
| Homo_sapiens.GRCh38.dna.chromosome.6.fa.gz | 49.1 MB | 25/02/2016, 04:27:00 |
| Homo_sapiens.GRCh38.dna.chromosome.7.fa.gz | 45.2 MB | 25/02/2016, 04:27:00 |
| Homo_sapiens.GRCh38.dna.chromosome.8.fa.gz | 41.6 MB | 25/02/2016, 04:27:00 |
| Homo_sapiens.GRCh38.dna.chromosome.9.fa.gz | 34.8 MB | 25/02/2016, 04:27:00 |
| Homo_sapiens.GRCh38.dna.chromosome.MT.fa.gz | 5.4 kB | 25/02/2016, 04:27:00 |
| Homo_sapiens.GRCh38.dna.chromosome.X.fa.gz | 44.0 MB | 25/02/2016, 04:27:00 |
| Homo_sapiens.GRCh38.dna.chromosome.Y.fa.gz | 6.8 MB | 25/02/2016, 04:27:00 |
| Homo_sapiens.GRCh38.dna.nonchromosomal.fa.gz | 2.9 MB | 25/02/2016, 04:27:00 |
| Homo_sapiens.GRCh38.dna.primary_assembly.fa.gz | 840 MB | 25/02/2016, 04:27:00 |
| Homo_sapiens.GRCh38.dna_rm.alt.fa.gz | 142 MB | 25/02/2016, 04:27:00 |
| Homo_sapiens.GRCh38.dna_rm.chromosome.10.fa.gz | 21.7 MB | 25/02/2016, 04:27:00 |
| Homo_sapiens.GRCh38.dna_rm.chromosome.11.fa.gz | 20.9 MB | 25/02/2016, 04:27:00 |
| Homo_sapiens.GRCh38.dna_rm.chromosome.12.fa.gz | 20.6 MB | 25/02/2016, 04:27:00 |
| Homo_sapiens.GRCh38.dna_rm.chromosome.13.fa.gz | 16.1 MB | 25/02/2016, 04:27:00 |
| Homo_sapiens.GRCh38.dna_rm.chromosome.14.fa.gz | 14.4 MB | 25/02/2016, 04:27:00 |
| Homo_sapiens.GRCh38.dna_rm.chromosome.15.fa.gz | 13.5 MB | 25/02/2016, 04:27:00 |
| Homo_sapiens.GRCh38.dna_rm.chromosome.16.fa.gz | 12.9 MB | 25/02/2016, 04:27:00 |
| Homo_sapiens.GRCh38.dna_rm.chromosome.17.fa.gz | 13.0 MB | 25/02/2016, 04:27:00 |
| Homo_sapiens.GRCh38.dna_rm.chromosome.18.fa.gz | 12.7 MB | 25/02/2016, 04:27:00 |
| Homo_sapiens.GRCh38.dna_rm.chromosome.19.fa.gz | 7.8 MB | 25/02/2016, 04:27:00 |
| Homo_sapiens.GRCh38.dna_rm.chromosome.1.fa.gz | 36.5 MB | 25/02/2016, 04:27:00 |
| Homo_sapiens.GRCh38.dna_rm.chromosome.20.fa.gz | 9.8 MB | 25/02/2016, 04:27:00 |
| Homo_sapiens.GRCh38.dna_rm.chromosome.21.fa.gz | 6.3 MB | 25/02/2016, 04:27:00 |
| Homo_sapiens.GRCh38.dna_rm.chromosome.22.fa.gz | 6.0 MB | 25/02/2016, 04:27:00 |

# STAR: Generate genome index

```
Terminal
anderss@lx5-fimm8:~/work/STAR_demo$ ls download
Homo_sapiens.GRCh38.84.gtf  Homo_sapiens.GRCh38.dna.primary_assembly.fa
anderss@lx5-fimm8:~/work/STAR_demo$ mkdir index_GRCh38.84
anderss@lx5-fimm8:~/work/STAR_demo$ STAR --runMode genomeGenerate --runThreadN
2 --genomeDir index_GRCh38.84 --genomeFastaFiles download/Homo_sapiens.GRCh38.d
na.primary_assembly.fa --sjdbGTFfile download/Homo_sapiens.GRCh38.84.gtf --sjdb
Overhang 99 --limitGenomeGenerateRAM 2500000000
Jul 06 17:03:26 ..... started STAR run
Jul 06 17:03:26 ... starting to generate Genome files
```

# STAR: Generate genome index

```
STAR  \
  --runMode genomeGenerate  \
  --runThreadN 2  \
  --genomeDir index_GRCh38.84  \
  --genomeFastaFiles download/Homo_sapiens.GRCh38.\
        dna.primary_assembly.fa  \
  --sjdbGTFfile download/Homo_sapiens.GRCh38.84.gtf
```

# STAR: Align the reads

```
STAR \
  --runThreadN 2  \
  --genomeDir index_GRCh38.84  \
  --readFilesIn sample1_1.fastq  \
        sample1_2.fastq  \
  --outFileNamePrefix sample1_ \
  --outSAMtype BAM SortedByCoordinate
```

and maybe:

```
  --quantMode GeneCounts
```

# STAR: output

Created files:

| size | file name |
|---|---|
| **34M** | sample1_Aligned.sortedByCoord.out.bam |
| 4.0K | sample1_Log.final.out |
| 16K | sample1_Log.out |
| 4.0K | sample1_Log.progress.out |
| 20K | sample1_ReadsPerGene.out.tab |
| 72K | sample1_SJ.out.tab |

# Alignment output: SAM files

Part of a SAM file:

- Header

  (lines start with @)

  - command line
  - chromosome names
  - etc

- Alignments
  - tab-separated table, at least 11 columns
  - one line per alignment

# SAM alignment section

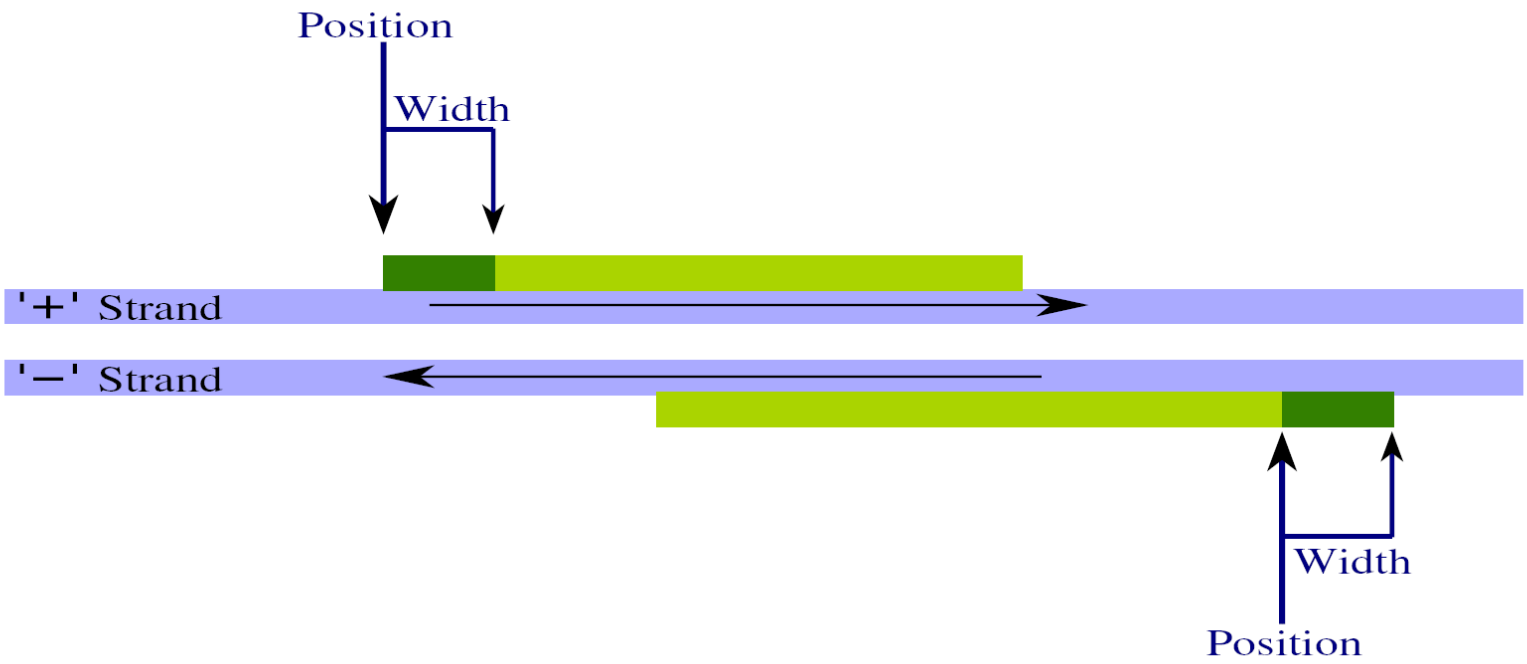
```
[...]

HWI-EAS225_309MTAAXX:5:1:689:1485  0          XIII       863564     25         36M        *
           0          0          GAAATATATACGTTTTTATCTATGTTACGTTATATA
           CCCCCCCCCCCCCCCCCCCCCCC4CCCB4CA?AAA<          NM:i:0     X0:i:1     MD:Z:36

HWI-EAS225_309MTAAXX:5:1:689:1485  16         XIII       863766     25         36M        *
           0          0          CTACAATTTTGCACATCAAAAAAGACCTCCAACTAC
           =8A=AA784A9AA5AAAAAAAAAAA=AAAAAAAAAA          NM:i:0     X0:i:1     MD:Z:36

HWI-EAS225_309MTAAXX:5:1:394:1171  0          XII        525532     25         36M        *
           0          0          GTTTACGGCGTTGCAAGAGGCCTACACGGGCTCATT
           CCCCCCCCCCCCCCCCCCCCCC?CCACCACA7?<???         NM:i:0     X0:i:1     MD:Z:36

HWI-EAS225_309MTAAXX:5:1:394:1171  16         XII        525689     25         36M        *
           0          0          GCTGTTATTTCTCCACAGTCTGGCAAAAAAAAGAAA
           7AAAAAA?AA<AA?AAAAA5AAA<AAAAAAAAAAAA          NM:i:0     X0:i:1     MD:Z:36

HWI-EAS225_309MTAAXX:5:1:393:671   0          XV         440012     25         36M        *
           0          0          TTTGGTGATTTTCCCGTCTTTATAATCTCGGATAAA
           AAAAAAAAAAAAAAAA<AAAAAAAA<AAAA5<AAAA3         NM:i:0     X0:i:1     MD:Z:36

HWI-EAS225_309MTAAXX:5:1:393:671   16         XV         440188     25         36M        *
           0          0          TCATAGATTCCATATGAGTATAGTTACCCCATAGCC
           ?9A?A?CC?<ACCCCCCCCCCCCCCCCCCACCCCCCC          NM:i:0     X0:i:1     MD:Z:36

[...]
```

# Columns in a SAM file

- QNAME ID of the read (“query”)
- FLAG alignment flags
- RNAME ID of the reference (typically: chromosome name)
- POS Position in reference (1-based, left side)
- MAPQ Mapping quality (as Phred score)
- CIGAR Alignment description (gaps etc.) in CIGAR format
- RNEXT Mate reference sequence name
- PNEXT Mate position
- ISIZE inferred insert size
- SEQ sequence of the read
- QUAL quality string of the read
- extra fields

for paired-end data

# Reads and fragments



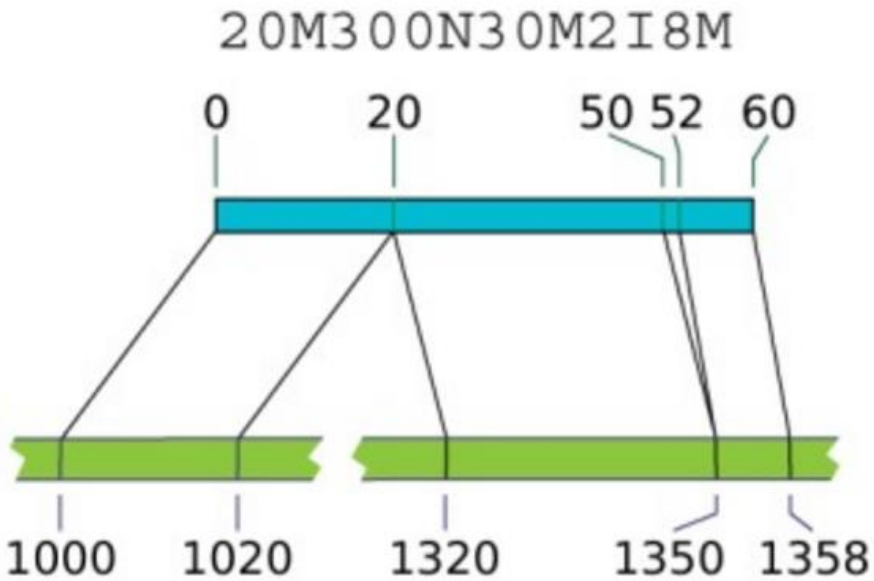
# SAM format: FLAG field

| bit | hex | decimal | |
|---|---|---|---|
| 0 | 00 01 | 1 | read is a paired-end read |
| 1 | 00 02 | 2 | read pair is properly matched |
| 2 | 00 04 | 4 | read has not been mapped |
| 3 | 00 08 | 8 | mate has not been mapped |
| 4 | 00 10 | 16 | read has been mapped to “-” strand |
| 5 | 00 20 | 32 | mate has been mapped to “-” strand |
| 6 | 00 40 | 64 | read is the first read in a pair |
| 7 | 00 80 | 128 | read is the second read in a pair |
| 8 | 01 00 | 256 | alignment is secondary |
| 9 | 02 00 | 512 | read did had not passed quality check |
| 10 | 04 00 | 1024 | read is a PCR or optical duplicate |

# SAM format: Optional fields

- last column
- triples of the format TAG : VTYPE : VALUE
- some important tag types:
  - NH: number of reported aligments
  - NM: number of mismatches
  - MD: positions of mismatches

# SAM format: CIGAR string




| type | size | query | reference |
|---|---|---|---|
| M | 20 | 0–20 | chrl: 1000–1020 |
| N | 300 | 20–20 | chrl: 1020–1320 |
| M | 30 | 20–50 | chrl: 1320–1350 |
| I | 2 | 50–52 | chrl: 1350–1350 |
| M | 8 | 52–60 | chrl: 1350–1358 |

# SAM format: Paired-end and multiple alignments

- Each line is one alignment for one read.
- Multiple alignments for a read take several lines
  - The Read ID groups them.
- Paired-end alignments take two lines.

- All these lines are not necessarily in adjacent lines!

# SAM and BAM

- Text SAM files (“.sam”):  human readable
- Binary SAM files (“.bam”): smaller, quicker to read

Use *samtools* for conversion.

# SAMtools

A small program to

- convert between .bam and .sam
- sort and merge SAM files
- index SAM and FASTA files for fast access
- calculate tallies (“flagstat”)
- …

The SAMtools C API is a library for tool development.

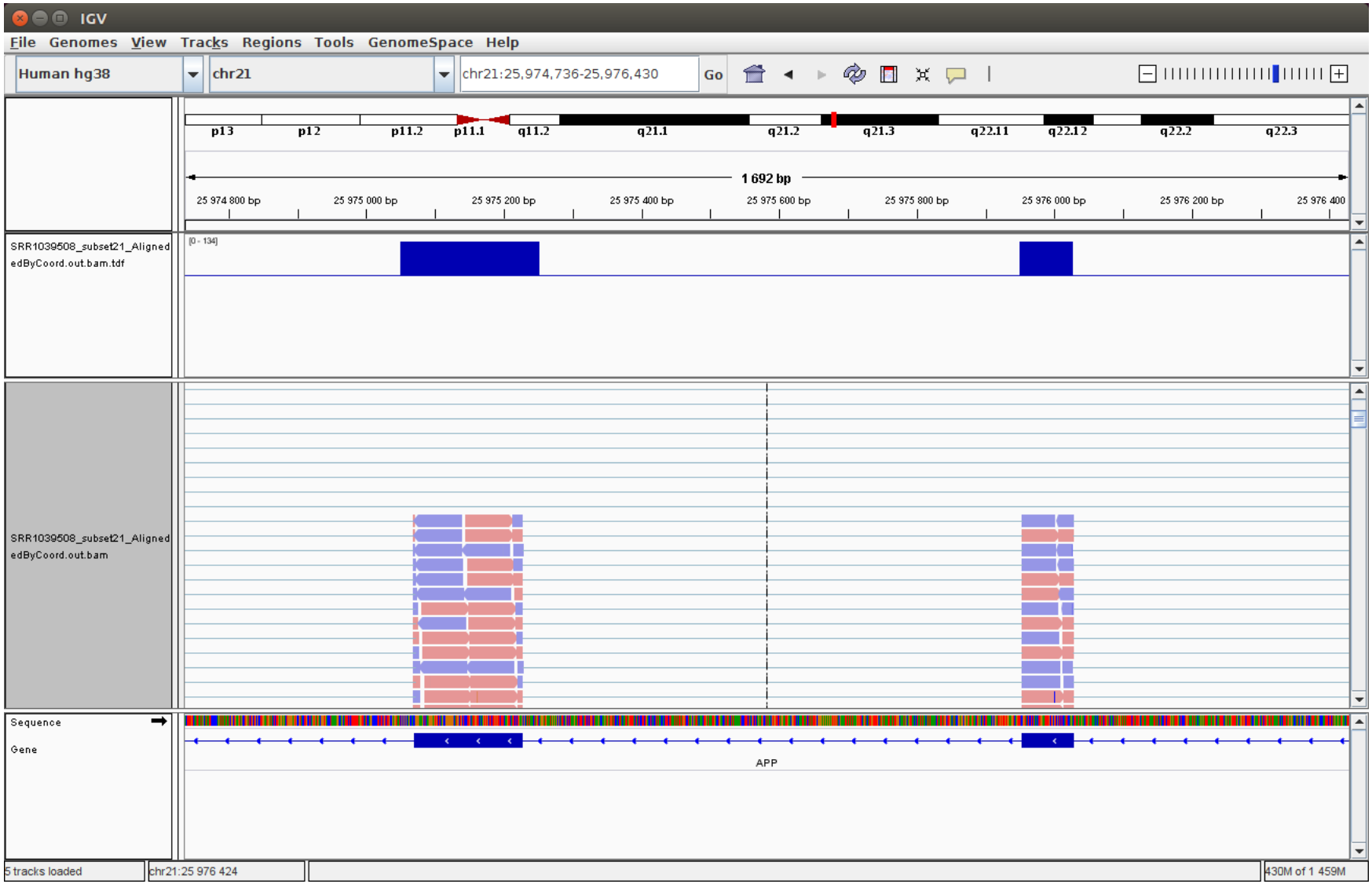
# Visually inspecting SAM files

*Look* at *all* your big files!

Learn to use text views *and* graphical tools.

For SAM, use e.g. the *Integrated Genome Viewer* (IGV).

# IGV


IGV
File Genomes View Tracks Regions Tools GenomeSpace Help
Human hg38
chr21
chr21:25,974,736-25,976,430
Go
p13 p12 p11.2 p11.1 q11.2 q21.1 q21.2 q21.3 q22.11 q22.12 q22.2 q22.3
1 692 bp
25 974 800 bp
25 975 000 bp
25 975 200 bp
25 975 400 bp
25 975 600 bp
25 975 800 bp
25 976 000 bp
25 976 200 bp
25 976 400
SRR1039508_subset21_AlignedByCoord.out.bam.tdf
[0 - 134]
SRR1039508_subset21_AlignedByCoord.out.bam
Sequence
Gene
APP
5 tracks loaded
chr21:25 976 424
430M of 1 459M

# Aligning multiple files

Sample table:

| | Library_ID | Cell_line | Treatment |
|---|---|---|---|
| 1 | SRR1039508 | N61311 | none |
| 2 | SRR1039509 | N61311 | dexa |
| 3 | SRR1039512 | N052611 | none |
| 4 | SRR1039513 | N052611 | dexa |
| 5 | SRR1039516 | N080611 | none |
| 6 | SRR1039517 | N080611 | dexa |
| 7 | SRR1039520 | N061011 | none |
| 8 | SRR1039521 | N061011 | dexa |

# Generating multiple alignment commands with R

```
sample_table <- read.csv( "reads/sample_table.csv" )

command <- paste0(
   "STAR --runThreadN 2 --genomeDir GRCh38.84 \\\n",
   "  --readFilesIn reads/###_1.fastq \\\n”,
   "      reads/###_2.fastq \\\n",
   "  --outSAMtype BAM SortedByCoordinate \\\n",
   "  --outFileNamePrefix star_out/###_ \\\n",
   "  --quantMode GeneCounts\n”,
   "\n" )

f <- file( "aligner_commands.sh", open="w" )
for( sampleID in sample_table$Library_ID ) {
   cat( gsub( "###", sampleID, command, fixed=TRUE ),
      file=f )
}
close( f )
```

# Next step: Using the alignments

RNA-Seq:

- What transcripts are there?
- Which genes change expression due to treatment?
  - Mike's talk and this afternoon's lab

ChIP-Seq:

- Peak finding and comparing
  - Thursday lab

In general:

- The real work starts only *now*.
  So far, this was just preparation.

*