

In-Person

*Foundations in  
Genomic Analyses*

# **Sequence Alignment and Mapping Reads**

Northwestern

INFORMATION TECHNOLOGY  
RESEARCH COMPUTING AND DATA SERVICES



### COMPUTING AND SOFTWARE

Access to high performance computing, research software, and global networks for conducting computationally intense research.



### DATA MANAGEMENT AND SHARING

Learn about data management planning and options for storing, securing, transferring, and sharing data.



### DATA SCIENCE, STATISTICS, AND VISUALIZATION

Support for collecting, analyzing, visualizing, and programming with research data.



### TRAINING AND CONSULTATION

Identify events, resources, and people to help you learn computational and data skills for your research.

# Research Computing and Data Services

*We're here to help after the workshop!*

**[quest-help@northwestern.edu](mailto:quest-help@northwestern.edu)**

**[bit.ly/rcdsconsult](https://bit.ly/rcdsconsult)**

**<https://sites.northwestern.edu/researchcomputing/>**

# Setup...

1. log onto Quest

```
ssh <netid>@quest.northwestern.edu # enter your netid password
```

2. move to our classroom folder

```
cd /projects/e32680
```

3. make your own subfolder (if you haven't in a previous week)

```
mkdir <folder_name>
```

# What is alignment?

Histone H1 (residues 120-180)	
HUMAN	KKASKPKKAASKAPTKKPKATPVKKAKKKLAATPKKAKKPKTVKAKPVKASKPKKAKPVK
CHIMP	KKASKPKKAASKAPTKKPKATPVKKAKKKLAATPKKAKKPKTVKAKPVKASKPKKAKPVK
MOUSE	KKAAKPKKAASKAPSKKPKATPVKKAKKKPAATPKKAKKPKVVKVPVKASKPKKAKTVK
RAT	KKAAKPKKAASKAPSKKPKATPVKKAKKKPAATPKKAKKPKIVKVPVKASKPKKAKPVK
COW	KKAAKPKKAASKAPSKKPKATPVKKAKKKPAATPKKTKKPKTVKAKPVKASKPKKTKPVK
	***.:*****.:***** *****.:***** **.:*****.:*
NON-CONSERVED AMINO ACIDS	Conservative Conservative Non-conservative Conservative Non-conservative Semi-conservative Conservative Non-conservative



# What is alignment? - hypothesis of homology

Histone H1 (residues 120-180)	
HUMAN	KKASKPKKAASKAPTKKPKATPVKKAKKKLAATPKKAKKPKTVKAKPVKASKPKKAKPVK
CHIMP	KKASKPKKAASKAPTKKPKATPVKKAKKKLAATPKKAKKPKTVKAKPVKASKPKKAKPVK
MOUSE	KKAAKPKKAASKAPSKKPKATPVKKAKKKPAATPKKAKKPKVVKVPVKASKPKKAKTVK
RAT	KKAAKPKKAASKAPSKKPKATPVKKAKKKPAATPKKAKKPKIVKVPVKASKPKKAKPVK
COW	KKAAKPKKAASKAPSKKPKATPVKKAKKKPAATPKKTKKPKTVKAKPVKASKPKKTKPVK
	***:*****:***** *****:***** **.******:*
NON-CONSERVED AMINO ACIDS	Conservative Conservative Non-conservative Conservative Non-conservative Semi-conservative Conservative Non-conservative

# How to pick an aligner?

- Type of data you have
  - aligning short read data (Illumina)
  - aligning long read data (PacBio, Nanopore)
  - aligning RNA-seq data - splice aware
- Accuracy
  - percent of reads aligned
  - estimated gene coverage
- Runtime

## Comparison of Short-Read Sequence Aligners Indicates Strengths and Weaknesses for Biologists to Consider

Ryan Musich<sup>1</sup>, Lance Cadle-Davidson<sup>1,2</sup> and Michael V. Osier<sup>1\*</sup>

<sup>1</sup>Thomas H. Gosnell School of Life Sciences, Rochester Institute of Technology, Rochester, NY, United States,

<sup>2</sup>USDA-Agricultural Research Service, Grape Genetics Research Unit, Geneva, NY, United States

### RESEARCH ARTICLE

Open Access

## Benchmarking short sequence mapping tools

Ayat Hatem<sup>1,2</sup>, Doruk Bozdağ<sup>2</sup>, Amanda E Toland<sup>3</sup> and Ümit V Çatalyürek<sup>1,2\*</sup>



A Recent (2020) Comparative Analysis of Genome Aligners Shows HISAT2 and BWA are Among the Best Tools

Article

## Aligning the Aligners: Comparison of RNA Sequencing Data Alignment and Gene Expression Quantification Tools for Clinical Breast Cancer Research

Isaac D. Raplee<sup>1</sup>, Alexei V. Evsikov<sup>2</sup> and Caralina Marín de Evsikova<sup>1,2,\*</sup>

# Aligners we will cover today

- BWA - short read DNA sequences
- bowtie2 - short read DNA sequences
- minimap2 - long read sequences

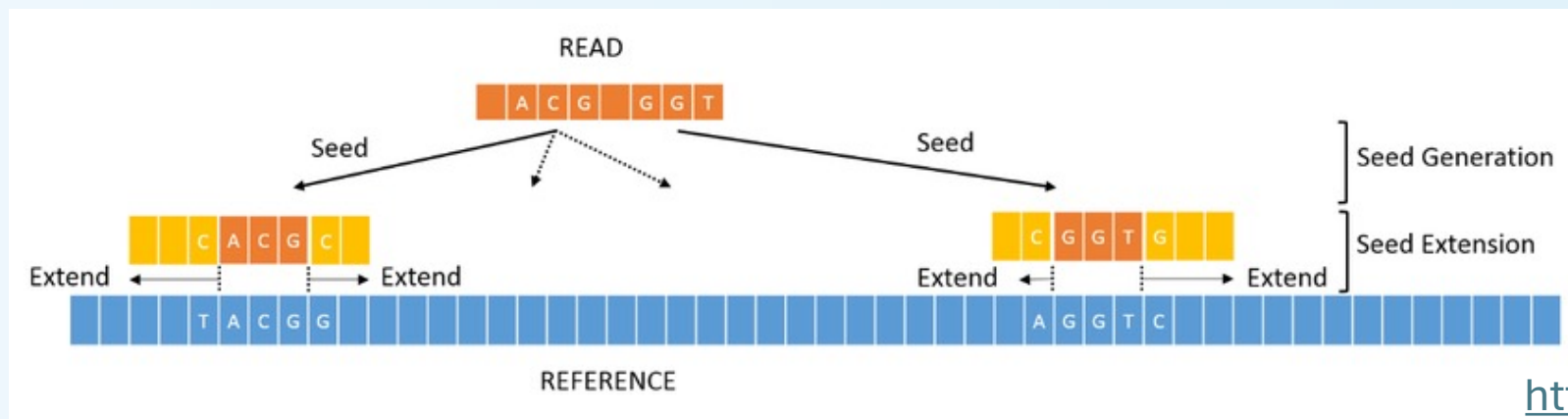
# Indexing

- Most tools require indexing your reference genome, and potentially also aligning reads too it
- Creates standardized map to use across alignments
- Generally, creates secondary files that need to be in the same folder as the genome, with the same name but difference file extension to be used properly



# BWA: Burrows-Wheeler Aligner

- Three algorithms: BWA-Backtrack, BWA-MEM, BWA-SW
  - BWA-Backtrack: designed for Illumina sequence reads up to 100bp
  - BWA-MEM and BWA-SW: for longer sequences ranged from 70bp to 1Mbp
- BWA-MEM and BWA-SW share similar features such as long-read support and split alignment
- BWA-MEM, which is the latest, is generally recommended for high-quality queries as it is faster and more accurate.
- BWA-MEM also has better performance than BWA-backtrack for 70-100bp Illumina reads.



Resources:

<https://github.com/lh3/bwa>

<https://bio-bwa.sourceforge.net/>

# BWA: Burrows-Wheeler Aligner

1. Check available version of bwa on Quest  
module spider bwa

2. Copy bwa.sh script to your folder

```
cp /projects/e32680/04_mapping_dnaseq/bwa.sh .
```

3. Print the contents of the script to the screen  
cat bwa.sh

# BWA: Burrows-Wheeler Aligner

```
#!/bin/bash
#SBATCH --account=p32300
#SBATCH --partition=normal
#SBATCH --job-name=bwa
#SBATCH --output=bwa.%A.out
#SBATCH --error=bwa.%A.err
#SBATCH --time=10:00:00
#SBATCH --mem=15G    ### edit this line to reserve more/less memory
#SBATCH --ntasks=8   ### edit this line to reserve different no. of cores
#SBATCH --nodes=1

module purge
module load bwa/0.7.17

export SAMPLE_DIR=/projects/e32680/04_mapping_dnaseq

bwa mem -t 8 $SAMPLE_DIR/oenotheraHarringtonii.softmasked.fasta \
$SAMPLE_DIR/BAC_12_09_S43_L001_R1_001.fastq.gz \
$SAMPLE_DIR/BAC_12_09_S43_L001_R2_001.fastq.gz > BAC_12_09_S43_L001_aln-bwa.sam
```

# BWA: Burrows-Wheeler Aligner

```
bwa mem -t 8 $SAMPLE_DIR/oenotheraHarringtonii.softmasked.fasta \  
$SAMPLE_DIR/BAC_12_09_S43_L001_R1_001.fastq.gz \  
$SAMPLE_DIR/BAC_12_09_S43_L001_R2_001.fastq.gz \  
> BAC_12_09_S43_L001_aln-bwa.sam
```

How many threads to parallelize over

# BWA: Burrows-Wheeler Aligner

```
bwa mem -t 8 $SAMPLE_DIR/oenotheraHarringtonii.softmasked.fasta \  
$SAMPLE_DIR/BAC_12_09_S43_L001_R1_001.fastq.gz \  
$SAMPLE_DIR/BAC_12_09_S43_L001_R2_001.fastq.gz \  
> BAC_12_09_S43_L001_aln-bwa.sam
```

Reference genome you're mapping to



# BWA: Burrows-Wheeler Aligner

```
bwa mem -t 8 $SAMPLE_DIR/oenotheraHarringtonii.softmasked.fasta \  
$SAMPLE_DIR/BAC_12_09_S43_L001_R1_001.fastq.gz \  
$SAMPLE_DIR/BAC_12_09_S43_L001_R2_001.fastq.gz \  
> BAC_12_09_S43_L001_aln-bwa.sam
```

Paired end reads you are mapping

# BWA: Burrows-Wheeler Aligner

```
bwa mem -t 8 $SAMPLE_DIR/oenotheraHarringtonii.softmasked.fasta \  
$SAMPLE_DIR/BAC_12_09_S43_L001_R1_001.fastq.gz \  
$SAMPLE_DIR/BAC_12_09_S43_L001_R2_001.fastq.gz \  
> BAC_12_09_S43_L001_aln-bwa.sam
```

Your output file

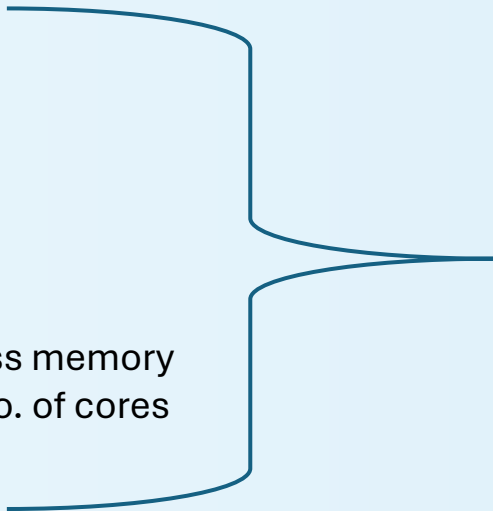
# BWA: Burrows-Wheeler Aligner

```
#!/bin/bash
#SBATCH --account=p32300
#SBATCH --partition=normal
#SBATCH --job-name=bwa
#SBATCH --output=bwa.%A.out
#SBATCH --error=bwa.%A.err
#SBATCH --time=10:00:00
#SBATCH --mem=15G    ### edit this line to reserve more/less memory
#SBATCH --ntasks=8   ### edit this line to reserve different no. of cores
#SBATCH --nodes=1
```

```
module purge
module load bwa/0.7.17
```

```
export SAMPLE_DIR=/projects/e32680/04_mapping_dnaseq
```

```
bwa mem -t 8 $SAMPLE_DIR/oenotheraHarringtonii.softmasked.fasta \
$SAMPLE_DIR/BAC_12_09_S43_L001_R1_001.fastq.gz \
$SAMPLE_DIR/BAC_12_09_S43_L001_R2_001.fastq.gz > BAC_12_09_S43_L001_aln-bwa.sam
```



Resources we're  
reserving to run  
this job

# BWA: Burrows-Wheeler Aligner

Job ID: 8070461

Nodes: 1

Cores per node: 8

CPU Efficiency: 96.28% of 04:25:28 core-walltime

Job Wall-clock time: 00:33:11

Memory Utilized: 6.16 GB

Job ID: 8080270

Nodes: 1

Cores per node: 4

CPU Efficiency: 99.24% of 02:35:56 core-walltime

Job Wall-clock time: 00:38:59

Memory Utilized: 4.14 GB

Job ID: 8080291

Nodes: 1

Cores per node: 16

CPU Efficiency: 97.93% of 03:08:16 core-walltime

Job Wall-clock time: 00:11:46

Memory Utilized: 9.96 GB

CPUUs	Time	RAM used
4	00:38:59	4.14 GB
8	00:33:11	6.16 GB
16	00:11:46	9.96 Gb

How many resources do you think we should ask for?

# BWA EXERCISE

1. Open bwa.sh for editing

```
nano bwa.sh
```

2. Set the following resources

```
#SBATCH --account = e32680
```

```
#SBATCH --partition = short
```

```
#SBATCH --mem = 7G
```

```
#SBATCH --ntasks = 4
```

```
#SBATCH --time = 01:00:00
```

3. Save your changes

```
ctrl^o
```

```
enter
```

```
ctrl^x
```

4. Submit the script

```
sbatch bwa.sh
```



# bowtie2

“an ultrafast and memory-efficient tool for aligning sequencing reads to long reference sequences. It is particularly good at aligning reads of about 50 up to 100s or 1,000s of characters, and particularly good at aligning to relatively long (e.g. mammalian) genomes. Bowtie 2 indexes the genome with an FM Index to keep its memory footprint small: for the human genome, its memory footprint is typically around 3.2 GB. Bowtie 2 supports gapped, local, and paired-end alignment modes.”

Resources:

<https://bowtie-bio.sourceforge.net/bowtie2/index.shtml>

<https://github.com/BenLangmead/bowtie2>

# End-to-end alignment versus local alignment

By default, Bowtie 2 performs end-to-end read alignment. That is, it searches for alignments involving all of the read characters. This is also called an "untrimmed" or "unclipped" alignment.

When the --local option is specified, Bowtie 2 performs local read alignment. In this mode, Bowtie 2 might "trim" or "clip" some read characters from one or both ends of the alignment if doing so maximizes the alignment score.

## End-to-end alignment example

Alignment:

Read:	GACTGGGCGATCTCGACTTCG
Reference:	GACTG - - CGATCTCGACATCG

## Local alignment example

Alignment:

Read:	ACGGTTGCGTTAA -TCCGCCACG
Reference:	TAACTTGCGTTAAATCCGCCTGG

# bowtie2 Exercise: Create a bowtie2.sh script

1. Check available version of bowtie2 on Quest  
    `module spider bowtie2`
2. Copy bwa.sh script to bowtie2.sh  
    `cp bwa.sh bowtie2.sh`
3. Convert the bwa command to a bowtie2 command...

# bowtie2 Exercise: Create a bowtie2.sh script

bowtie2 \

-p <n> \	# number of processors
-x \$SAMPLE_DIR/bowtie2_index \	# genome index name
-1 <reads_1.fastq> \	# read 1 of pair
-2 <reads_2.fastq> \	# read 2 of pair
-S <output>-aln-bowtie2.sam	# output file

Don't forget to load the module!

# bowtie2 Exercise: Create a bowtie2.sh script

#SBATCH --job-name=bowtie	# update job name
#SBATCH --output=bowtie.%A.out	# log file name
#SBATCH --error=bowtie.%A.err	# error file name

Don't forget to load the module!



# bowtie2 Exercise: Create a bowtie2.sh script

1. Check available version of bowtie2 on Quest  
    `module spider bowtie2`
2. Copy bwa.sh script to bowtie2.sh  
    `cp bwa.sh bowtie2.sh`
3. Convert the bwa command to a bowtie2 command...
4. Submit your bowtie2 script  
    `sbatch bowtie2.sh`
5. Check on your jobs

# Check on jobs:

`queue -u <netid>`

# shows all running and pending jobs

`sacct -X`

# shows all jobs from today

`sacct -X -S 100125`

# shows all jobs from this month

# Debates/comparisons of aligners

- <https://www.biostars.org/p/125020/>
- <https://www.seqanswers.com/forum/bioinformatics/bioinformatics-aa/12790-bowtie-2-versus-bwa/page4>
- <https://help.galaxyproject.org/t/bowtie2-and-bwa-aligners/11953/3>

# minimap2

Typical use cases include:

1. mapping PacBio or Oxford Nanopore genomic reads to the human genome
2. finding overlaps between long reads with error rate up to ~15%
3. splice-aware alignment of PacBio Iso-Seq or Nanopore cDNA or Direct RNA reads against a reference genome
4. aligning Illumina single- or paired-end reads
5. assembly-to-assembly alignment
6. full-genome alignment between two closely related species with divergence below ~15%

Resources:

<https://lh3.github.io/minimap2/minimap2.html>

<https://github.com/lh3/minimap2>

# minimap2

```
module load minimap2/2.24
```

```
# latest version is 2.28
```

Without any options, minimap2 takes a reference database and a query sequence file as input and produce approximate mapping, without base-level alignment (i.e. coordinates are only approximate and no CIGAR in output), in the [PAF format](#):

- `minimap2 ref.fa query.fq > approx-mapping.paf`
- You can ask minimap2 to generate CIGAR at the cg tag of PAF with:
- `minimap2 -c ref.fa query.fq > alignment.paf`
- or to output alignments in the [SAM format](#):
- `minimap2 -a ref.fa query.fq > alignment.sam`





### COMPUTING AND SOFTWARE

Access to high performance computing, research software, and global networks for conducting computationally intense research.



### DATA MANAGEMENT AND SHARING

Learn about data management planning and options for storing, securing, transferring, and sharing data.



### DATA SCIENCE, STATISTICS, AND VISUALIZATION

Support for collecting, analyzing, visualizing, and programming with research data.



### TRAINING AND CONSULTATION

Identify events, resources, and people to help you learn computational and data skills for your research.

# Research Computing and Data Services

*We're here to help after the workshop!*

**[quest-help@northwestern.edu](mailto:quest-help@northwestern.edu)**

**[bit.ly/rcdsconsult](https://bit.ly/rcdsconsult)**

**<https://sites.northwestern.edu/researchcomputing/>**