

# Mapping Tutorial

For this tutorial we will work in a folder called **mapping**

```
$ cd
$ mkdir mapping
$ cd mapping
$ mkdir 1347
$ mkdir 2087
$ mkdir 5035
$ cp ~/Course_Data/{PG15-BW001347.R1.fastq.gz,PG15-BW001347.R2.fastq.gz,All_HIV_Ref.fas,HIVHXB2CG.fa} ~/mapping/1347
$ cp ~/Course_Data/{PG16-BW002087.R1.fastq.gz,PG16-BW002087.R2.fastq.gz,All_HIV_Ref.fas} ~/mapping/2087
$ cp ~/Course_Data/{5035_R1.fastq.gz,5035_R2.fastq.gz,All_NORO_Ref.fas} ~/mapping/5035
$ ls -lh
$ trim_galore ~/mapping/1347/*.gz -q 30 --phred33 -o ~/mapping/1347/ --illumina --max_n 2 --paired
$ trim_galore ~/mapping/2087/*.gz -q 30 --phred33 -o ~/mapping/2087/ --illumina --max_n 2 --paired
$ trim_galore ~/mapping/5035/*.gz -q 30 --phred33 -o ~/mapping/5035/ --illumina --max_n 2 --paired
$ cd ~/mapping
```

## Mapping to A Reference

So we know we have sequenced HIV from a clinical sample and want to assemble the genome to extract the consensus sequence and identify variant position. The most well known HIV1 sequence in the database is the HXB2 strain which is a B type genotype. Will this be an ideal reference sequence?

For mapping we will start with **BWA**. It uses a Burrowes Wheeler based algorithm to align. BWA included three different algorithm out of which we will use **BWA mem**. We will test the effect of other alignment tools later in the tutorial.

```
$ bwa mem
```

We will work in the subfolder **1347**

```
$ cd ~/mapping/1347
$ ls -lh
```

We first index the HXB2 reference sequence before mapping using bwa index. The reference sequence is in HIVHXB2CG.fa

```
$ bwa index HIVHXB2CG.fa
$ ls -lh
```

You can see multiple files have been created that contains the index information. We can now map the trimmed file to the HXB2 reference

```
$ bwa mem -t 4 -v 1 HIVHXB2CG.fai PG15-BW001347.R1_val_1.fq.gz PG15-BW001347.R2_val_2.fq.gz > 1347HXB2ref.sam
```

**-t** : Number of threads

**-v** : Verbosity level of the output

Reference file

Input file 1 and 2

**'>'** : Output file

Let us look at how many reads mapped to the HXB2 genome. For most manipulation of alignment files we use **samtools**

```
$ samtools --help
$ samtools flagstat 1347HXB2ref.sam
```

**How many reads map to HXB2 genome?**

**Do you think this is a good choice of a reference?**

We can now map to a curated dataset of HIV1 references. If the B type is indeed the correct reference all reads that map to the curated database should have also mapped to the HXB2 reference. Let's check if that is true

```
$ bwa index All_HIV_Ref.fas
$ bwa mem -t 4 -v 1 All_HIV_Ref.fai PG15-BW001347.R1_val_1.fq.gz PG15-BW001347.R2_val_2.fq.gz > 1347multiref.sam
$ samtools flagstat 1347multiref.sam
```

**How many reads map to the curated genomes?**

**Based on this information do you think HXB2 was a good reference?**

**Note:** In the second case we will have a lower proportion of the reads mapped in proper pairs. That is because we are using multiple references where one of the pair maps better to a different genome than the other.

So how do we identify the best reference to map our reads to? Let us look at the mapping to identify the reference sequence that has the maximum number of reads mapping to it.

```
$ sed -e '1,171d' 1347multiref.sam | cut -f3 | sed -e '/*d' | sort | uniq -c |
sort -rn | head -n 10 > top10hits.txt
$ cat top10hits.txt
```

This shell command counts the number of read that map to each reference in the multi-reference files and sort them from maximum to minimum and picks the top 10 hits. 171 is the number of references in the All\_HIV\_Ref.fas + 1 (170+1)

**What do you think is the correct reference sequence to map the reads?**

**Extracting the best Reference**

We are going to extract the C type Reference from our All\_HIV\_Ref.fas file

```
$ samtools faidx All_HIV_Ref.fas Ref.C.IN.95.95IN21068.AF067155 > besthit.fasta
$ cat besthit.fasta | head
```

**faidx** : Indexes and Extracts sequence from a multi-reference FASTA file

Let us now map to the top hit and check if we have picked the correct reference sequence

```
$ bwa index besthit.fasta
$ bwa mem -t 4 -v 1 besthit.fasta PG15-BW001347.R1_val_1.fq.gz PG15-
BW001347.R2_val_2.fq.gz > 1347singleref.sam
$ samtools flagstat 1347multiref.sam
$ samtools flagstat 1347HXB2ref.sam
$ samtools flagstat 1347singleref.sam
```

**Do you think Ref.C.IN.95 is a better reference for our sample?**

**Note:** In this case there is a lot of crossover mapping from C type to B but there are some HIV genotypes that are so distinct that the proportion of reads mapping to the incorrect reference will be significantly lower.

Now repeat this with the sample in the subfolder **2087**

```
$ cd ~/mapping/2087
$ ls -lh
```

**What is the best reference for this sequence?**

**What proportion of reads map to the reference?**

**Is that enough to give a good mapping across the genome?**

**Note:** We will be looking at visualization of these alignments later on in the course.

## Identifying Mixed Infections

---

We are going to define mixed infection as infection with two different genotypes. Multiple infection with the same genotype is usually difficult to tease out by mapping. We can look at the change in proportion of reads mapped to give us a clue if there is another genotype in our sample. For this part of the tutorial we will work in the subfolder **5035**.

**Note :** The example is done with Norovirus but holds true for HIV

```
$ cd ~/mapping/5035
$ ls -lh
```

We will first identify the top hit and see if it best explains the data

```

$ bwa index All_NORO_Ref.fas
$ bwa mem -t 4 -v 1 All_NORO_Ref.fas 5035_R1_val_1.fq.gz 5035_R2_val_2.fq.gz >
5035multiref.sam
$ sed -e '1,91d' 5035multiref.sam | cut -f3 | sed -e '/*d' | sort | uniq -c |
sort -rn | head -n 10 > top10hits.txt
$ cat top10hits.txt
$ samtools faidx All_NORO_Ref.fas AB983218_Kawasaki323_GIIP17_GII17 >
besthit.fasta
$ bwa index besthit.fasta
$ bwa mem -t 4 -v 1 besthit.fasta 5035_R1_val_1.fq.gz 5035_R2_val_2.fq.gz >
5035singleref.sam
$ samtools flagstat 5035multiref.sam
$ samtools flagstat 5035singleref.sam

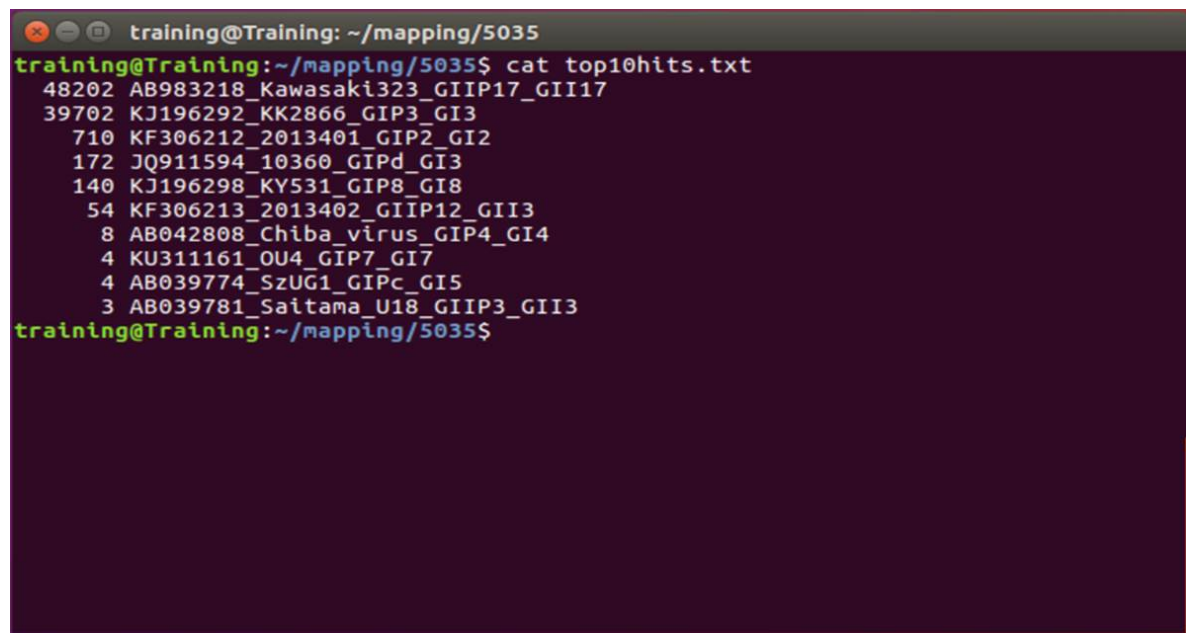
```

**How many reads map to the multiref vs singleref file?**

**How do you explain a massive difference in reads between the two even though we have selected the top hit?**

In this case we go back to the top10hits file and see if we can identify other references that may be present.

```
$ cat top10hits.txt
```



```

training@Training: ~/mapping/5035
training@Training:~/mapping/5035$ cat top10hits.txt
48202 AB983218_Kawasaki323_GIIP17_GII17
39702 KJ196292_KK2866_GIP3_GI3
710 KF306212_2013401_GIP2_GI2
172 JQ911594_10360_GIPd_GI3
140 KJ196298_KY531_GIP8_GI8
54 KF306213_2013402_GIIP12_GII3
8 AB042808_Chiba_virus_GIP4_GI4
4 KU311161_OU4_GIP7_GI7
4 AB039774_SzUG1_GIPc_GI5
3 AB039781_Saitama_U18_GIIP3_GII3
training@Training:~/mapping/5035$

```

The two strains that could be present in the sample are **AB983218\_Kawasaki323\_GIIP17\_GII17** (best hit) and **KJ196292\_KK2866\_GIP3\_GI3**. Let's map back to these two references to see if we can get recover all the reads that map to the multi-reference file

```

$ samtools faidx All_NORO_Ref.fas AB983218_Kawasaki323_GIIP17_GII17
KJ196292_KK2866_GIP3_GI3 > besthitdual.fasta
$ bwa index besthitdual.fasta
$ bwa mem -t 4 -v 1 besthitdual.fasta 5035_R1_val_1.fq.gz 5035_R2_val_2.fq.gz >
5035dualref.sam
$ samtools flagstat 5035multiref.sam
$ samtools flagstat 5035dualref.sam

```

**Do you think this sample was a mixed infection?**

This is a very simple check to identify mixed infections in our samples

## Aligners

For this section we will work in the subfolder **1347**

```
$ cd ~/mapping/1347
```

### Global vs Local Alignment

Global	Local
Read: GACTGGGCGATCTCGACTTCG	Read: ACGGTTGCGTTAATCCGCCACG
Reference: GACTGCGATCTCGACATCG	Reference: TAACTTGCGTTAAATCCGCCTGG
Alignment:	
Read: GACTGGGCGATCTCGACTTCG	Read: ACGGTTGCGTTAA-TCCGCCACG
Reference: GACTG--CGATCTCGACATCG	Reference: TAACTTGCGTTAAATCCGCCTGG

BWA mem was an example of a local aligner. We are going to use **Bowtie2** as an example of both global vs local alignment. Bowtie2 by default does an global alignment but has a local option too. Let us test how this effects the number of reads mapping when all other parameters are kept the same. Bowtie2 can be found in

```
$ ~/software/bowtie2-2.3.5.1-linux-x86_64/bowtie2 -h
or
$ bowtie2 -h
```

We will first index the reference file and then map the trimmed reads using bowtie2. In this case we already know the most suitable reference (besthit.fasta) and we can use it directly

```
$ bowtie2-build besthit.fasta besthit.fasta
$ bowtie2 -x besthit.fasta -1 PG15-BW001347.R1_val_1.fq.gz -2 PG15-
BW001347.R2_val_2.fq.gz -S 1347global.sam
$ bowtie2 --local -x besthit.fasta -1 PG15-BW001347.R1_val_1.fq.gz -2 PG15-
BW001347.R2_val_2.fq.gz -S 1347local.sam
$ samtools flagstat 1347global.sam
$ samtools flagstat 1347local.sam
```

bowtie2-build indexes the reference file. In this case we have named the input reference file and the indexed reference file the same

**-x** : Indexed file

**-1** : First FASTQ file

**-2** : Second FASTQ file

**-S** : Output SAM file

**--local** : Local alignment. The default is global

**How many reads map to each?**

**How does this difference change downstream analysis?**

## Other Commonly used Aligners

We are going to use two more aligners in this section. **Tanoti** which is a hash based aligner designed primarily for assembling viral sequences and **bbMap** which is a splice aware global aligner. We will continue working in the same folder.

**Note:** Tanoti does not accept gzipped FASTQ files so we have to unzip the files before running. We will use the same input for bbMap too.

```
$ gunzip *.gz
```

**Tanoti** can be found in

```
$ ~/software/Tanoti-1.2-Linux/tanoti -h  
or  
$ tanoti -h
```

**Note:** Tanoti does not accept gzipped FASTQ files so we have to unzip the files before running

```
$ tanoti -r besthit.fasta -i PG15-BW001347.R1_val_1.fq PG15-BW001347.R2_val_2.fq  
-p 1 -o 1347tanoti.sam
```

**-r** : Reference file

**-i** : Input files (Unzipped)

**-p** : 1 = Paired 0 = Single

**-o** : Output

**bbMap** can be found in

```
$ ~/software/bbmap/bbmap.sh
```

```
$ ~/software/bbmap/bbmap.sh -Xmx2G nodisk ref=besthit.fasta in=PG15-  
BW001347.R1_val_1.fq in2=PG15-BW001347.R2_val_2.fq out=1347bbmapglobal.sam
```

nodisk : Indexes the reference fresh every time

ref : Reference sequence

in : First input read file

in2 : Second input read file

**bbMap** can also be run as an local aligner (default is global) by adding the 'local' flag

**Note:** It is not a true local aligner as it first makes a global alignment and then optimizes it to be a local alignment

```
$ ~/software/bbmap/bbmap.sh -Xmx2G nodisk local ref=besthit.fasta in=PG15-  
BW001347.R1_val_1.fq in2=PG15-BW001347.R2_val_2.fq out=1347bbmaplocal.sam
```

## Comparison of different aligners

```
$ samtools flagstat 1347singleref.sam
$ samtools flagstat 1347local.sam
$ samtools flagstat 1347global.sam
$ samtools flagstat 1347tanoti.sam
$ samtools flagstat 1347bbmaplocal.sam
$ samtools flagstat 1347bbmapglobal.sam
```

Aligners	File	Number of Reads
Bwa mem	1347singleref.sam	900062
Bowtie2 Local	1347local.sam	718755
Bowtie2 Global	1347global.sam	336948
Tanoti	1347tanoti.sam	736060
BBMap Local	1347bbmaplocal.sam	703910
BBMap Global	1347bbmapglobal.sam	703911

Here we see that different aligners can give different results when we run them. At the consensus level usually there are no effects. Some aligners can have effects on low frequency variants. **The main point to remember is we usually cannot compare numbers across aligners so it is best to stick to one aligner when running a dataset where you expect to compare the outputs from them.**