Lecture 5

Genome data analysis 2: Mapping

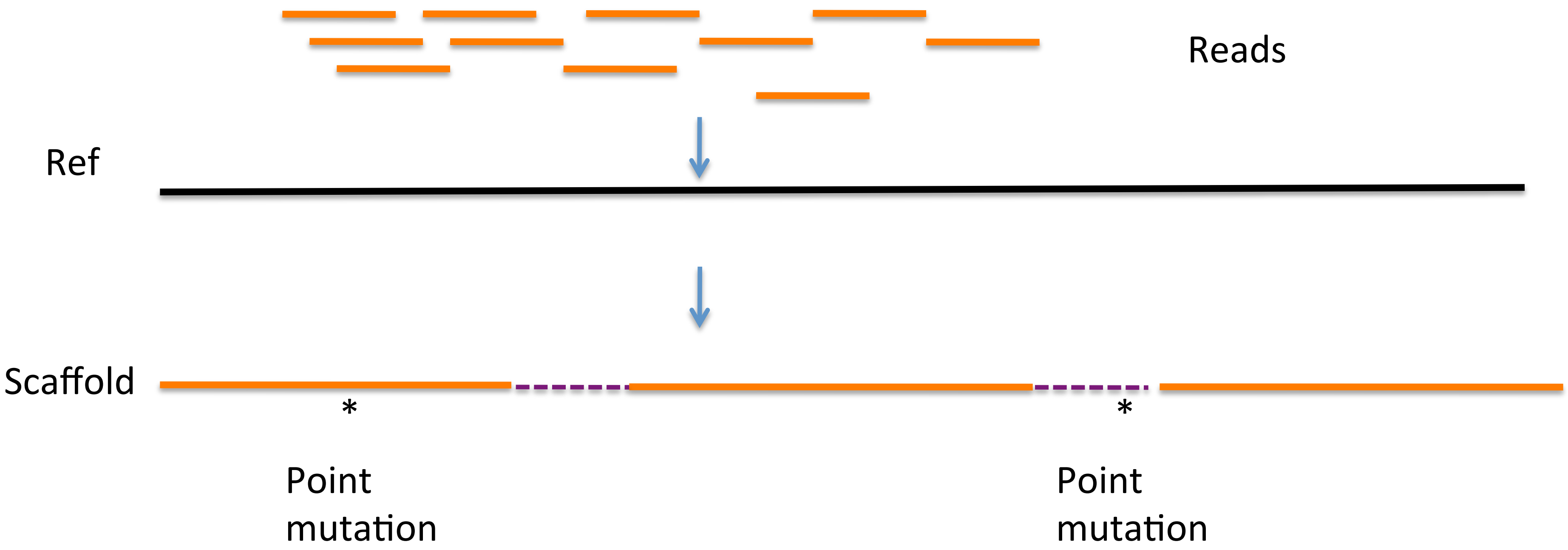
Dealing with millions of small reads

Museum of the
Inquisition – Lima

- Dealing with millions of small reads

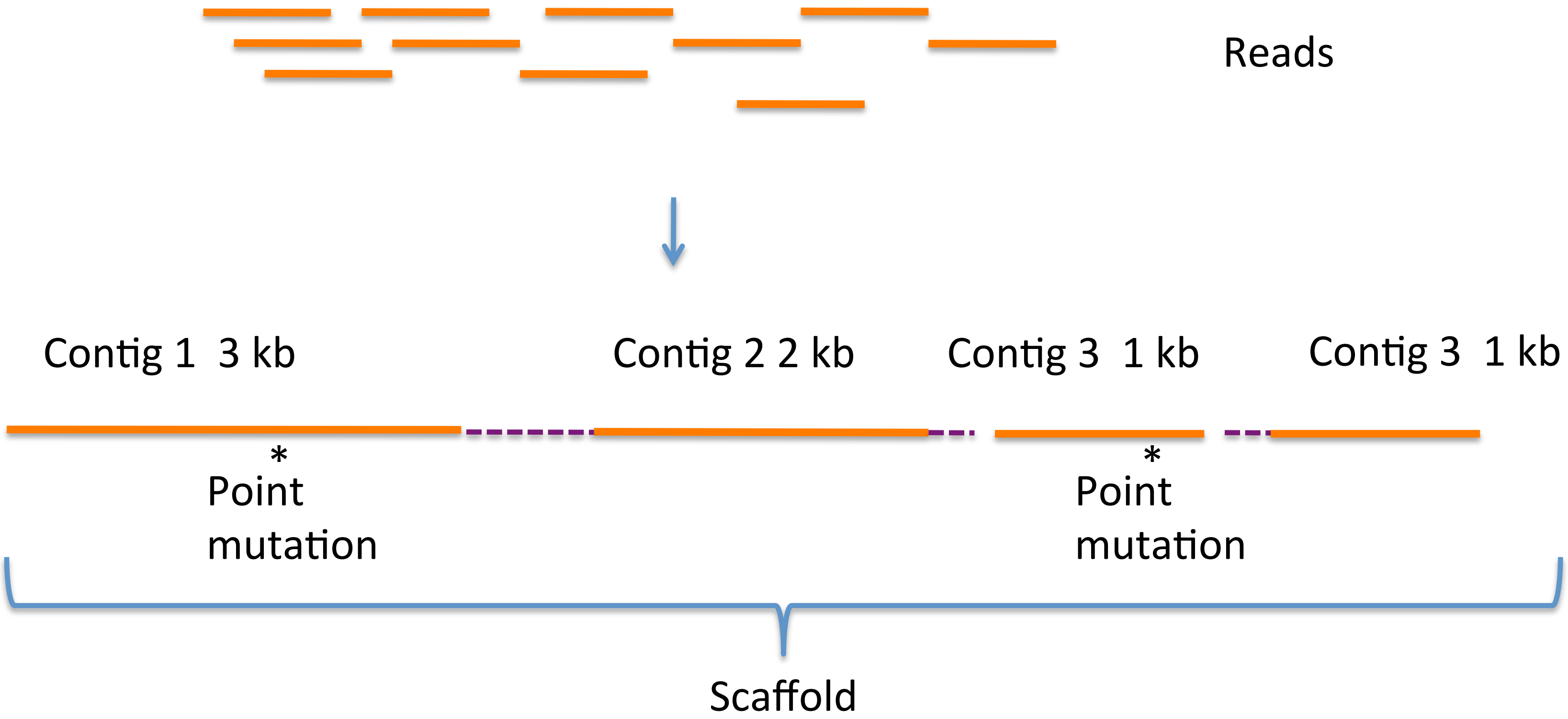
Dealing with millions of small reads

Mapping



Dealing with millions of small reads

de novo assembly

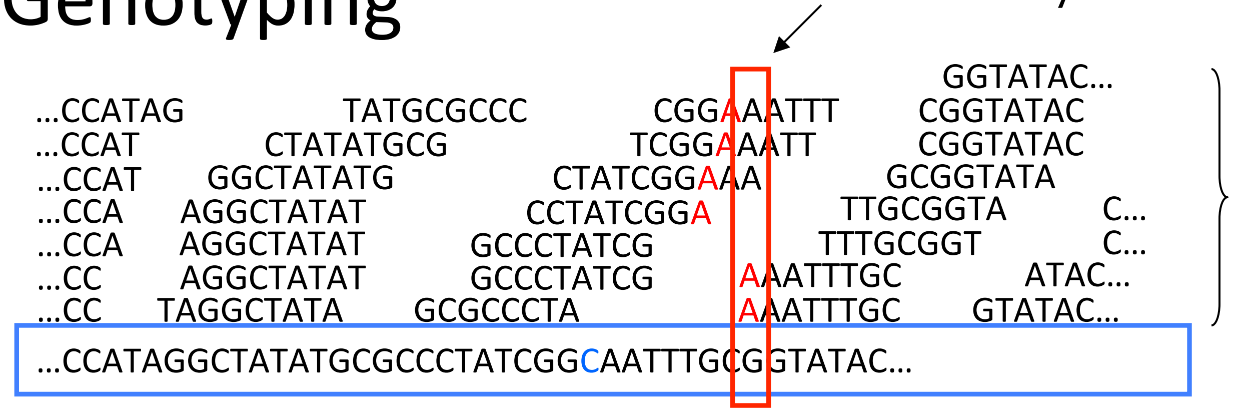


N50 is the contig length such that using equal or longer contigs produces half the bases of the assembly. This can be thought of as the point of half of the mass of the distribution.

Coverage = L x N/ G
L = Length reads
N = number of reads
G = Length of the genome

Short Read Applications

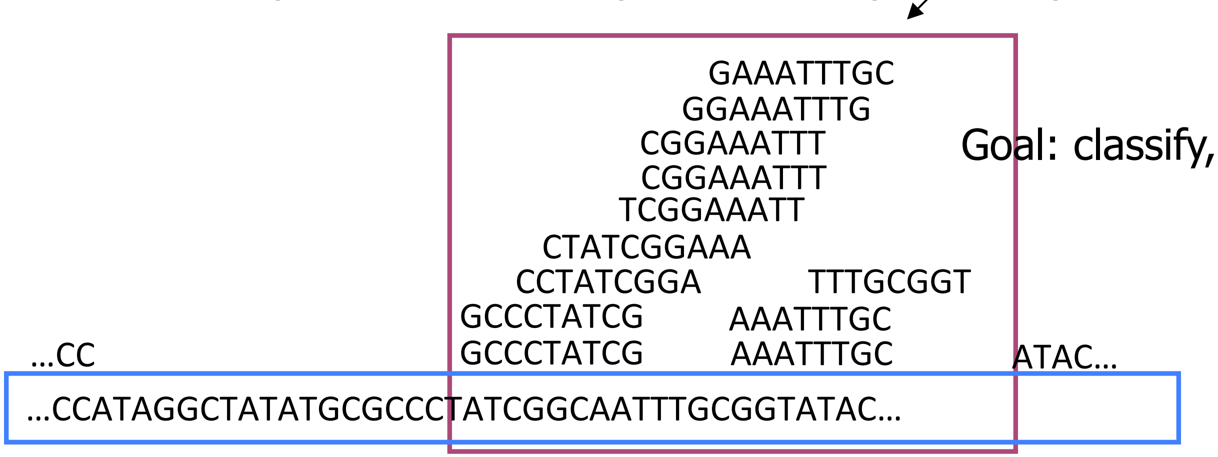
- Genotyping



Goal: identify variations

```
                                                       GGTATAC...
...CCATAG          TATGCGCCC          CGGAAATTT      CGGTATAC
...CCAT         CTATATGCG           TCGGAAATT        CGGTATAC
...CCAT      GGCTATATG          CTATCGGAAA         GCGGTATA
...CCA     AGGCTATAT          CCTATCGGA          TTGCGGTA          C...
...CCA     AGGCTATAT        GCCCTATCG           TTTGCGGT           C...
...CC      AGGCTATAT        GCCCTATCG         AAATTTGC           ATAC...
...CC     TAGGCTATA       GCGCCCTA            AAATTTGC      GTATAC...
...CCATAGGCTATATGCGCCCTATCGGCAATTTGCGGTATAC...
```

- RNA-seq, ChIP-seq, Methyl-seq

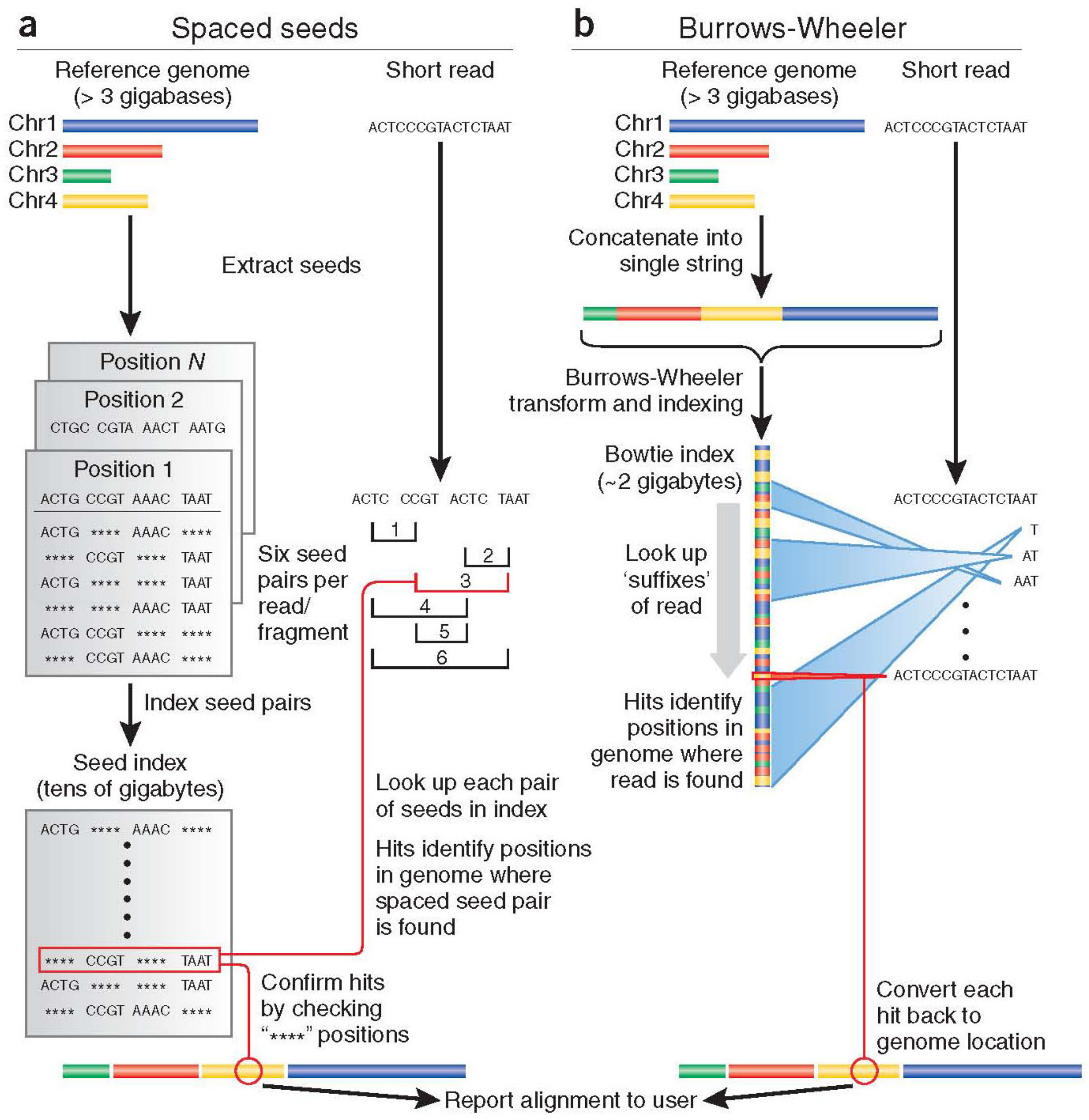


Goal: classify, measure significant peaks

```
                                  GAAATTTGC
                                 GGAAATTTG
                               CGGAAATTT
                               CGGAAATTT
                              TCGGAAATT
                          CTATCGGAAA
                         CCTATCGGA        TTTGCGGT
                       GCCCTATCG         AAATTTGC
...CC                  GCCCTATCG         AAATTTGC          ATAC...
...CCATAGGCTATATGCGCCCTATCGGCAATTTGCGGTATAC...
```

CBCB
CENTER FOR BIOINFORMATICS & COMPUTATIONAL BIOLOGY

Bowtie: A Highly Scalable Tool for Post-Genomic Datasets


a
Spaced seeds
Reference genome
(> 3 gigabases)
Short read
Chr1
Chr2
Chr3
Chr4
ACTCCCGTACTCTAAT
Extract seeds
Position N
Position 2
CTGC CGTA AACT AATG
Position 1
ACTG CCGT AAAC TAAT
ACTG **** AAAC ****
**** CCGT **** TAAT
ACTG **** **** TAAT
**** **** AAAC TAAT
ACTG CCGT **** ****
**** CCGT AAAC ****
ACTC CCGT ACTC TAAT
1
2
3
4
5
6
Six seed pairs per read/ fragment
Index seed pairs
Seed index
(tens of gigabytes)
ACTG **** AAAC ****
**** CCGT **** TAAT
ACTG **** **** TAAT
**** CCGT AAAC ****
Look up each pair of seeds in index
Hits identify positions in genome where spaced seed pair is found
Confirm hits by checking "****" positions
Report alignment to user
b
Burrows-Wheeler
Reference genome
(> 3 gigabases)
Short read
Chr1
Chr2
Chr3
Chr4
ACTCCCGTACTCTAAT
Concatenate into single string
Burrows-Wheeler transform and indexing
Bowtie index
(~2 gigabytes)
ACTCCCGTACTCTAAT
T
AT
AAT
Look up 'suffixes' of read
ACTCCCGTACTCTAAT
Hits identify positions in genome where read is found
Convert each hit back to genome location

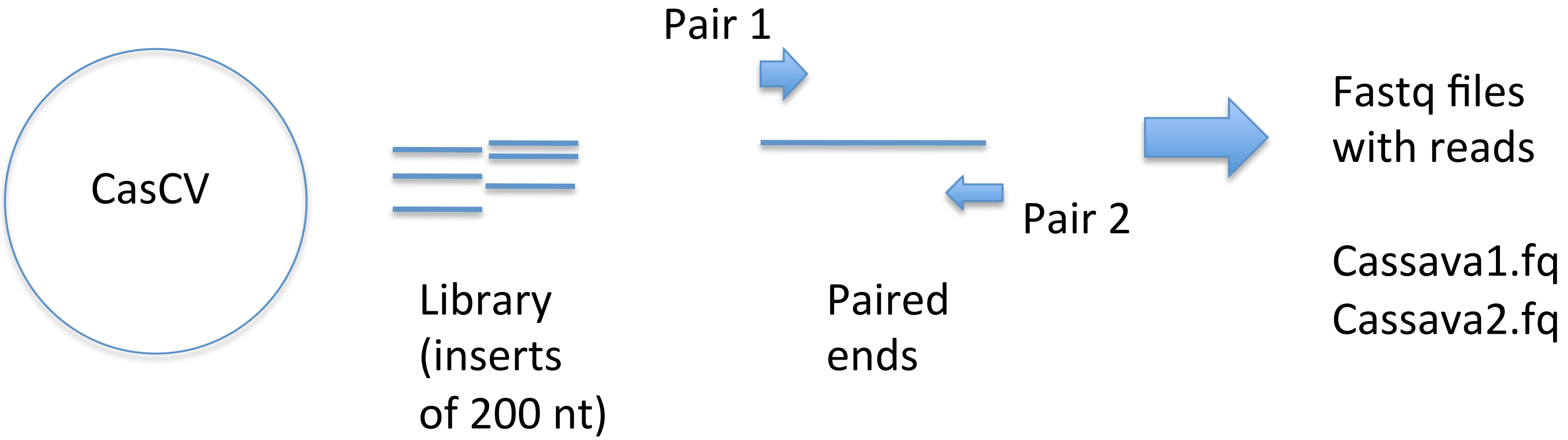
The Bowtie output

- A set of alignments for each short read
- This output can be parsed using a second tool: SAMTOOLS

- EXAMPLE:

Identification of a new variant of a cassava virus

- Dayaram et. al. Reported a novel virus (CasCV) isolated from cassava (2012). This is a circular ssDNA virus of 2220 nt.
- You have illumina sequences (paired end) from a pool of new CasCV isolates. Using Bowtie map these reads in CasCV and find a possible point mutations in your isolate.



Mapping pipeline

- Check your fastq files with FastaQC
- Use these files to map in cassava genome: File: cassava_virus.fna
- First create a index for the target for cassava_virus.fna
- bowtie2-build cassava_virus.fna cassava_virus
- Map cassava_virus1.fq and cassava_virus2.fq
- bowtie2 -x cassava_virus -1 cassava1.fq -2 cassava2.fq results.sam

Mapping pipeline

- SAM FORMAT:
- Use less to inspect the results.sam file
- Describe the file

- 1 Read name
- 2- SAm flag
- 3- target
- 4-position
- 5-mapping quality based on the scores, 0= non-unique, > probably unique
- 6-CIGAr string (describes position of insertion/deletions/matches) For example 35M 35 matches
- 7-name of the mate (often =)
- 8-postion of the mate (mate info)
- 9-template length
- 10-read sequence
- 11-read quality
- 12- program specififc flags

SNP calling (formatting first)

- Use samtools to parse results
- #The results are in sam format and have to be compacted in bam format using samtools
- samtools view -bS -o results.bam results.sam
- #bam file have to be sorted
- samtools sort results.bam -o results.sorted.bam
- #and indexed
- samtools index results.sorted.bam

SNP calling (use freebayes)

- By default freebayes ‘thinks’ that your reference is a diploid. But we are working with a virus

- freebayes -f cassava_virus.fna results.sorted.bam > possible_SNPs.vcf

Parse the vcf file

```
grep 'gi|' possible_SNPs.vcf | cut -f 1,2,3,4,5,6
```

Lets stop and think ….Check the results and the SNPs. How do you explain these results?

Check

```
samtools tview results.sorted.bam cassava_virus.fna
```

To move type g and them fill the name of the genome : position

Other tools

- BWA
- Maq