Multi-genome mapping: Are short-read mapping tools influenced by the order of reference sequences?

Sarah Krautwurst, FSU Jena

39th TBI Winterseminar, Bled, Feb 12, 2024

FRIEDRICH-SCHILLER-
UNIVERSITÄT
JENA

Project background

- two widespread bee viruses: *Deformed Wing Virus* (DWV-A) and *Varroa destructor virus-1* (DWV-B)
- approx. 84% sequence similarity, mismatches distributed across the whole sequence alignment
- recombination between DWV-A and DWV-B possible in case of co-infection → new viral strains with potentially altered virulence and host range

Data and experimental setup

Apis mellifera, honey bee

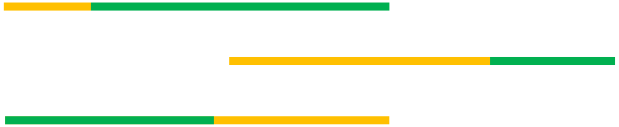
Bombus terrestris, bumble bee

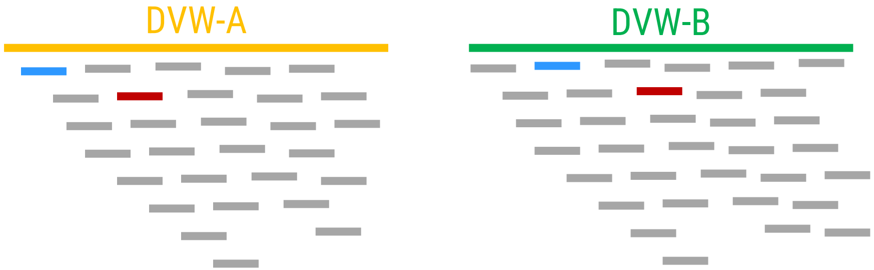
- both bee species infected with *Deformed Wing Virus* (DWV-A) or *Varroa destructor virus-1* (DWV-B) or co-infected with both viruses (6 conditions total)
- 10 replicates, 10 passagings
- Illumina sequencing, done by Robert Paxton lab (Halle)

~80 samples

Goal of the project

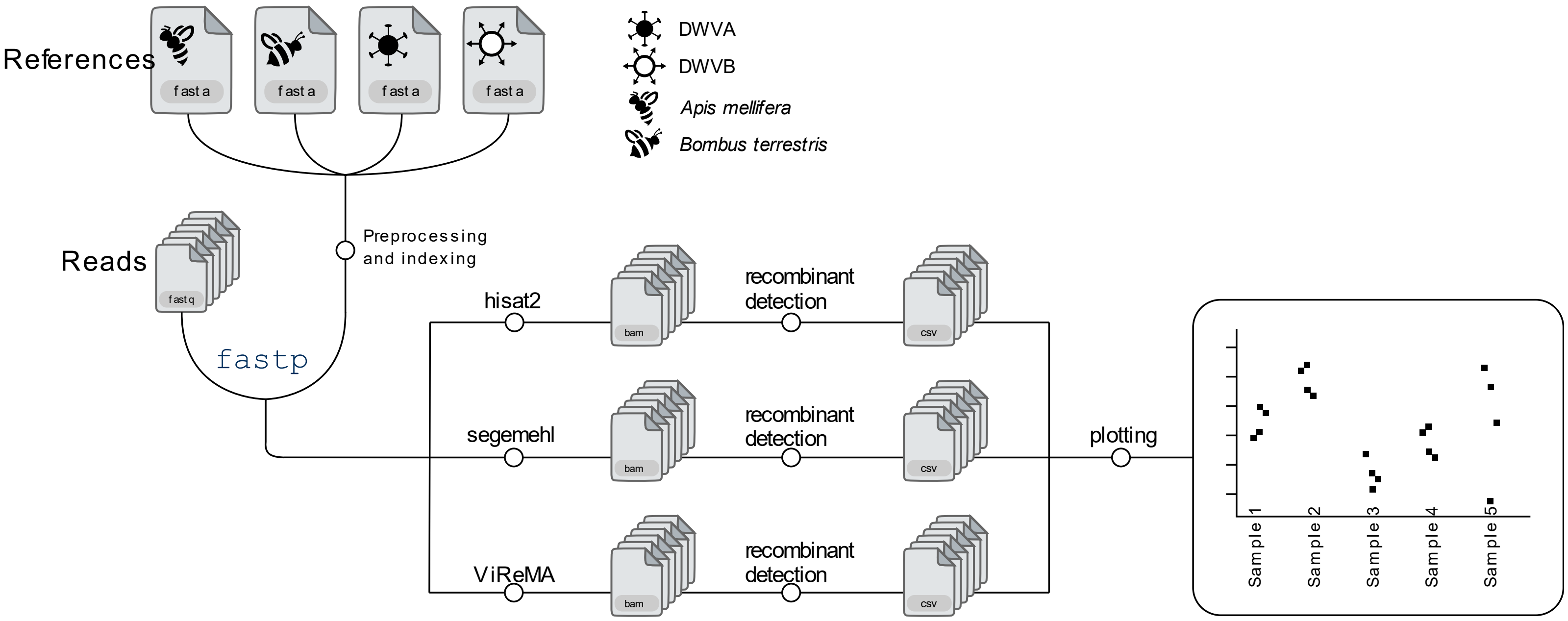
analyze potential **recombination events** between DWV-A and DWV-B and find the breaking points





count and investigate paired-end reads that map on both genomes

Approach for recombinant detection

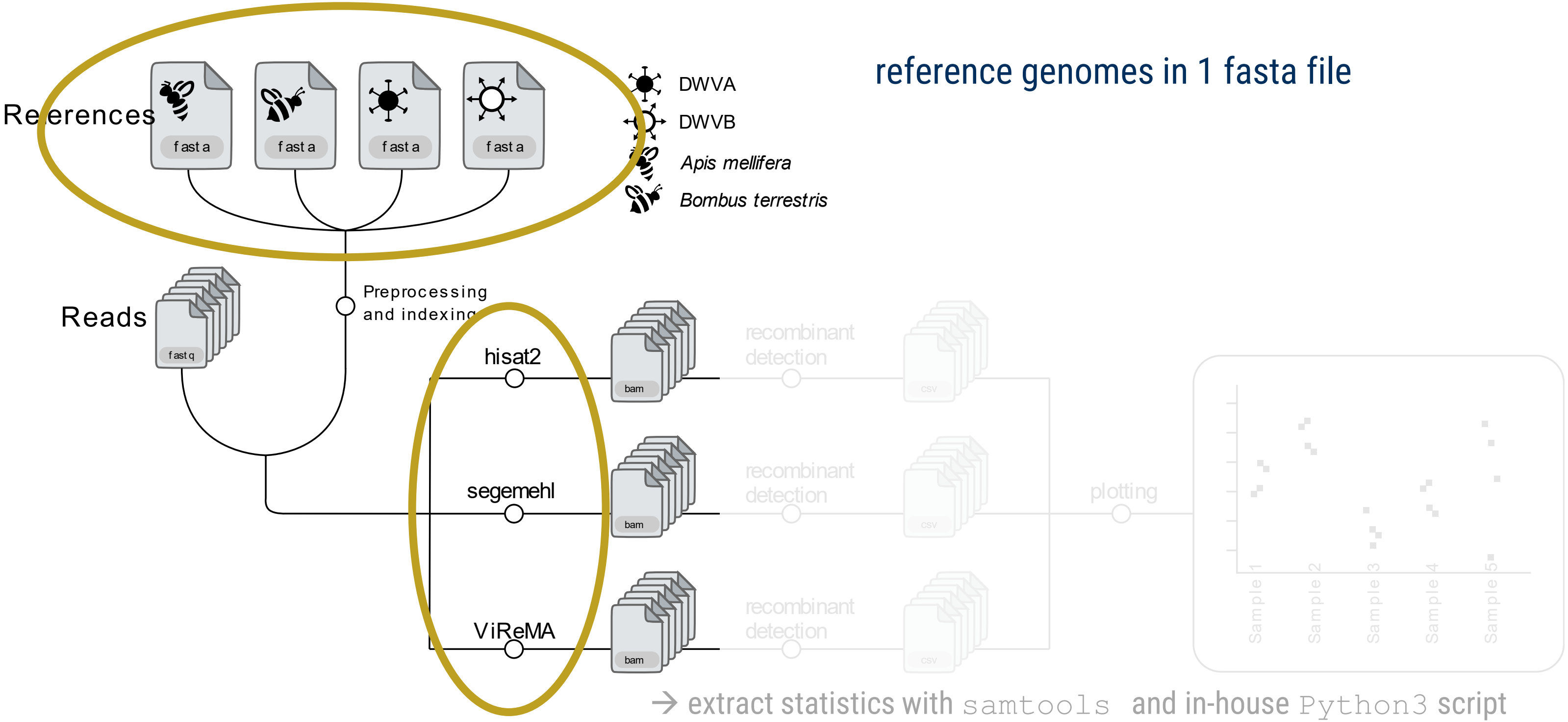


Technical side quest

- strategy for multi-genome mapping:
  map reads to all reference genomes at once instead of iteratively
  → Does the order of the reference genomes matter?

- aim: **Which mapping tool and which reference genome combination leads to the most robust mapping results?**

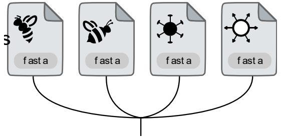
Technical side quest: reference genome order



reference genomes in 1 fasta file

→ extract statistics with `samtools` and in-house `Python3` script

Reference genomes



- viral genomes: *de novo* assembly from Paxton lab (inoculum samples)
- host genomes: GCF_003254395.2 (*A. mellifera*), GCF_000214255.1 (*B. terrestris*)

- in theory: 24 combinations in order

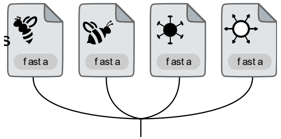
DVW-A

DVW-B

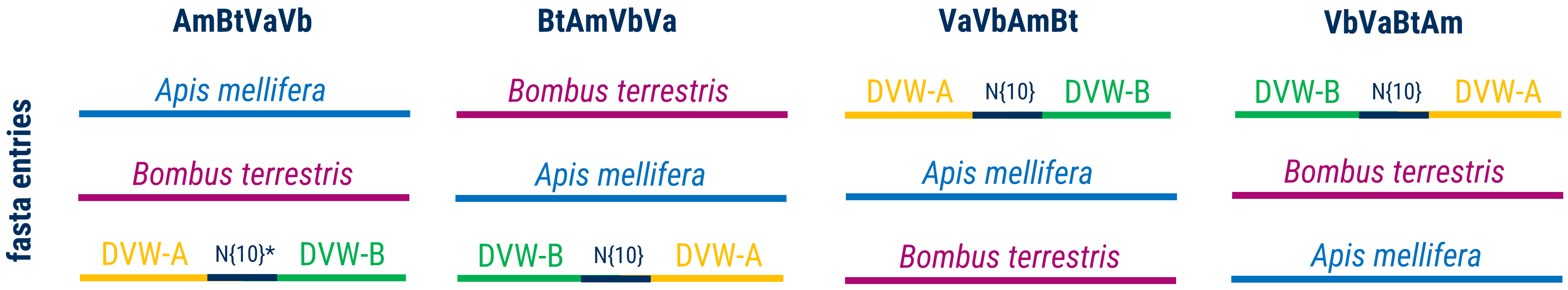
Apis mellifera

Bombus terrestris

Reference genomes



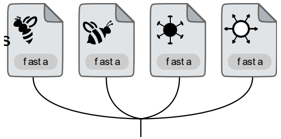
- in theory: 24 combinations in order → start with 4 selected combinations



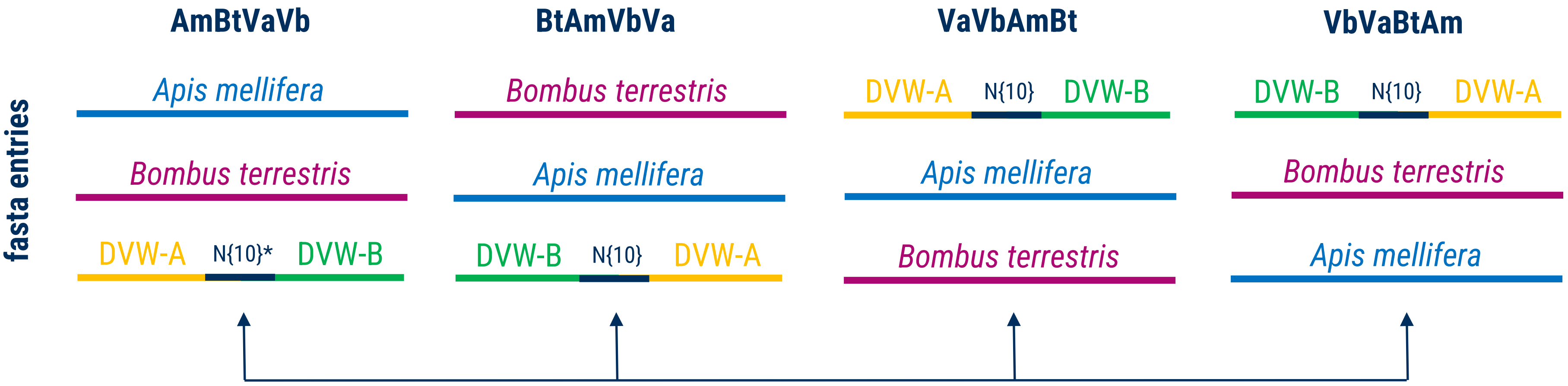
fasta entries	AmBtVaVb	BtAmVbVa	VaVbAmBt	VbVaBtAm
	Apis mellifera	*Bombus terrestris*	DVW-A N{10} DVW-B	DVW-B N{10} DVW-A
	Bombus terrestris	*Apis mellifera*	*Apis mellifera*	*Bombus terrestris*
	DVW-A N{10}* DVW-B	DVW-B N{10} DVW-A	*Bombus terrestris*	*Apis mellifera*

*concatenated "pseudogenome"

Reference genomes



- in theory: 24 combinations in order → start with 4 selected combinations



fasta entries

AmBtVaVb	BtAmVbVa	VaVbAmBt	VbVaBtAm
Apis mellifera	*Bombus terrestris*	DVW-A N{10} DVW-B	DVW-B N{10} DVW-A
Bombus terrestris	*Apis mellifera*	*Apis mellifera*	*Bombus terrestris*
DVW-A N{10}* DVW-B	DVW-B N{10} DVW-A	*Bombus terrestris*	*Apis mellifera*

apply to all samples and for each mapping tool (`hisat2, segemehl, ViReMa`)

*concatenated "pseudogenome"

Current progress

- all samples mapped with hisat2 for first 4 reference combinations
- overall mapping rates seem similar per sample, e.g.:

MBombABG10L08	mapped	unmapped
AmBtVaVb	51,447,551 (93.52%)	3,567,077
BtAmVbVa	51,446,596 (93.51%)	3,568,032
VaVbAmBt	51,447,548 (93.52%)	3,567,080
VbVaBtAm	51,446,596 (93.51%)	3,568,032

Current progress

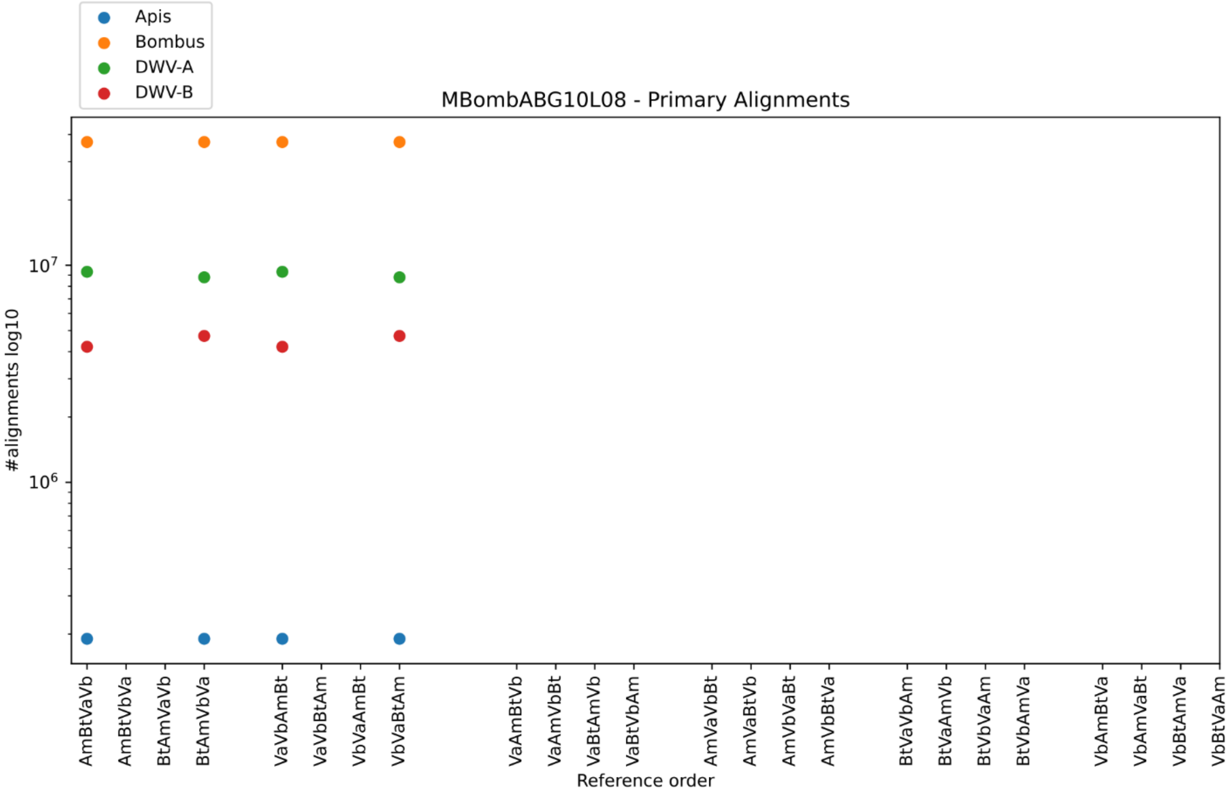
- extract further statistics for all mappings: Where do reads map on the different reference genomes?

→ plot per sample: comparison between reference combinations

Current progress

- extract further statis reference genomes?

→ plot per sample: cor



Next steps & Outlook

- continue mappings for reference combinations with segemehl and ViReMa
- evaluation of mappings → all 24 reference genome combinations necessary?

Next steps & Outlook

- continue mappings for reference combinations with segemehl and ViReMa
- evaluation of mappings → all 24 reference genome combinations necessary?

- analyze available ONT samples: quality, mappings, recombination events
- Is short-read data sufficient for detecting recombination events in closely related viruses?

→ combine sequencing data as a hybrid approach?

ONT data
longer reads: coverage of breaking points in genomes

+

Illumina data
lower error rate: confidence of recombination events

- ideas: connect recombination hotspots to base modifications or RNA secondary structures

FRIEDRICH-SCHILLER-UNIVERSITÄT JENA

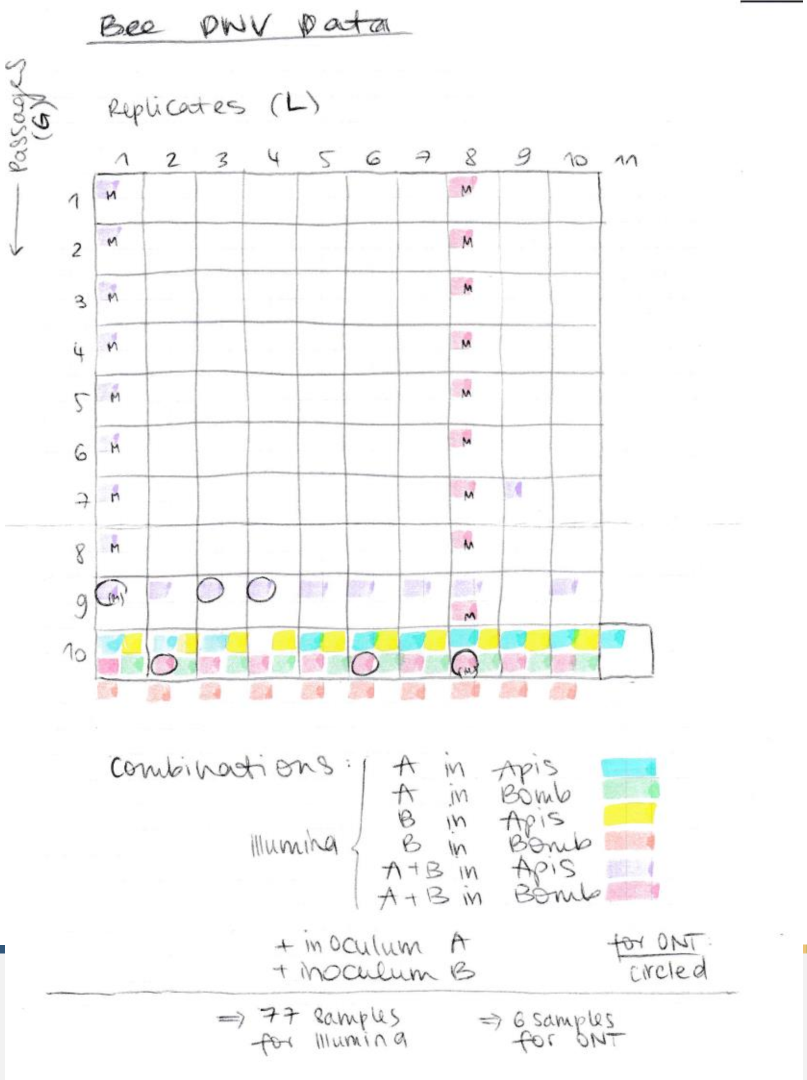
NFDI4 MICROBIOTA

RNA BIOINFORMATICS & HIGH-THROUGHPUT ANALYSIS

Thank you!

MARTIN-LUTHER-UNIVERSITÄT
HALLE-WITTENBERG

Sample overview



Bee DWV Data

Replicates (L)
Passages (G)
1
2
3
4
5
6
7
8
9
10
11
M
(M)

Combinations:

Illumina { A in Apis
A in Bomb
B in Apis
B in Bomb
A+B in Apis
A+B in Bomb

+ inoculum A
+ inoculum B

for ONT circled

⇒ 77 samples for Illumina

⇒ 6 samples for ONT

Bioinformatic approach

- reads pre-processed with `fastp`
- reference genomes in 1 fasta file

DVW-A

DVW-B

Apis mellifera

Bombus terrestris

- mapping with 3 tools: `hisat2, segemehl, ViReMa`
- extract statistics with `samtools` and in-house `Python3` script

Mappings

```
hisat2 -x "INDEX" -1 "$READL" -2 "$READR" --summary-file "$SAMPLE".log -
-new-summary | samtools sort > "$SAMPLE"_hisat2.bam
samtools index "$SAMPLE"_hisat2.bam

segemehl.x -t 12 -S -i "$INDEX".idx -d "$REF".fasta -q "$READL" -p
"$READR" -o "$SAMPLE".log | samtools view -b | samtools sort >
"$SAMPLE"_segemehl.bam
samtools index "$SAMPLE"_segemehl.bam

samtools view -@ 8 -f 0x40 -F 0x4 "$MAP" | cut -f1 | sort -T ./ | uniq |
wc -l
samtools view -@ 8 -f 0x80 -F 0x4 "$MAP" | cut -f1 | sort -T ./ | uniq |
wc -l
samtools view -@ 8 -f 0x4 -c "$MAP"
```