Supervisor:
Ekaterina Chernyaeva,
Theodosius Dobzhansky Center for Genome Bioinformatics

Students:
Yaroslav Baranov, Bioinformatics Institute
Svyatoslav Sidorov, St. Petersburg University of the RAS

Study of structural variations

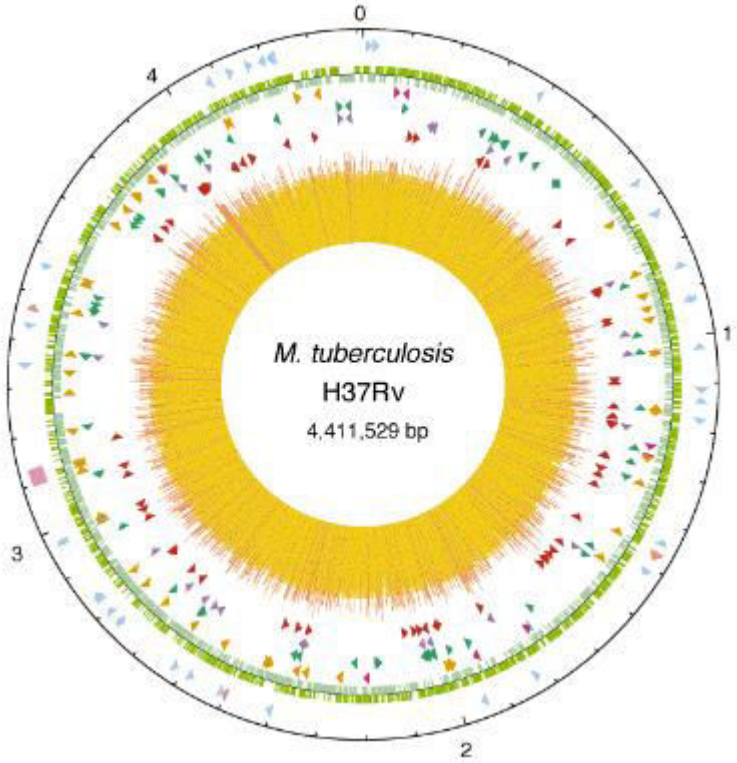
in *Mycobacterium tuberculosis* genome

Problem statement

- In 2012 8.6 million people fell ill with TB
- A total of 1.3 million people died from TB in 2012
- TB is the leading killer of people living with HIV
- TB occurs in every part of the world. Nearly 60% of new TB cases occurred in Asia in 2012. The greatest rate of new cases per capita was in sub-Saharan Africa.
- **No country has ever eliminated this disease.**

www.who.int

M. tuberculosis H37Rv genome



Size: 4411532 b.p.
Genes: 3993 protein coding
50 RNA coding
GC content: 65%

Insertion elements: 56 copies of IS-elements (Families IS3, IS5, IS21, IS30, IS110, IS256, ISL3, IS1535)

First genome was published in 1998 (Sanger)

Aim

Bioinformatical analysis of M. tuberculosis genome sequences for detection structual variations and further in vitro confirmation

What do we have?

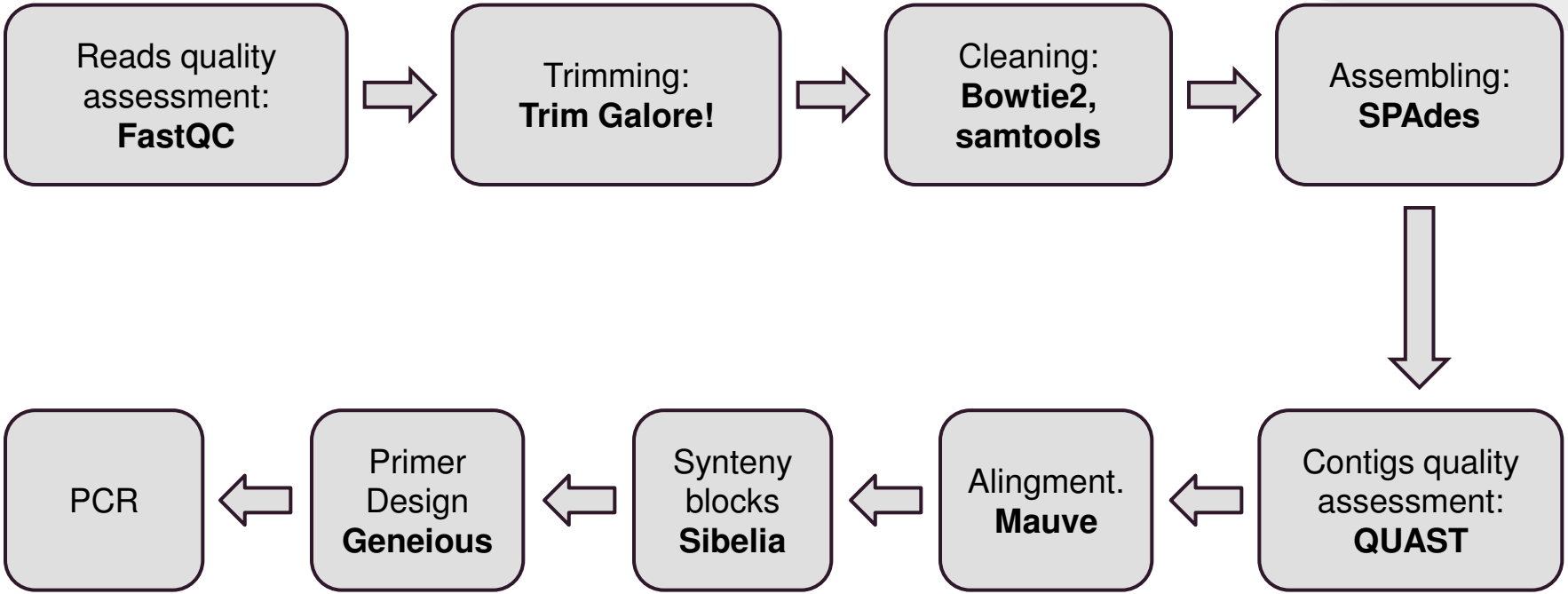
1. 33 paired-end read libraries from various *Mycobacterium tuberculosis* strains sequenced at Dobzhansky Center by Ekaterina Chernyaeva (one or two libraries per strain).

2. Median insertion size ~250 - 270 bp (libraries for sequencing were prepared with Nextera Kit).

3. Read length: 150 bp, 250 bp.

4. 0.5 - 2.5 millions of paired-end reads per library.

Pipeline


Reads quality assessment: **FastQC**
Trimming: **Trim Galore!**
Cleaning: **Bowtie2, samtools**
Assembling: **SPAdes**
Contigs quality assessment: **QUAST**
Alingment. **Mauve**
Synteny blocks **Sibelia**
Primer Design **Geneious**
PCR

Results

1. Total quality assessments of all read libraries.
2. 6 libraries were assembled to contigs in various ways of preceding trimming and cleaning.
3. 1 contaminated library trimmed, partially cleaned and assembled.
4. Structural variations.

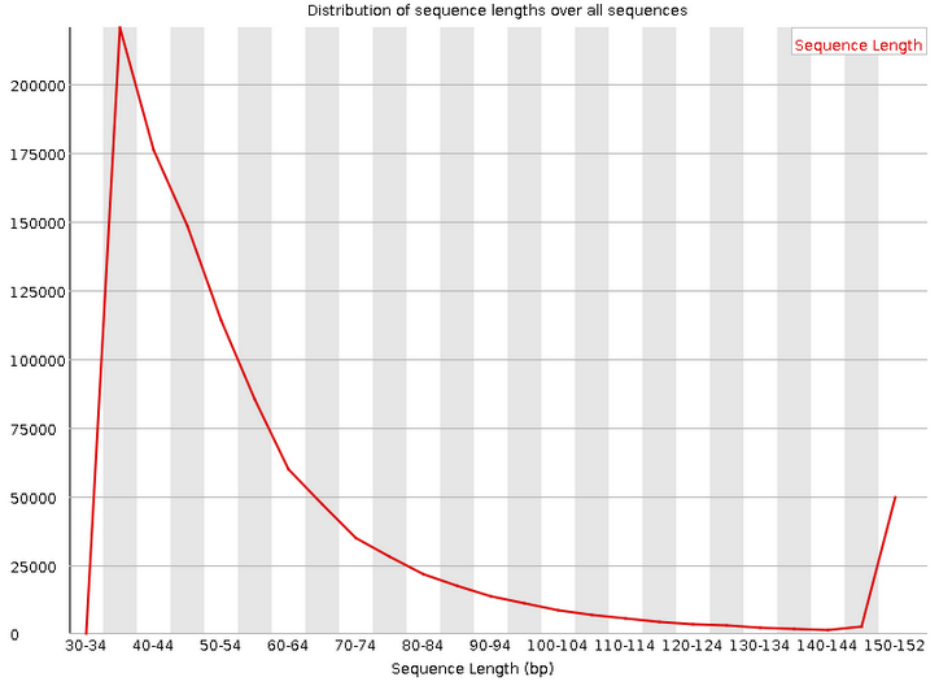
Reads quality assessment

Main FastQC metrics:

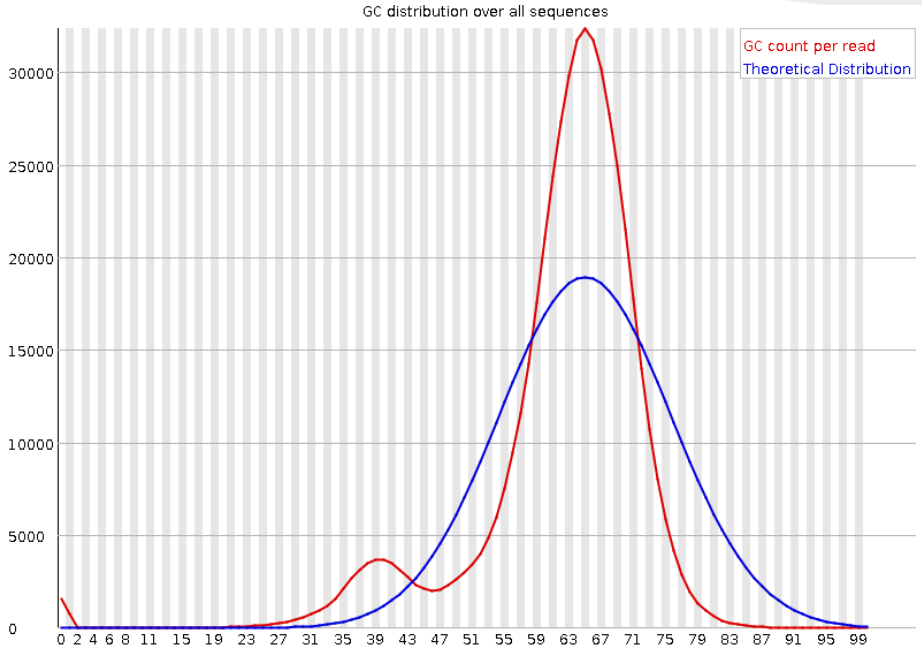
- Sequence Length Distribution,
- Sequence GC content,
- Per base sequence quality

Library quality problems

Exclude libraries with great quantity of short reads



Separate libraries with contamination



Trimming and cleaning

1. Adapters trimming (overlap with reads >= 5 bp).

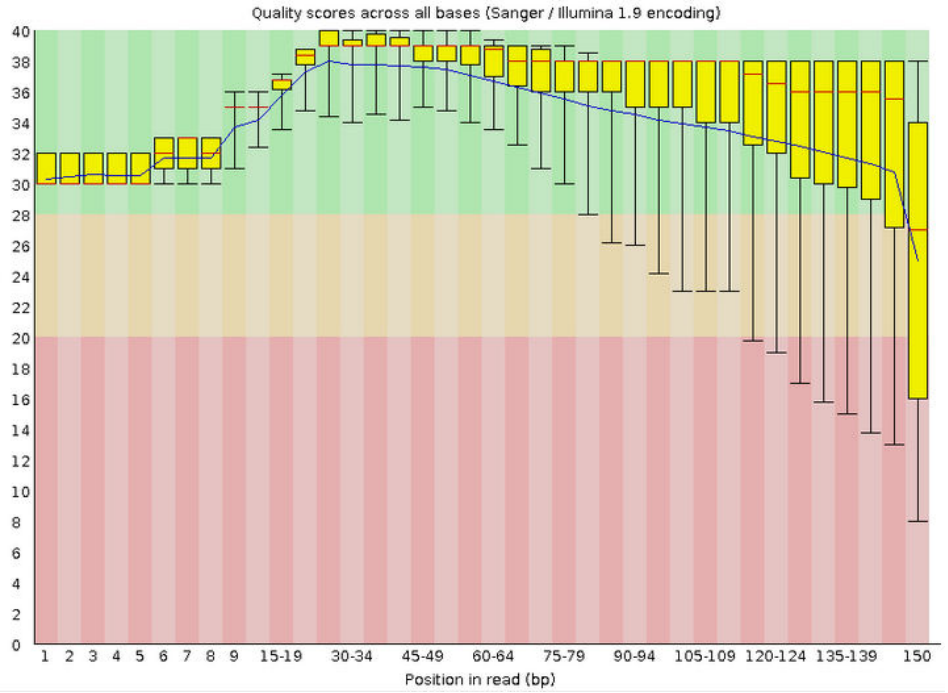
2. Quality trimming at score 20 (Trim Galore!).

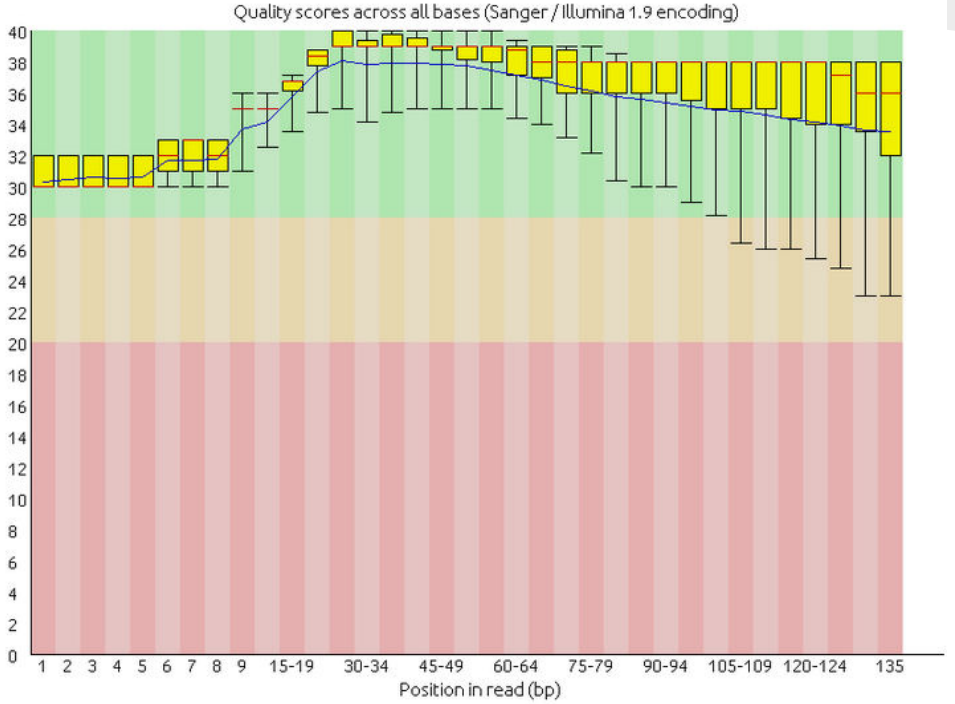
3. If library is contaminated (*H. sapiens*) then align reads to *H. sapiens* genome (hg19) and exclude aligned ones.

Other options:

1. Cut 15 - 17 bp from 3'-end.
2. Exclude reads with |poly-G| and |poly-C| >= 17 bp.
3. Quality trimming at score 2 (Trim Galore!).
4. Adapter trimming with overlap >=7, >= 10.

Trimming

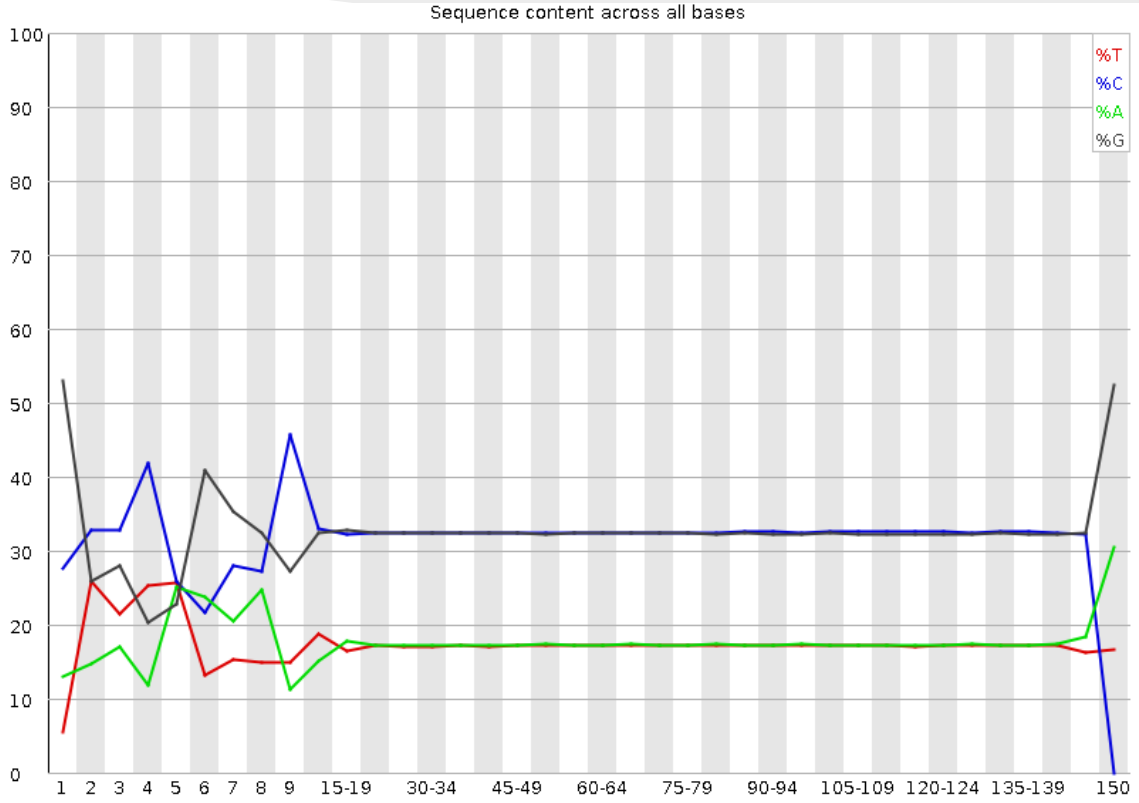

Quality scores across all bases (Sanger / Illumina 1.9 encoding)
Position in read (bp)


Quality scores across all bases (Sanger / Illumina 1.9 encoding)
Position in read (bp)

Trimming

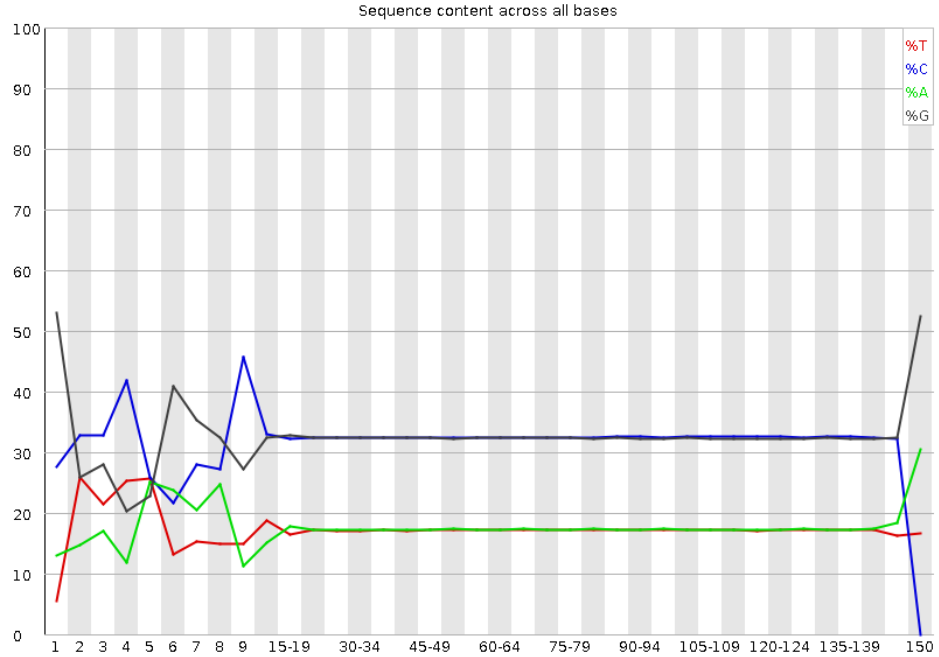
5'-end: DNA was cutted nonrandomly by enzymes during library preparation

3'-end: adapter sequences with many errors and fake subsequences generated by MiSeq when it tried to read sequence shorter than 150 / 250 bp **or** consequences of nonrandom DNA cutting

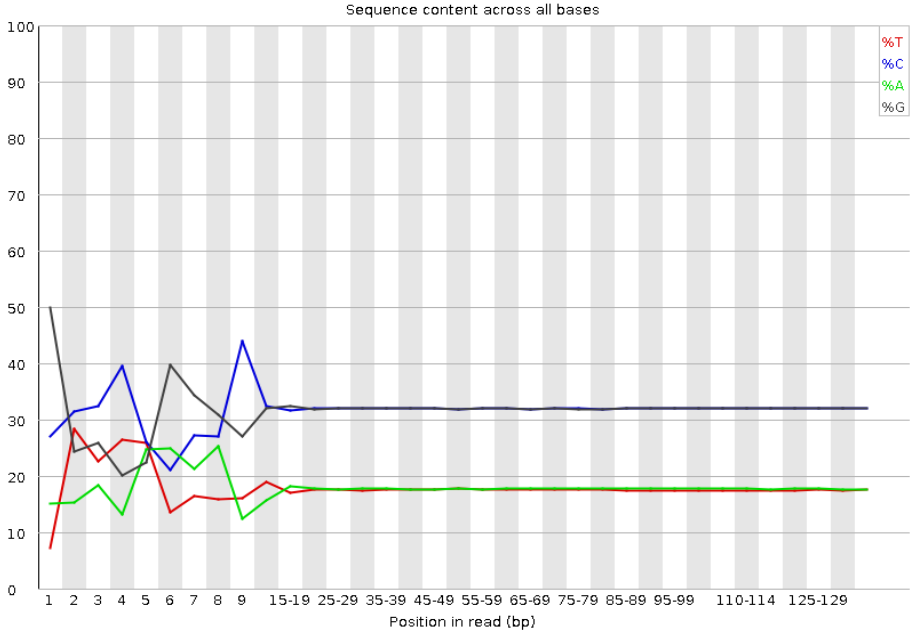


Trimming

Adapter and quality trimming only



0 - 15 bp are trimmed from 3'-end



Poly-G and Poly-C problem

Poly-G / Poly-C in *M. tuberculosis* reference genomes (NCBI):

1. **H37Rv:** max(|poly-G|) = **9**, max(|poly-C|) = **9**
2. **7199-99**: max(|poly-G|) = 10, max(|poly-C|) = 8
3. **CAS_NITR204**: max(|poly-G|) = 11, max(|poly-C|) = 10
4. **CCDC5079**: max(|poly-G|) = **16**, max(|poly-C|) = 11
5. **F11**: max(|poly-G|) = 10, max(|poly-C|) = 9

In TB0001 contigs: max(|polyG|) = 100, max(|polyC|) = 27.

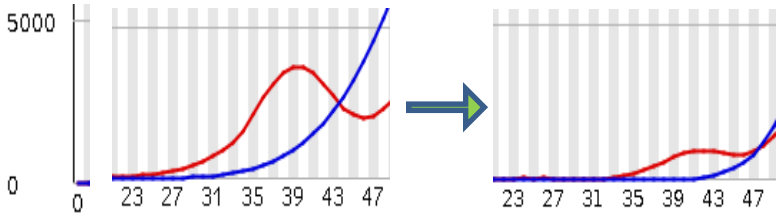
In TB0001 contigs after excluding reads with |polyG| and |polyC| >= 17:

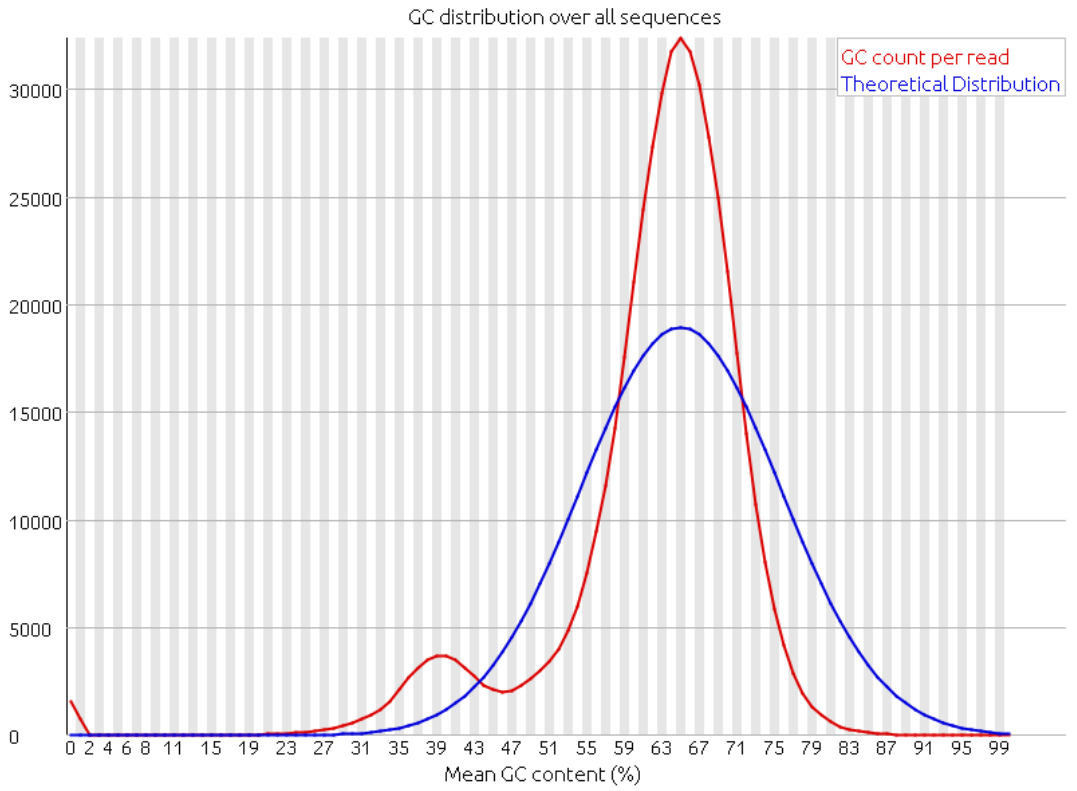
max(|polyG|) = 16, max(|polyC|) = 14

Cleaning

Library with *Homo sapiens* contamination:
~ 36 - 47 GC content

Exclude reads mapped to hg19:



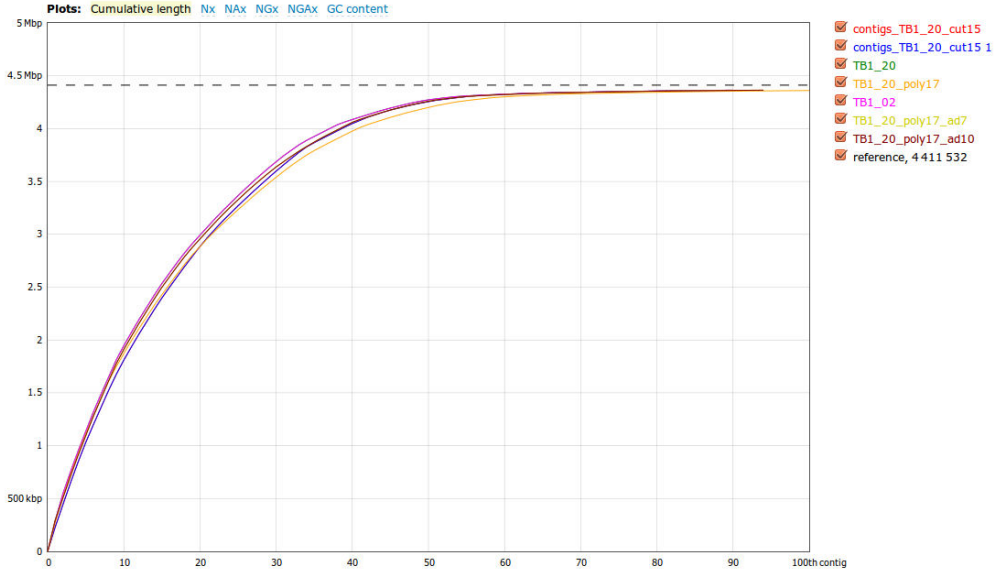


TB0001 contig assemblies

Statistics without reference	contigs_TB1_20_cut15	TB1_20	TB1_20_poly17	TB1_02	TB1_20_poly17_ad7	TB1_20_poly17_ad10
# contigs	90	91	100	91	94	94
Largest contig	232 368	295 445	273 583	295 445	295 445	295 445
Total length	4 363 917	4 362 857	4 360 849	4 362 948	4 364 278	4 364 269
N50	114 954	122 403	110 555	122 403	115 278	115 278
Genome statistics						
Genome fraction (%)	98.213	98.222	98.119	98.212	98.188	98.188

1. SPAdes: BayesHammer, k = 21, 33, 55, 77, (99)

2. QUAST: skip contigs shorter than 500 bp

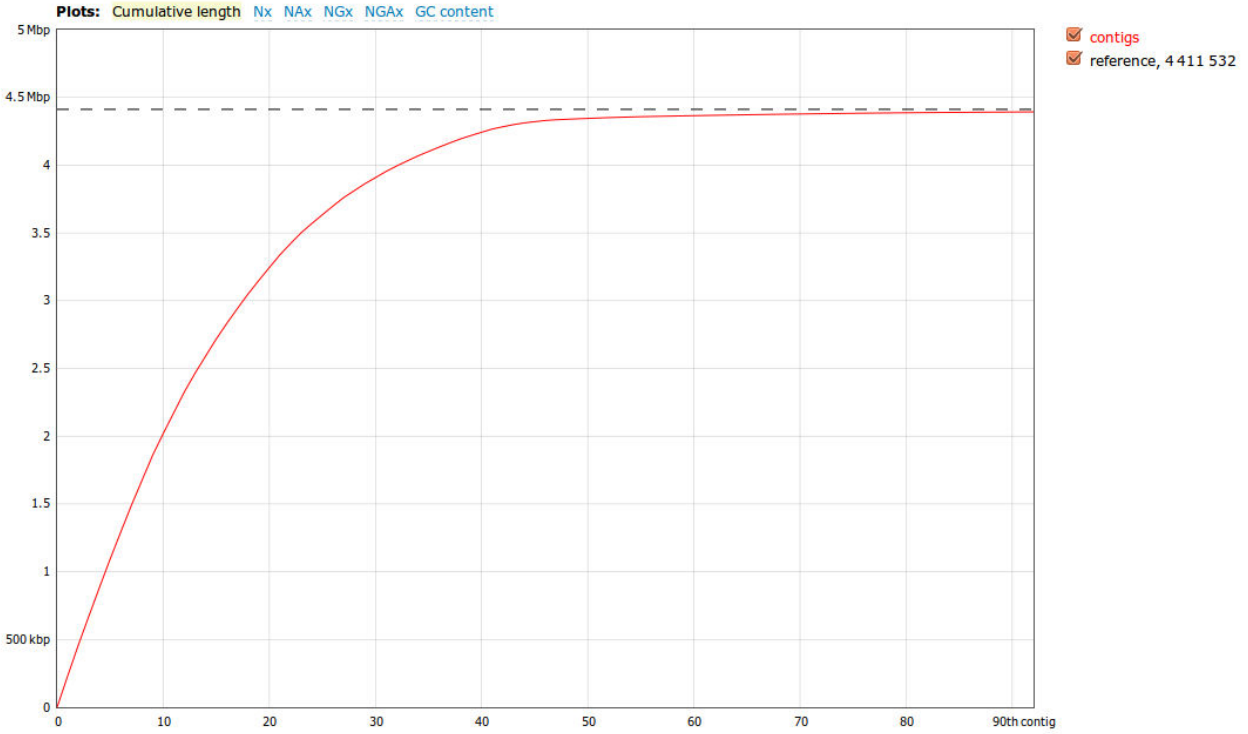


TB0002 contig assembly

Statistics without reference	contigs
# contigs	92
Largest contig	231 720
Total length	4 393 075
N50	154 214
Genome statistics	
Genome fraction (%)	98.682

1. SPAdes: BayesHammer, k = 21, 33, 55, 77, (99)

2. QUAST: skip contigs shorter than 500 bp

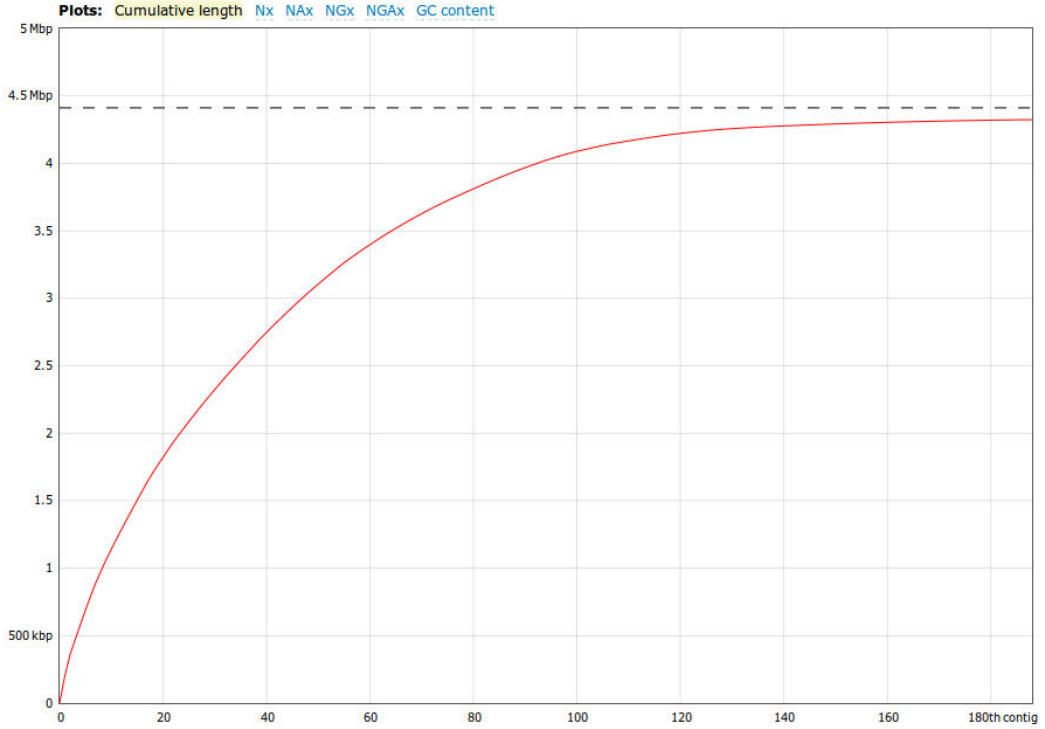


TB0005_1 contig assembly

Statistics without reference	contigs
# contigs	188
Largest contig	199 052
Total length	4 324 315
N50	48 179
Genome statistics	
Genome fraction (%)	97.039

1. SPAdes: BayesHammer, k = 21, 33, 55, 77

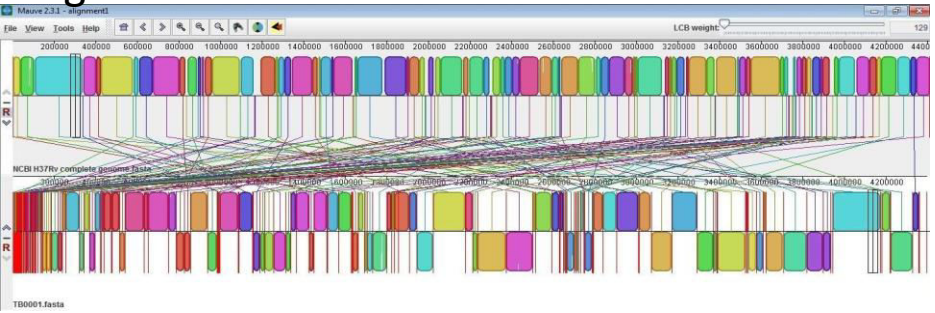
2. QUAST: skip contigs shorter than 500 bp



Contig alignment

- Mauve 2.3.1

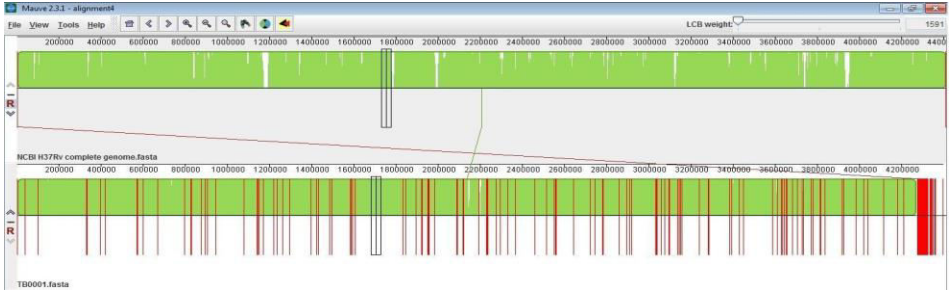
alignment1



alignment2

alignment3

alignment4

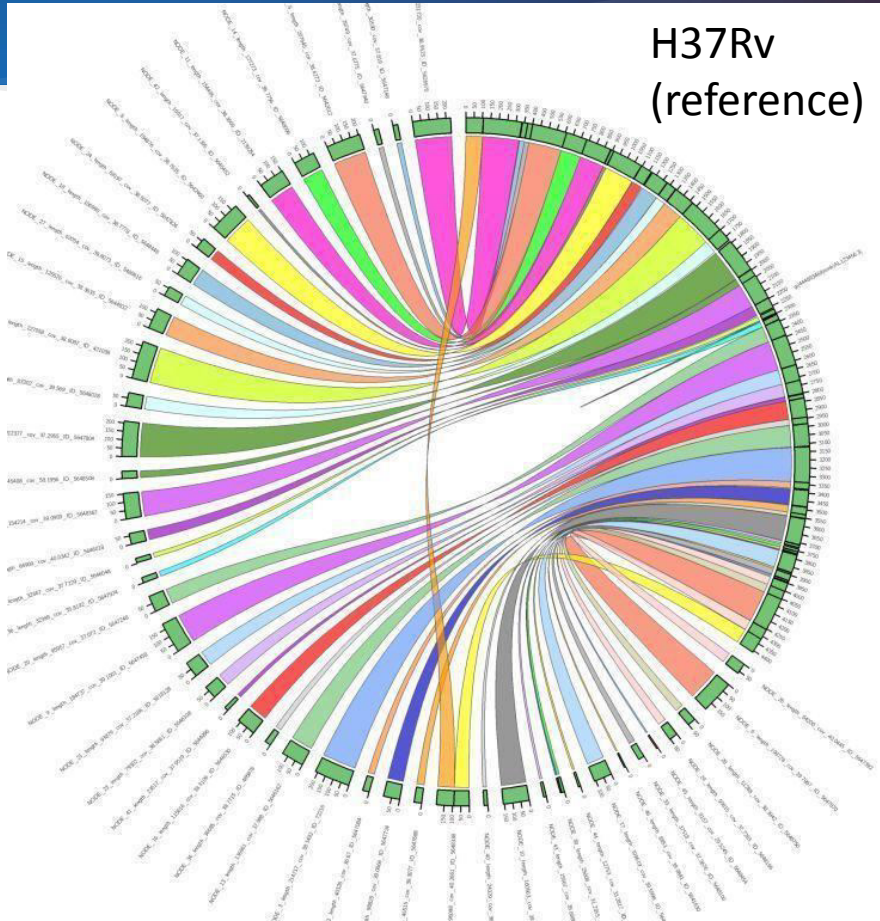


Synteny blocks

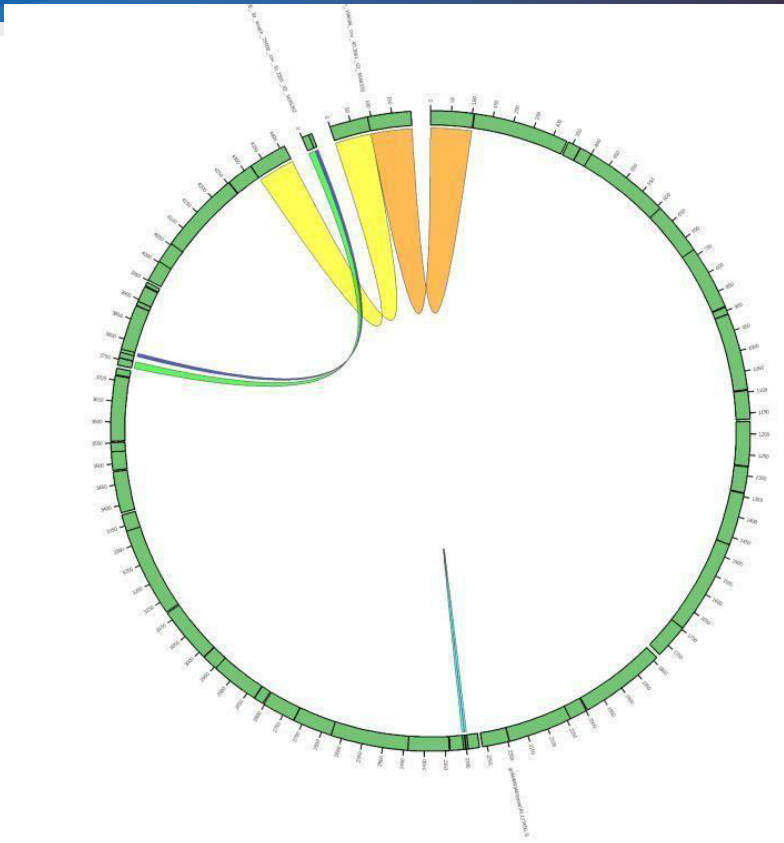
- Sibelia

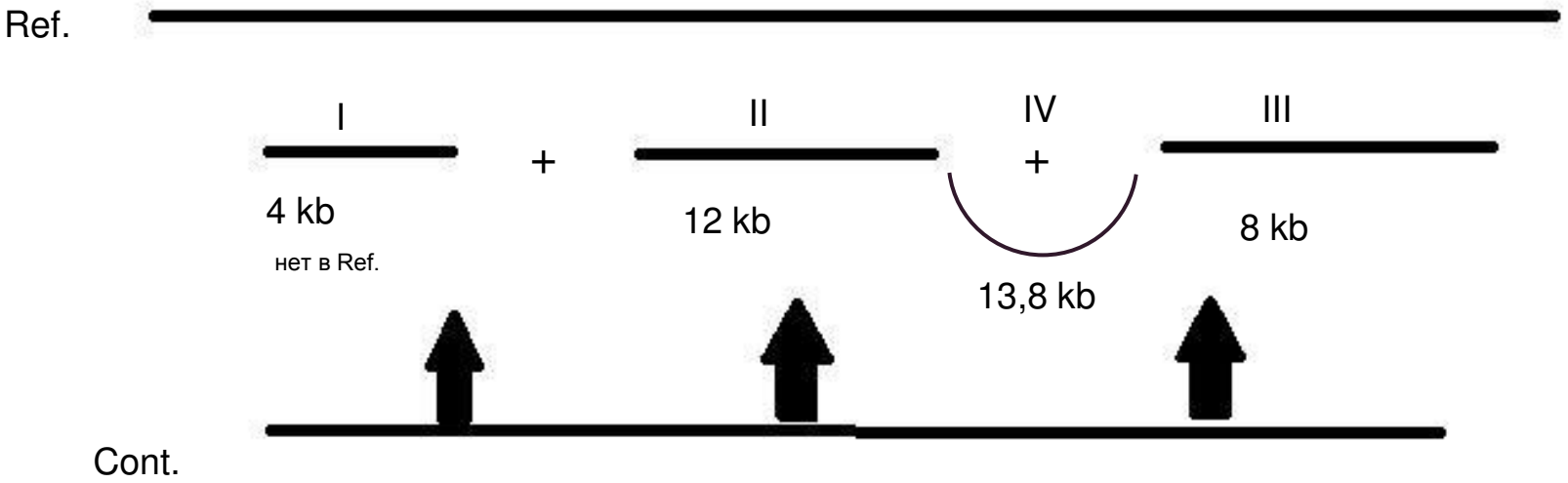
All contigs

TB0002



Rearranged contigs only




Ref.
I
II
IV
III
+
+
4 kb
нет в Ref.
12 kb
13,8 kb
8 kb
Cont.

I - contain in other ref. of MTBC

II - PE/PPE genes

III - conserved hypotheticals,
virulence, detoxification,
adaptation, cell wall and cell processes

IV - Rv3348, Rv3349c - insertion seqs and phages

Further research

- QC and filtration the rest of the strains
- Assemble *M. tuberculosis* genomes to contigs
- Compare sequnced genomes
- Confirm structural variations by PCR

Thank you!

* School of BEES Research Blog