MUNI FI

Comparing subtelomeric sequences in human genomes in terms of sequence (and methylation) similarity

Jindřich Matuška

Faculty of Informatics, Masaryk University

14. května 2025

Subtelomeres

- Region ca. 500000 bp beside telomeres
- Frequent modifications
- Transposable element

Why do that

- Modifications (duplication) may lead to:
  - creation of new genes
  - diversification of chromosomes

Pipeline

1. Extraction of subtelomeres
2. Similiarity by ModDotPlot
3. Further analysis

Extraction of subtelomeres

1. Extraction of buffered sequence from ends (SeqKit, script)
2. Extraction of telomere regions (Seqtk telo)
3. Inversion, cropping subtelomeres to length (bedtools, script, SeqKit)
4. Reversion of end sequences (SeqKit)

Similiarity by ModDotPlot

1. Selection of subtelomeres (SeqKit)
2. Split by N-sequences (script, SeqKit)
3. Sort, merge (SeqKit)
4. Similiarity analysis (ModDotPlot)
5. Finish?

Further analysis

1. Parsing .bedpe files into heatmap (Python, Matplotlib)
2. Selection of groups (SeqKit)
3. Similiarity analysis (ModDotPlot)

Running pipeline

2 slow parts

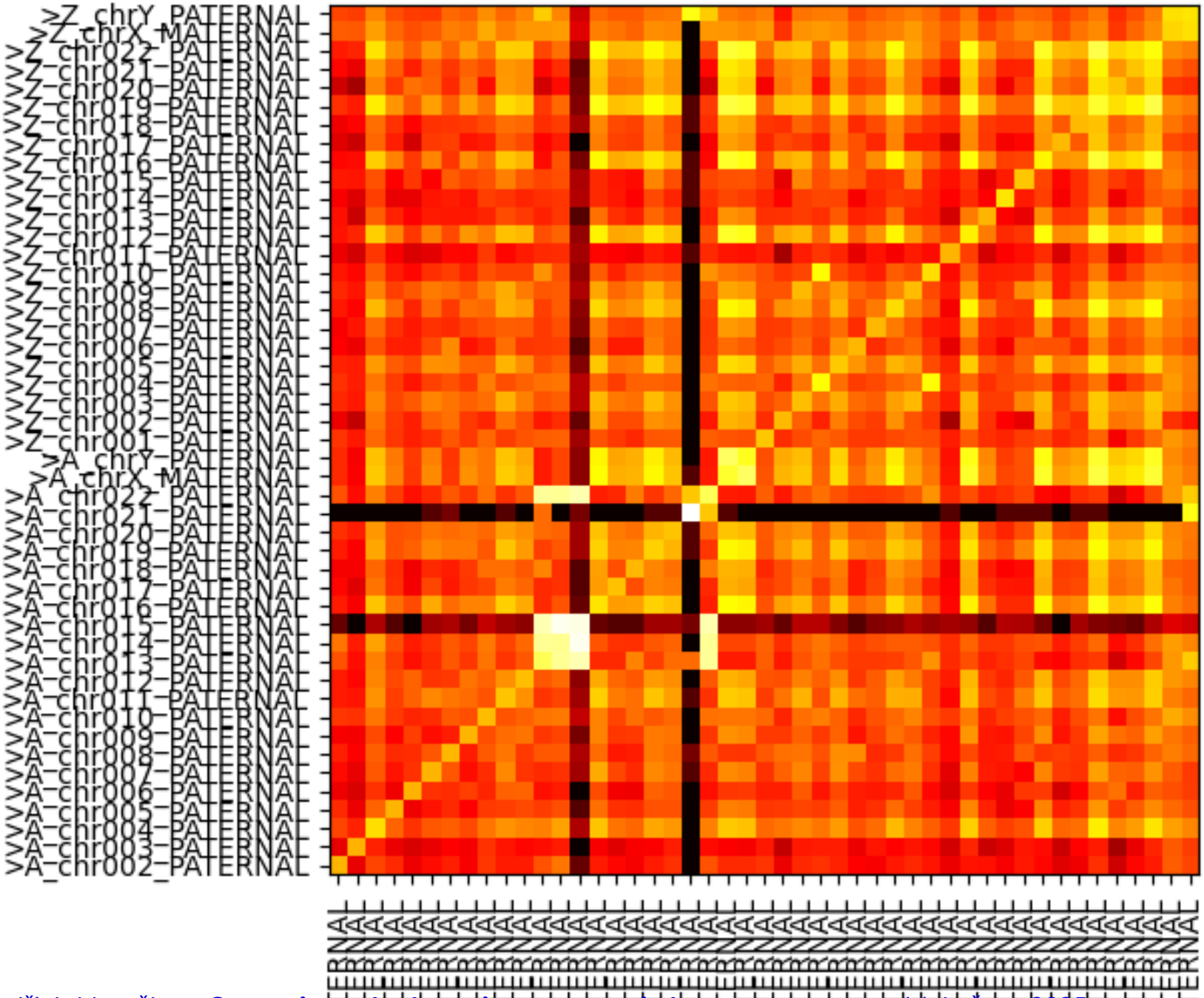
- Extraction of buffered sequences from ends
- ModDotPlot

Otherwise quite fast

Analysis of heatmap

3 groups with high similiarity:

- Big tandem repeats
- Singleton chr21-A
- Small similiar subsequences

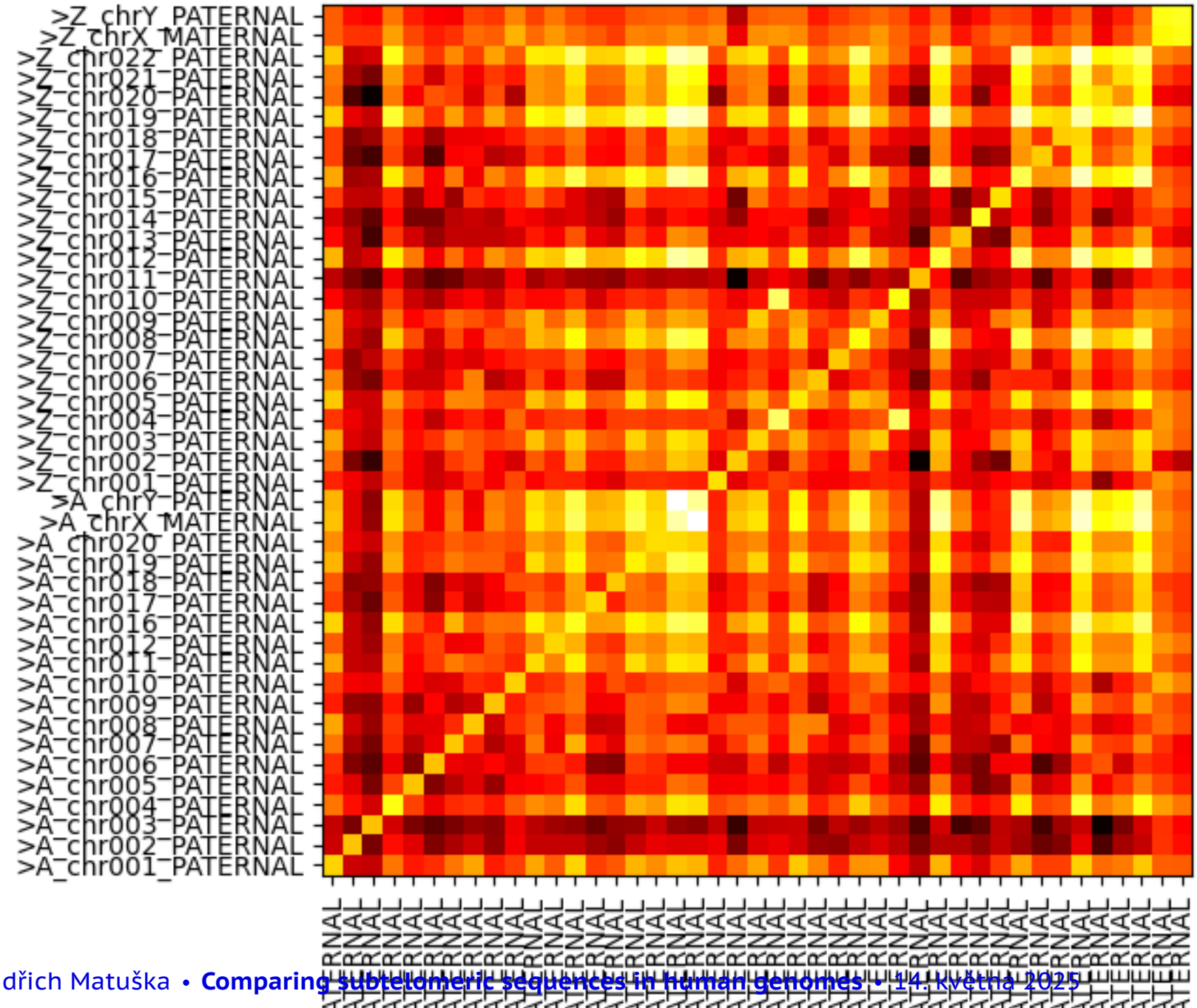

>Z_chrY_PATERNAL
>Z_chrX_MATERNAL
>Z_chr022_PATERNAL
>Z_chr021_PATERNAL
>Z_chr020_PATERNAL
>Z_chr019_PATERNAL
>Z_chr018_PATERNAL
>Z_chr017_PATERNAL
>Z_chr016_PATERNAL
>Z_chr015_PATERNAL
>Z_chr014_PATERNAL
>Z_chr013_PATERNAL
>Z_chr012_PATERNAL
>Z_chr011_PATERNAL
>Z_chr010_PATERNAL
>Z_chr009_PATERNAL
>Z_chr008_PATERNAL
>Z_chr007_PATERNAL
>Z_chr006_PATERNAL
>Z_chr005_PATERNAL
>Z_chr004_PATERNAL
>Z_chr003_PATERNAL
>Z_chr002_PATERNAL
>Z_chr001_PATERNAL
>A_chrY_PATERNAL
>A_chrX_MATERNAL
>A_chr022_PATERNAL
>A_chr021_PATERNAL
>A_chr020_PATERNAL
>A_chr019_PATERNAL
>A_chr018_PATERNAL
>A_chr017_PATERNAL
>A_chr016_PATERNAL
>A_chr015_PATERNAL
>A_chr014_PATERNAL
>A_chr013_PATERNAL
>A_chr012_PATERNAL
>A_chr011_PATERNAL
>A_chr010_PATERNAL
>A_chr009_PATERNAL
>A_chr008_PATERNAL
>A_chr007_PATERNAL
>A_chr006_PATERNAL
>A_chr005_PATERNAL
>A_chr004_PATERNAL
>A_chr003_PATERNAL
>A_chr002_PATERNAL

Groups

Big tandem repeats

- chr13-A
- chr14-A
- chr15-A
- chr22-A

Singleton chr21-A


>Z_chrY_PATERNAL
>Z_chrX_MATERNAL
>Z_chr022_PATERNAL
>Z_chr021_PATERNAL
>Z_chr020_PATERNAL
>Z_chr019_PATERNAL
>Z_chr018_PATERNAL
>Z_chr017_PATERNAL
>Z_chr016_PATERNAL
>Z_chr015_PATERNAL
>Z_chr014_PATERNAL
>Z_chr013_PATERNAL
>Z_chr012_PATERNAL
>Z_chr011_PATERNAL
>Z_chr010_PATERNAL
>Z_chr009_PATERNAL
>Z_chr008_PATERNAL
>Z_chr007_PATERNAL
>Z_chr006_PATERNAL
>Z_chr005_PATERNAL
>Z_chr004_PATERNAL
>Z_chr003_PATERNAL
>Z_chr002_PATERNAL
>Z_chr001_PATERNAL
>A_chrY_PATERNAL
>A_chrX_MATERNAL
>A_chr020_PATERNAL
>A_chr019_PATERNAL
>A_chr018_PATERNAL
>A_chr017_PATERNAL
>A_chr016_PATERNAL
>A_chr012_PATERNAL
>A_chr011_PATERNAL
>A_chr010_PATERNAL
>A_chr009_PATERNAL
>A_chr008_PATERNAL
>A_chr007_PATERNAL
>A_chr006_PATERNAL
>A_chr005_PATERNAL
>A_chr004_PATERNAL
>A_chr003_PATERNAL
>A_chr002_PATERNAL
>A_chr001_PATERNAL

Groups 2

Small similiar subsequences

- chr16-A
- chrX-A
- chrY-A
- chr8-Z
- chr12-Z
- chr16-Z
- chr19-Z
- chr20-Z
- chr22-Z

Further work

- Extraction of sequences
- Methylation data
- Multiple sources of data