Investigating the heterogeneity of repeat regions in the human genome

Students: N.Gulyaeva,
P. Sinitsyn
Tutors: I. Pulyakhina,
P.A.C.'t Hoen
J.F.J. Laros
M. Schaap

Overview

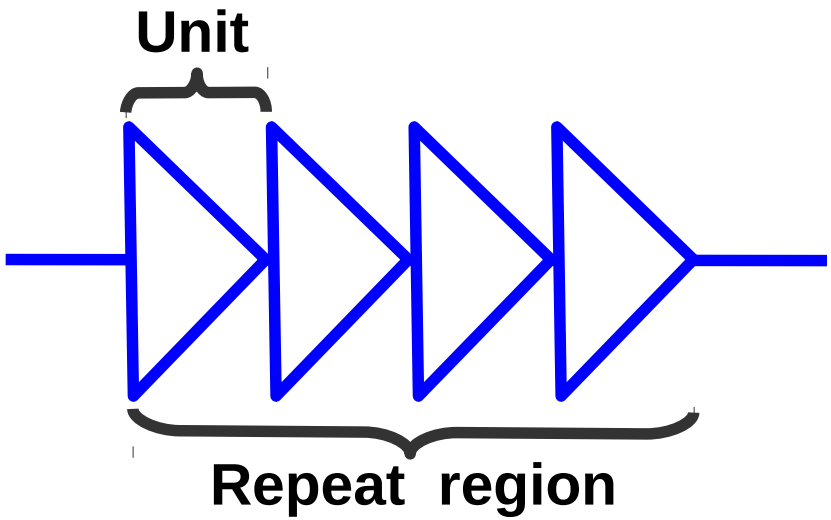
Classification of repeated sequences:

- Microsatellites (1-7bp)
- Minisatellites (<100bp)
- Macrosatellites (>100bp)

Overview

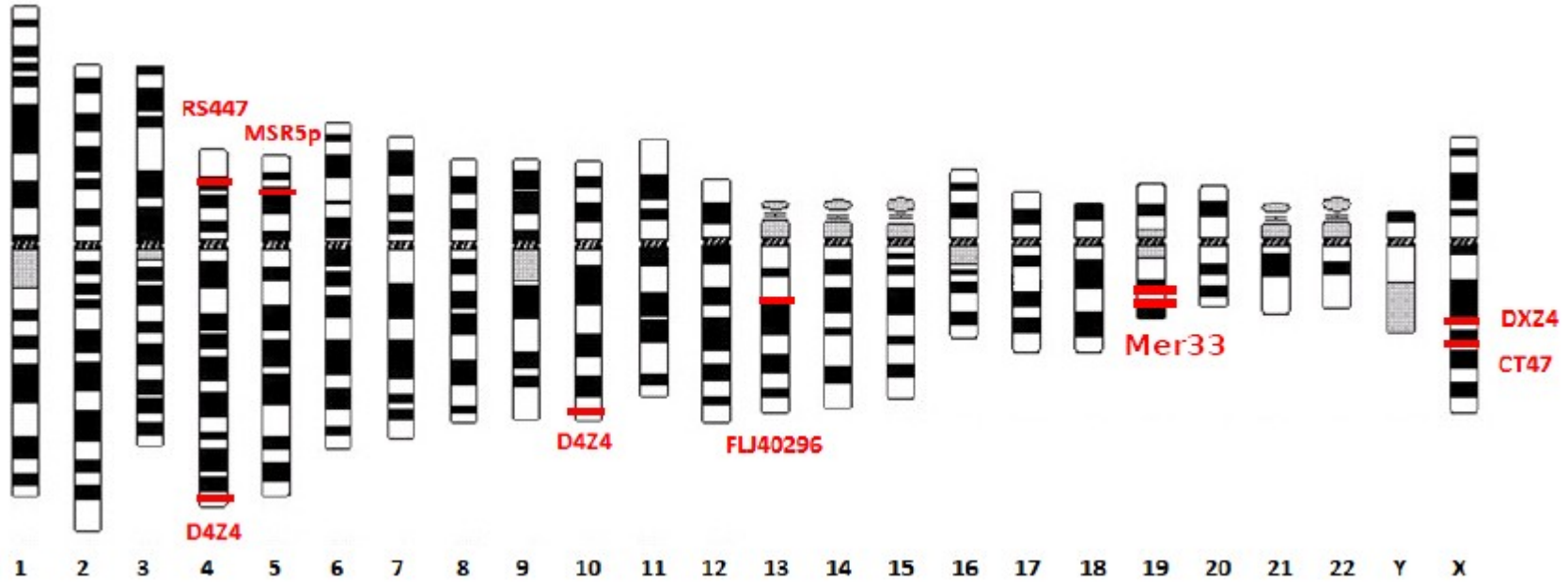
Classification repeated sequences:

- Microsatellites (1-7bp)
- Minisatellites (<100bp)
- **Macrosatellites (>100bp)**



Overview

- 9 MSRs on different chromosome
- 3.0-6.6 kb
- Units within an array: 95-99% homologous



Relevance of project

- ***Medical relevance***

The development of the FSHD-disease depend on amount of units in repeat region D4Z4 chr4 (<11 units FSHD)

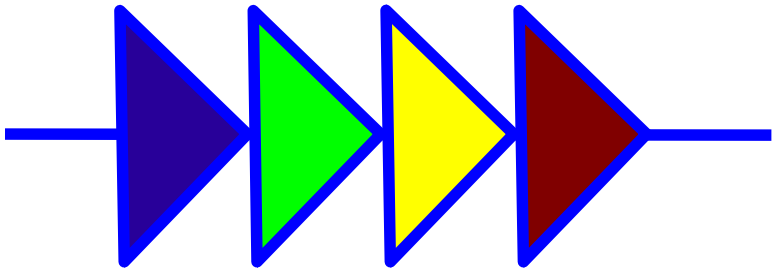
- ***Bioinformatic relevance***

The complexity of genome assembling in the repeat regions leads to necessity of development new approaches to their study

Aims

Investigating the heterogeneity of repeat regions:

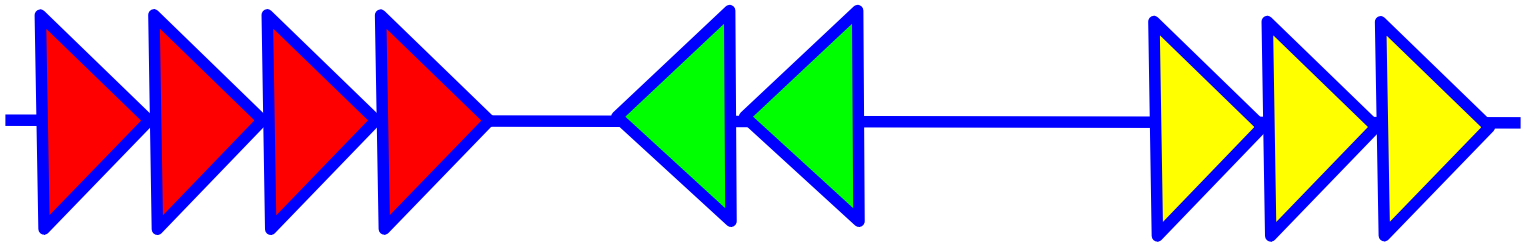
- Heterogeneity among units of one repeat region



Aims

Investigating the heterogeneity of repeat regions:

- Heterogeneity among units of one repeat region
- Heterogeneity among repeat regions of one genome



Aims

Investigating the heterogeneity of repeat regions:

- Heterogeneity among units of one repeat region
- Heterogeneity among repeat regions of one genome
- Heterogeneity of repeat regions among genomes

Next Generation Sequencing data

- Millions of short fragments (reads)
- Length of reads 36-100 nt

Typical fastq-file (file containing reads):

```
……
@ERR000624.129 EAS254_13:8:1:7:874/2
AGGCAAAAAGAAACAGGCAAAACTAATATTTTAAATATCCCAACGTAACCA
+
,I3$5=243&);3.8=(41315+91."84B)7"""/"+-3)+3/+0+*,'*
……
```

Two different approaches

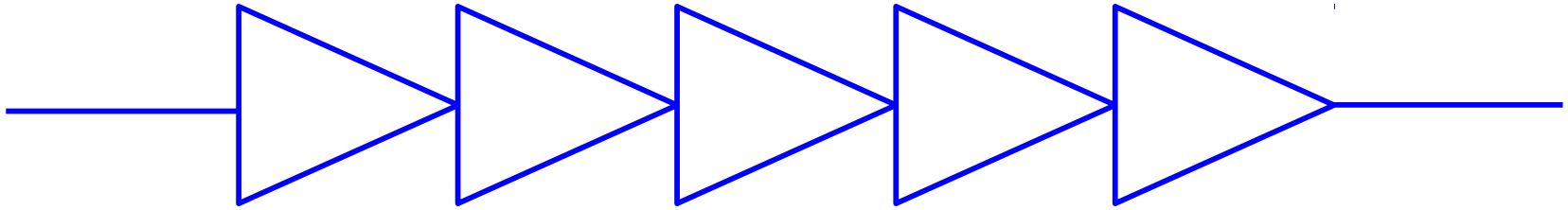
- ***Work with consensuses of repeat units***

(N. Gulyaeva)

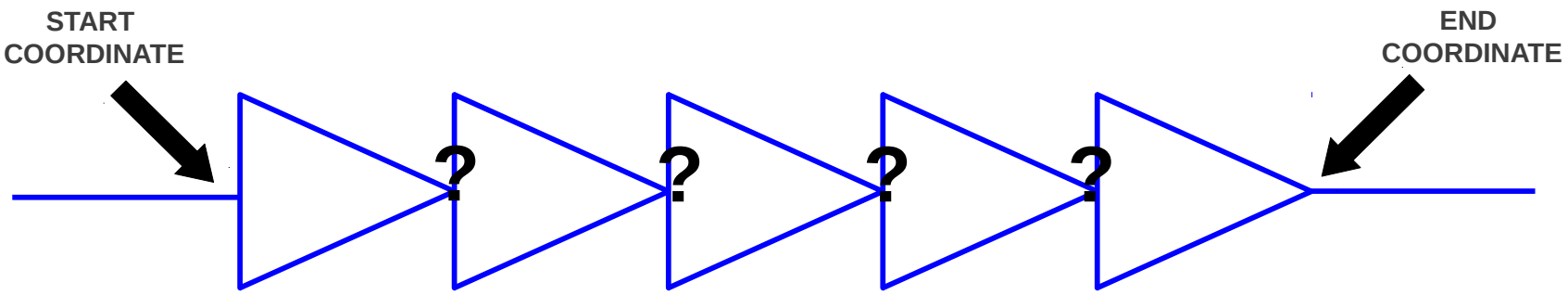
- ***K-mer profiling***

(P. Sinitsyn)

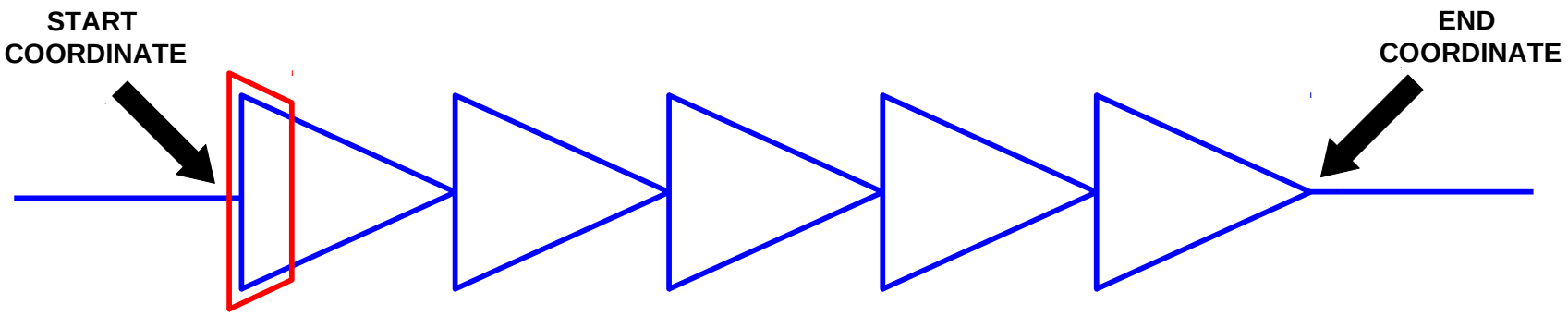
Find boundaries of units of MSRs in the reference genome



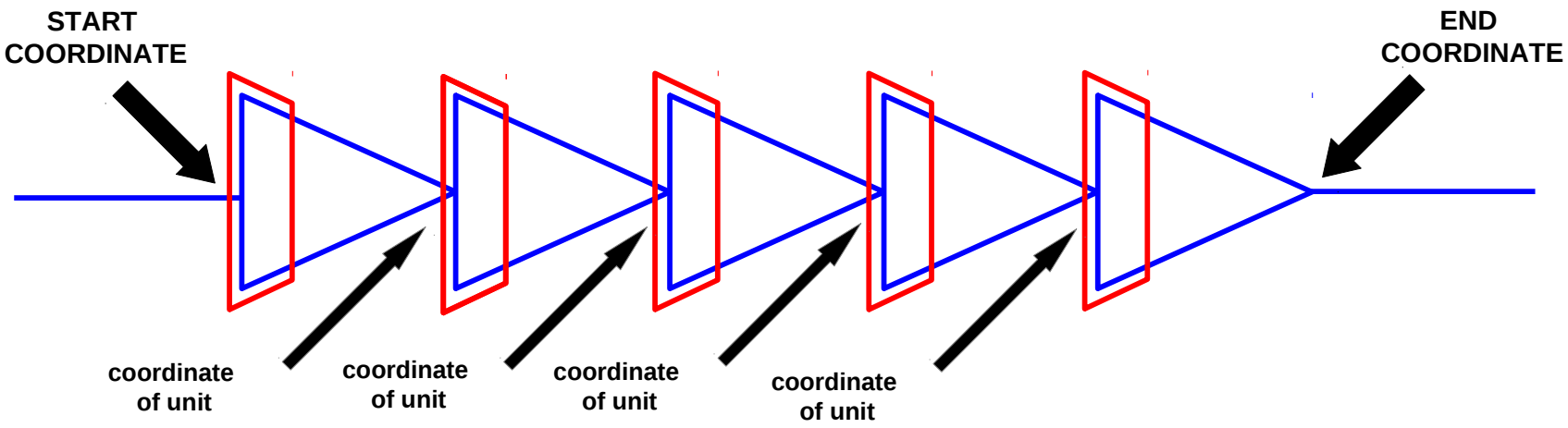
Find boundaries of units of MSRs in the reference genome

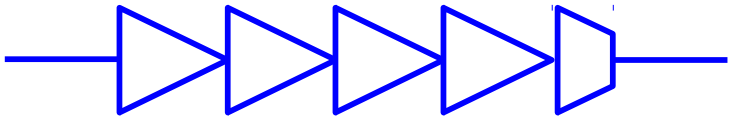
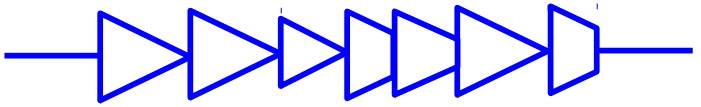


Find boundaries of units of MSRs in the reference genome

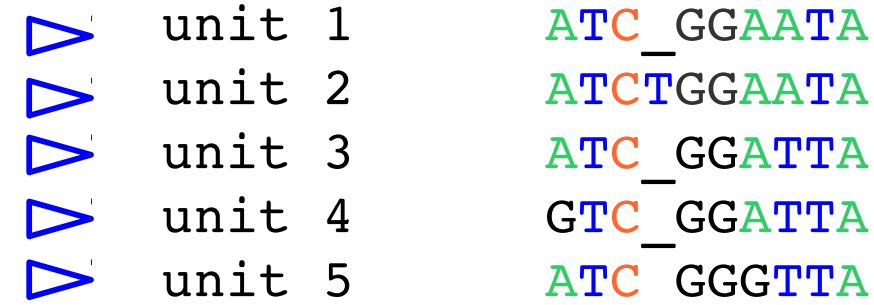


Find boundaries of units of MSRs in the reference genome

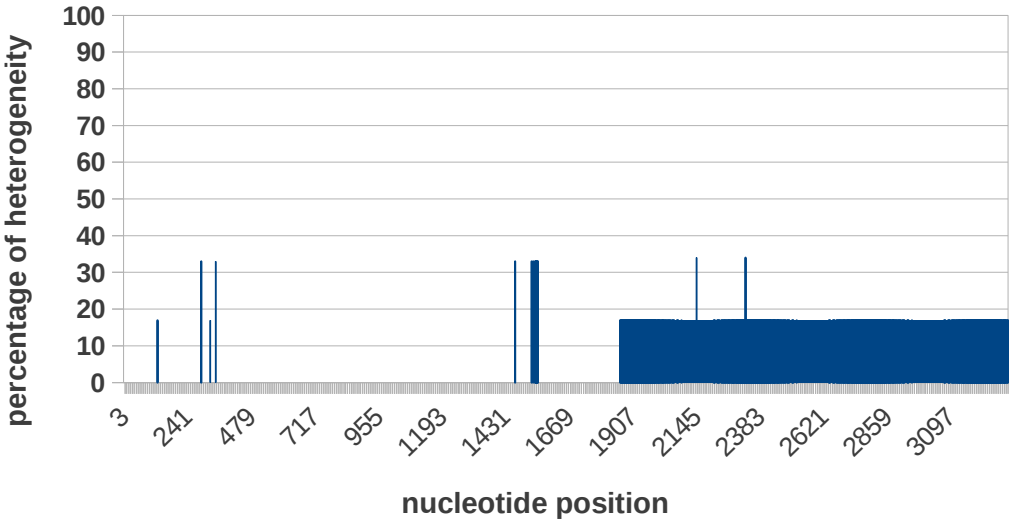


MSR	Number of units	Sequence structure
D4Z4_4q	8	
D4Z4_10q	6	
FLJ40296	5	
RS447	10	
Mer33_1	11	
CT47	14	
MSR5p	11	
DXZ4	14	
Mer33_2	12	

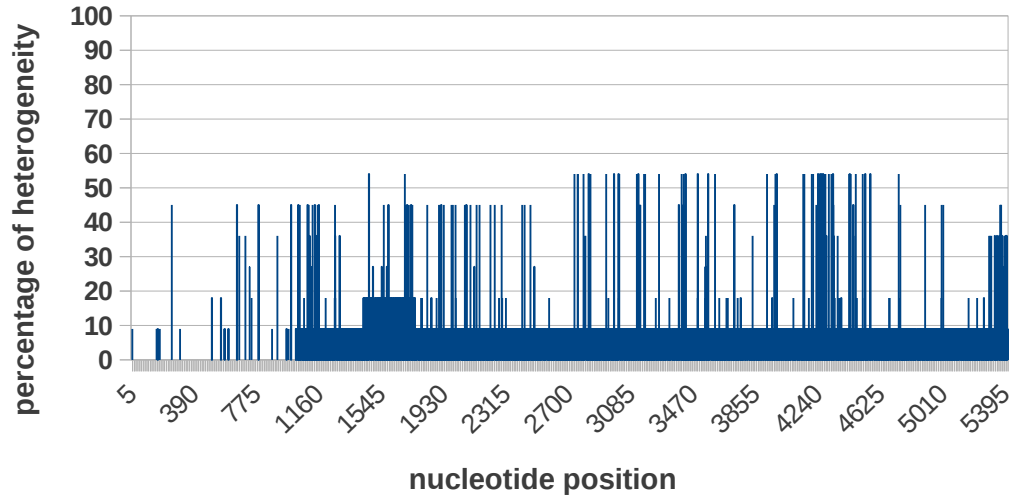
Multiple alignment of units for each MSR (clustalw)



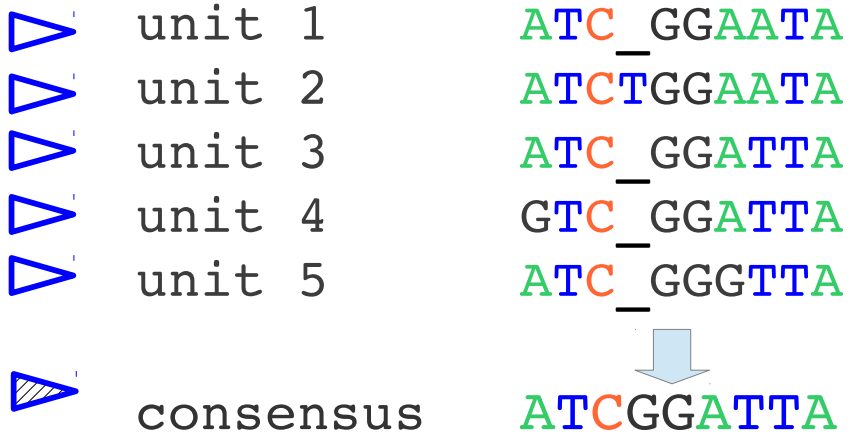
D4Z4_10q



Mer33_1



Build consensus unit for each MSR

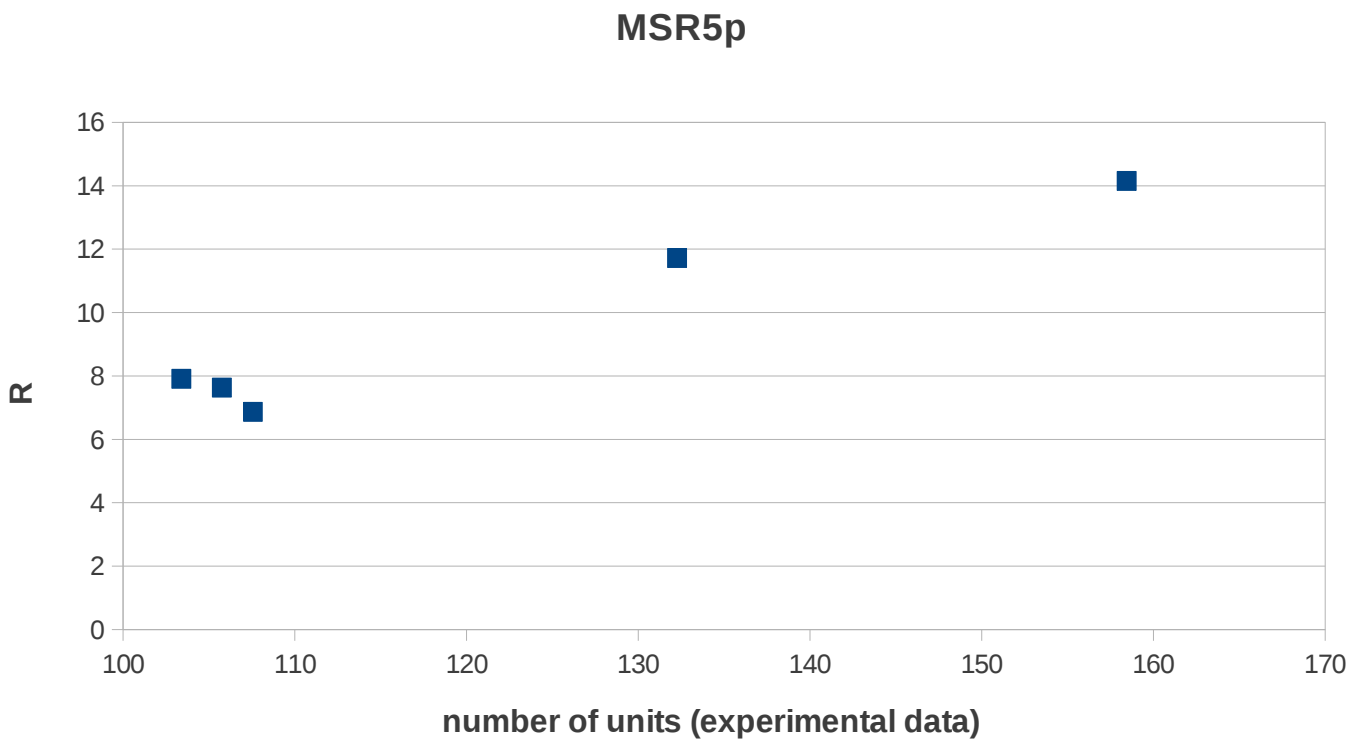


Calculations based on alignment results

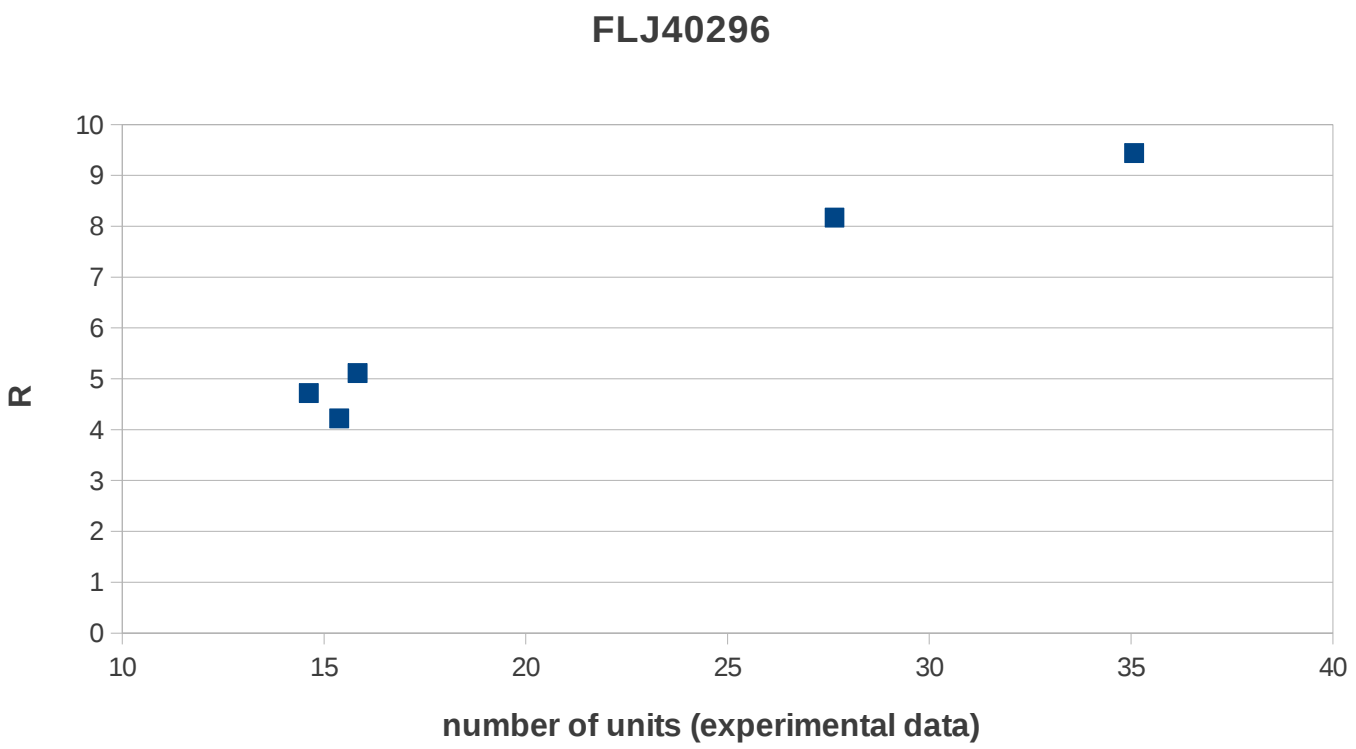
$$R = \frac{n}{T \times l} \times 10^9$$

- R – proportion of reads per nucleotide of MSR
- n – number of reads that map to consensus unit of MSR
- T – total number of reads
- l – length of consensus unit

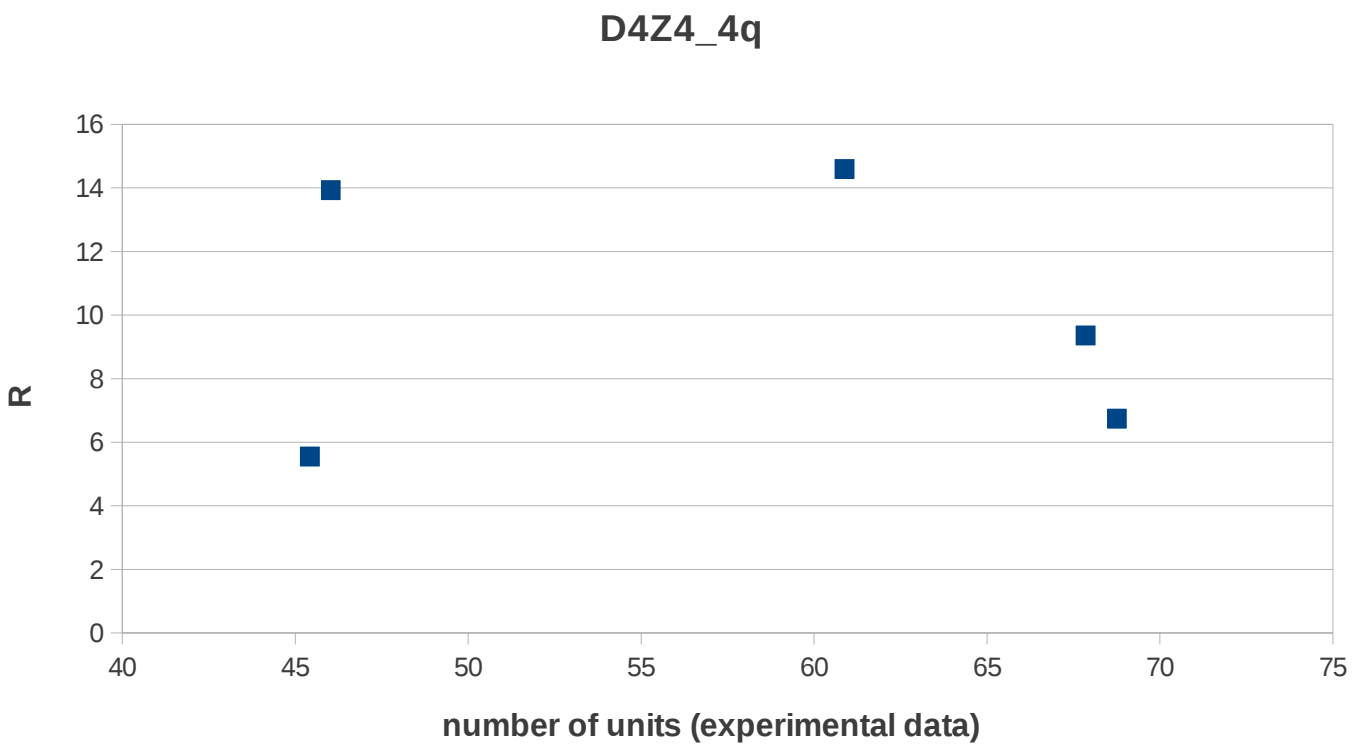
Calculations based on alignment results



Calculations based on alignment results



Calculations based on alignment results

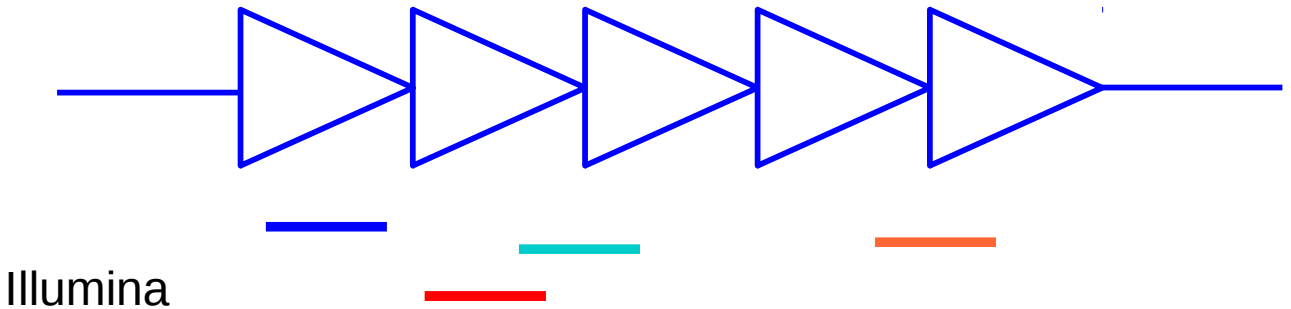


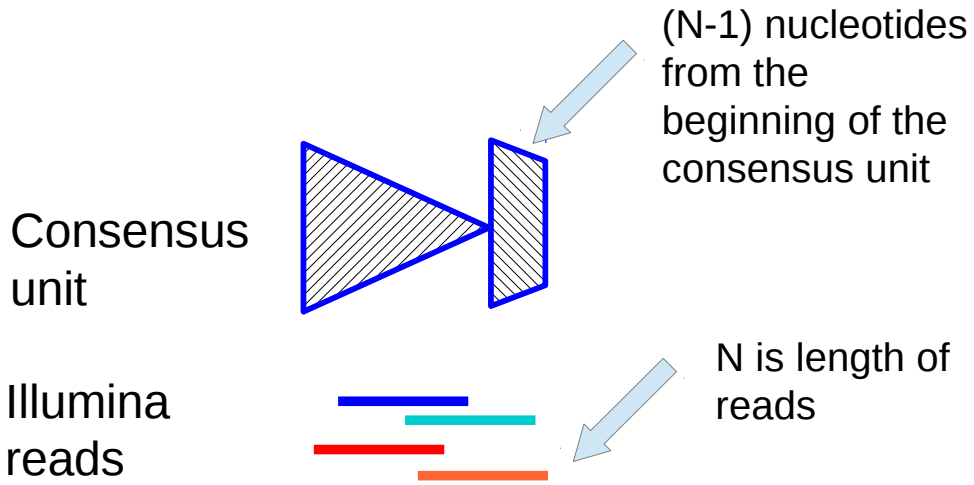
Calculations based on alignment results

Sample GM12878

	CT47	D4Z4_4q	D4Z4_10q	DXZ4	FLJ40296	Mer33_1	Mer33_2	MSR5p	RS447
reads – 36 nucleotides	87.72	6.74	6.63	20.9	4.72	151.76	176.27	7.63	9.41
reads – 37 nucleotides	86.83	7.59	7.55	16.81	4.08	146.14	175.73	6.54	8.97
reads – 67 nucleotides	5.99	1.63	1.82	8.1	1.68	10.33	17.44	5.73	14.83

Alignment of reads on consensus units (bwa)





Aims

Investigating the heterogeneity of repeat regions:

- Heterogeneity among units of one repeat region
- Heterogeneity among repeat regions of one genome
- Heterogeneity of repeat regions among genomes

Aims

Investigating the heterogeneity of repeat regions:

- Heterogeneity among units of one repeat region
- Heterogeneity among repeat regions of one genome
- Heterogeneity of repeat regions among genomes

How to do: k-mer profiling

Sequence of repeat unit X:

AATGC

Sequence of genome with 2 repeat region X: AAA AATGC AATGC AA

How to do: k-mer profiling

Sequence of repeat unit X:

AATGC

Sequence of genome with 2 repeat region
X: AAA AATGC AATGC AA

Look at the frequency all possible

sub-sequences length 2nt for whole genome and reapeat region

How to do: k-mer profiling

Sequence of repeat unit X:

AATGC

Sequence of genome with 2 repeat region X: AAA AATGC AATGC AA

Look at the frequency all possible sub-sequences length 2nt for whole genome and reapeat region

***Scientific Slang**: create k-mer profiling (k=2) for whole genome and repeat region*

K-mer profiling (k=2)

Forward strand

AAA AATGCAATGC AA

Reverse strand

TT GCATTGCATT TTT

Repeat region

	forward	reverse	Sum
AA	2		
AT			
TG			
GC			
CA			
TT			

K-mer profiling (k=2)

Forward strand

AAA AATGCAATGC AA

Reverse strand

TT GCATTGCATT TTT

Repeat region

	forward	reverse	Sum
AA	2		
AT	2		
TG			
GC			
CA			
TT			

K-mer profiling (k=2)

Forward strand

AAA AATGCAATGC AA

Reverse strand

TT GCATTGCATT TTT

	Repeat region		
	forward	reverse	Sum
AA	2	0	2
AT	2	2	4
TG	2	2	4
GC	2	2	4
CA	2	2	4
TT	0	2	2

K-mer profiling (k=2)

Forward strand

AAA AATGCAATGC AA

Reverse strand

TT GCATTGCATT TTT

	Repeat region/Genome		
	forward	reverse	Sum
AA	2/6	0/0	2/6
AT	2/2	2/2	4/4
TG	2/2	2/3	4/5
GC	2/2	2/2	4/4
CA	2/3	2/2	4/5
TT	0/0	2/6	2/6

Extraction of unique k-mers (k=2)

Unique k-mer is a sequence whose frequency at genome is equal to frequency at repeat region

Repeat region/Genome

	forward	reverse	Sum
AA	2/6	0/0	2/6
AT	2/2	2/2	4/4
TG	2/2	2/3	4/5
GC	2/2	2/2	4/4
CA	2/3	2/2	4/5
TT	0/0	2/6	2/6

Extraction of unique k-mers (k=2)

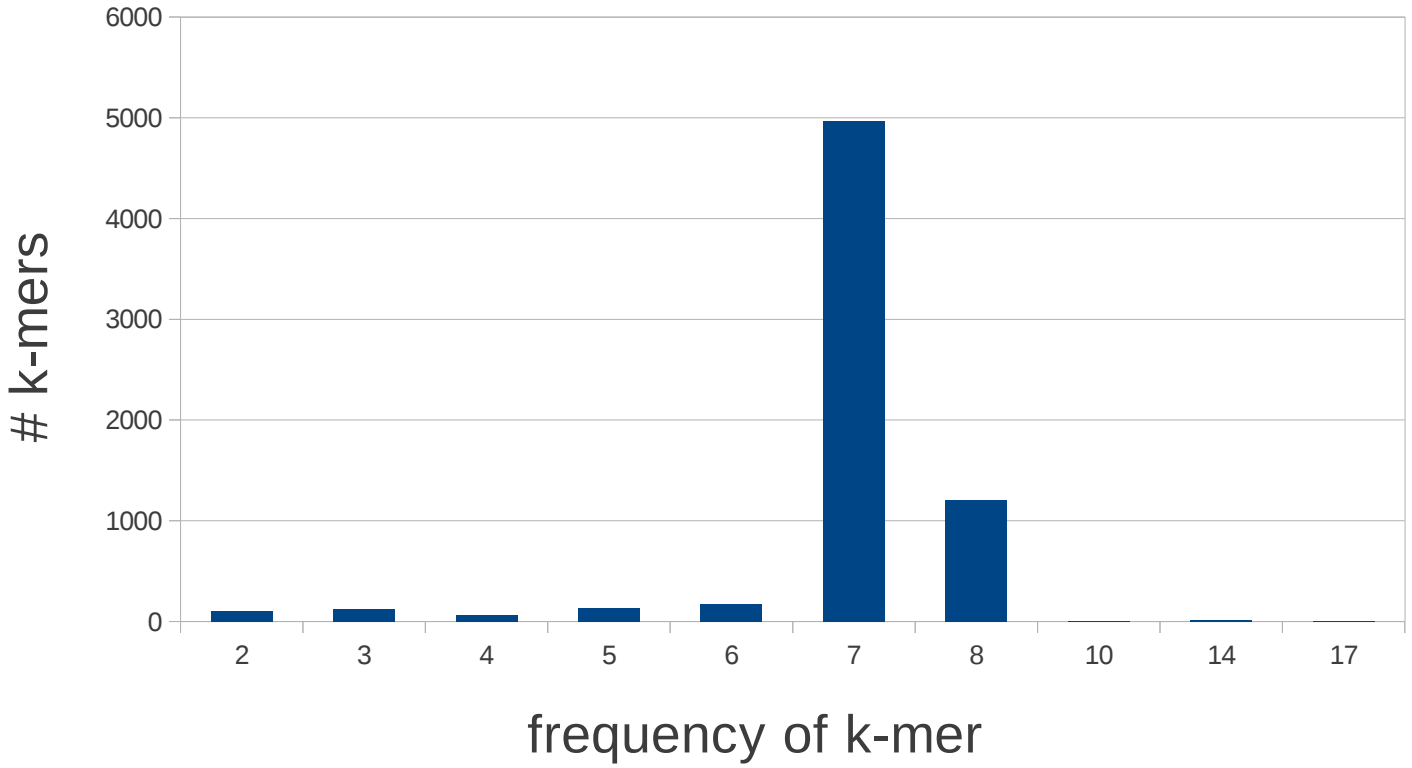
Unique k-mer is a sequence whose frequency at genome is equal to frequency at repeat region

	Repeat region/Genome		
	forward	reverse	Sum
AA	2/6	0/0	2/6
AT	2/2	2/2	4/4
TG	2/2	2/3	4/5
GC	2/2	2/2	4/4
CA	2/3	2/2	4/5
TT	0/0	2/6	2/6

Unique 2-mers for repeat region X is:
AT and GC

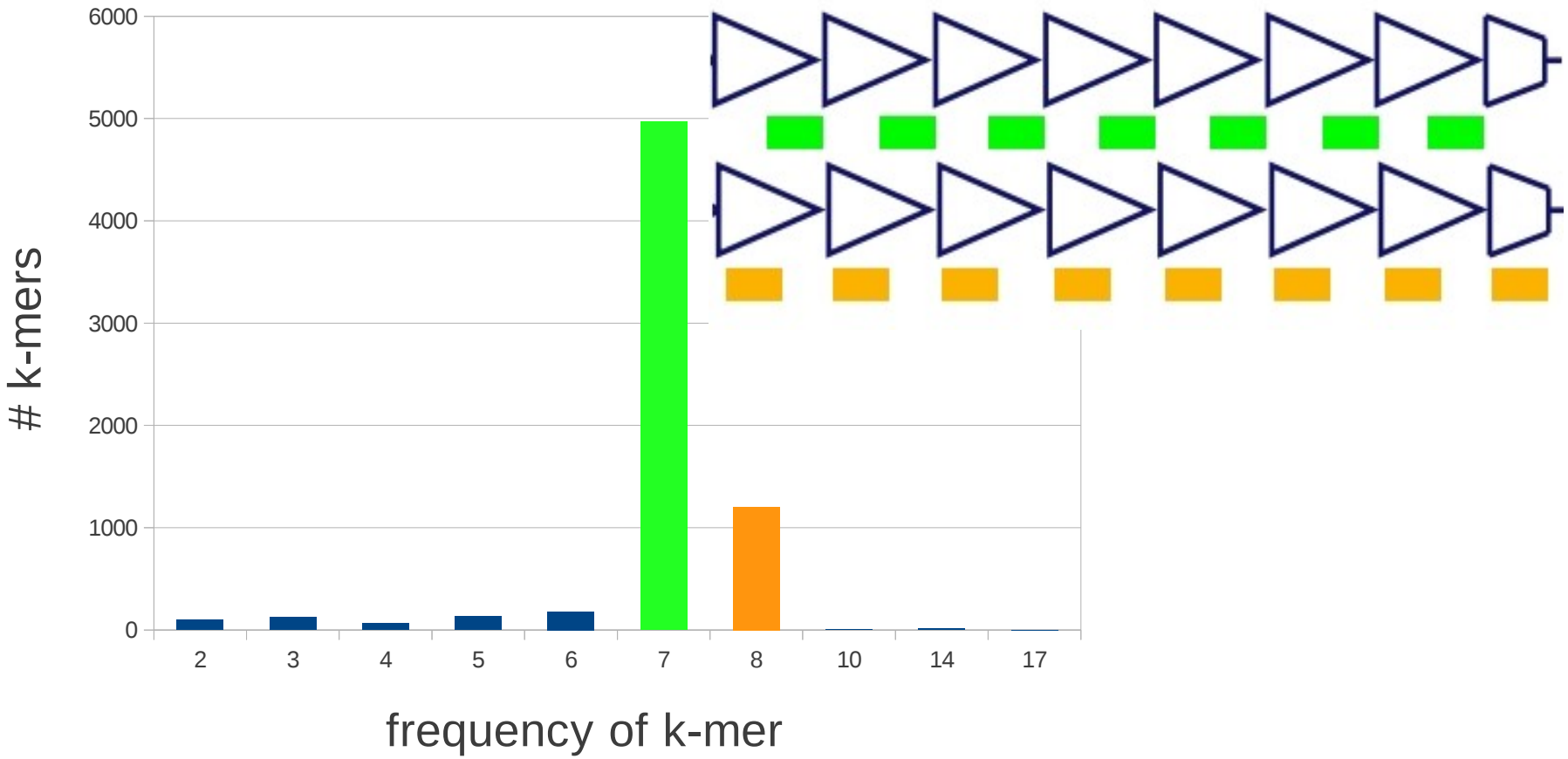
Result: k-mer profiling (k=17)

17-mer profiling for D4Z4_4q repeat region



Result: k-mer profiling (k=17)

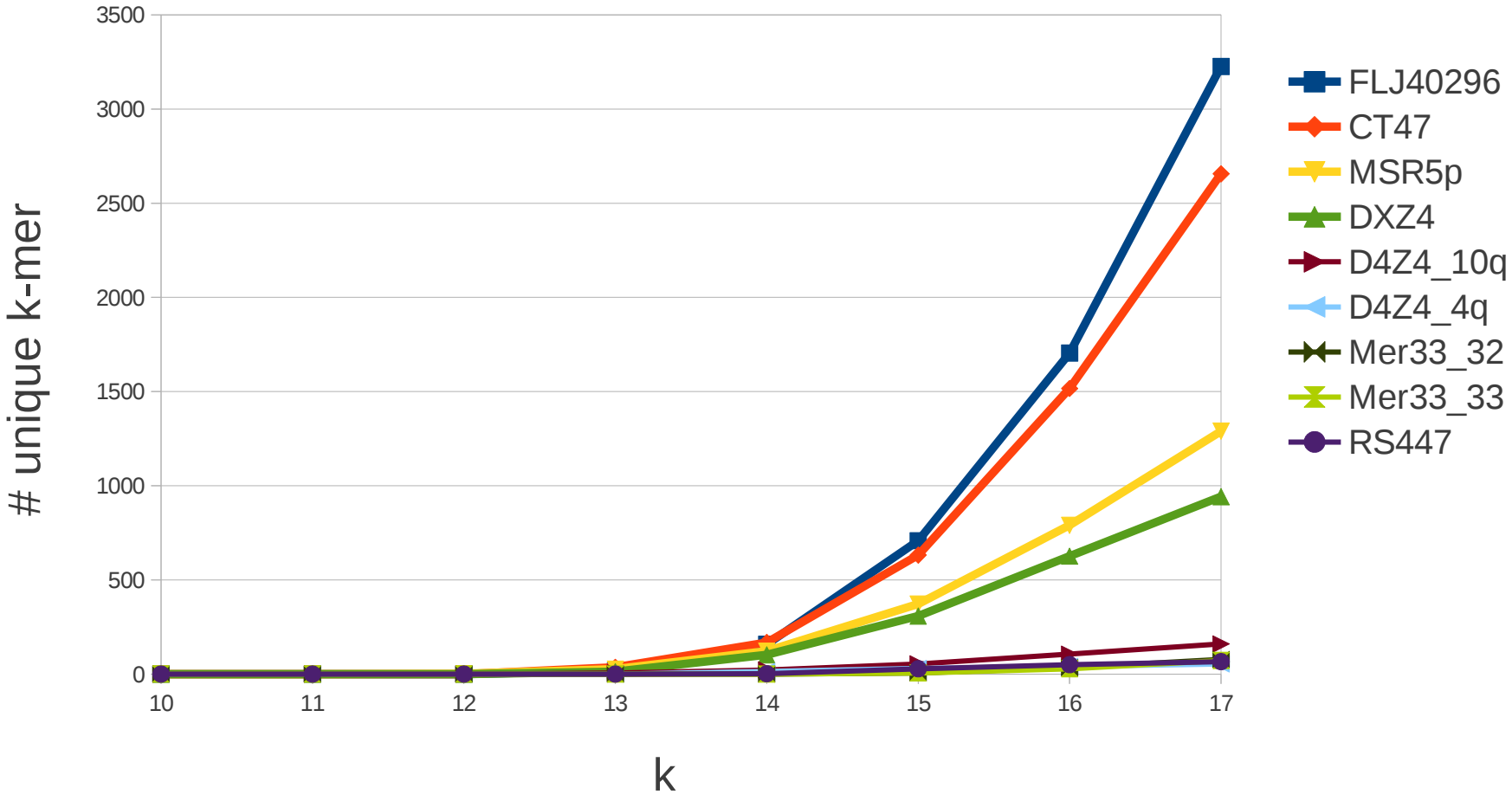
17-mer profiling for D4Z4_4q repeat region



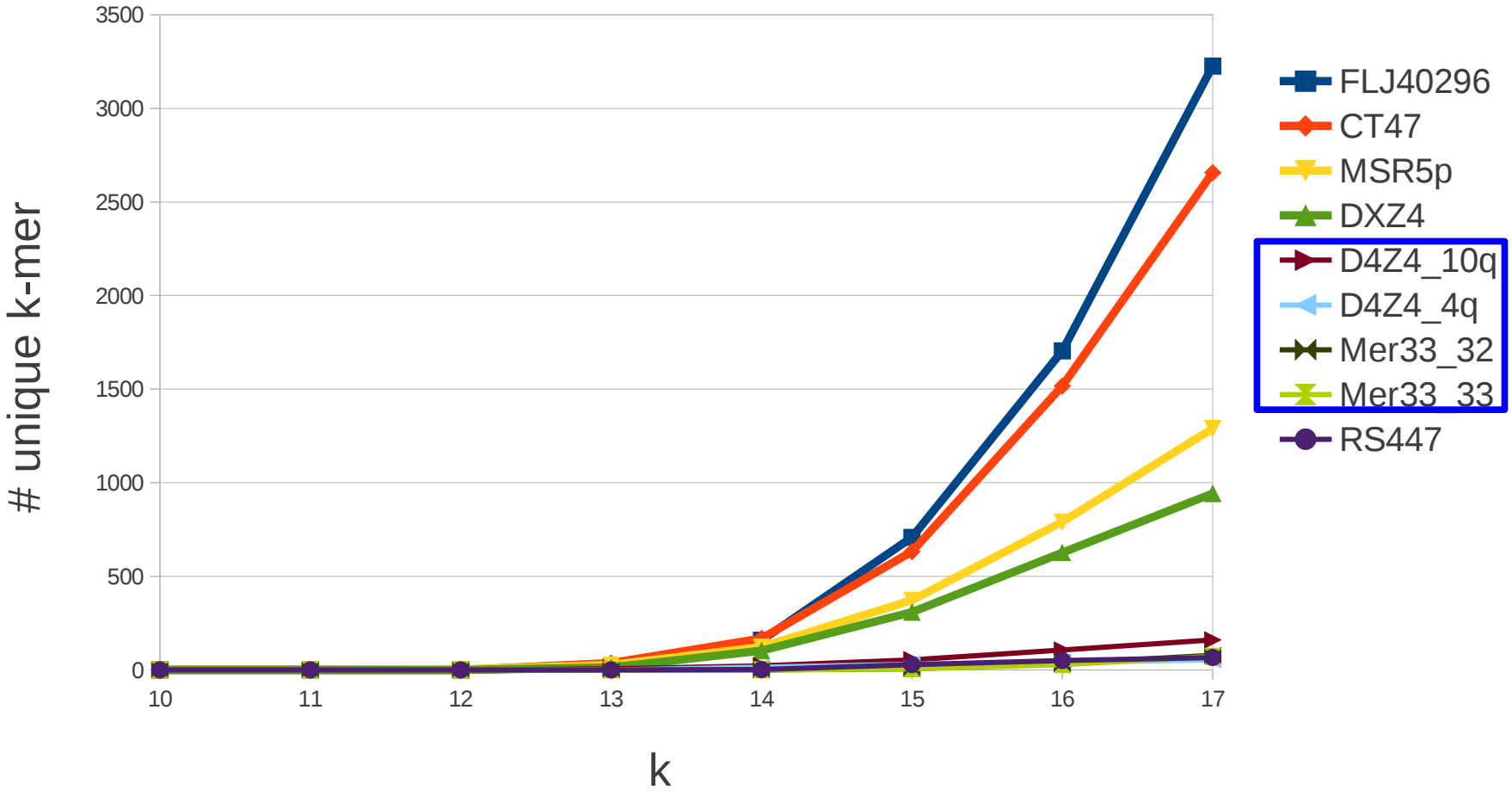
Result: k-mer profiling (k=17)

rep.region	*# units*	*Nastya's #units*
CT47	**11-13**	**14**
D4Z4_4q	**6-8**	**8**
D4Z4_10q	**5-6**	**6**
DXZ4	**14-15**	**14**
FLJ40296	**4-5**	**5**
RS447	**9**	**10**
MSR5p	**7-11**	**11**
Mer33_32	**9-11**	**11**
Mer33_33	**7-11**	**12**

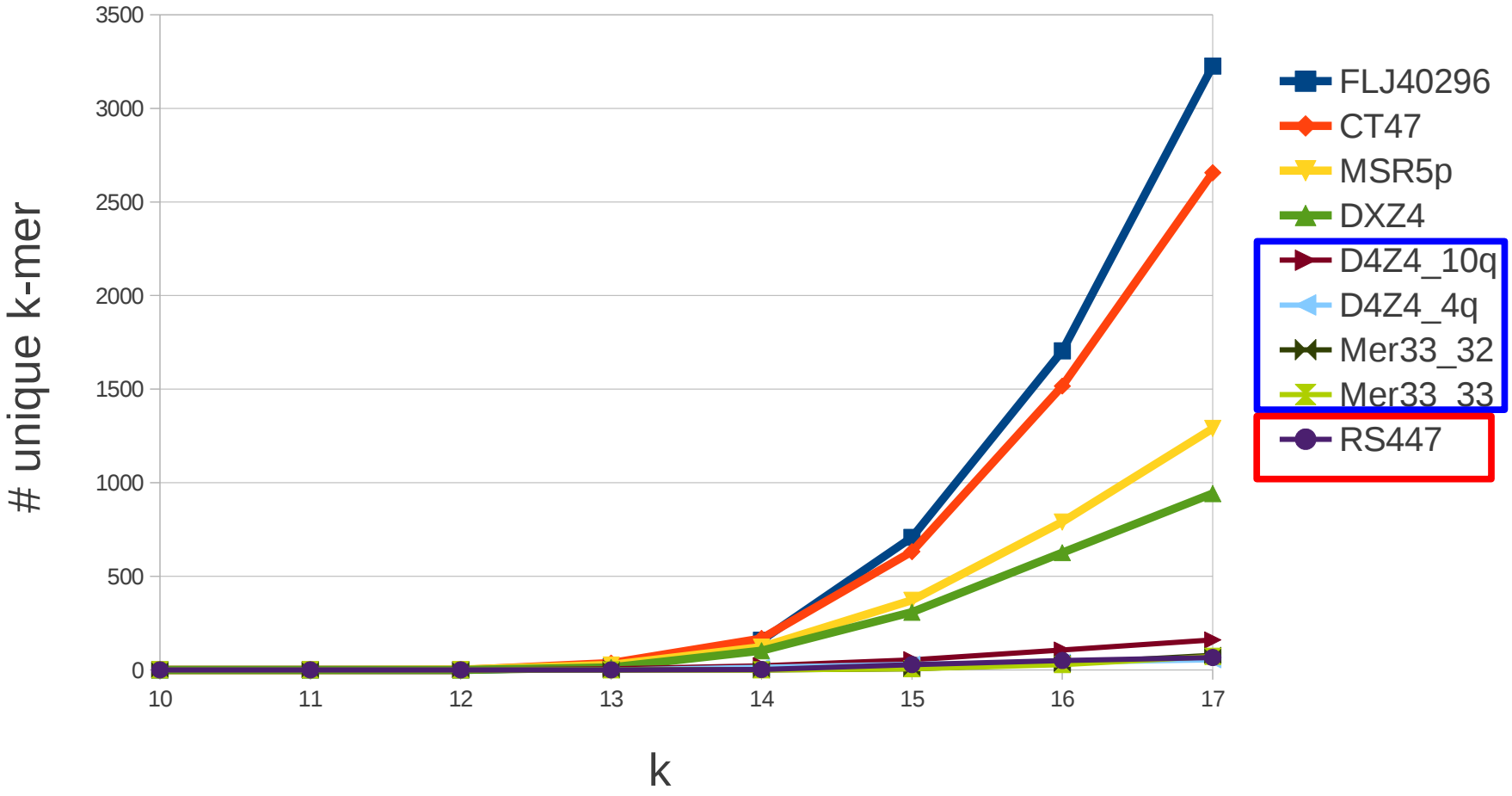
Result: unique k-mer (k=10..17)



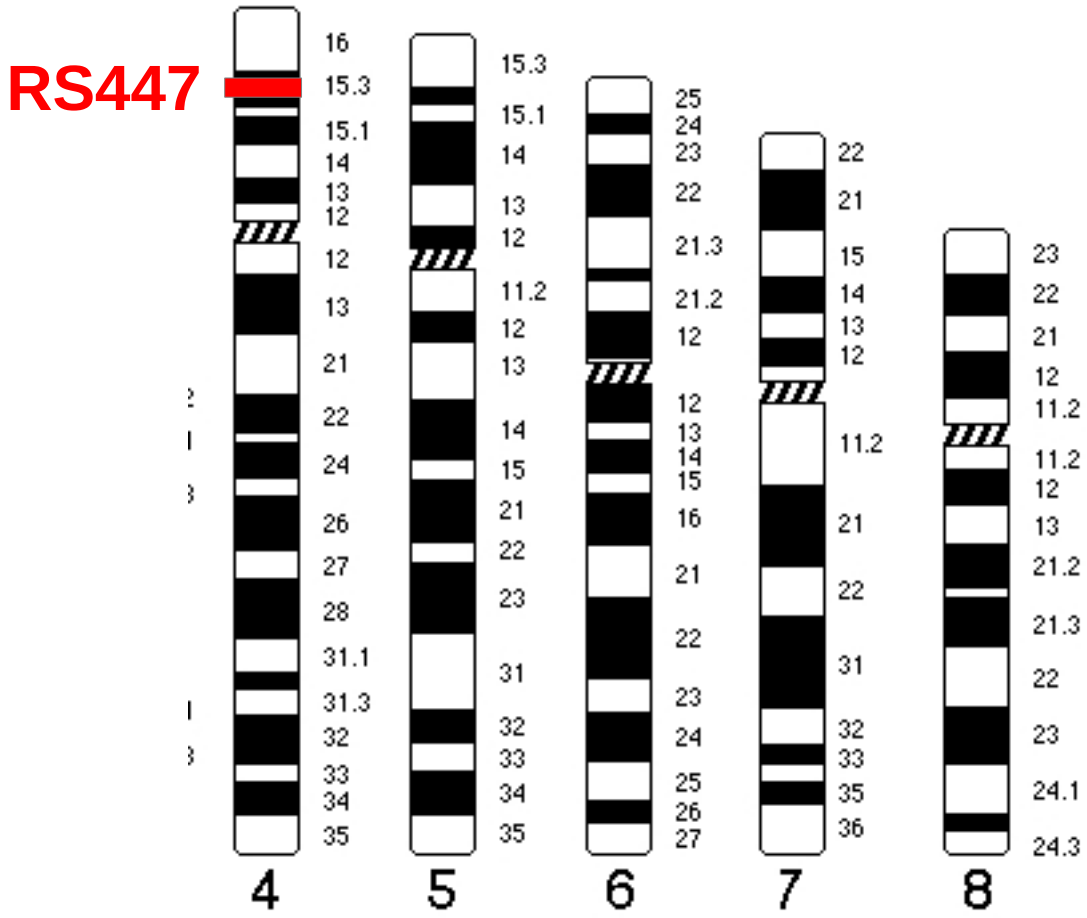
Result: unique k-mer (k=10..17)



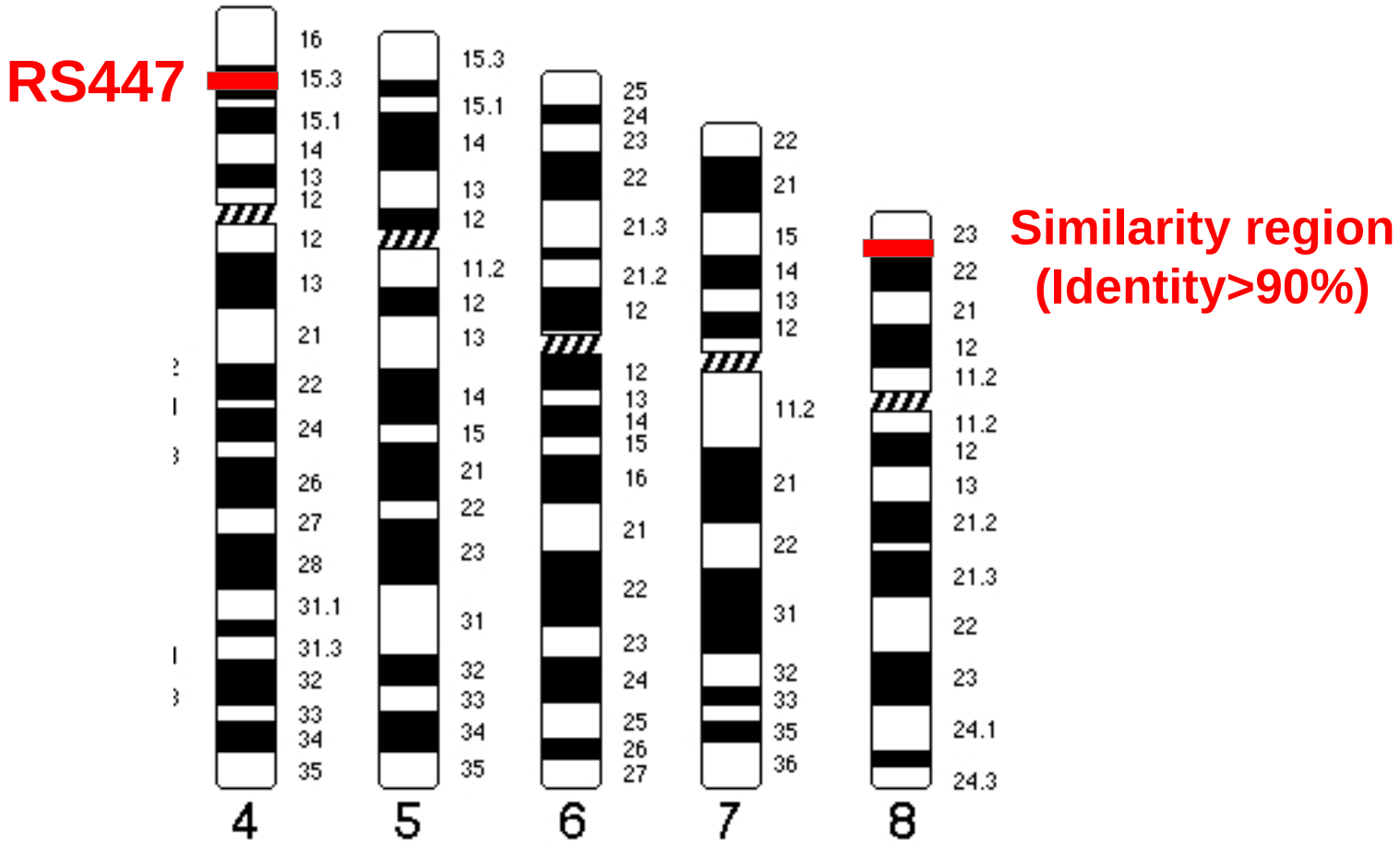
Result: unique k-mer (k=10..17)



Result: RS447



Result: RS447



Aims

Investigating the heterogeneity of repeat regions:

- Heterogeneity among units of one repeat region
- Heterogeneity among repeat regions of one genome
- Heterogeneity of repeat regions among genomes

Future

- There are NextGenerationSequencing sample from 80 different individuals
- Number of units in repeat-region (experimental results for all 80 ind.) and unique k-mers for repeat regions (previous result) are known

Future

- we'll catch reads containing those unique k-mers
- we correlate this result with experimental data

Result: 'predictive power' to infer how many units are in repeate region

Acknowledgment

Irina Pulyakhina

Peter-Bram 't Hoen

Jeroen Laros

Mireille Schaap