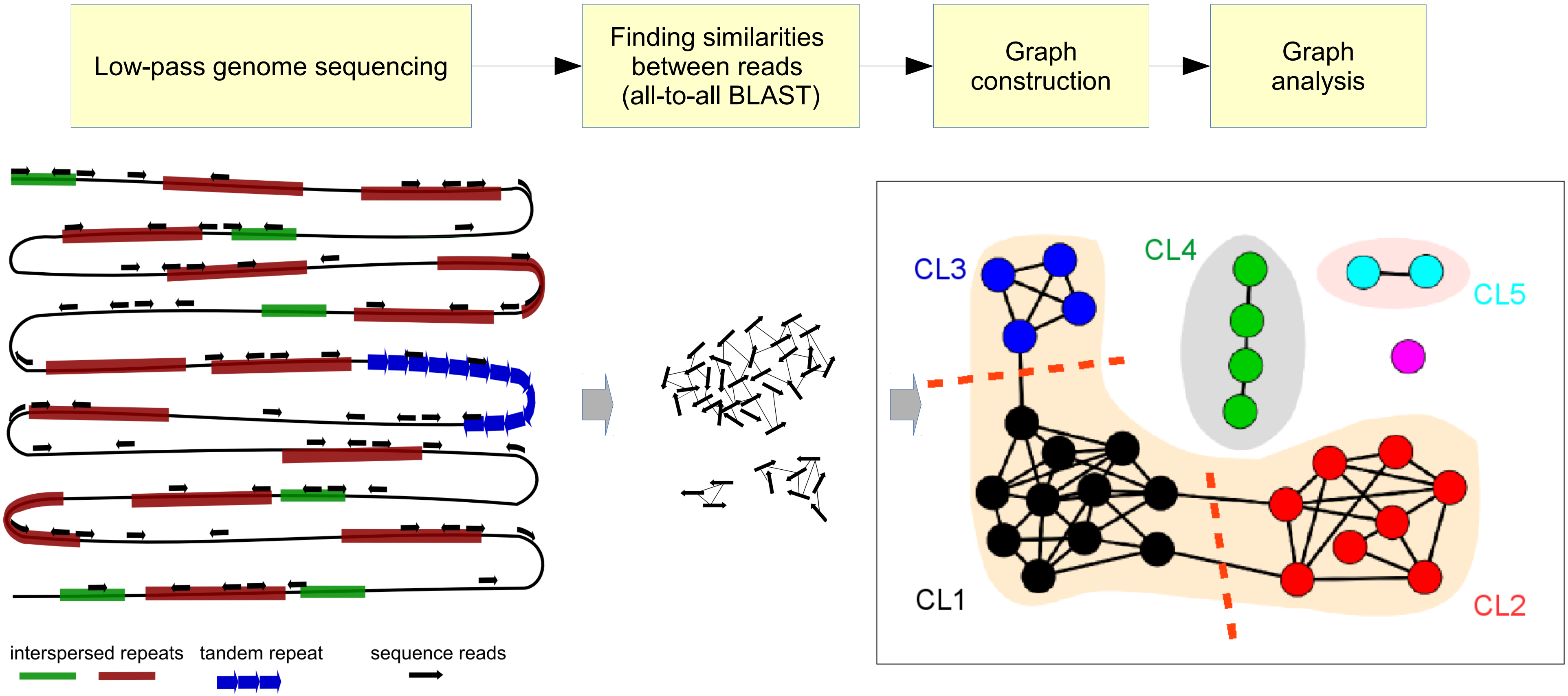
Design of sequencing and repeat analysis experiments

Platform and coverage

- Illumina is preferred, use **paired-end reads**

- avoid coverage > 1x (→ similarity hits from single/low copy sequences) , **0.1-0.5x is optimal**



Low-pass genome sequencing → Finding similarities between reads (all-to-all BLAST) → Graph construction → Graph analysis

Design of sequencing and repeat analysis experiments

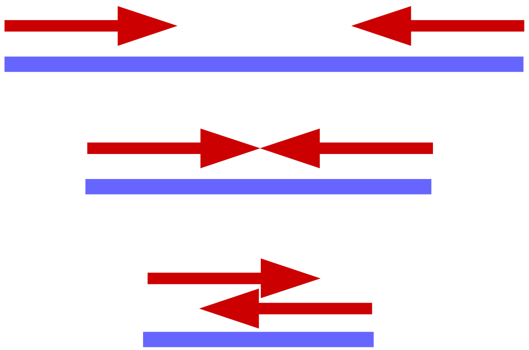
Platform and coverage

- Illumina is preferred, use **paired-end reads**
- avoid coverage > 1x (→ similarity hits from single/low copy sequences) , **0.1-0.5x is optimal**

Read length vs. fragment length

- sequenced fragments should be > 2 x read length

paired-end reads



Consider eventual bias in template preparation

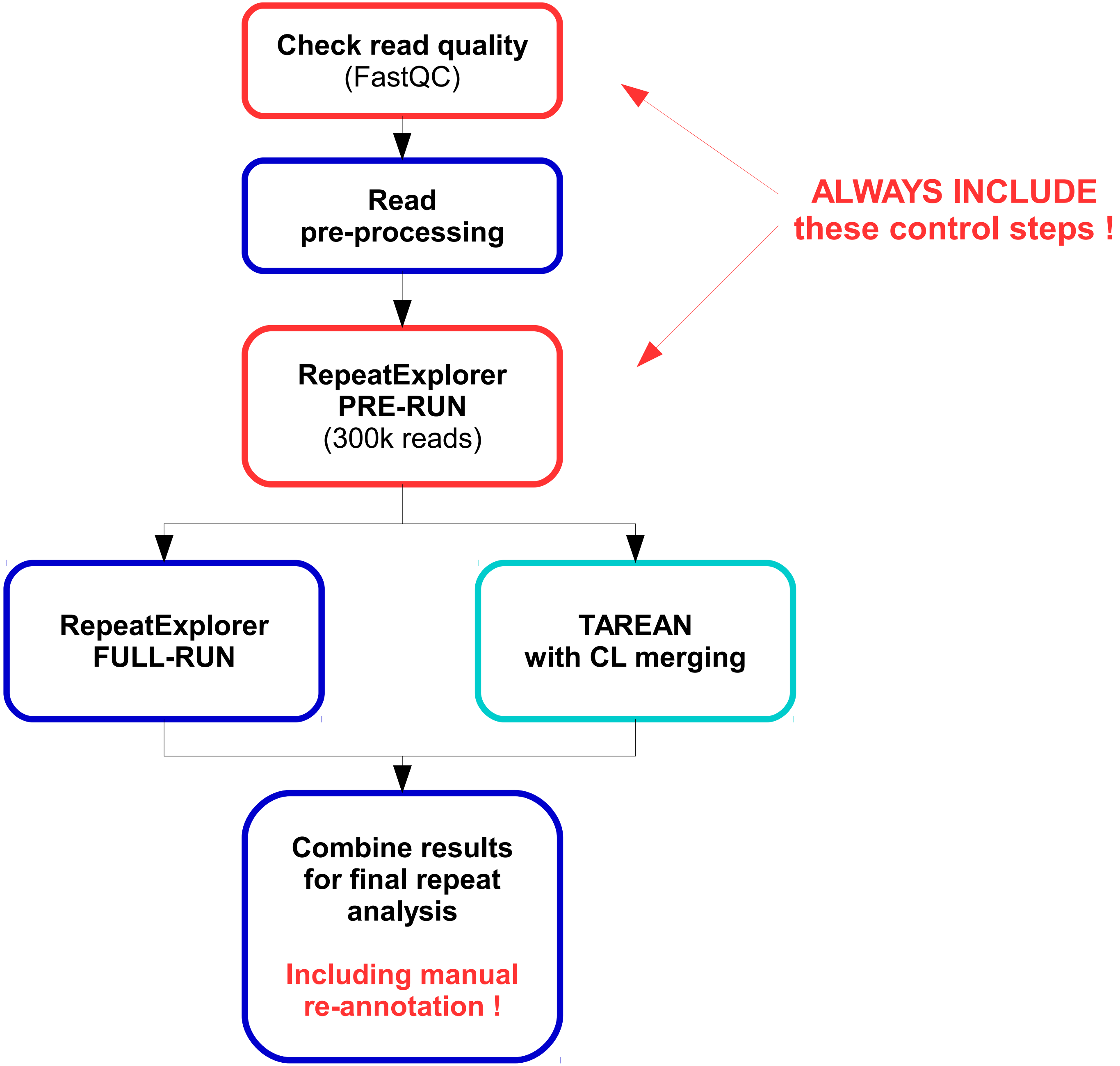
- PCR-based   x   PCR-free kits
- avoid transposon-based kits

Spend some more money to make experimental replicates for quantification

- repeating library preparations (not re-sequencing from the same library)

CHECK YOUR INPUT DATA ! (and do not trust public sequence archives)

Analysis workflow



Check read quality (FastQC)

Read pre-processing

RepeatExplorer PRE-RUN (300k reads)

RepeatExplorer FULL-RUN

TAREAN with CL merging

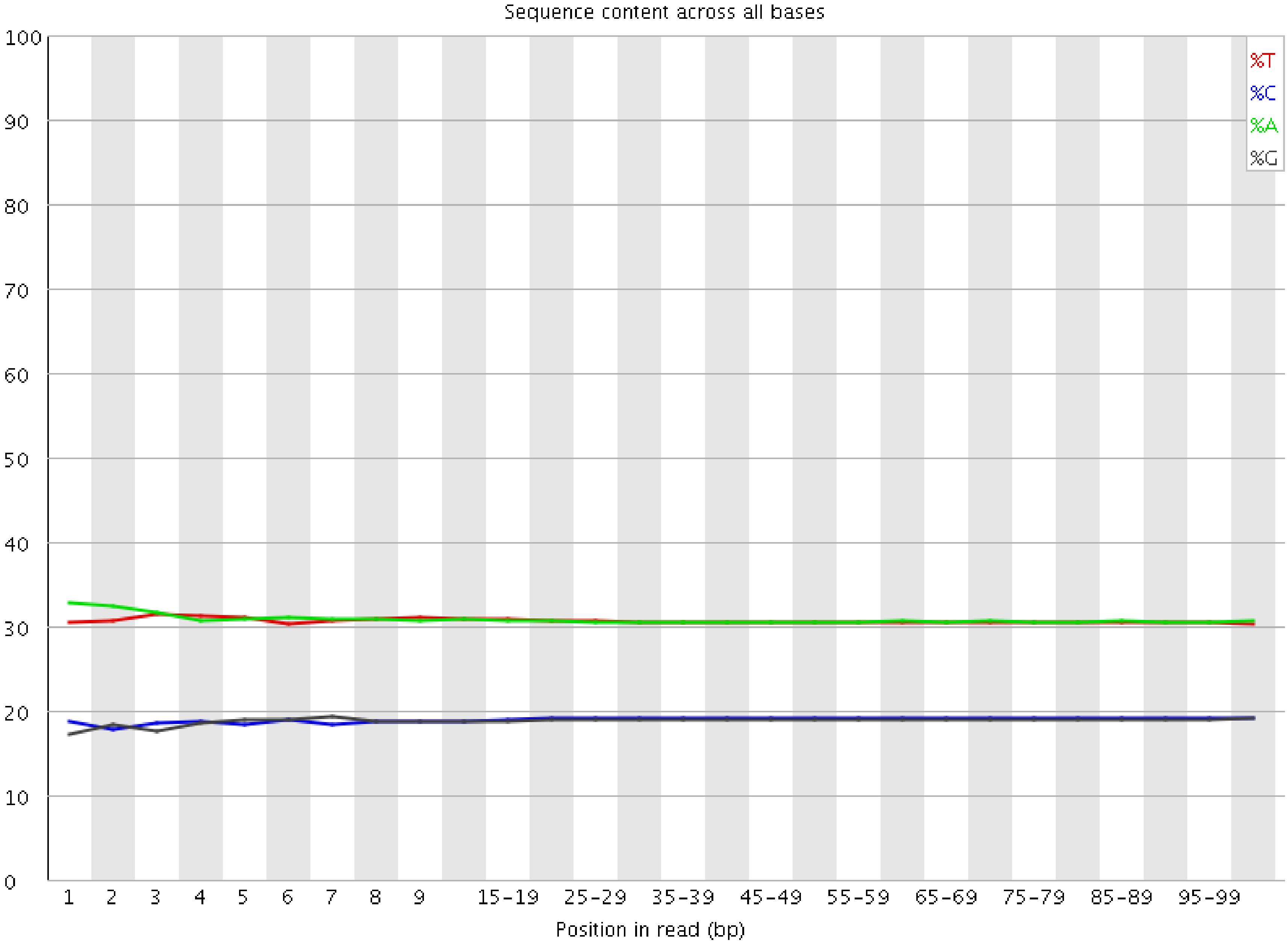
Combine results for final repeat analysis

Including manual re-annotation !

ALWAYS INCLUDE these control steps !

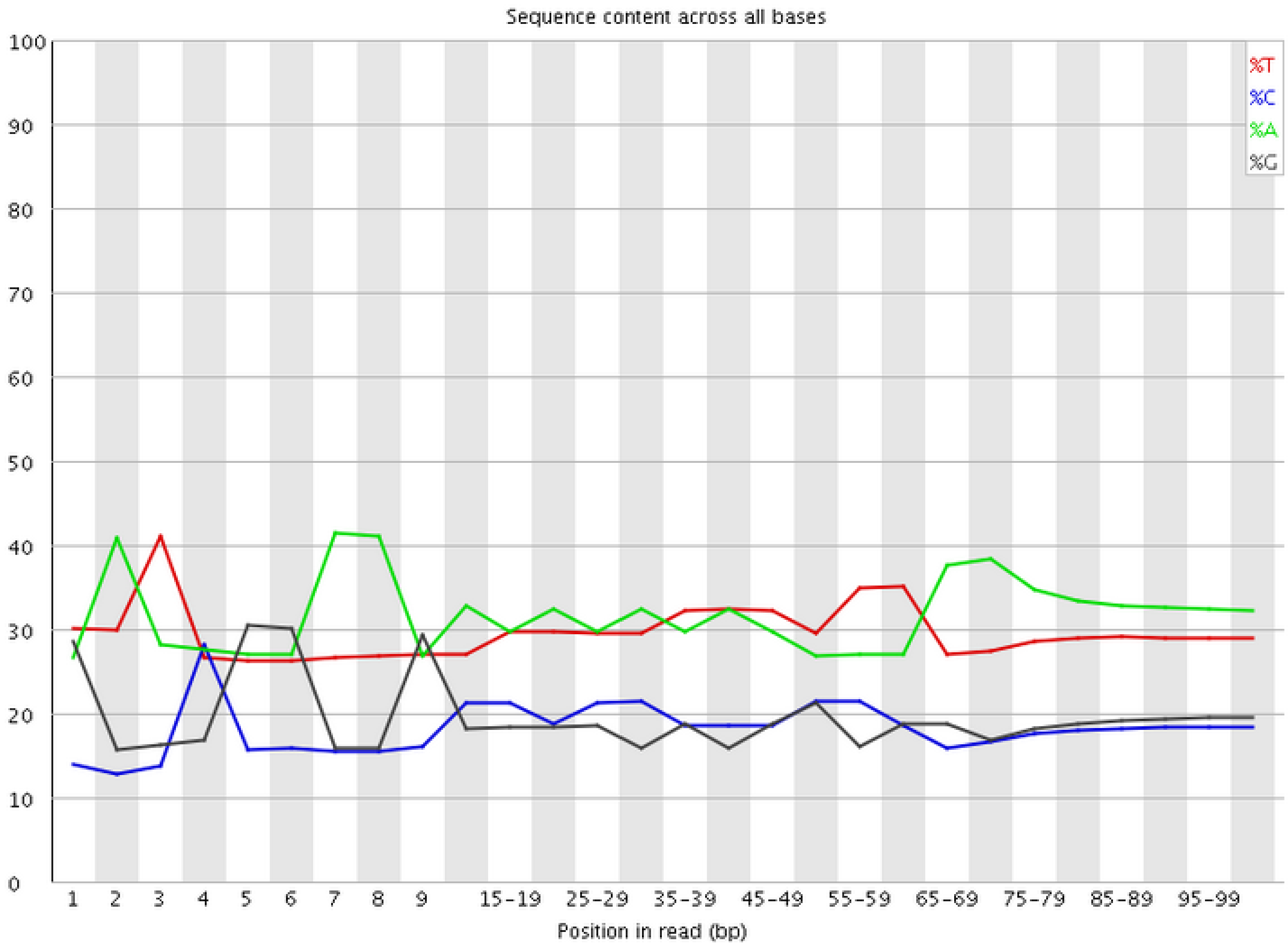
OK

Per base sequence content

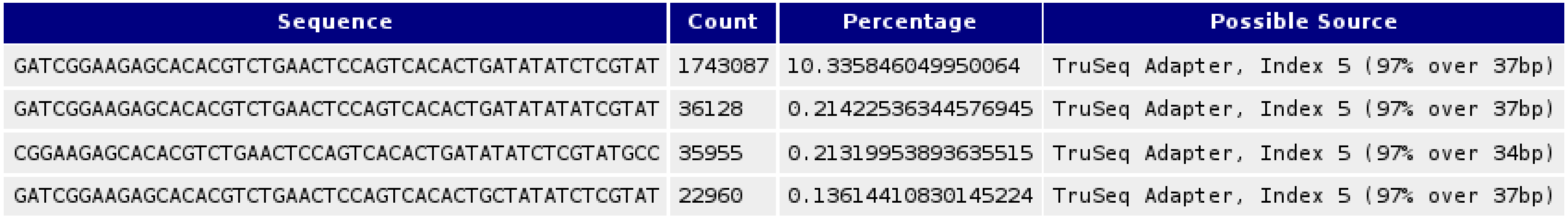


Per base sequence content

BAD !



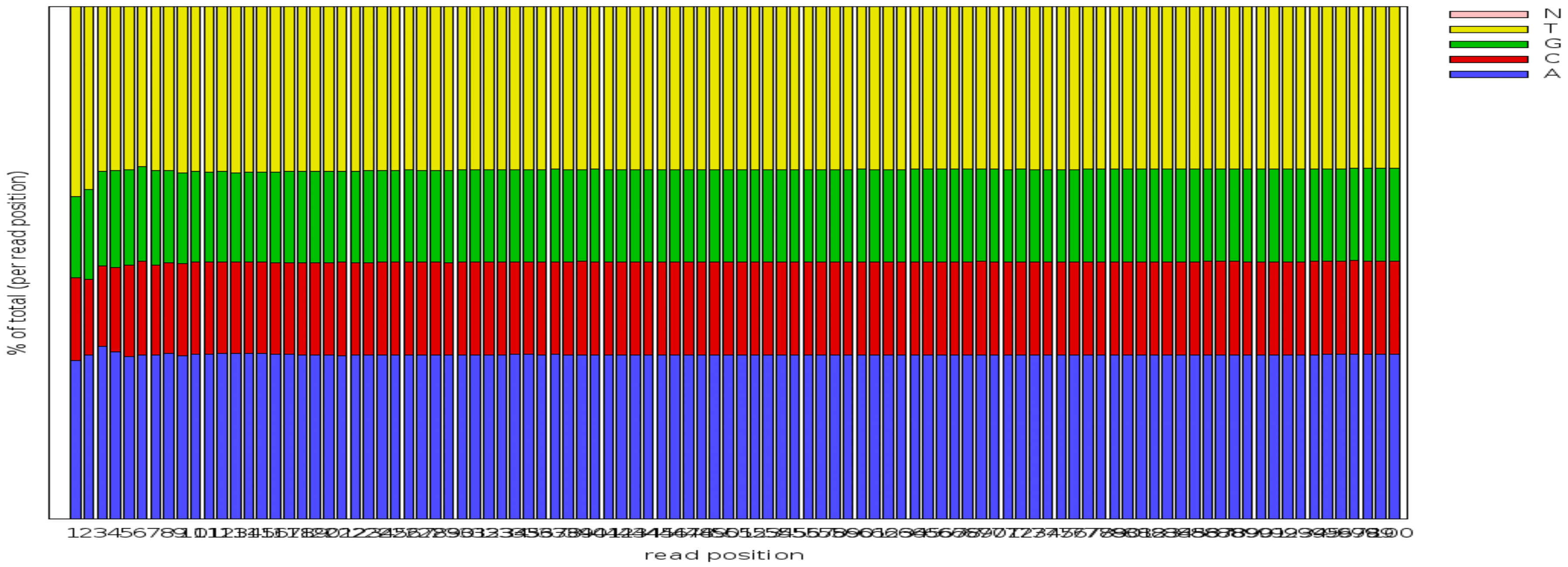
Overrepresented sequences



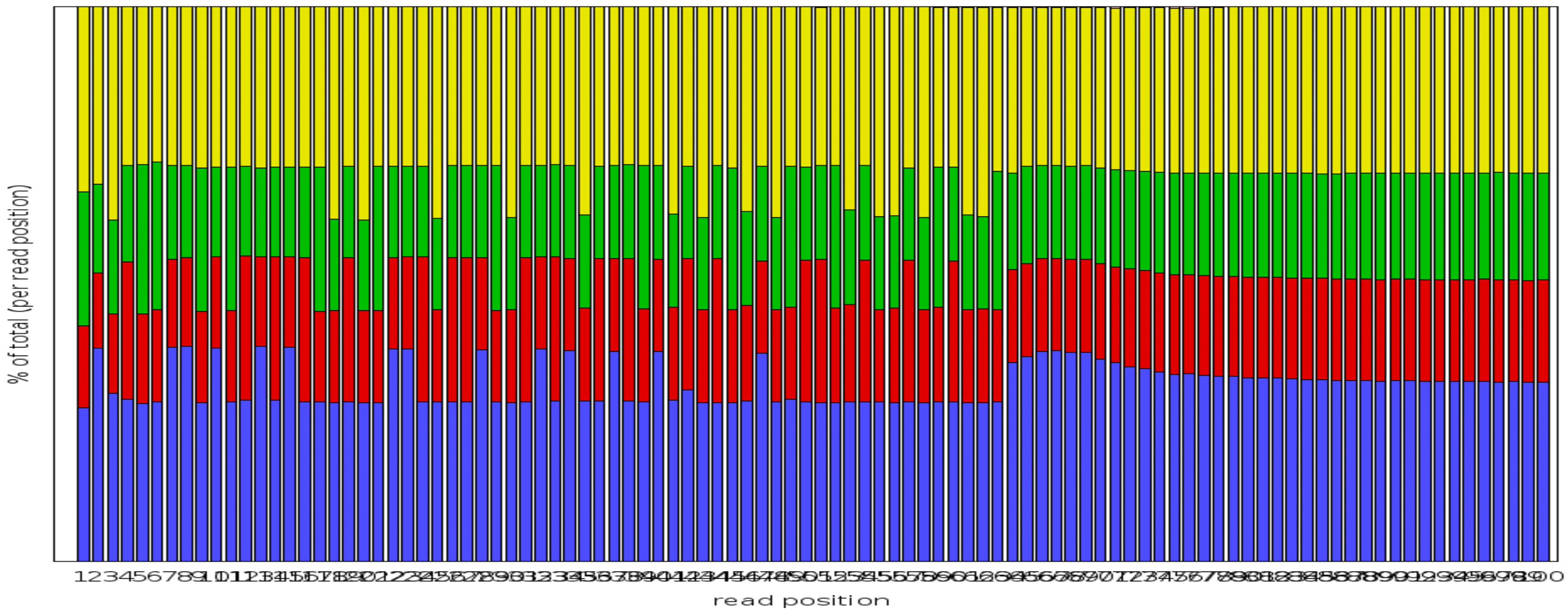
Sequence	Count	Percentage	Possible Source
GATCGGAAGAGCACACGTCTGAACTCCAGTCACACTGATATATCTCGTAT	1743087	10.335846049950064	TruSeq Adapter, Index 5 (97% over 37bp)
GATCGGAAGAGCACACGTCTGAACTCCAGTCACACTGATATATATCGTAT	36128	0.21422536344576945	TruSeq Adapter, Index 5 (97% over 37bp)
CGGAAGAGCACACGTCTGAACTCCAGTCACACTGATATATCTCGTATGCC	35955	0.21319953893635515	TruSeq Adapter, Index 5 (97% over 34bp)
GATCGGAAGAGCACACGTCTGAACTCCAGTCACACTGCTATATCTCGTAT	22960	0.13614410830145224	TruSeq Adapter, Index 5 (97% over 37bp)

FASTX-Toolkit: Compute quality statistics -> Draw nucleotides distribution chart
(alternative to FastQC)

OK

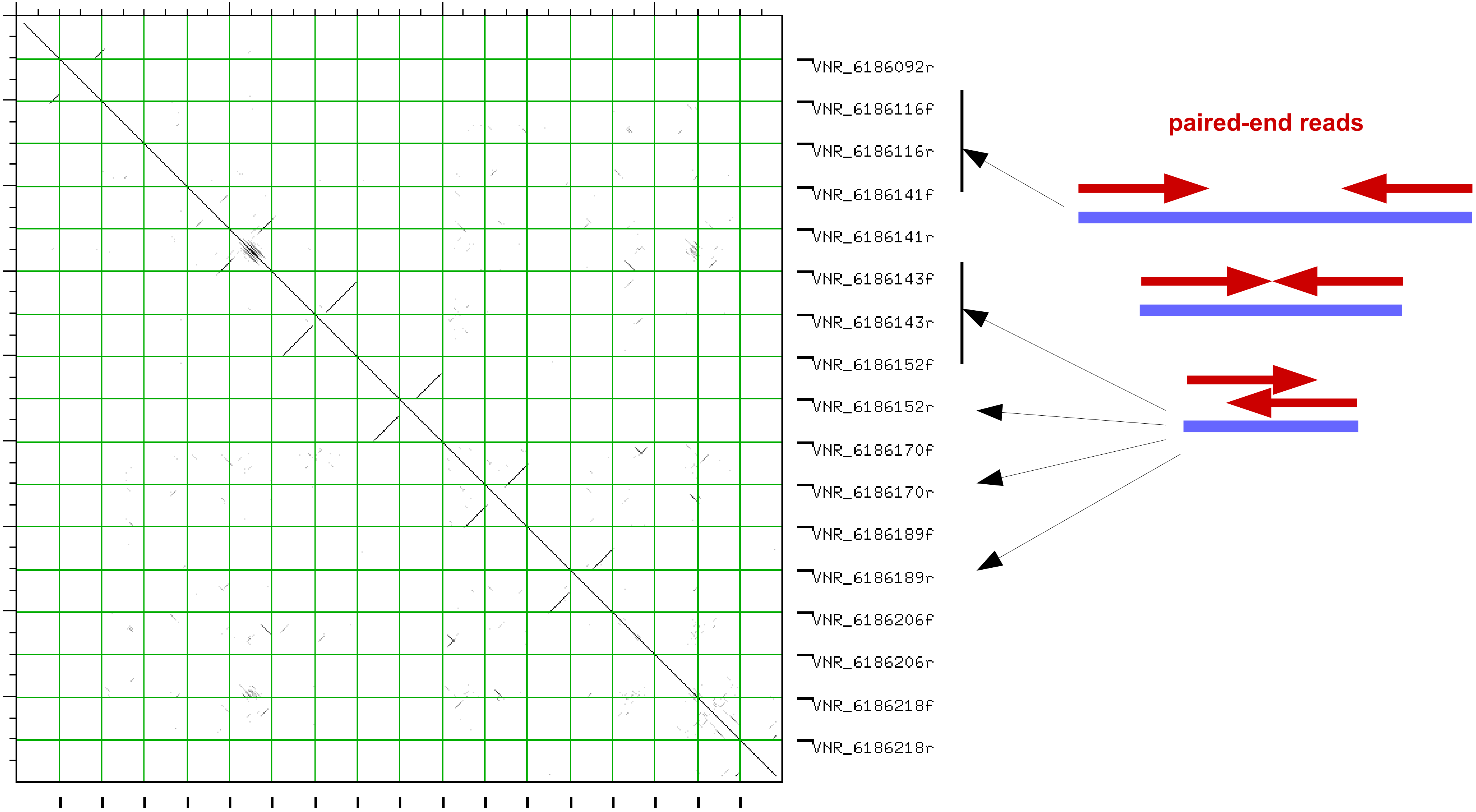


Adaptor contam.



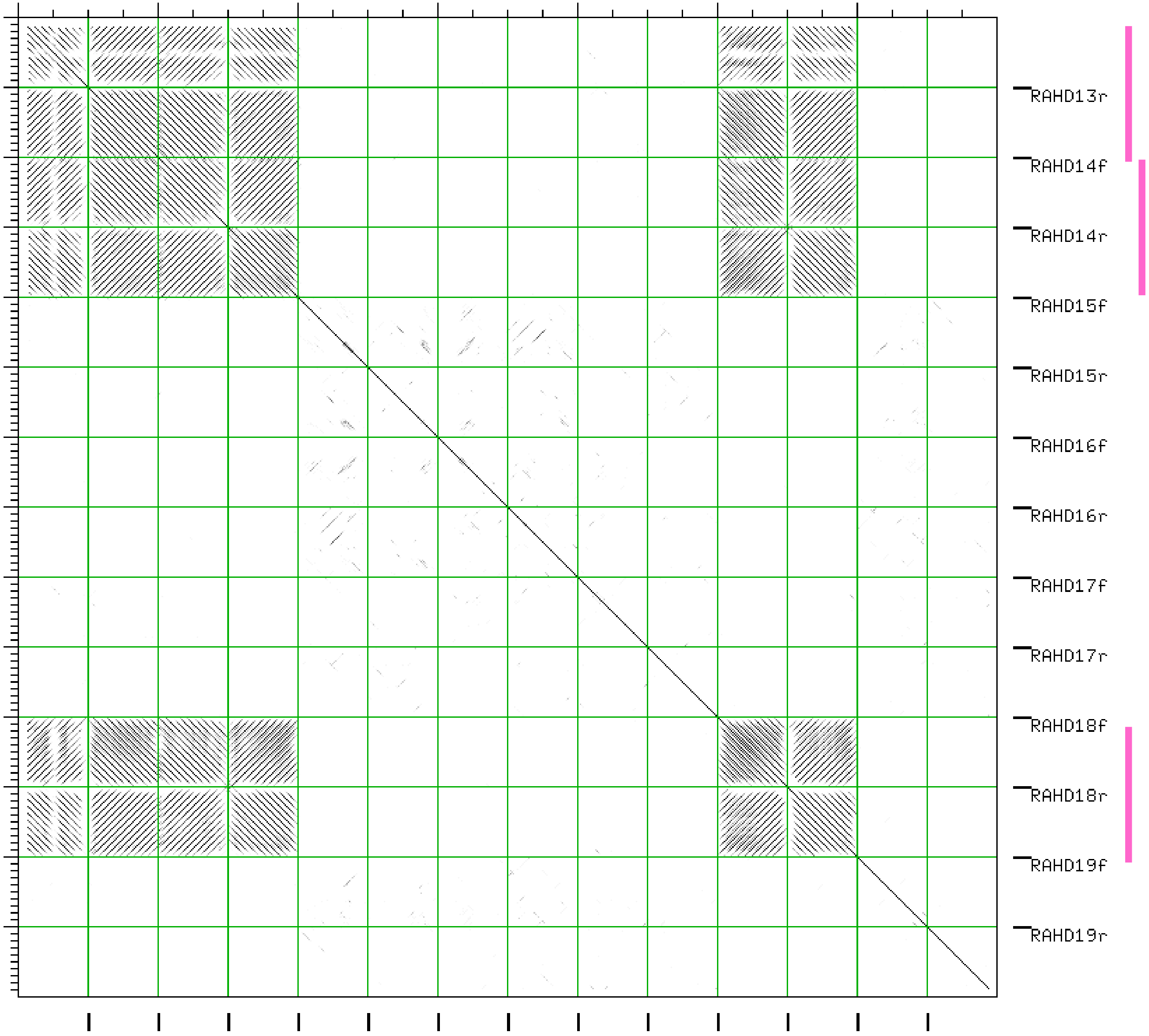
Read length vs. fragment length

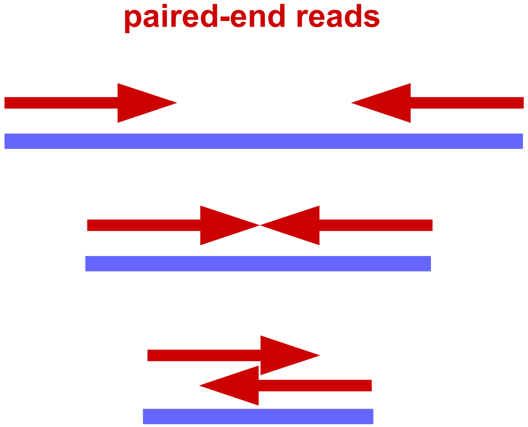
- sequenced fragments should be > 2 x read length



Read length vs. fragment length

- sequenced fragments should be > 2 x read length





Reads containing tandem repeats with short monomers

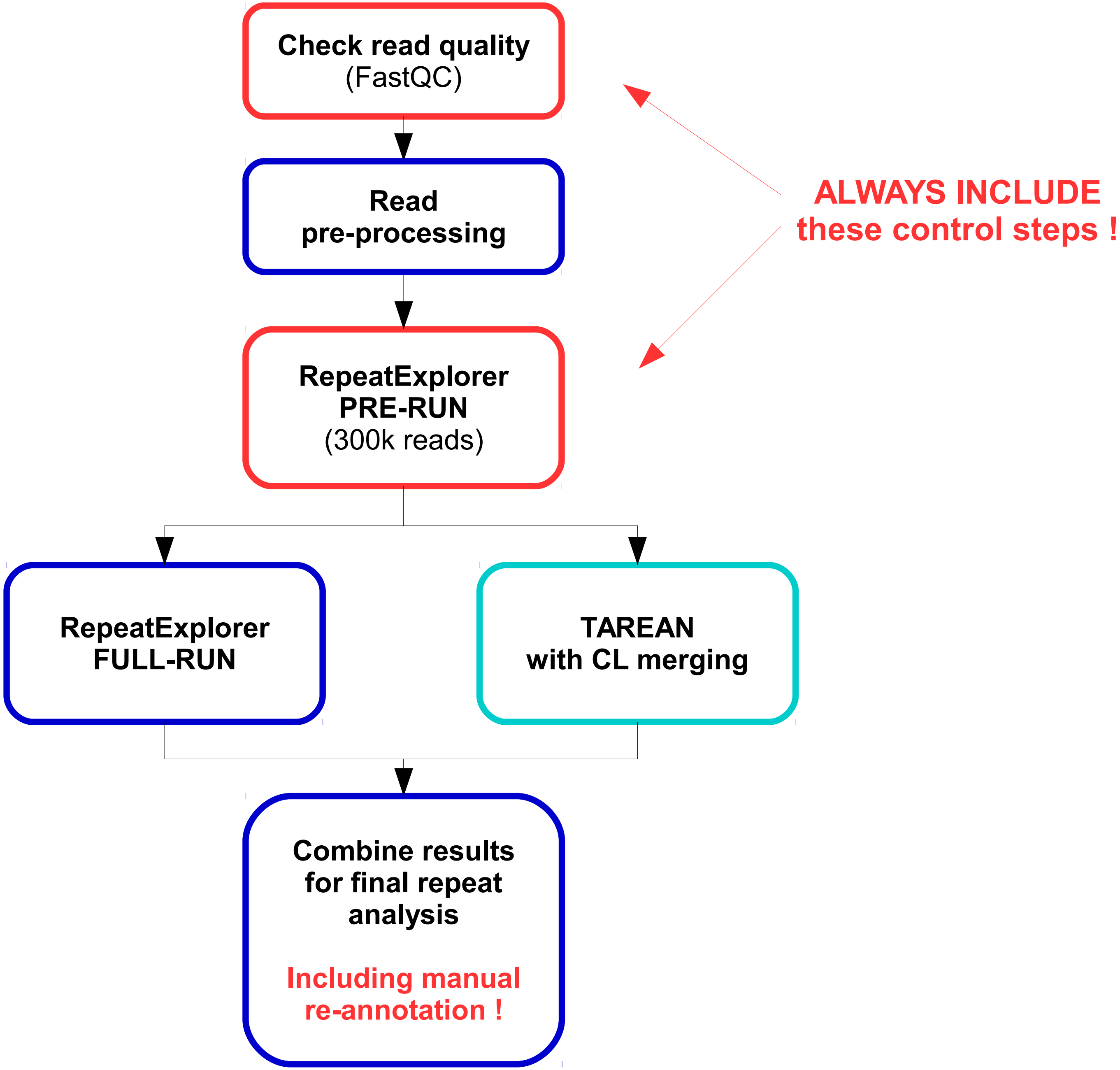
!!! can be detected as overlapping !!!

Adela (*Adelhaide kratzmarii*)

- Fictional plant from a popular Czech movie
- We reconstructed its genome ;-)
- It turned out to be very small but made only by repeats !
- EXCELLENT for training at this workshop

- Do not expect such nice results with your real data
- On the other hand, your real model will not eat you

Analysis workflow



Check read quality
(FastQC)

Read
pre-processing

RepeatExplorer
PRE-RUN
(300k reads)

RepeatExplorer
FULL-RUN

TAREAN
with CL merging

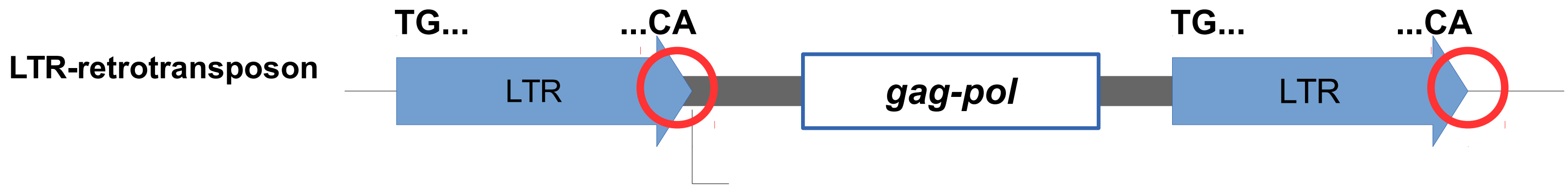
Combine results
for final repeat
analysis

Including manual
re-annotation !

ALWAYS INCLUDE
these control steps !

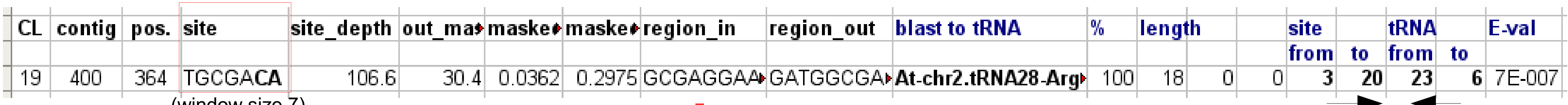
Insertion sites of mobile elements

A new tool for detection of LTR / PBS sites



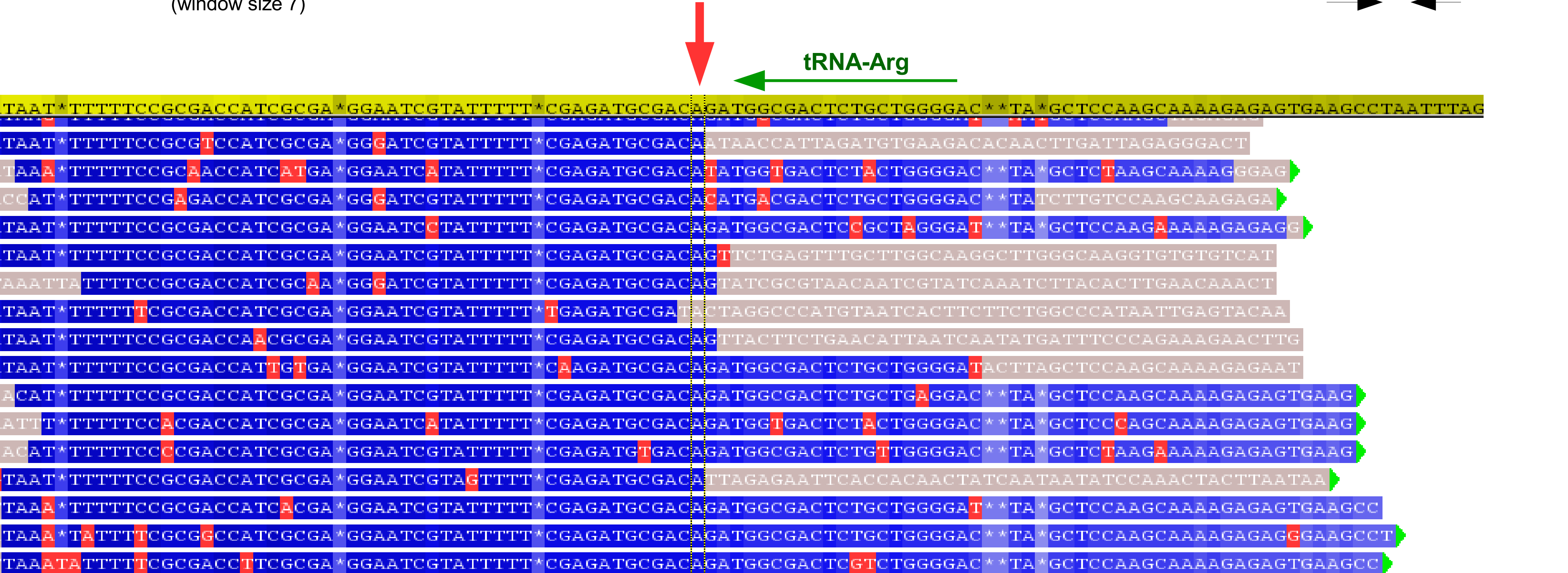
PBS (primer binding site)
similarity to tRNA

Program output:



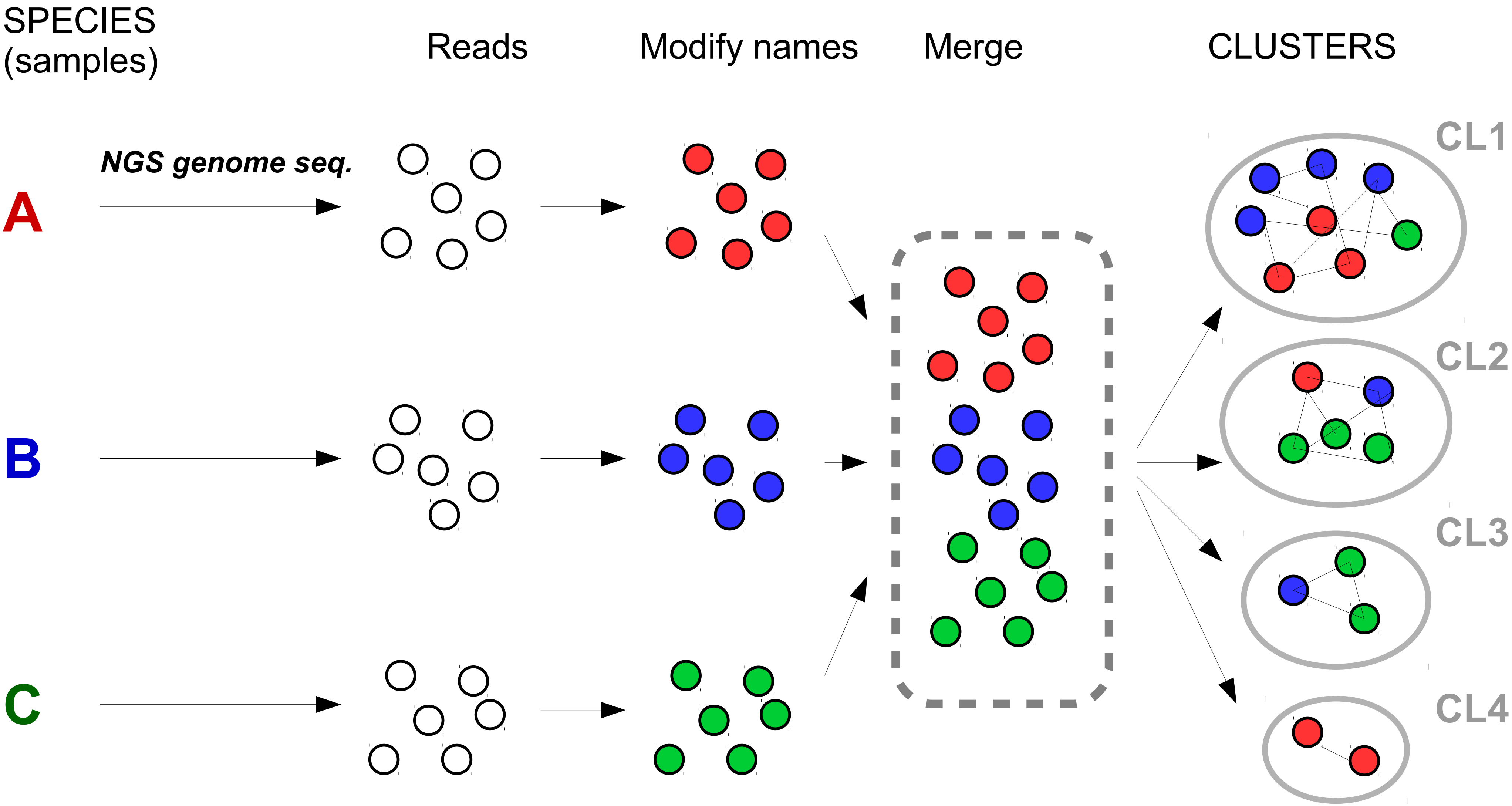
CL	contig	pos.	site	site_depth	out_ma▸	maske▸	maske▸	region_in	region_out	blast to tRNA	%	length			site		tRNA		E-val
															from	to	from	to	
19	400	364	TGCGA**CA**	106.6	30.4	0.0362	0.2975	GCGAGGAA▸	GATGGCGA▸	**At-chr2.tRNA28-Arg▸**	100	18	0	0	**3**	**20**	**23**	**6**	7E-007

(window size 7)



```
TAAT*TTTTTCCGCGACCATCGCGA*GGAATCGTATTTTT*CGAGATGCGACAGATGGCGACTCTGCTGGGGAC**TA*GCTCCAAGCAAAAGAGAGTGAAGCCTAATTTAG
TAAT*TTTTTCCGCGTCCATCGCGA*GGGATCGTATTTTT*CGAGATGCGACAATAACCATTAGATGTGAAGACACAACTTGATTAGAGGGACT
TAAA*TTTTTCCGCAACCATCATGA*GGAATCATATTTTT*CGAGATGCGACATATGGTGACTCTACTGGGGAC**TA*GCTCTAAGCAAAAGGGAG
CCAT*TTTTTCCGAGACCATCGCGA*GGGATCGTATTTTT*CGAGATGCGACACATGACGACTCTGCTGGGGAC**TATCTTGTCCAAGCAAGAGA
TAAT*TTTTTCCGCGACCATCGCGA*GGAATCCTATTTTT*CGAGATGCGACAGATGGCGACTCCGCTAGGGAT**TA*GCTCCAAGAAAAAGAGAGG
TAAT*TTTTTCCGCGACCATCGCGA*GGAATCGTATTTTT*CGAGATGCGACAGTTCTGAGTTTGCTTGGCAAGGCTTGGGCAAGGTGTGTGTCAT
AAATTATTTTCCGCGACCATCGCAA*GGGATCGTATTTTT*CGAGATGCGACAGTATCGCGTAACAATCGTATCAAATCTTACACTTGAACAAACT
TAAT*TTTTTTCGCGACCATCGCGA*GGAATCGTATTTTT*TGAGATGCGATACTAGGCCCATGTAATCACTTCTTCTGGCCCATAATTGAGTACAA
TAAT*TTTTTCCGCGACCAACGCGA*GGAATCGTATTTTT*CGAGATGCGACAGTTACTTCTGAACATTAATCAATATGATTTCCCAGAAAGAACTTG
TAAT*TTTTTCCGCGACCATTGTGA*GGAATCGTATTTTT*CAAGATGCGACAGATGGCGACTCTGCTGGGGATACTTAGCTCCAAGCAAAAGAGAAT
ACAT*TTTTTCCGCGACCATCGCGA*GGAATCGTATTTTT*CGAGATGCGACAGATGGCGACTCTGCTGAGGAC**TA*GCTCCAAGCAAAAGAGAGTGAAG
ATTT*TTTTTCCACGACCATCGCGA*GGAATCATATTTTT*CGAGATGCGACAGATGGTGACTCTACTGGGGAC**TA*GCTCCCAGCAAAAGAGAGTGAAG
ACAT*TTTTTCCCCGACCATCGCGA*GGAATCGTATTTTT*CGAGATGTGACAGATGGCGACTCTGTTGGGGAC**TA*GCTCTAAGAAAAAGAGAGTGAAG
TAAT*TTTTTCCGCGACCATCGCGA*GGAATCGTAGTTTT*CGAGATGCGACATTAGAGAATTCACCACAACTATCAATAATATCCAAACTACTTAATAA
TAAA*TTTTTCCGCGACCATCACGA*GGAATCGTATTTTT*CGAGATGCGACAGATGGCGACTCTGCTGGGGAT**TA*GCTCCAAGCAAAAGAGAGTGAAGCC
TAAA*TATTTTCGCGGCCATCGCGA*GGAATCGTATTTTT*CGAGATGCGACAGATGGCGACTCTGCTGGGGAC**TA*GCTCCAAGCAAAAGAGAGGGAAGCCT
TAAATATTTTTCGCGACCTTCGCGA*GGAATCGTATTTTT*CGAGATGCGACAGATGGCGACTCGTCTGGGGAC**TA*GCTCCAAGCAAAAGAGAGTGAAGCC
```

Comparative analysis - principle

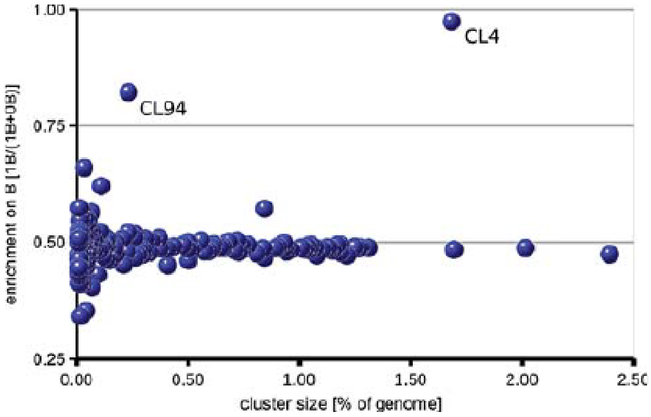


Comparative analysis

Two samples only (e.g. genotypes of the same species +/- B chromosomes)

- simultaneous clustering of +B and 0B reads, same genome coverage

- evaluate ratios of +B/0B reads in each cluster



Multiple samples (e.g. a set of species differing in genome size)

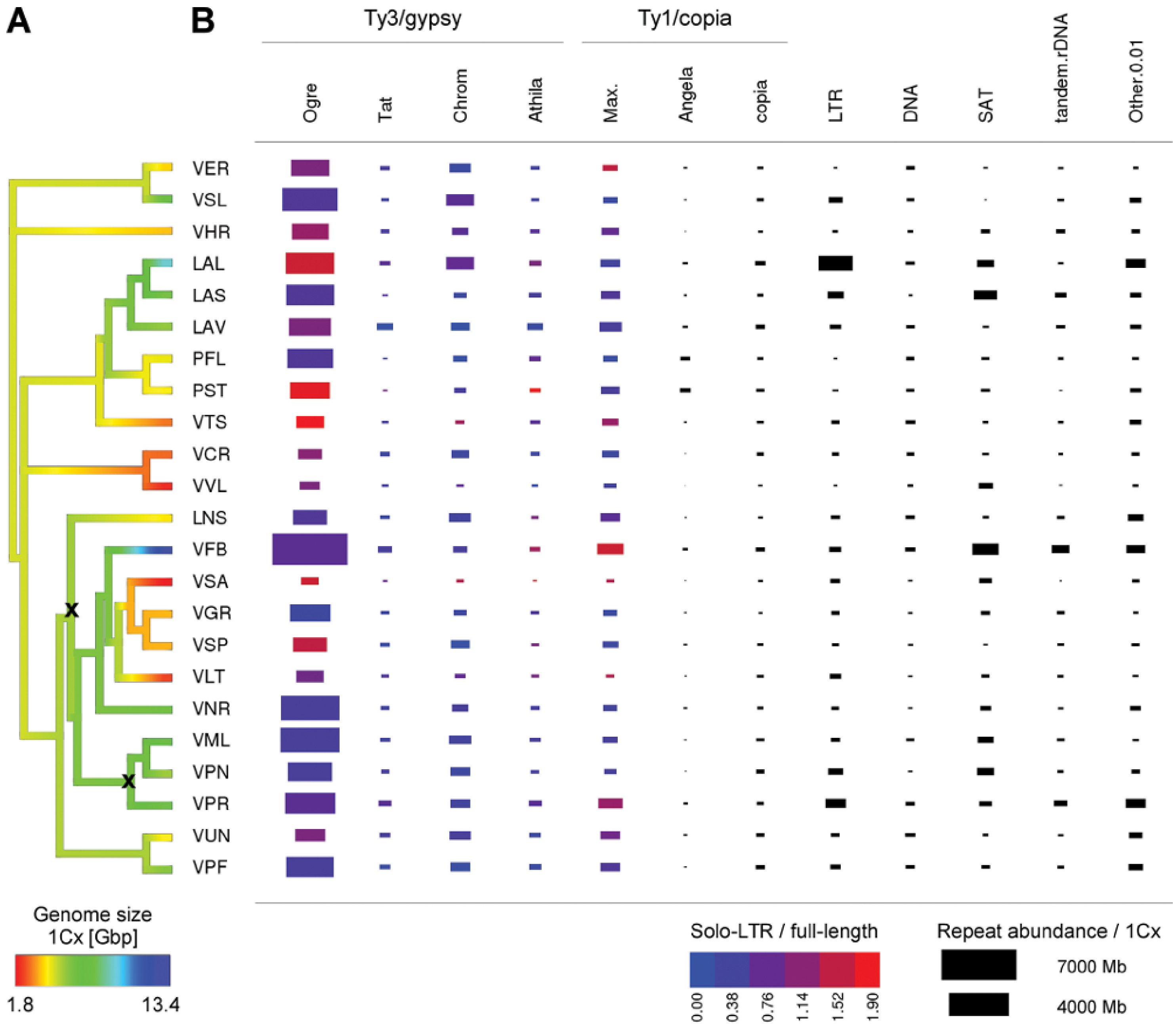
- comparative clustering – equal read numbers of genome coverages ?
  - problems with species with big variations in genome sizes
  - problems when analyzing large numbers of samples

- two-step approach

  1./ perform repeat analysis in each species separately

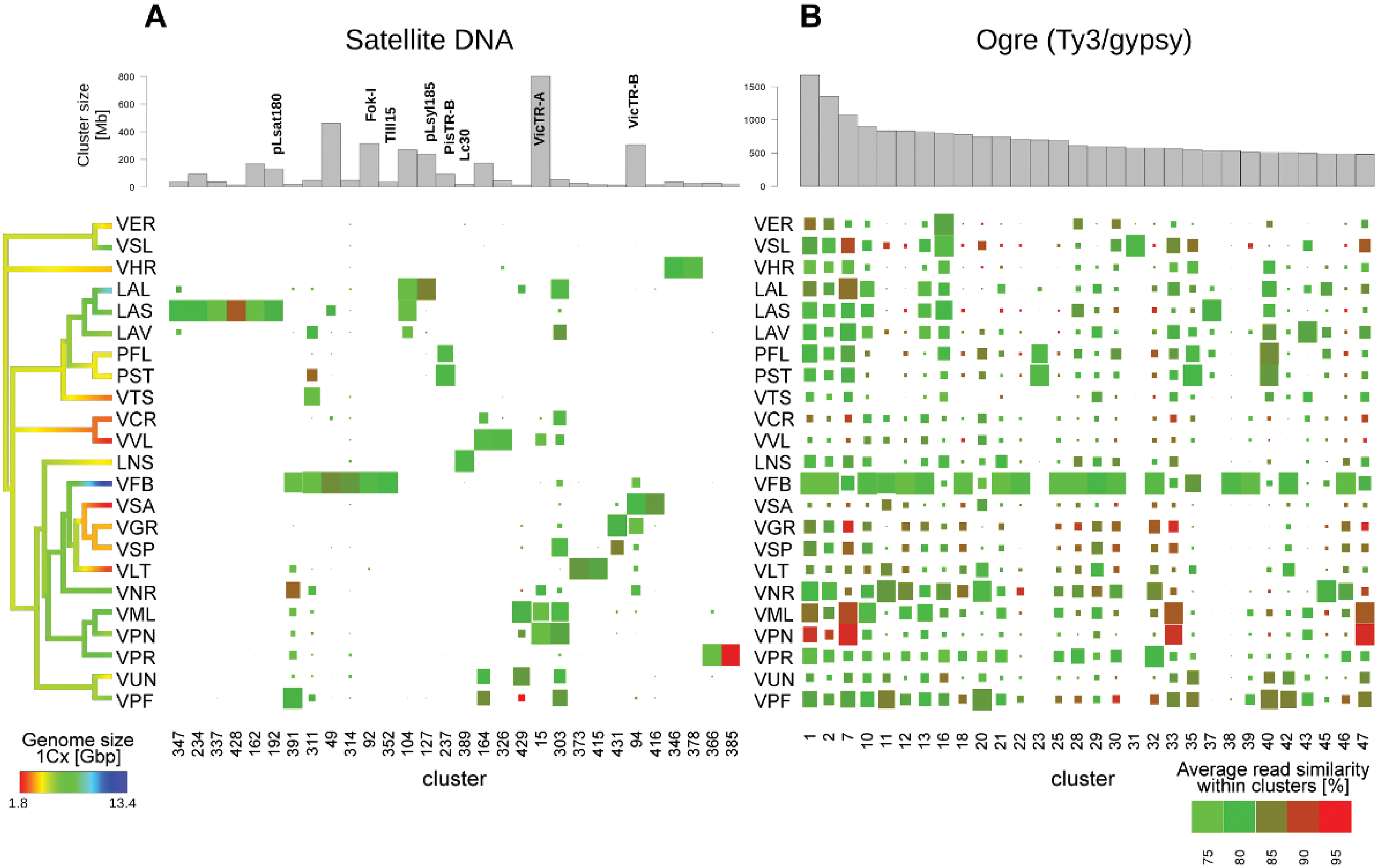
  2./ comparative clustering with reads sampled from (1) – finding “orthologous” repeats

Comparative study of repeats in 23 species of *Fabeae*



Macas et al. (2015) PLoS ONE 10(11): e0143424

Comparative study of repeats in 23 species of *Fabeae*



Macas et al. (2015) PLoS ONE 10(11): e0143424