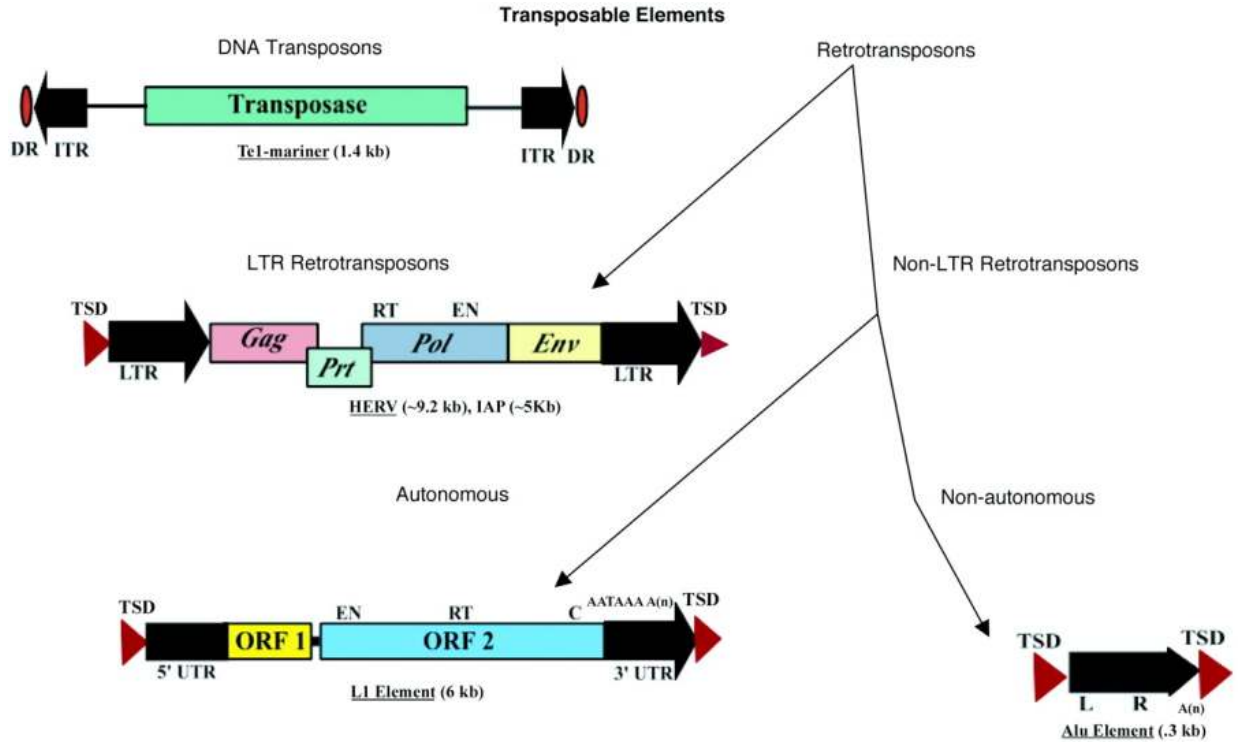
Repeat classification in mammalian genomes

Dmitrii Meleshko, Oleg Yasnev
Advisor: Son Pham
Joint collaborators: Benedict Paten (UCSC),
Thomas Keane (Welcome Trust Sanger Institute)

Big goal

- Human genomes have huge number of repetitive elements.
- Most of their roles are believed to be unknown.
- **Big goal**: characterize and explain the mechanism of repeats in mammalian genomes.

Transposable elements



Target site duplication (TSD)

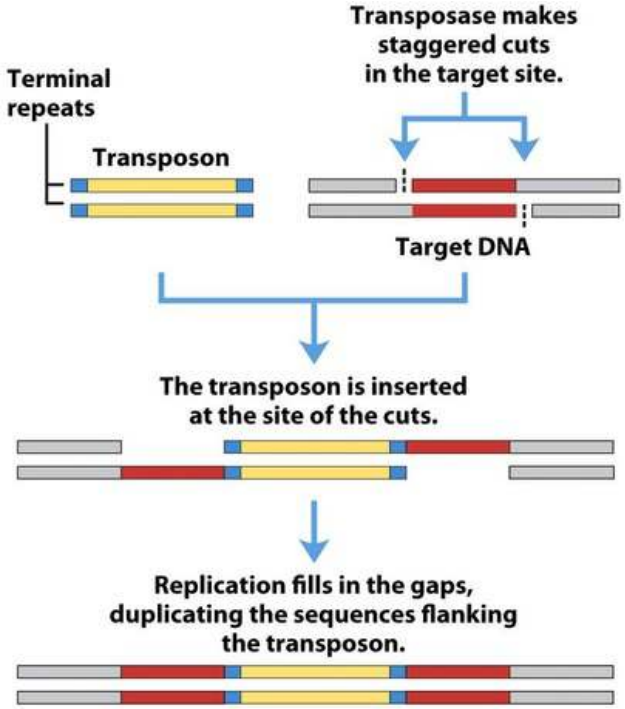
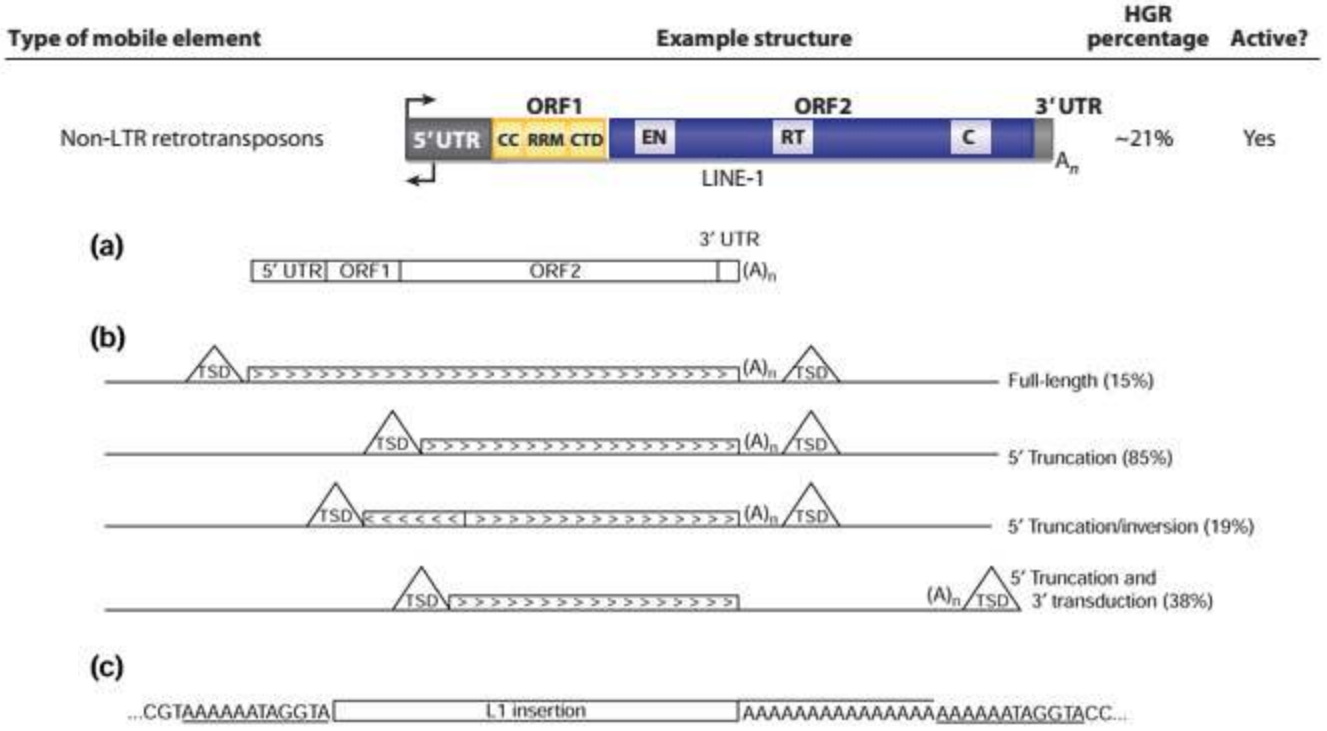


Figure 25-44
Lehninger Principles of Biochemistry, Fifth Edition

L1 – the best TE in the world



First task

Efficient algorithm to find L1-transposition events with additional information about TSD structure.

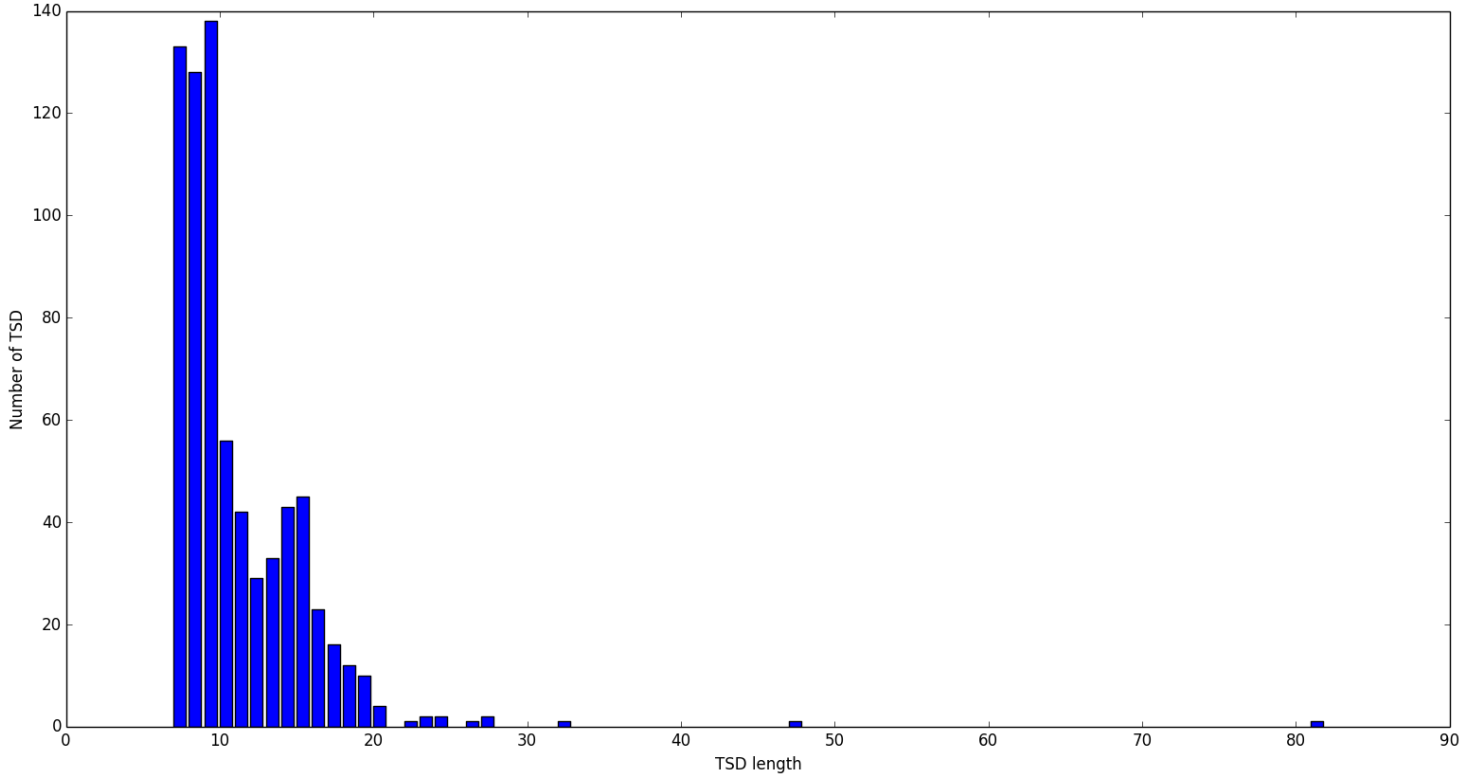
Our approach

1. Use RepeatMasker to find repeats.
2. Use local alignment and heuristic scoring function to find TSD.
3. Try to find some specific properties about TSD.

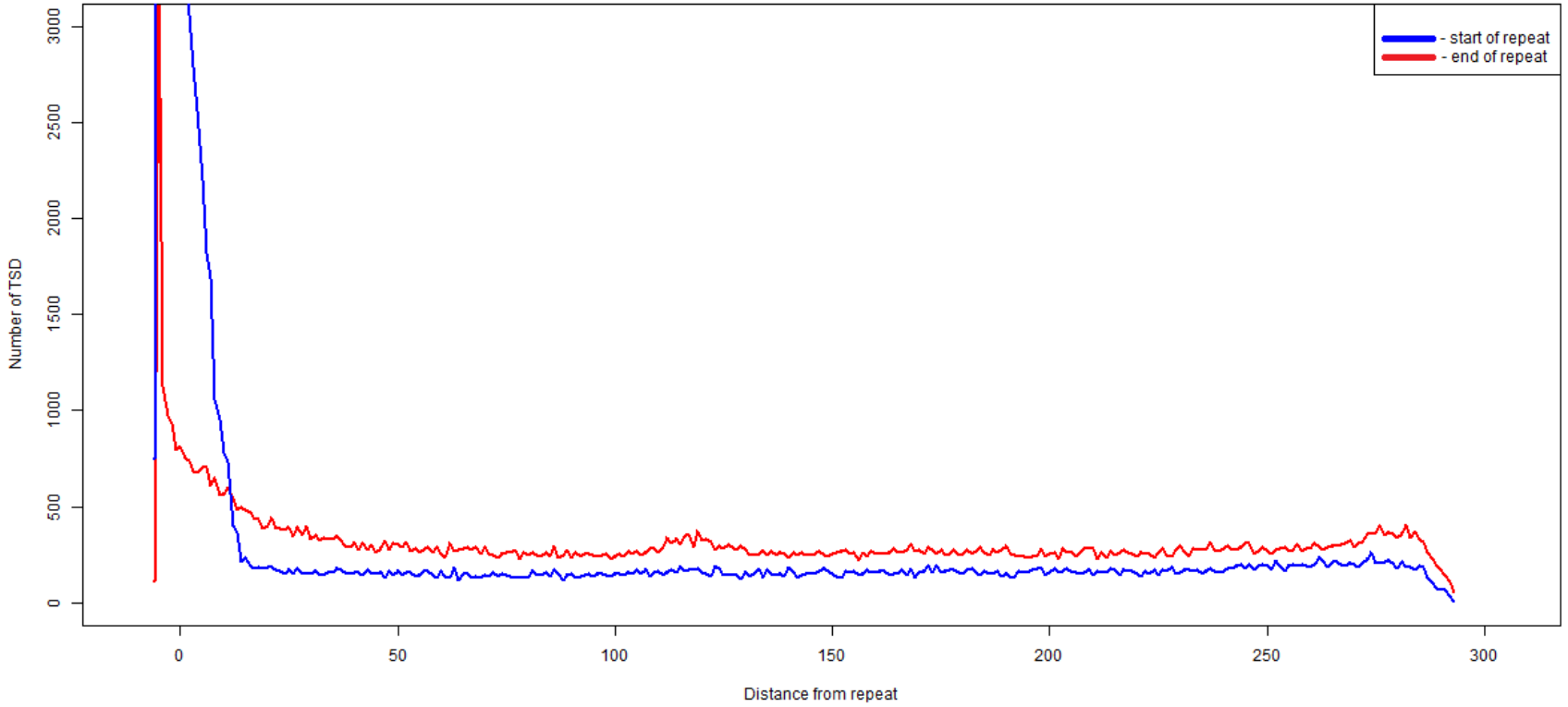
Our results

- Developed reliable algorithm to locate nearly all TSD in genome.
- Gather statistics about different properties of TSD in human genome.
- Developed algorithm to locate poly-A tails of repeats and gather statistics.

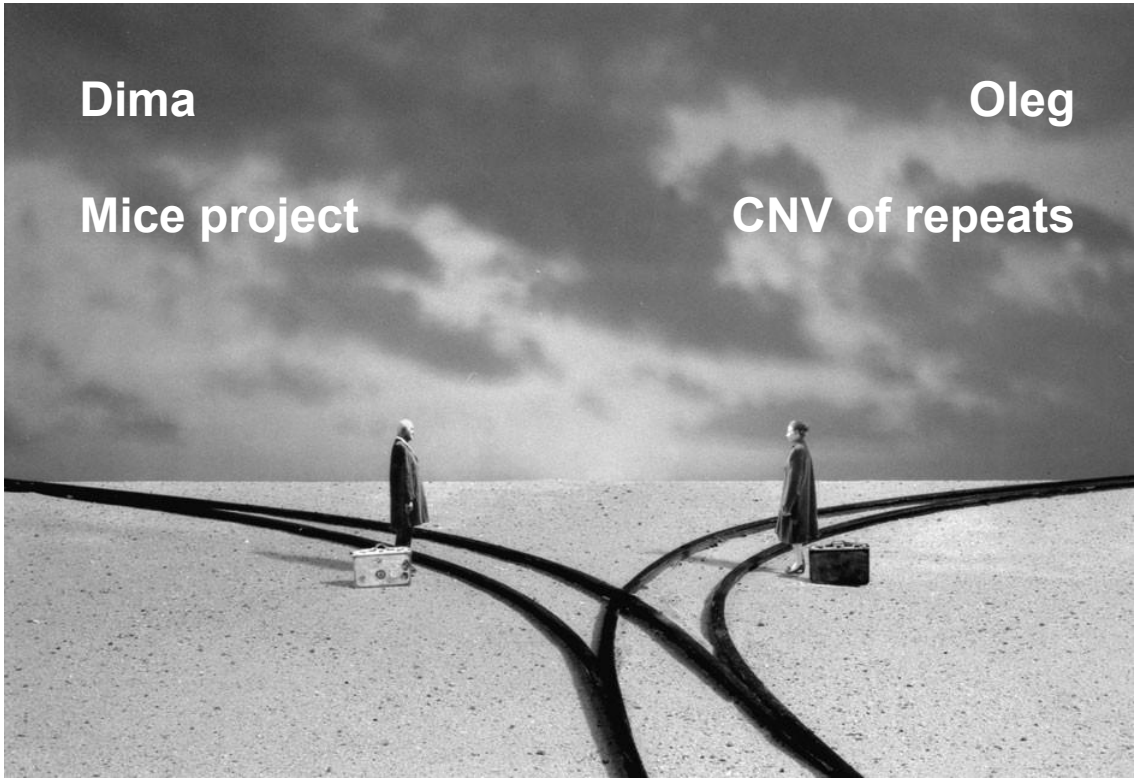
TSD length distribution



TSD distance distribution



And then...



Mouse Genomes Project

Sequencing 18 laboratory mouse strains

- Largest effort to date to sequence genomes of laboratory mouse strains
- 129P2/OlaHsd, 129S1/SvImJ, 129S5SvEvBrd, A/J, AKR/J, BALB/cJ, C3H/HeJ, C57BL/6NJ, CAST/EiJ, CBA/J, DBA/2J, FVB/NJ, LP/J, NOD/ShiLtJ, NZO/HlLtJ, PWK/PhJ, SPRET/EiJ, WSB/EiJ

Phase 1 (2009-2011)

- Deep sequencing of each strain (>25x)
- Illumina GAII (54-108bp reads)
- Comprehensive catalog sequence variation
  - Quantify effects of sequence variation on phenotypes

Phase 2 (2012-2013)

- Draft genome sequence and annotation of each strain
- Laboratory mouse pan genome
- Strain specific gene prediction/annotation
- Reference-free representations of multiple mouse genomes

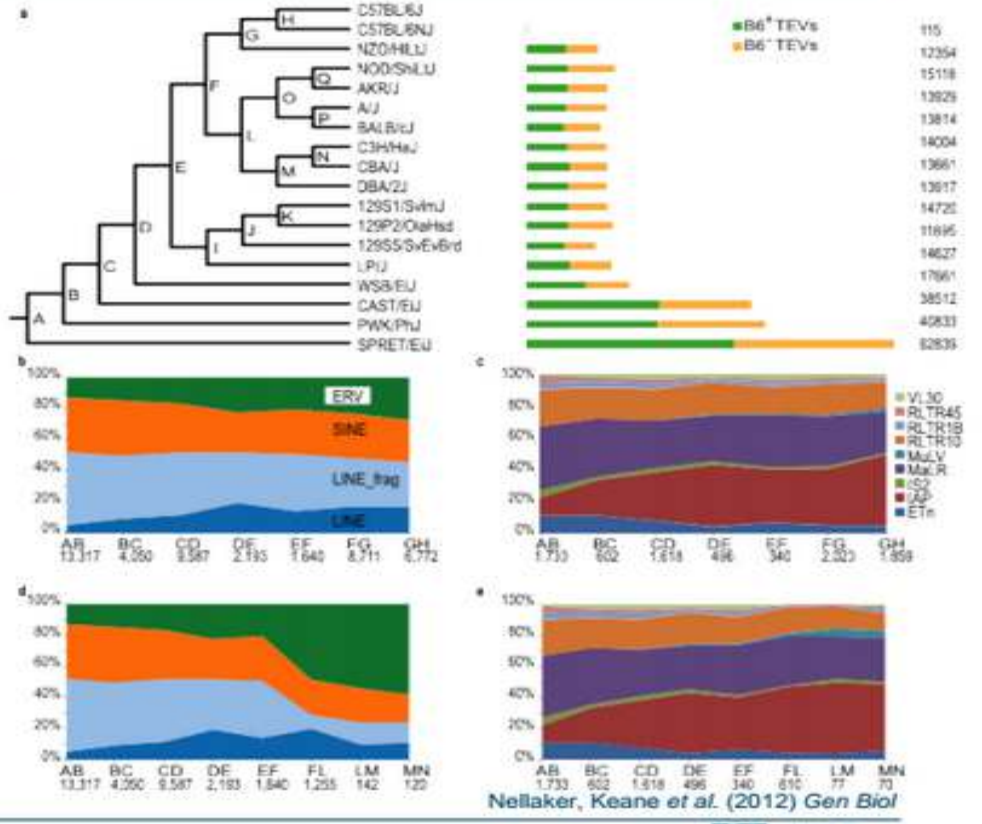
~100k Transposable Element Variants

Transposable elements (TE)

- Mobile DNA elements
- 38-69% of genomic sequence
- Can modulate gene formation, function and regulation

Three Distinct classes

- Short interspersed nuclear elements (SINEs) ~28K
- Long interspersed nuclear elements (LINEs) ~40K
- Endogenous retroviruses (ERV) ~34.7K



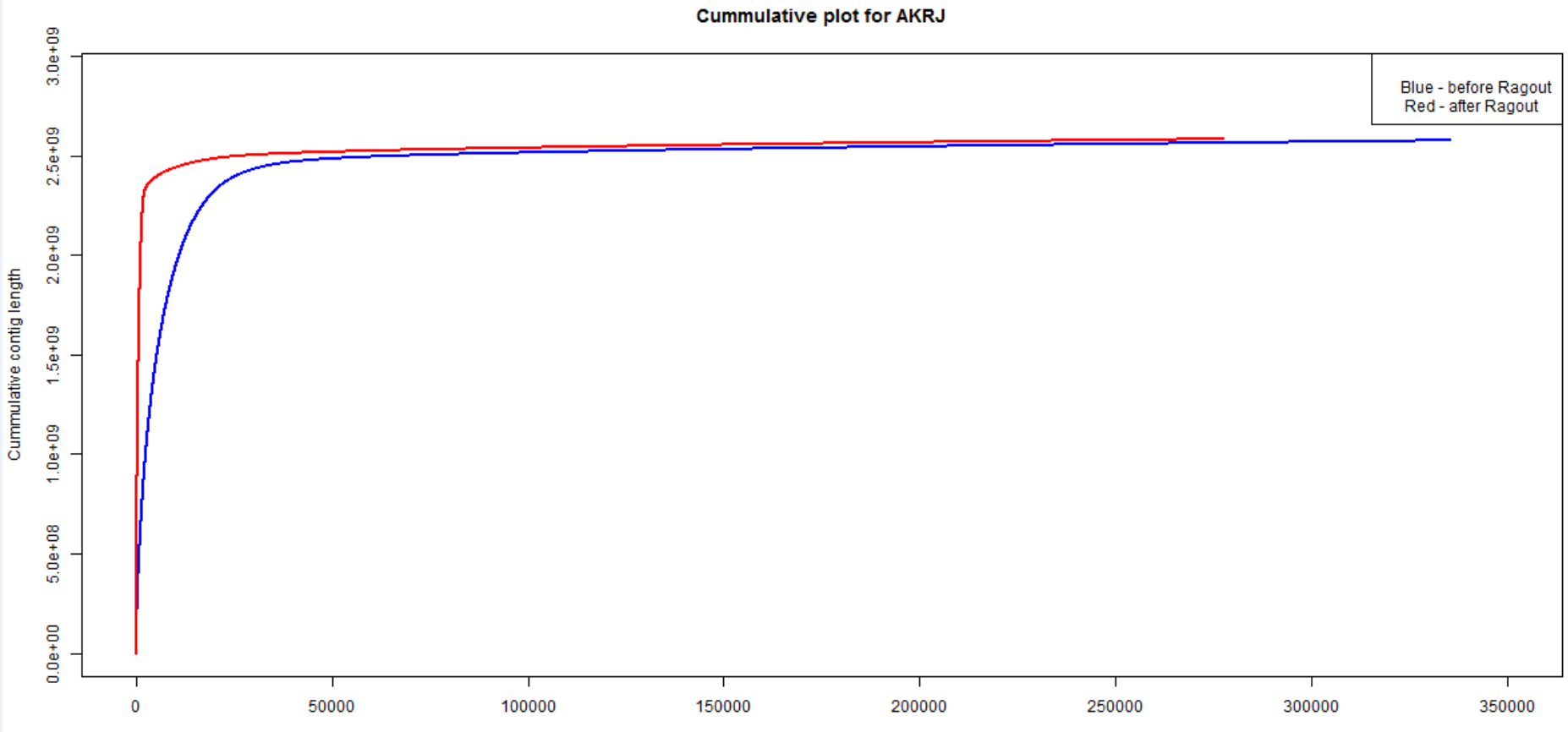
Nellaker, Keane *et al.* (2012) *Gen Biol*

sanger

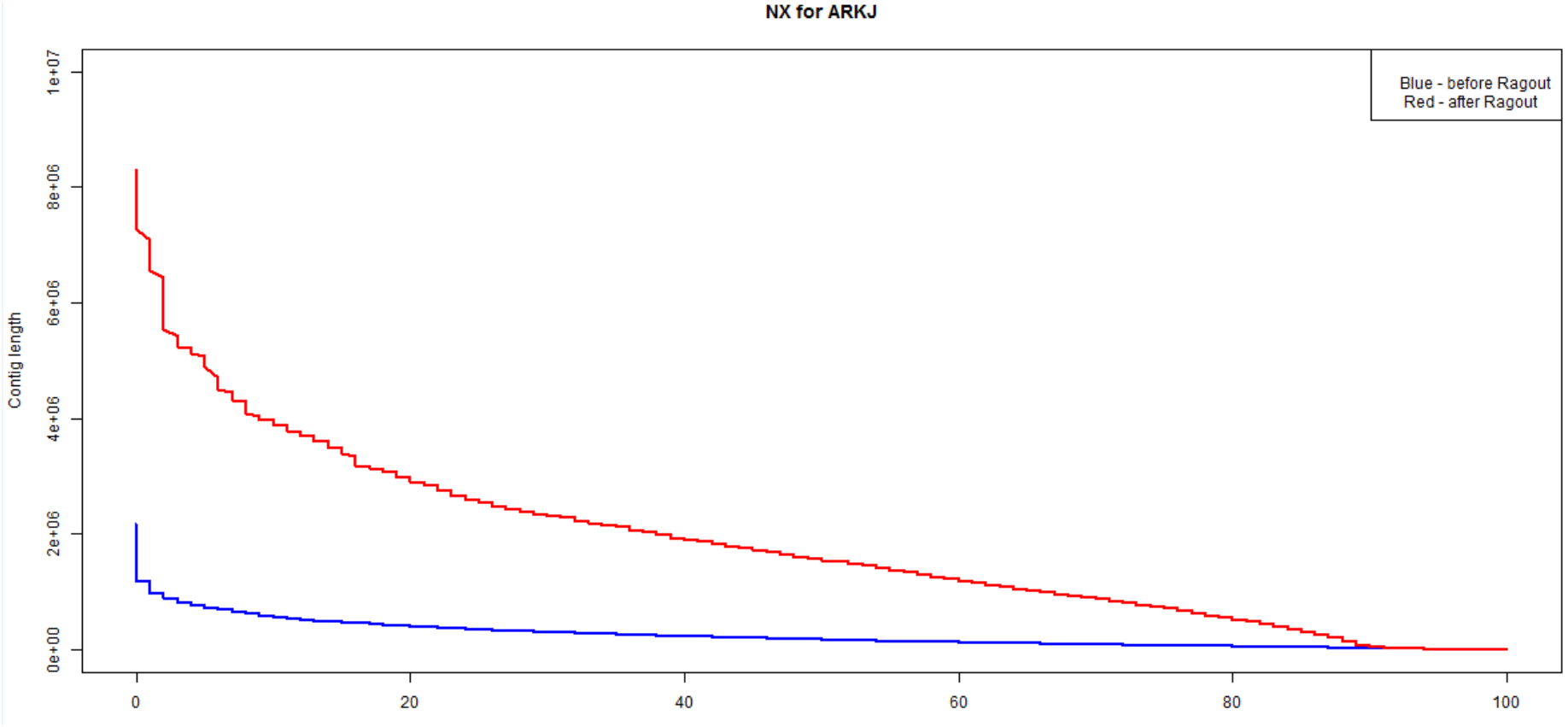
Mice project

- We use Ragout to improve mice assembly.
- I’ve implemented fast overlap graph construction.
- Now it is possible to operate with large genomes more efficiently.
- I also count a lot of statistics for Ragout.

Cummulative - AKRJ



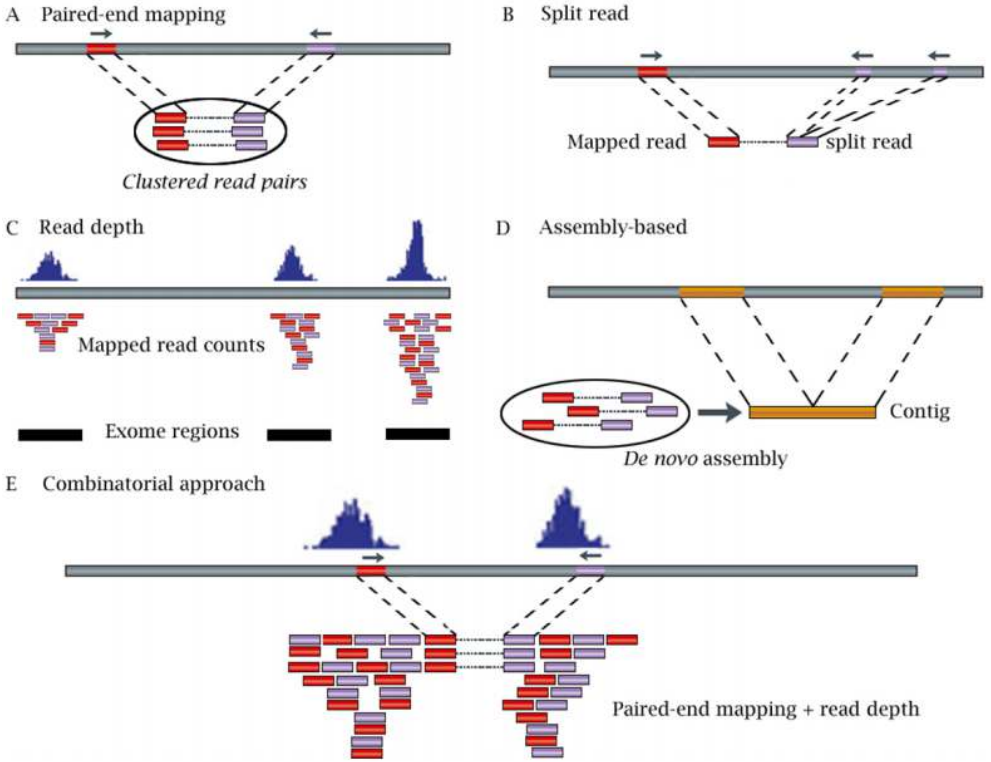
Nx - ARKJ


NX for ARKJ
Blue - before Ragout
Red - after Ragout
Contig length
0e+00
2e+06
4e+06
6e+06
8e+06
1e+07
0
20
40
60
80
100

CNV of repeats

Approaches:

1. Pair-end mapping
2. Split read
3. Read depth
4. Assembly-based
5. Combinatorial



Task

- All of the approaches compare a particular genome with reference.

- Our task: we have two sets of reads from two very resembling genomes.
  Can we find repeats movement?

Current progress

- Reading articles about CNV detection of repeats.
- Coming up with ideas of how to solve our task.
- Given reads of twin pairs: one has disease, the other does not.
  Try to find TSD/new insertion of retrotransposon.

Thank you!


How I feel about my research
One minute after creation:
I'm... A GENIUS
One day later...
One week later...
ONE MONTH LATER