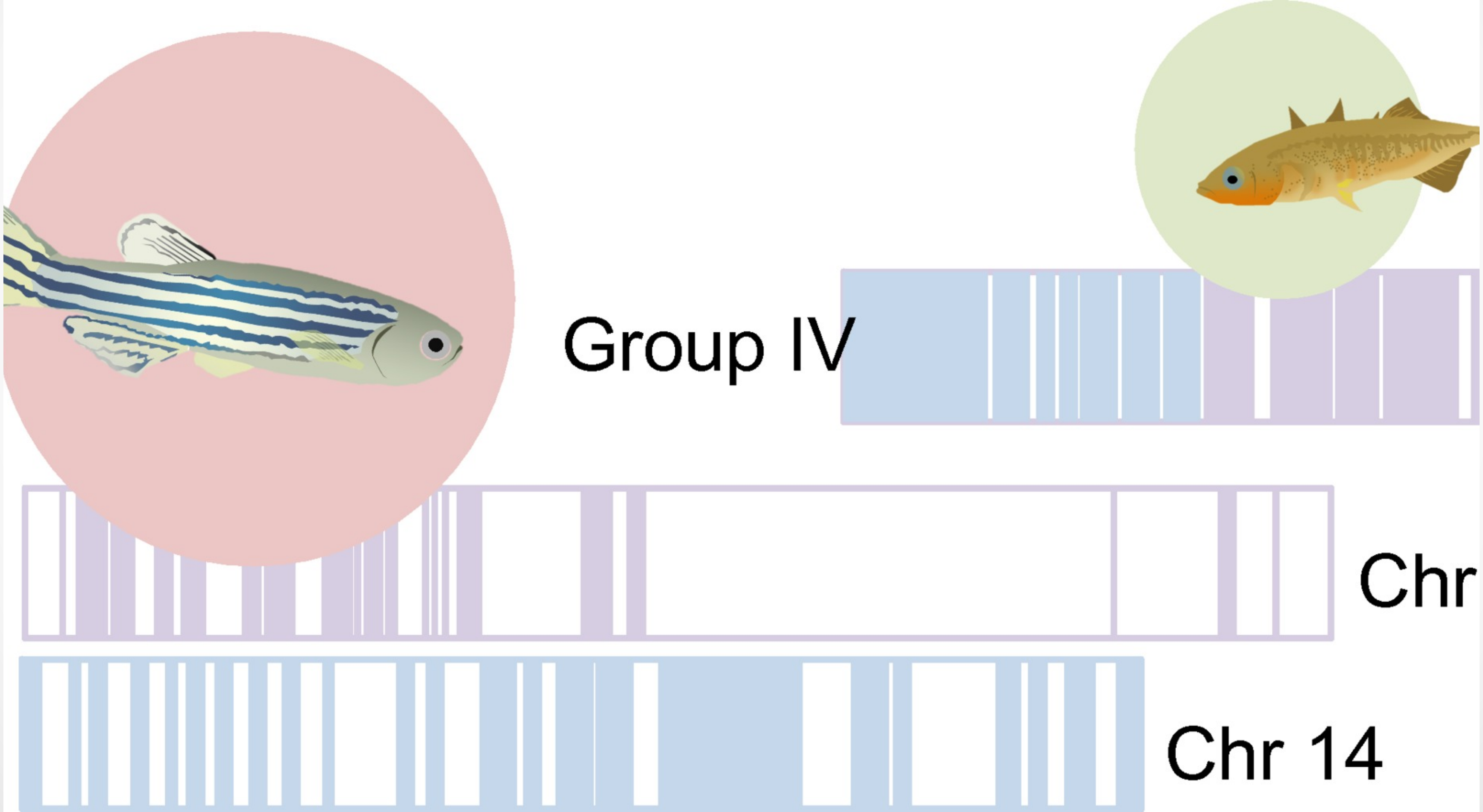
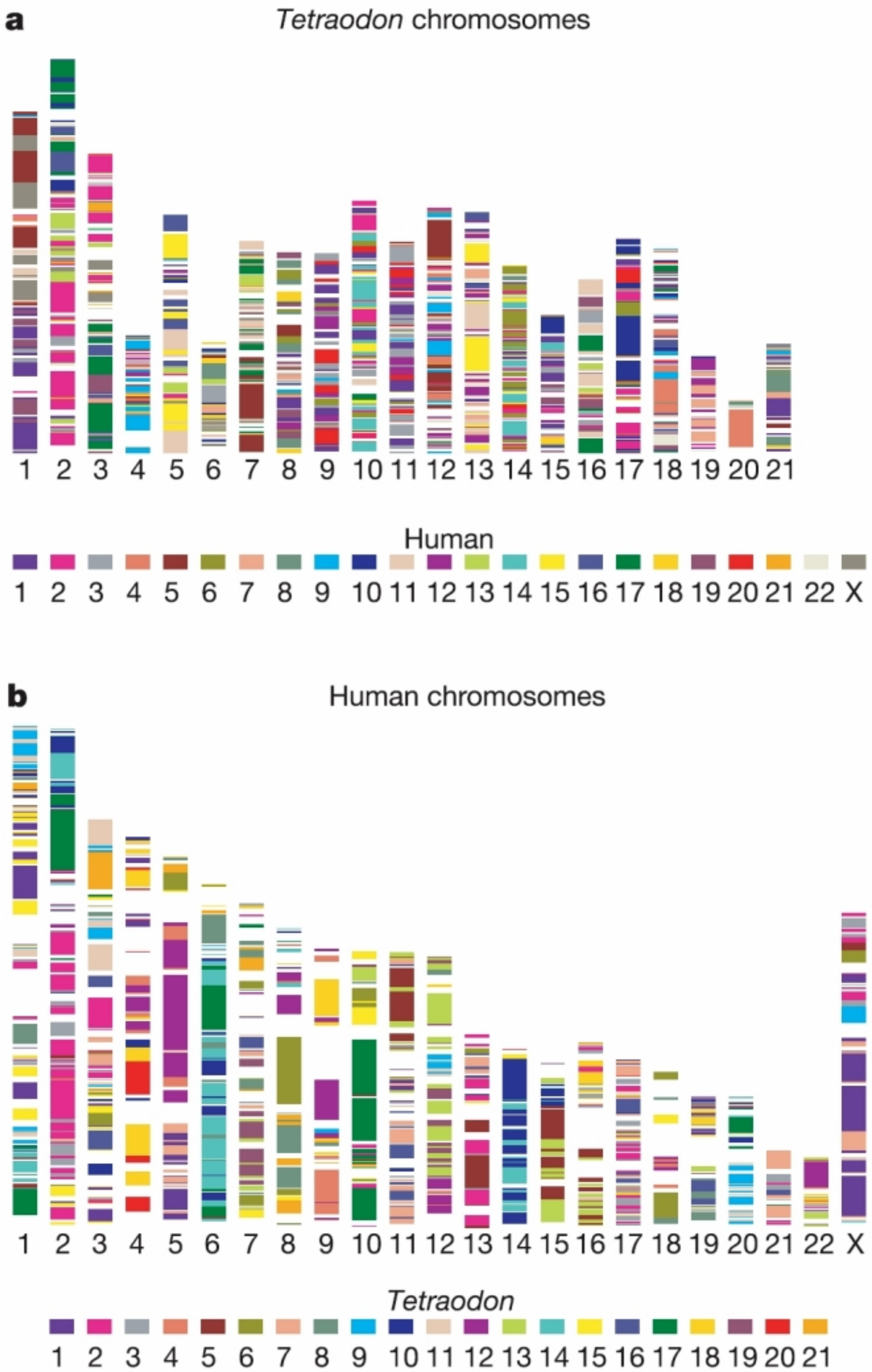
Comparing genomes

A universal genomic coordinate translator for comparative genomics

Inspiration from SocBiN



Architecture and evolution of Neopterygii genomes
Görel Sundström, Neda Zamini and Manfred Grabherr

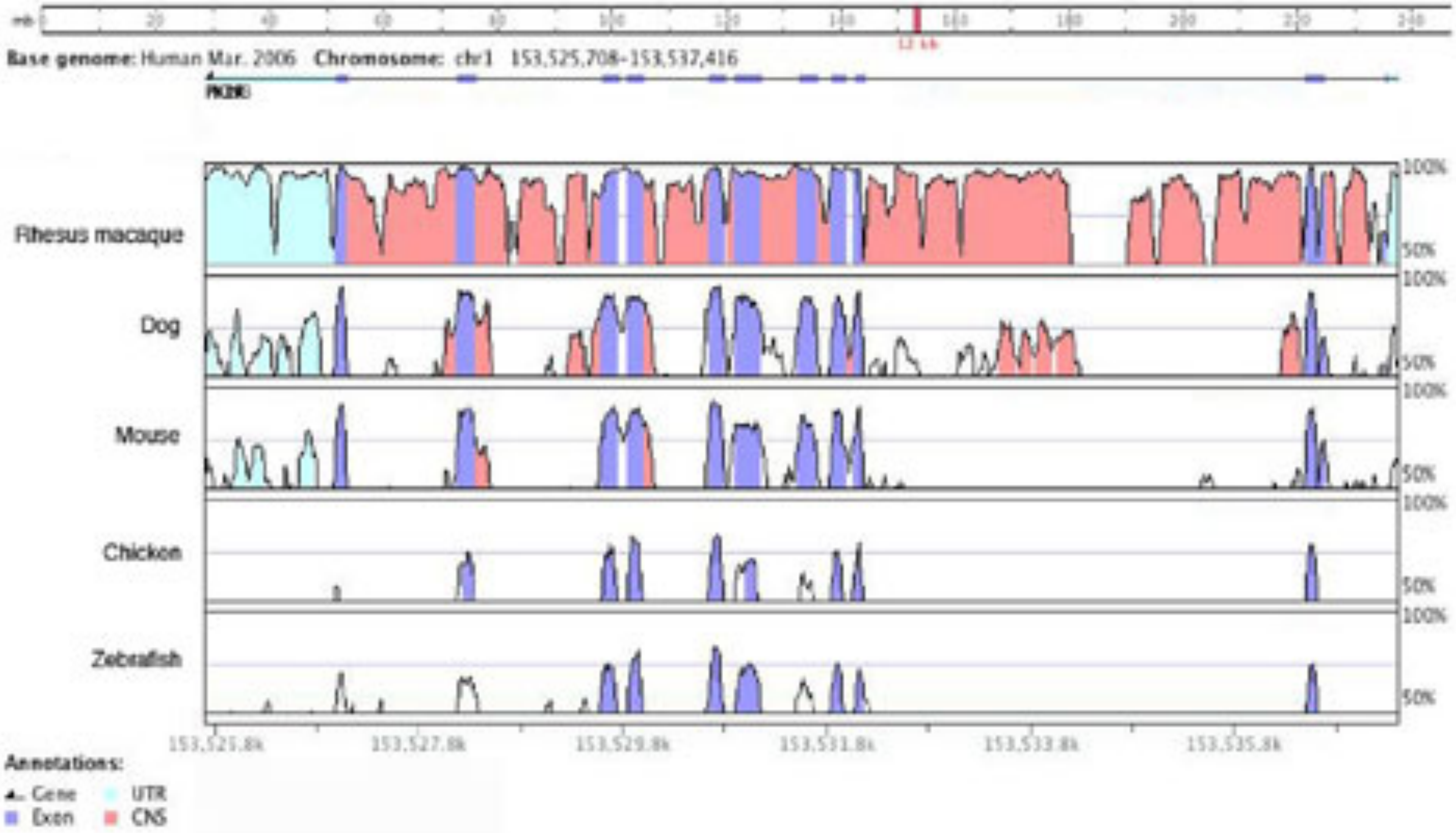


Synteny

- Conservation of blocks of order between sets of chromosomes.
- Conserved synteny - locally conserved order and orientation of features.

Jaillon, O. et al. Genome duplication in the teleost fish Tetraodon nigroviridis reveals the early vertebrate proto-karyotype. Nature 431, 951 (2004).

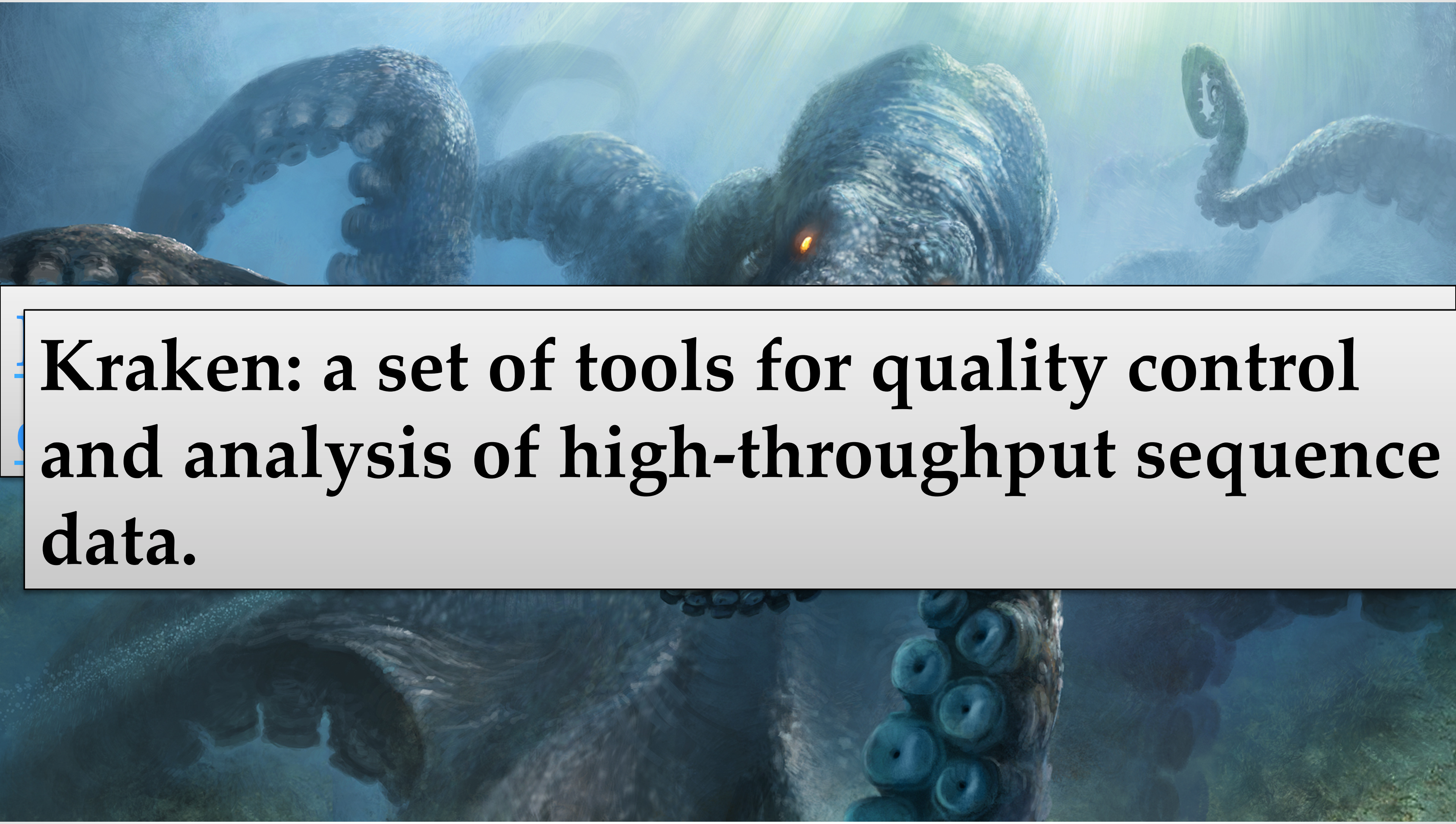
Comparative genomics



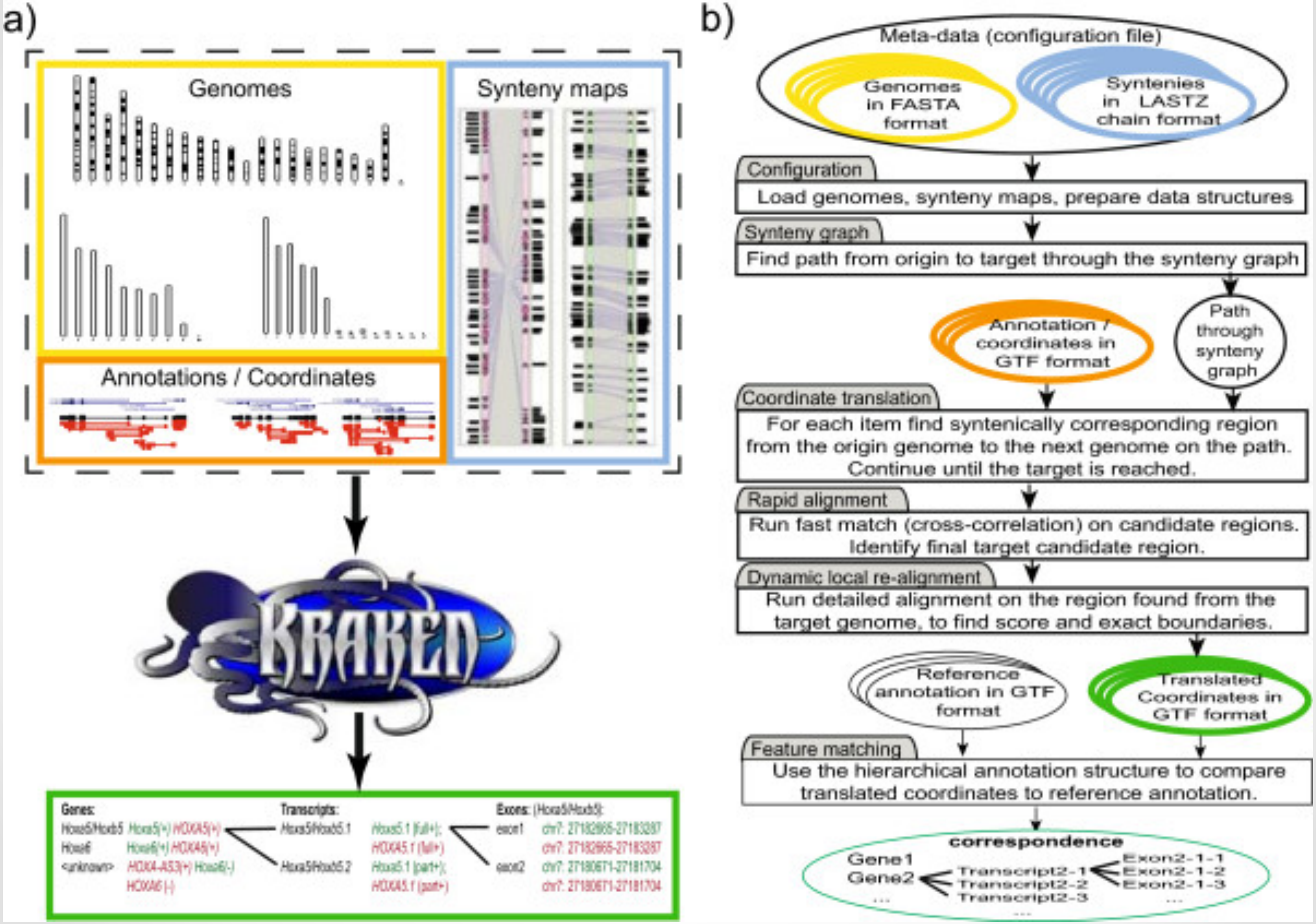
Synteny alignments

- Relative to one central genome
- Complete set of pairwise comparisons (computational time for N genomes is O*N^2)

- LASTZ: The reference genome is aligned with all others with LASTZ.
- MUMmer is a system for rapidly aligning entire genomes, whether in complete or draft form.
- Satsuma finds sequence matches through cross-correlation like comparing audio signals.



Kraken: a set of tools for quality control and analysis of high-throughput sequence data.


a)
Genomes
Synteny maps
Annotations / Coordinates
KRAKEN
Genes:
Hoxa5/Hoxb5 Hoxa5(+) HOXA5(+)
Hoxa6 Hoxa6(+) HOXA6(+)
<unknown> HOXA-AS3(+) Hoxa6(-)
HOXA6 (-)
Transcripts:
Hoxa5/Hoxb5.1 Hoxa5.1 (full+); HOXA5.1 (full+)
Hoxa5/Hoxb5.2 Hoxa5.1 (part+); HOXA5.1 (part+)
Exons: (Hoxa5/Hoxb5):
exon1 chr7: 27182865-27183287
chr7: 27182865-27183287
exon2 chr7: 27180671-27181704
chr7: 27180671-27181704
b)

Meta-data (configuration file)

Genomes in FASTA format

Synteniesin LASTZ chain format

Configuration

Load genomes, synteny maps, prepare data structures

Synteny graph

Find path from origin to target through the synteny graph

Annotation / coordinates in GTF format

Path through synteny graph

Coordinate translation

For each item find syntenically corresponding region from the origin genome to the next genome on the path. Continue until the target is reached.

Rapid alignment

Run fast match (cross-correlation) on candidate regions. Identify final target candidate region.

Dynamic local re-alignment

Run detailed alignment on the region found from the target genome, to find score and exact boundaries.

Reference annotation in GTF format

Translated Coordinates in GTF format

Feature matching

Use the hierarchical annotation structure to compare translated coordinates to reference annotation.

Workflow

1. Estimate candidate locations of orthologous coordinates through the synteny graph
2. Alignment of the input sequence against the target region based on cross-correlation algorithm
3. Compute local alignment to determine exact target coordinates.
4. Coordinates in target coordinates are compared against the reference GTF (optional)

Configuration file

- File locations of the genomes (FASTA)
- Synteny maps (LASTZ and Satsuma format) and which genomes in what direction are connected.
- Pairwise synteny maps in one direction (genome A to genome B).

Synteny graph

- Exhaustive search through all possibilities from source to target genomes.

- Selects a path:
  - with lowest number of indirect mappings

  or

  - By minimizing the accumulated genomic distances

Coordinate translation

- Each interval in the source GTF is translated individually

- Translated target coordinates on the same chromosome or scaffold with syntenic flanks in consistent orientation of up to 100,000 nt are passed to next step
- Otherwise the region is split into two target intervals for searching the boundaries of syntenic breaks.

Rapid alignment

- Quick search of source sequence against the target interval using approximate cross-correlation alignment.
- The target interval is broken into blocks of 2^{14} nt and the block with highest cross-correlation signal is computed.
- Kraken determines a candidate region the size of the source interval plus flanks on each end.

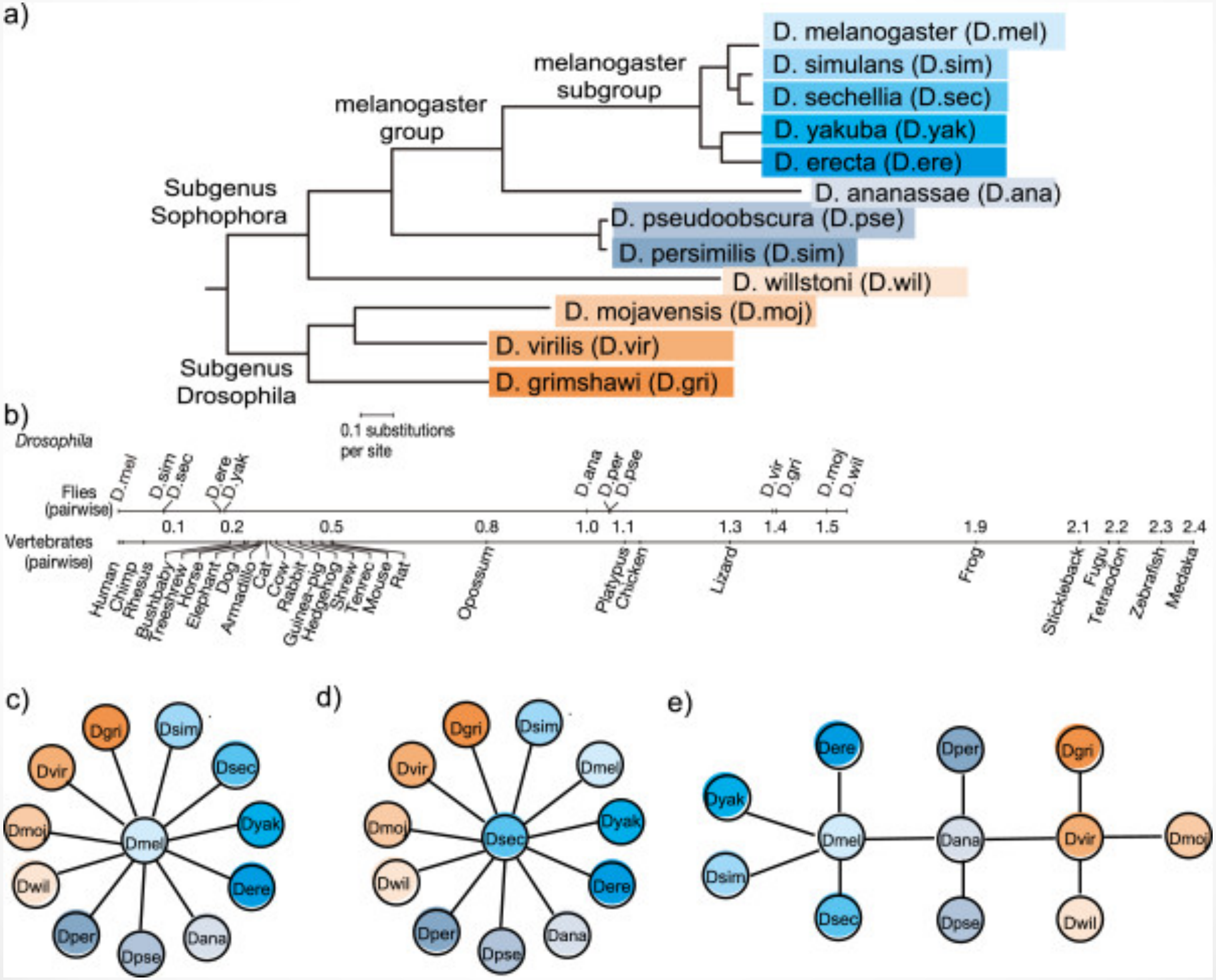
Dynamic local re-alignment

- Detailed alignment of the source with Cola against the target subsequence defined by the rapid alignment.
- For source intervals >100nt the sequence is split into two 100nt chunks covering the start and end regions.
- For all the items that were successfully translated an output entry is produced containing the translated coordinates in GTF format.

Feature matching

- GTF coordinates are stored as exons, transcripts or loci.
- Inferring the spatial relationship of genomic features taking into account the multi-exonic structures.
- Matches are classified as full sense overlap, partial sense overlap, intronic or antisense.
- Coordinates of translated features, relationships and overlapping target annotations are reported in human-readable outputs that are also parsing-friendly.

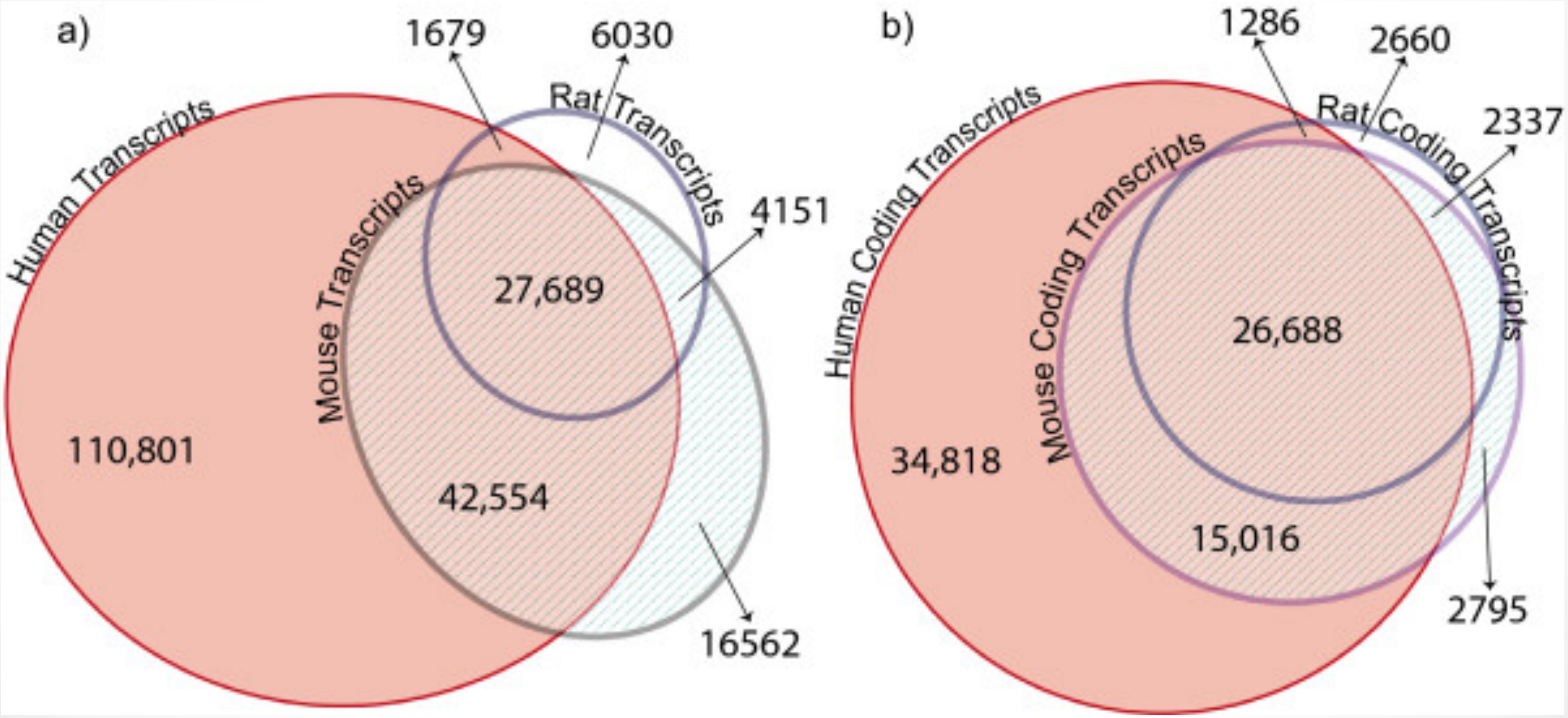
Evaluating synteny graphs



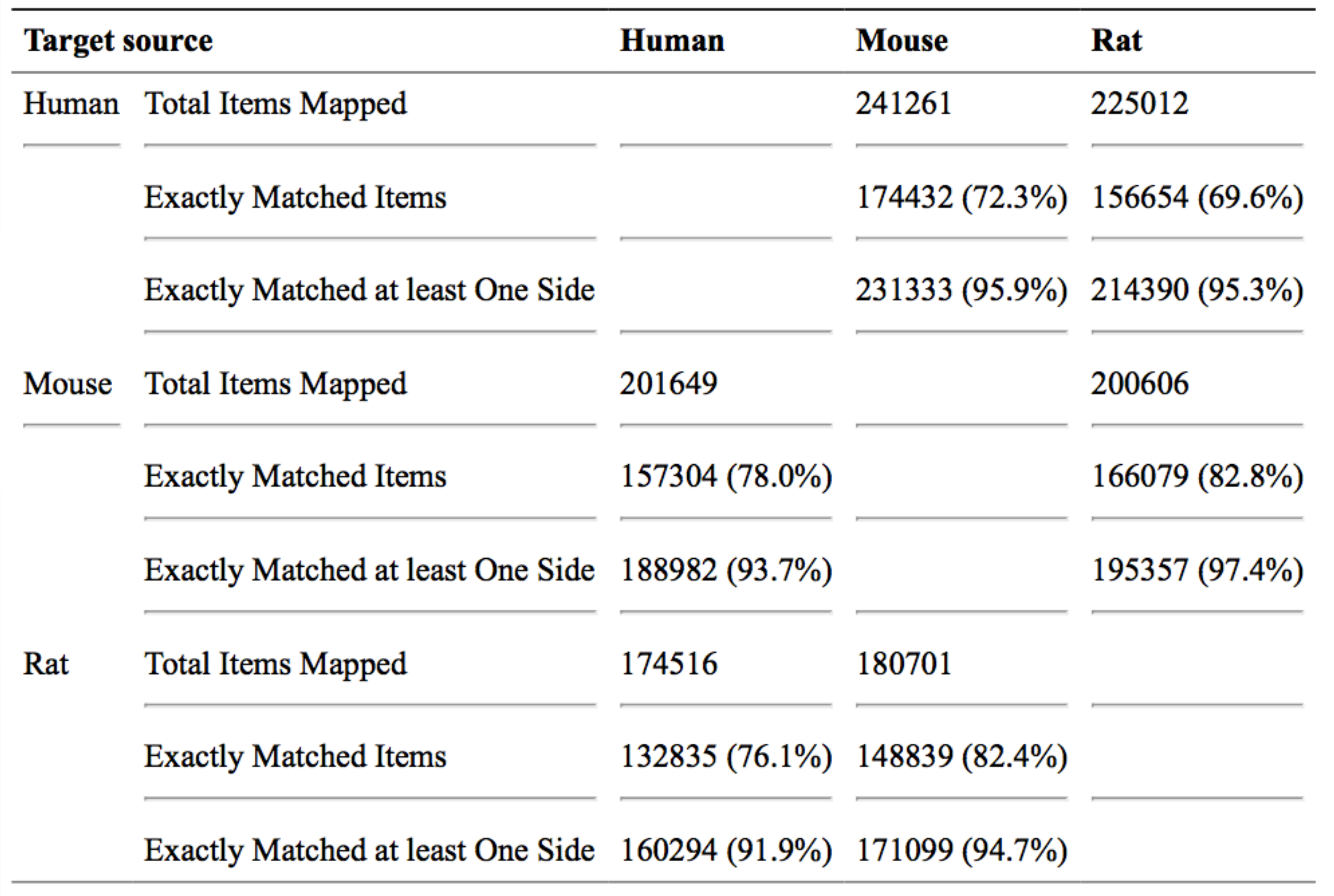
Comparison of direct and indirect pairwise coordinate translations

Source - target	Direct mapped	Direct mappings matched by indirect translations		
		Melanogaster star configuration	*Sechellia* star configuration	Clade center configuration
D.ana - D.ere	43%	99.0%	98.3%	99.0%
D.ana – D.gri	16%	96.3%	89.9%	93.5%
D.ere – D.sim	76%	98.4%	98.3%	98.4%
D.moj – D-per	13%	94.4%	88.1%	88.8%
D.pse – D.sim	27%	97.7%	97.0%	96.6%
Median		97.4%	93.6%	91.7%

Mathcing genes between human, rat and mouse

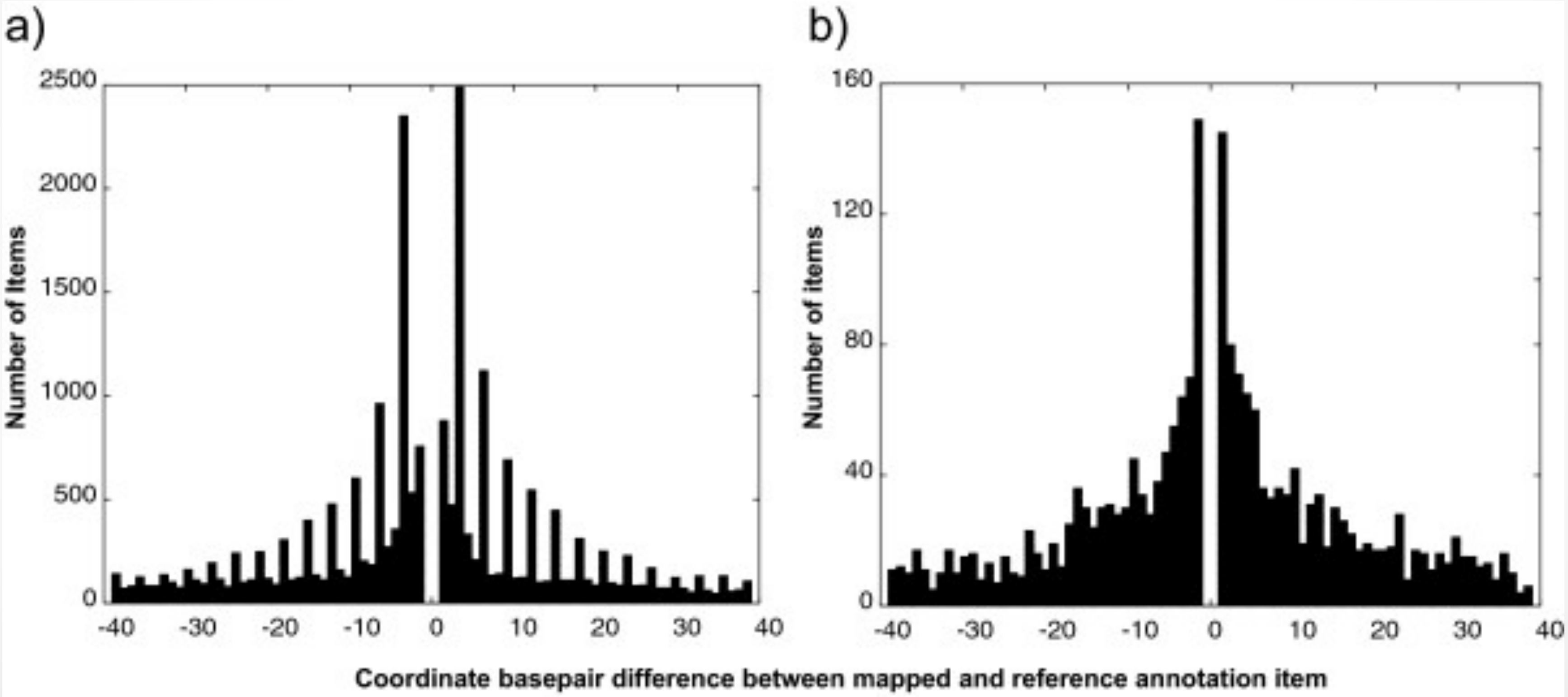


Accuracy on nucleotide level



Target source		Human	Mouse	Rat
Human	Total Items Mapped		241261	225012
	Exactly Matched Items		174432 (72.3%)	156654 (69.6%)
	Exactly Matched at least One Side		231333 (95.9%)	214390 (95.3%)
Mouse	Total Items Mapped	201649		200606
	Exactly Matched Items	157304 (78.0%)		166079 (82.8%)
	Exactly Matched at least One Side	188982 (93.7%)		195357 (97.4%)
Rat	Total Items Mapped	174516	180701	
	Exactly Matched Items	132835 (76.1%)	148839 (82.4%)	
	Exactly Matched at least One Side	160294 (91.9%)	171099 (94.7%)	

Histogram of nucleotide differences

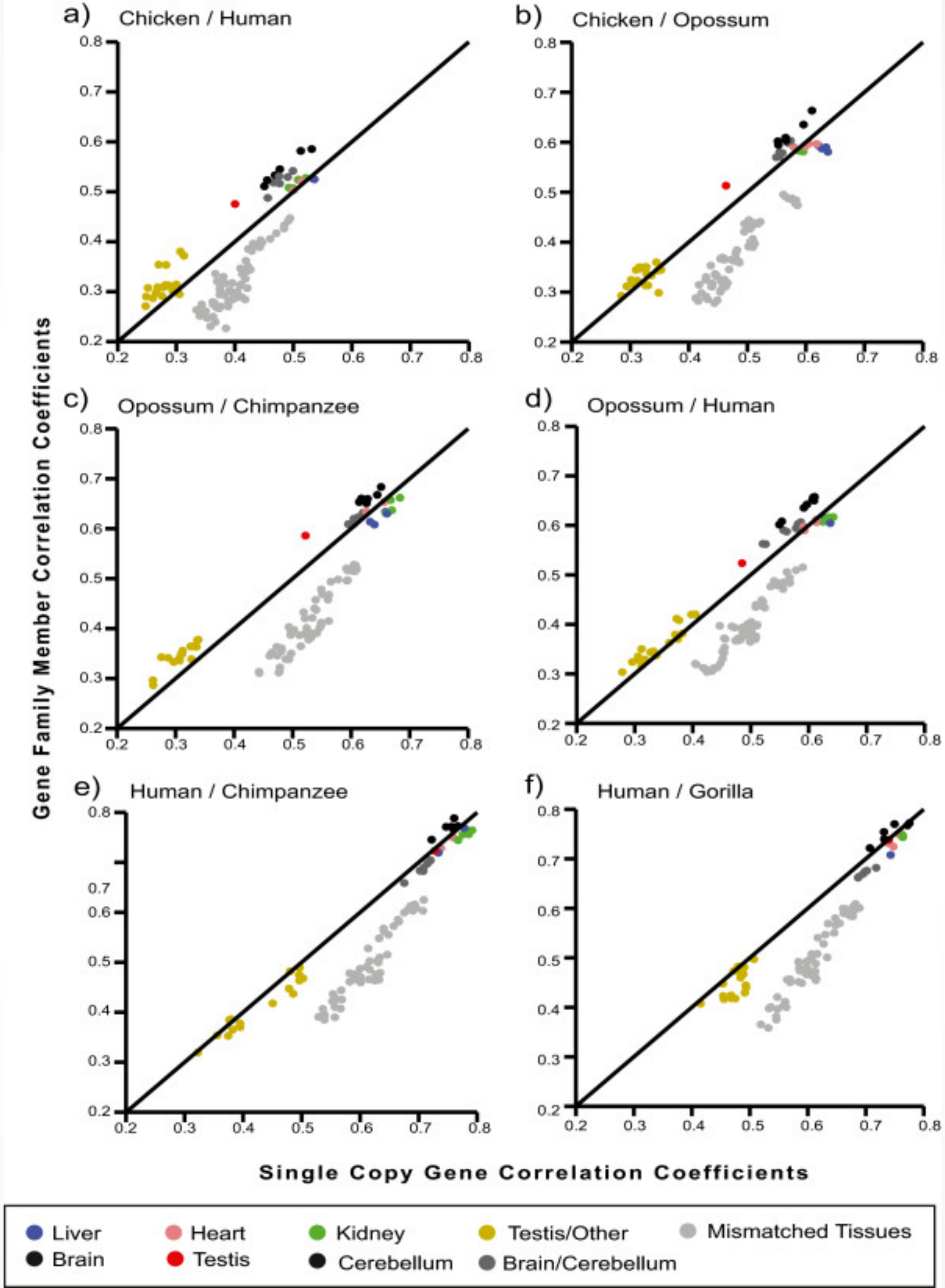


Coding genes show periodicity of 3 nt

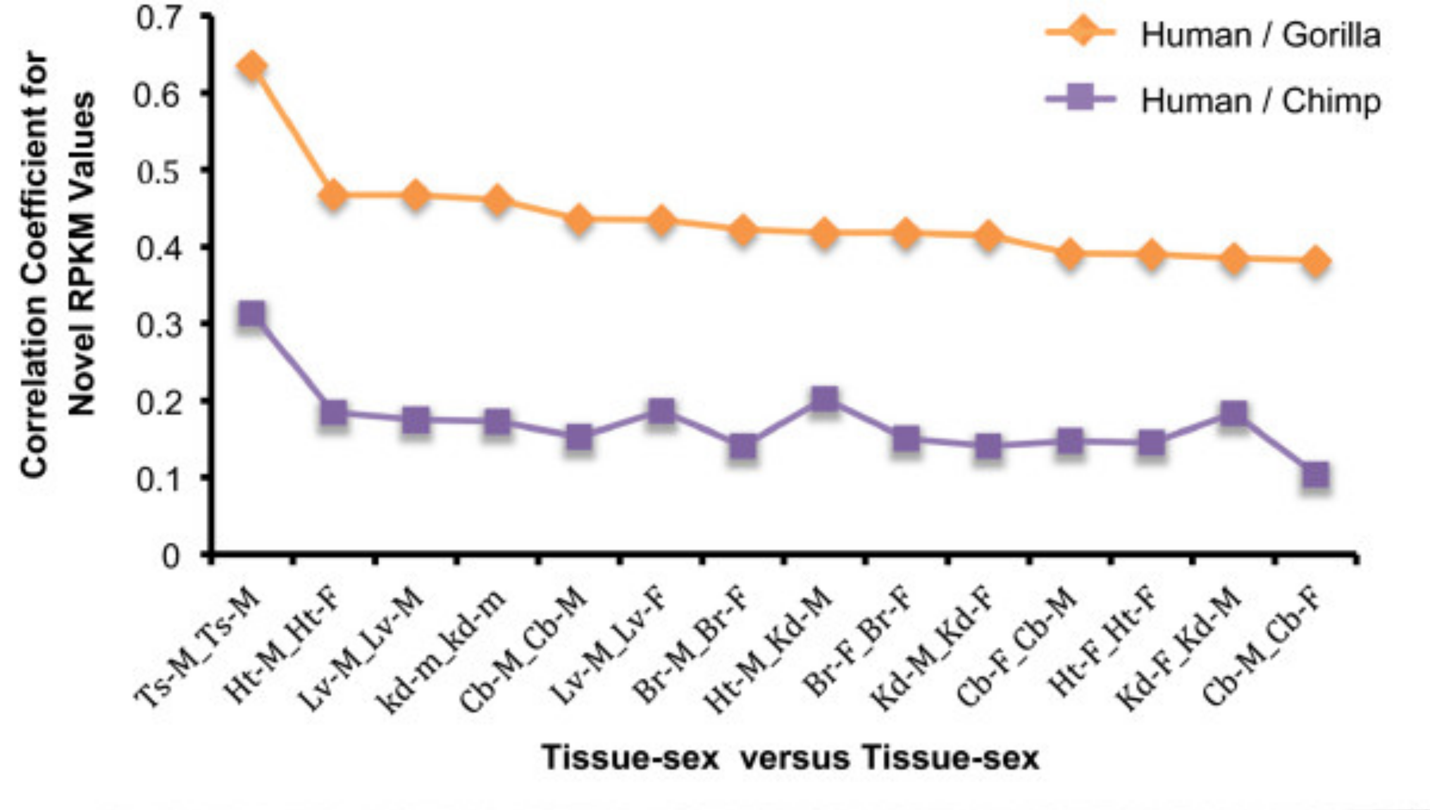
Non-coding RNAs don't have periodicity

Analyzing large RNA-seq data set

- Translating features of human, chimp, gorilla, opossum and chicken RNA-seq reads.
- 250 000 reads translated with Kraken human-chimp
- 100 000 reads translated with Kraken chicken-human


a)
Chicken / Human
b)
Chicken / Opossum
c)
Opossum / Chimpanzee
d)
Opossum / Human
e)
Human / Chimpanzee
f)
Human / Gorilla
0.2
0.3
0.4
0.5
0.6
0.7
0.8
Gene Family Member Correlation Coefficients
Single Copy Gene Correlation Coefficients
Liver
Brain
Heart
Testis
Kidney
Cerebellum
Testis/Other
Brain/Cerebellum
Mismatched Tissues

Correlation of un-annotated transcribed features



Ts : Testis Ht : Heart Lv: Liver Cb: Cerebellum Br: Brain Kd: Kidney
M: Male F: Female

Conclusions

- Indirect translation with synteny graphs scales linearly with the N of genomes analyzed. Marginal cost in sensitivity gives substantial gain in computational efficiency.
- Mapping orthologous sequences is highly accurate in predicting the precise boundaries of genomic features. Kraken can be used to create annotations through orthology.
- Analysis of RNA-seq data from 6 species and 8 tissues each was done in a few hours.

Future

- Authors expect Kraken to reduce computational analysis time for future large-scale comparative studies
- For a newly sequenced mammal genome generating synteny map for only one other mammal gives the possibility to compare it with dozens of others.