Chromosome communities in the human pangenome

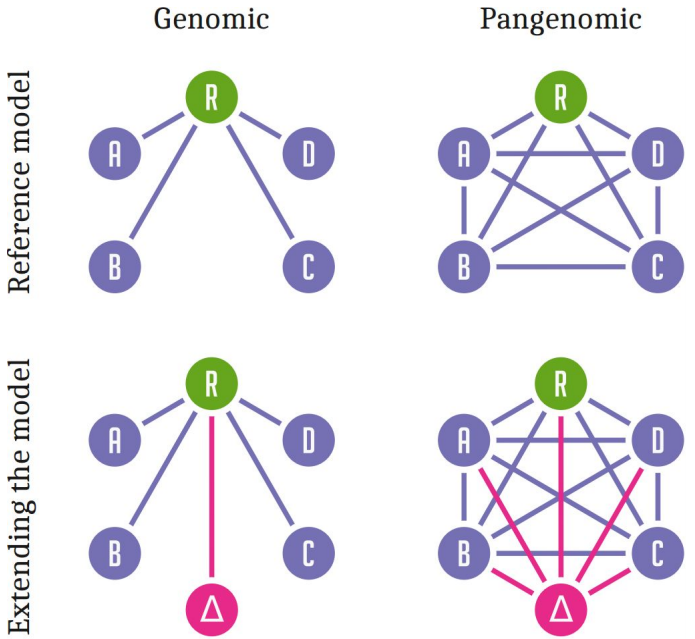
Biodiversity Genomics 2021
From species to ecosystems: Integrated population and speciation genomics
30 September 2021

Andrea Guarracino and Erik Garrison

De novo assembly and pangenomes

Thanks to advances in sequencing technology, new **telomere-to-telomere** genome assemblies are produced at a high rate.

Pangenomes can **model** the full set of genomic elements in a given species or clade, reducing the **reference-bias**.



Δ: new genome; R: reference genome.
Figure from Eizenga et al., 2020.

Pangenome graphs

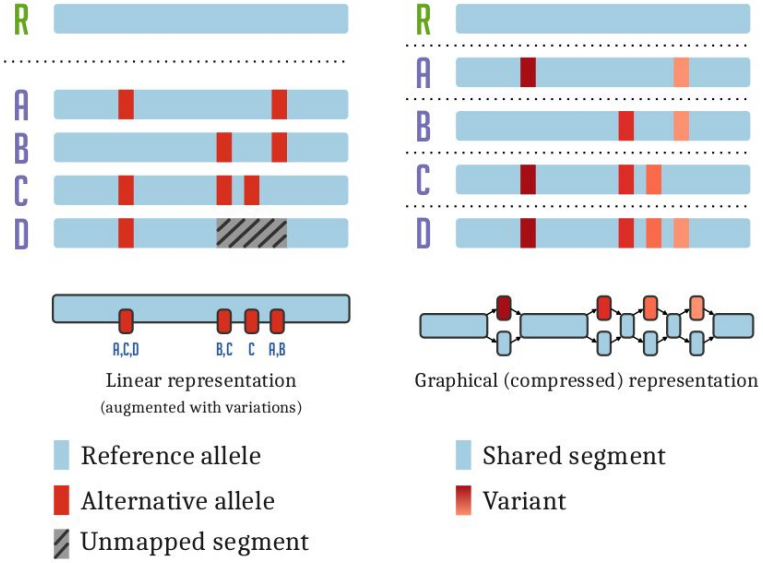


Figure from Eizenga et al., 2020.

Pangenomes can take many forms, including **graph-based** data structures.

Pangenome graphs compress redundant sequences into a smaller data structure that is still representative of the full set.

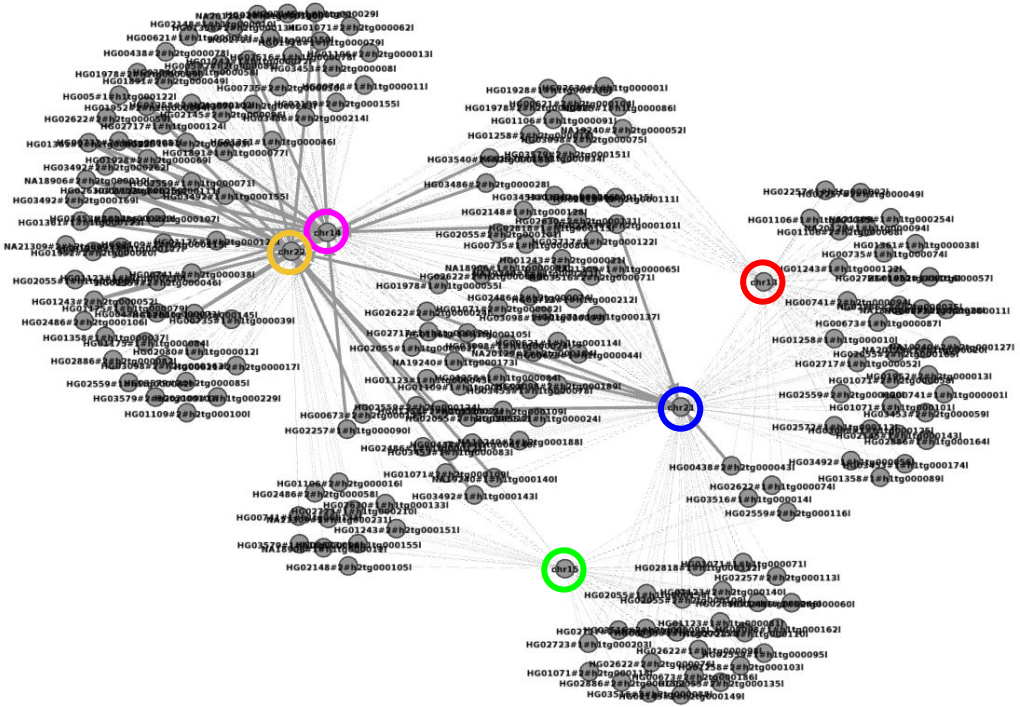
Human Pangenome Graph Consortium data

Phased diploid *de novo* human assemblies from 47 individuals, for a total of 94 **full** (telomere-to-telomere) haplotypes.

Contigs are partitioned by chromosome by mapping each *de novo* assembly against the GRCH38 and CHM13 reference genomes.

HPRC *de novo* assemblies

Misjoins identification



Alignment graph of the misjoins. Every node is a contig and every edge represents the number of mapping between nodes. Alignment graph obtained with pafnet and visualized with gephi. Color code: **chr13**, **chr14**, **chr15**, **chr21**, **chr22**.

Using centromeric repeat annotation, Heng Li identified (code) the **misjoins** in the whole HPRC dataset.

Misjoins involve only* **acrocentric chromosomes**, and they are present in most of the assemblies.

* With the exception of assembly errors in one haplotype (HG02080 paternal).

Louvain chromosome communities

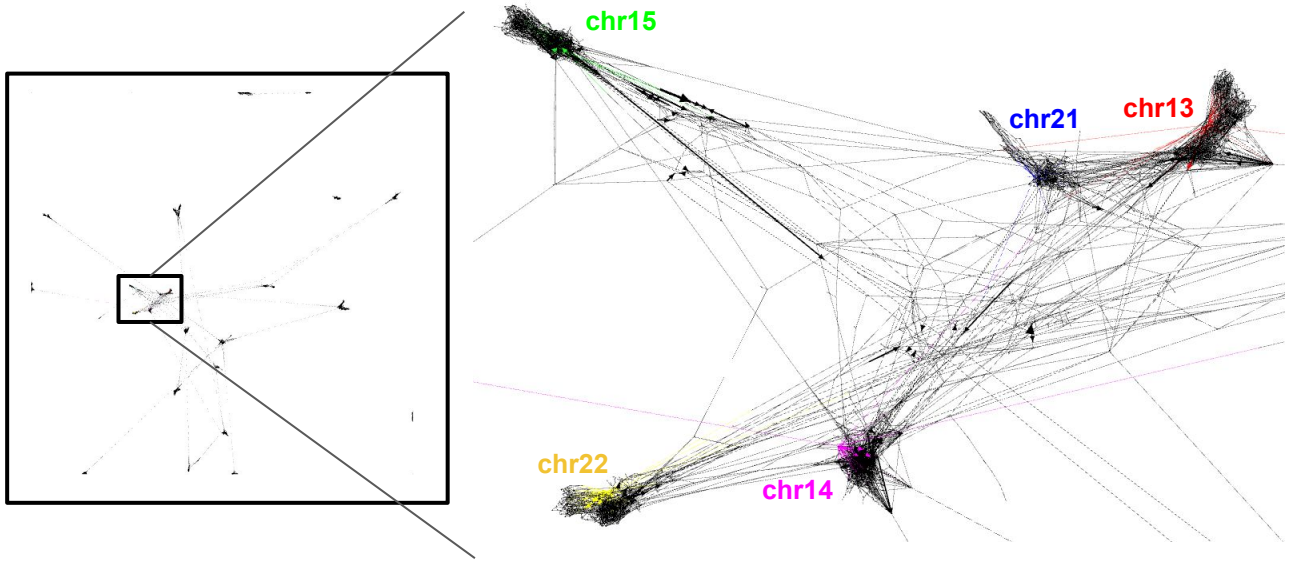
We apply the Louvain method (code) to detect **communities** from the alignment graph made with all contigs.

Chromosome communities:

1. chm13#chr1 grch38#chr1
2. chm13#chr2 grch38#chr2
3. chm13#chr3 grch38#chr3
4. chm13#chr4 grch38#chr4
5. chm13#chr5 grch38#chr5
6. chm13#chr6 grch38#chr6
7. chm13#chr7 grch38#chr7
8. chm13#chr8 grch38#chr8
9. chm13#chr10 grch38#chr10
10. chm13#chr11 grch38#chr11
11. chm13#chr12 grch38#chr12
12. **chm13#chr13 chm13#chr21 grch38#chr13 grch38#chr21**
13. **chm13#chr14 chm13#chr22 grch38#chr14 grch38#chr22**
14. chm13#chr15 grch38#chr15
15. chm13#chr16 grch38#chr16
16. chm13#chr17 grch38#chr17
17. chm13#chr18 grch38#chr18
18. chm13#chr19 grch38#chr19
19. chm13#chr20 grch38#chr20
20. **chm13#chrX grch38#chrX grch38#chrY**

Louvain chromosome communities

We apply the Louvain method (code) to detect **communities** from the alignment graph made with all contigs.

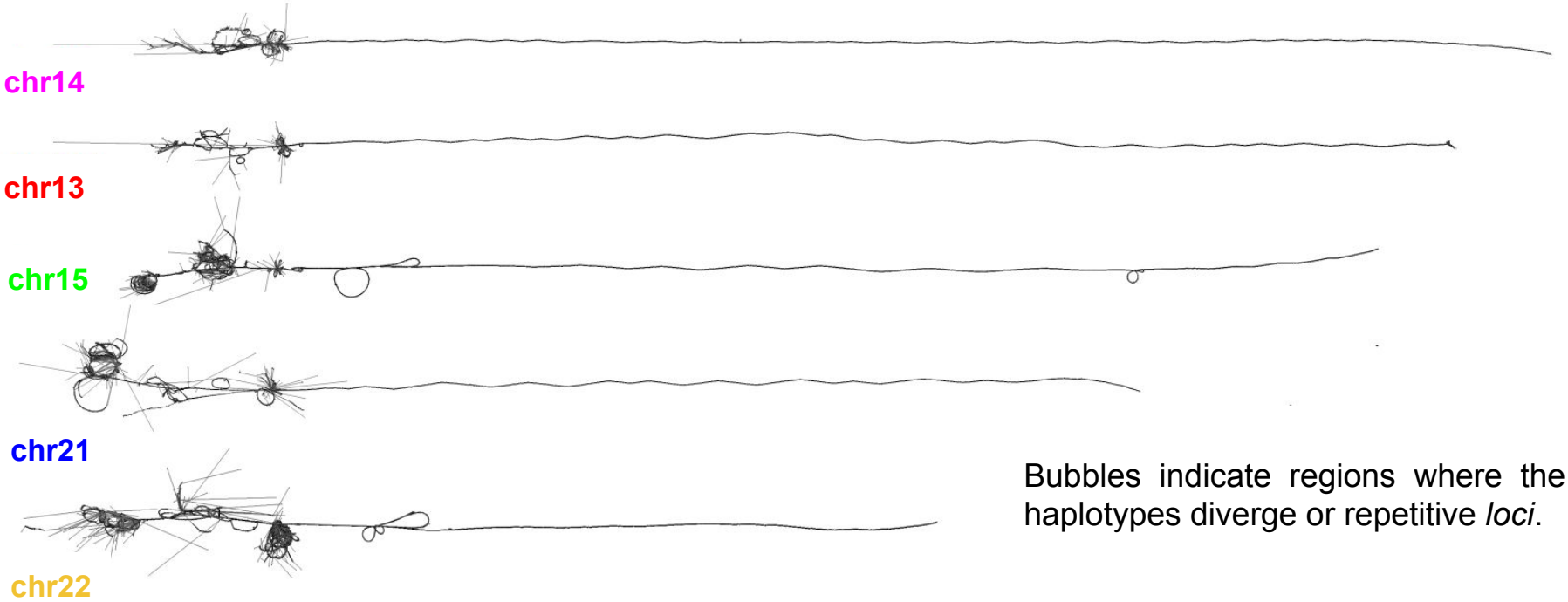


(More stringent) alignment graph from the all-vs-all mappings. The colors show the chm13 chromosomes and their immediate neighbors. Alignment graph obtained with pafnet and visualized with gephi. Color code: **chr13**, **chr14**, **chr15**, **chr21**, **chr22**.

Chromosome communities:

1. chm13#chr1 grch38#chr1
2. chm13#chr2 grch38#chr2
3. chm13#chr3 grch38#chr3
4. chm13#chr4 grch38#chr4
5. chm13#chr5 grch38#chr5
6. chm13#chr6 grch38#chr6
7. chm13#chr7 grch38#chr7
8. chm13#chr8 grch38#chr8
9. chm13#chr10 grch38#chr10
10. chm13#chr11 grch38#chr11
11. chm13#chr12 grch38#chr12
12. **chm13#chr13 chm13#chr21 grch38#chr13 grch38#chr21**
13. **chm13#chr14 chm13#chr22 grch38#chr14 grch38#chr22**
14. chm13#chr15 grch38#chr15
15. chm13#chr16 grch38#chr16
16. chm13#chr17 grch38#chr17
17. chm13#chr18 grch38#chr18
18. chm13#chr19 grch38#chr19
19. chm13#chr20 grch38#chr20
20. **chm13#chrX grch38#chrX grch38#chrY**

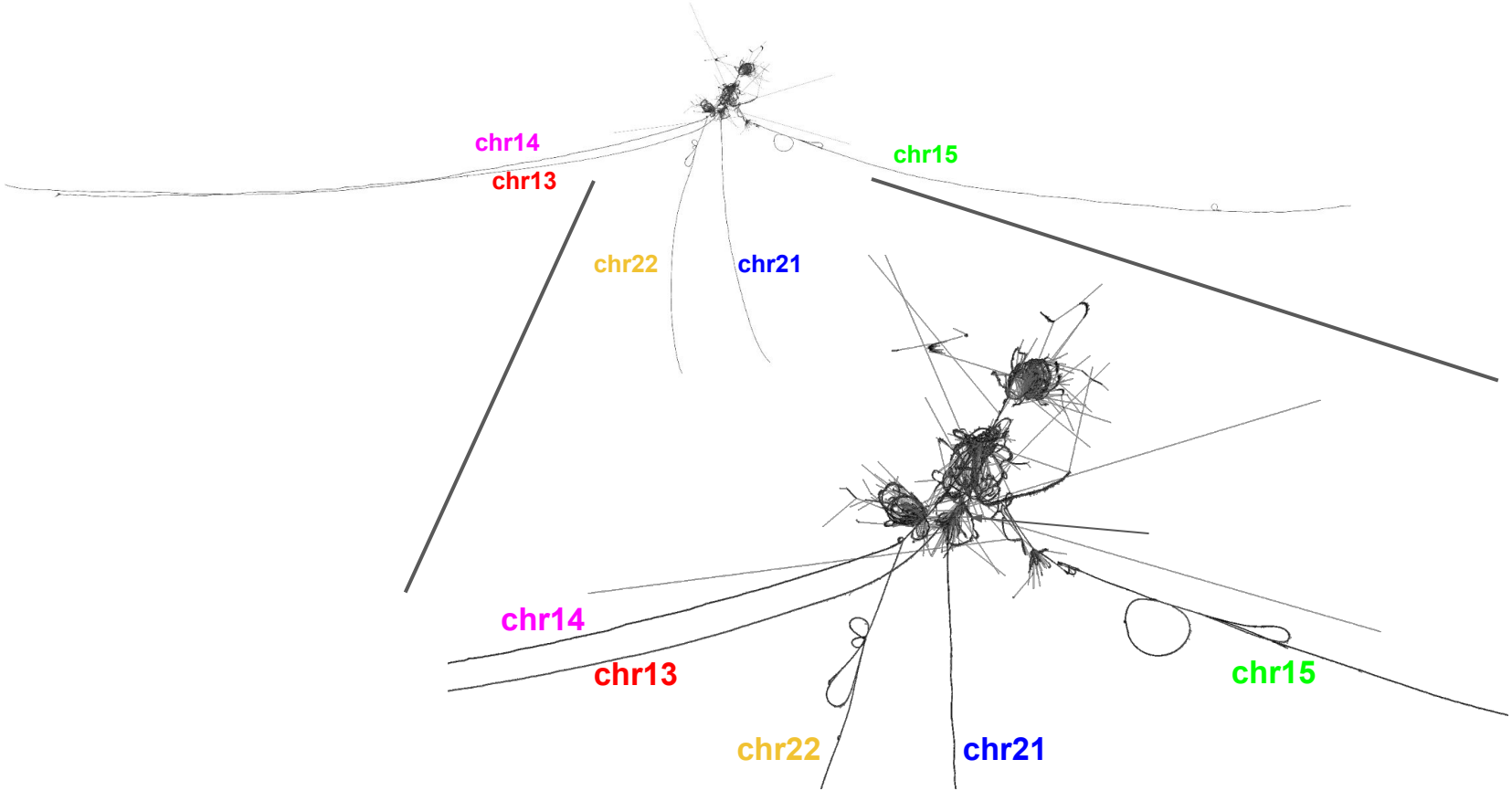
Acrocentric chromosomes processed separately



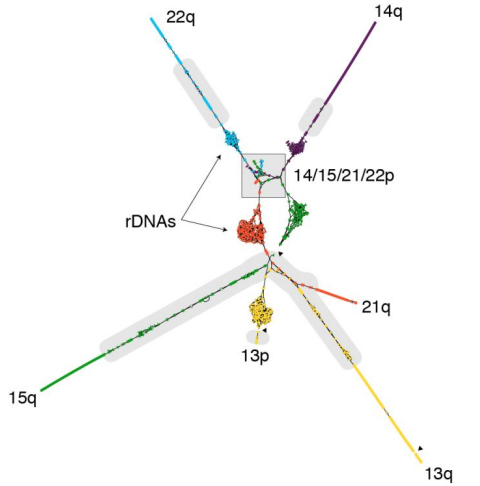
Layout visualizations of the pangenome graphs of the acrocentric chromosomes. Images obtained with odgi draw. Pangenome graphs obtained with pggb.

Bubbles indicate regions where the haplotypes diverge or repetitive *loci*.

Acrocentric chromosomes processed jointly

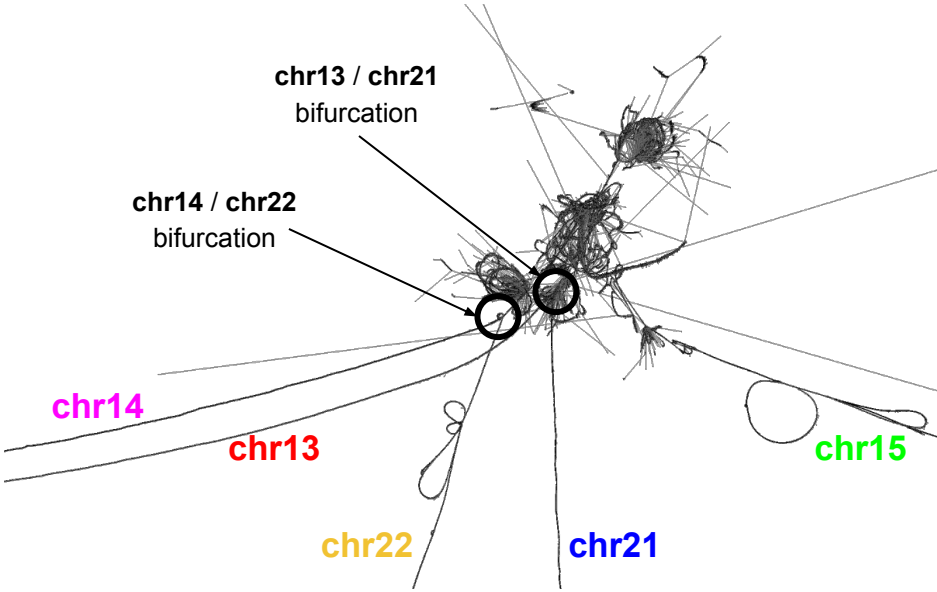


Acrocentric chromosomes processed jointly

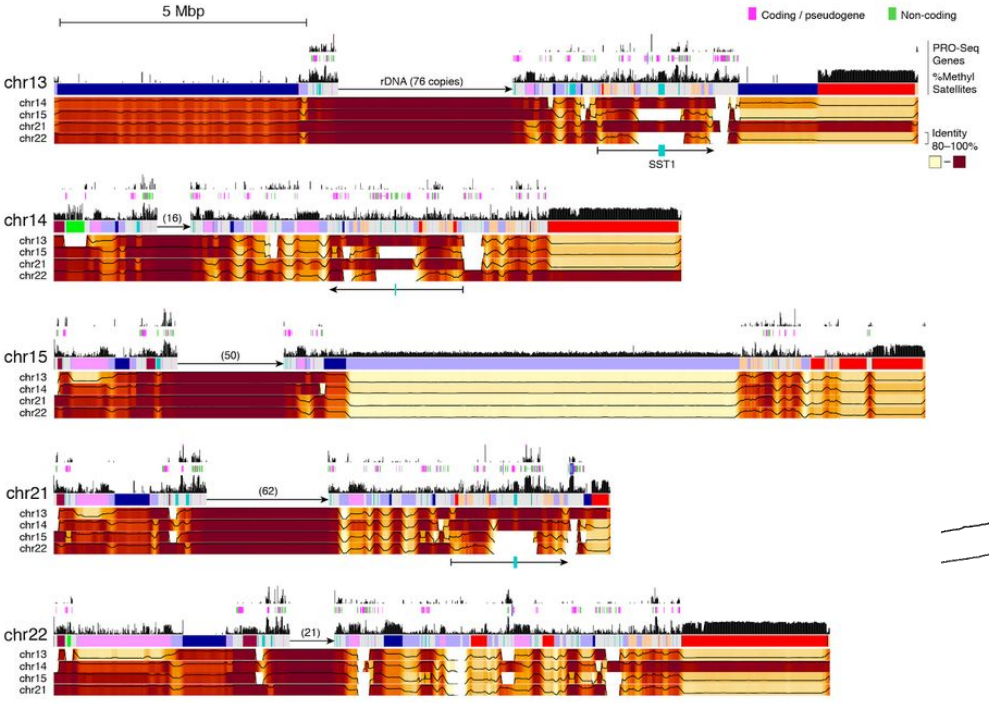


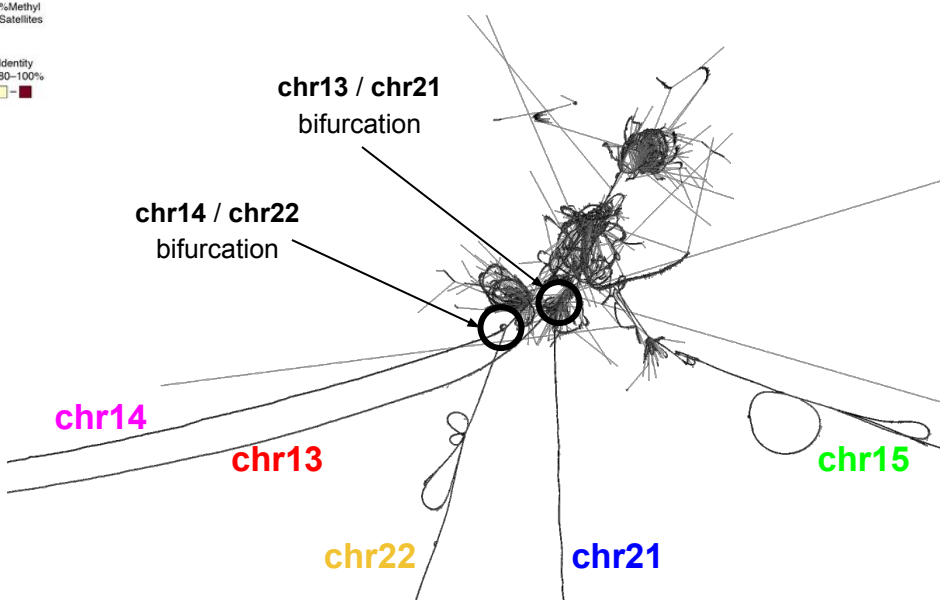
HiFi-based assembly string graph of the acrocentric chromosome of the CHM13 genome. Figure from Nurk, Koren, Rhie, Rautiainen et al., 2021.

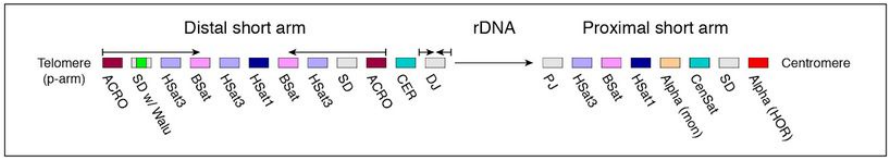
Acrocentric chromosomes processed jointly



Acrocentric chromosomes processed jointly

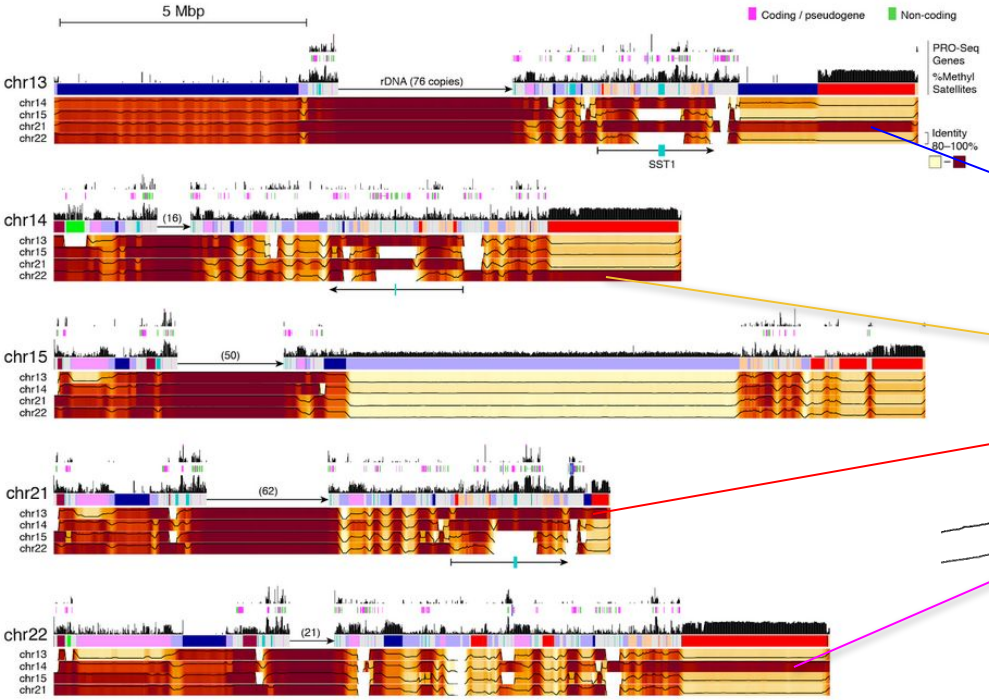


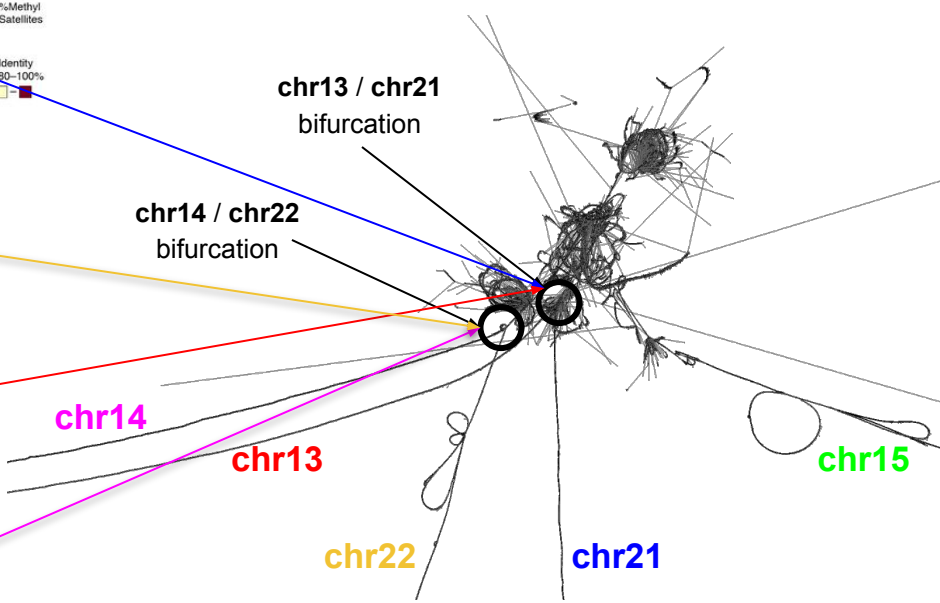


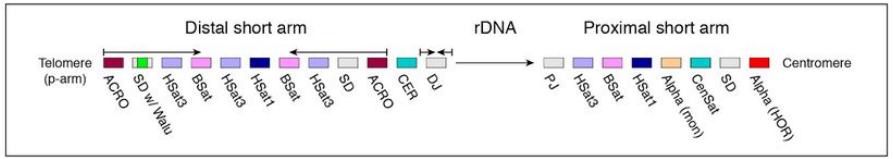


Short arms of the human acrocentric chromosomes in CHM13.
Figure from [Nurk, Koren, Rhie, Rautiainen et al., 2021](#).

Acrocentric chromosomes processed jointly







Short arms of the human acrocentric chromosomes in CHM13.
Figure from Nurk, Koren, Rhie, Rautiainen et al., 2021.

Towards traces of recombination

The high level of homology of the acrocentric chromosomes could be due to **non-homologous recombination**.

High-quality *de novo* assemblies and pangenomic approaches will shed light on the most difficult regions of the human genomes.

Volume 16 Number 4 1988 Nucleic Acids Research

Homologous alpha satellite sequences on human acrocentric chromosomes with selectivity for chromosomes 13, 14 and 21: implications for recombination between nonhomologues and Robertsonian translocations

K.H.Choo*, B.Vissel, R.Brown, R.G.Filby and E.Earle

ABSTRACT

We report a new subfamily of alpha satellite DNA (pTRA-2) which is found on all the human acrocentric chromosomes. The alphoid nature of the cloned DNA was established by partial sequencing. Southern analysis of restriction enzyme-digested DNA fragments from mouse/human hybrid cells containing only human chromosome 21 showed that the predominant higher-order repeating unit for pTRA-2 is a 3.9 kb structure. Analysis of a "consensus" in situ hybridisation profile derived from 13 normal individuals revealed the localisation of 73% of all centromeric autoradiographic grains over the five acrocentric chromosomes, with the following distribution: 20.4%, 21.5%, 17.1%, 7.3% and 6.5% on chromosomes 13, 14, 21, 15 and 22 respectively. An average of 1.4% of grains was found on the centromere of each of the remaining 19 nonacrocentric chromosomes. These results indicate the presence of a common subfamily of alpha satellite DNA on the five acrocentric chromosomes and suggest an evolutionary process consistent with recombination exchange of sequences between the nonhomologues. The results further suggests that such exchanges are more selective for chromosomes 13, 14 and 21 than for chromosomes 15 and 22. The possible role of centromeric alpha satellite DNA in the aetiology of 13q14q and 14q21q Robertsonian translocations involving the common and nonrandom association of chromosomes 13 and 14, and 14 and 21 is discussed.

Chroo et al., 1988.