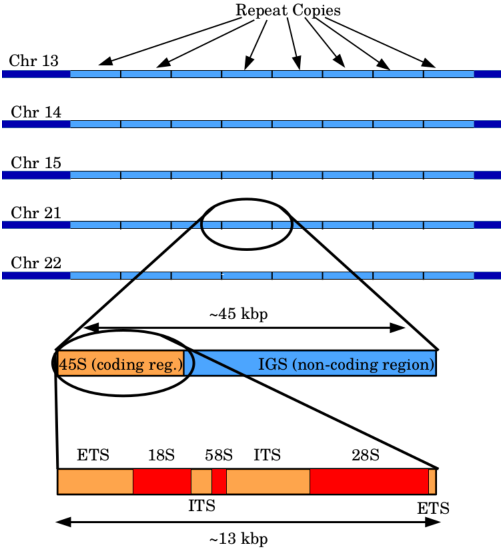
A combinatorial approach for reconstructing rDNA repeats

Frederik Oehl

Heinrich-Heine-Universität Düsseldorf

Introduction



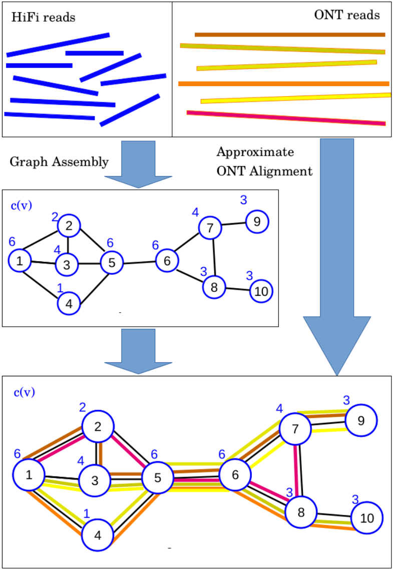
Content

- Combinatorial method for resolving the individual rDNA repeat copies from any given human sample
- Assembly of the rDNA repeat copies from six samples → CHM13, HPRC (Human Pangenome Reference Consortium)
- CHM13: Comparison with T2T assembly → Appendix

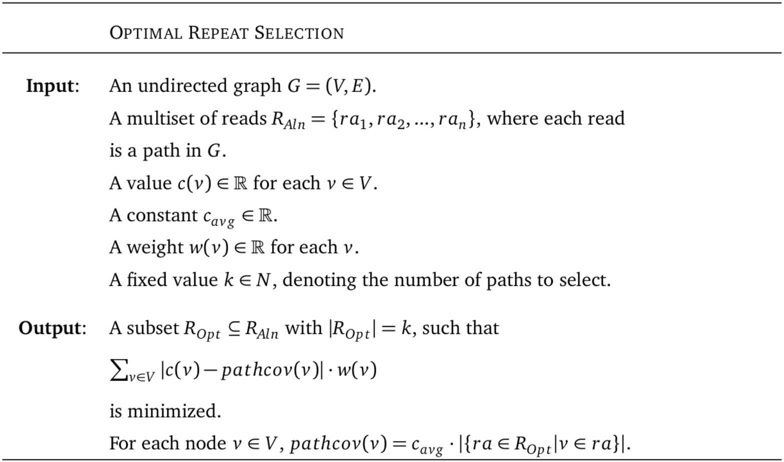
Related Work

- Telomere-to-telomere (T2T) Consortium [1] → CHM13 reference genome, rDNA assembly
- Methods for graph-based assembly and sequence-to-graph-alignment → MBG [2] and GraphAligner [3]
- Viral quasispecies assembly by Baaijens et al. [4] → We use a similar optimization approach

Model: Preprocessing



Model



OPTIMAL REPEAT SELECTION

Input: An undirected graph $G = (V, E)$.

A multiset of reads $R_{Aln} = \{ra_1, ra_2, ..., ra_n\}$, where each read is a path in G.

A value $c(v) \in \mathbb{R}$ for each $v \in V$.

A constant $c_{avg} \in \mathbb{R}$.

A weight $w(v) \in \mathbb{R}$ for each v.

A fixed value $k \in N$, denoting the number of paths to select.

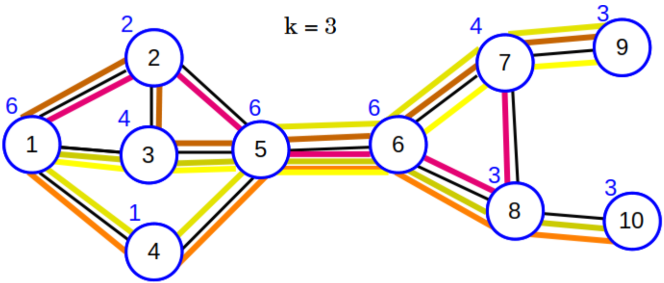
Output: A subset $R_{Opt} \subseteq R_{Aln}$ with $|R_{Opt}| = k$, such that

$\sum_{v \in V} |c(v) - pathcov(v)| \cdot w(v)$

is minimized.

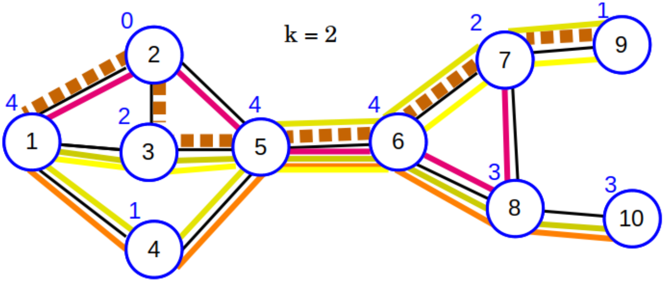
For each node $v \in V$, $pathcov(v) = c_{avg} \cdot |\{ra \in R_{Opt} | v \in ra\}|$.

Model: Example



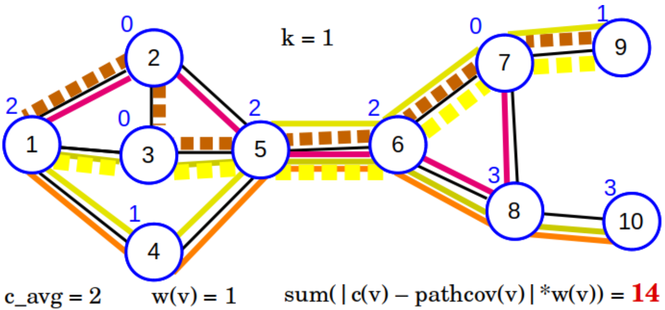
c_avg = 2 w(v) = 1 sum(| c(v) – pathcov(v) | *w(v)) = **38**

Model: Example

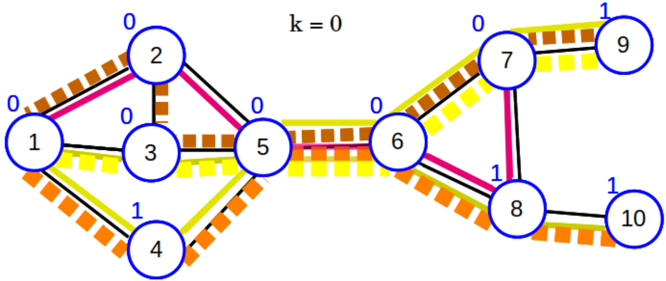


c_avg = 2 w(v) = 1 sum(| c(v) – pathcov(v) | *w(v)) = **24**

Model: Example



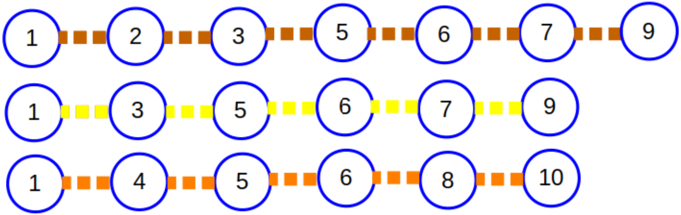
Model: Example



c_avg = 2 w(v) = 1 sum(| c(v) – pathcov(v) | *w(v)) = 4

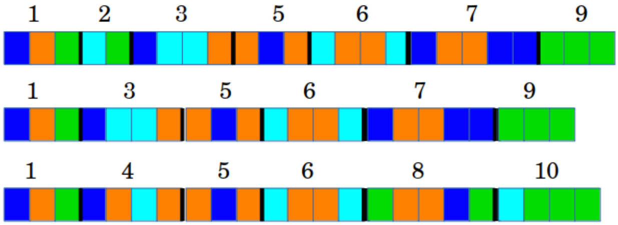
Model: Example

Chosen paths:



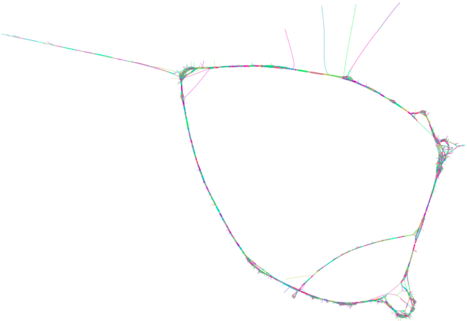
Model: Example

Reconstructed Repeats:



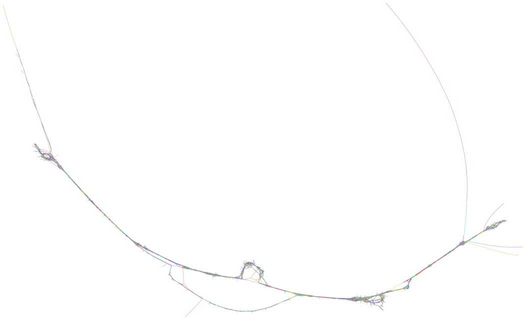
Results: CHM13

- Assembly graph

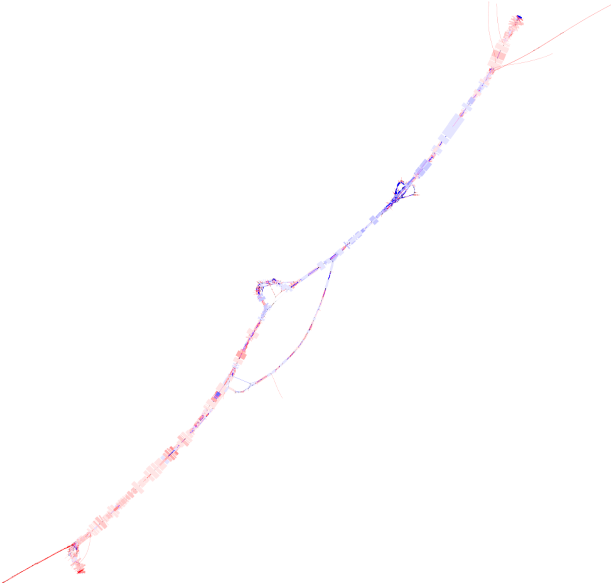


Results: CHM13

- Assembly graph



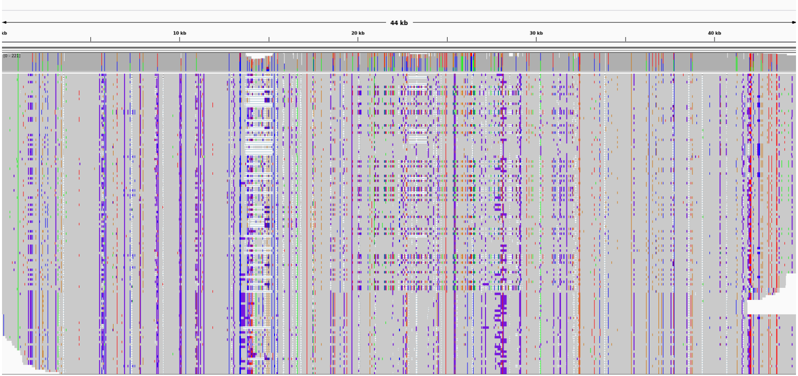
Results: CHM13



- Over- and underexplanation of coverage by the model

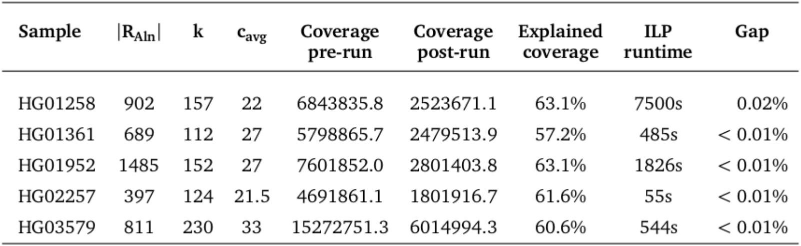
Results: CHM13

- Repeat sequences from our model, aligned against the canonical rDNA unit KY962518.1



Results: HPRC samples

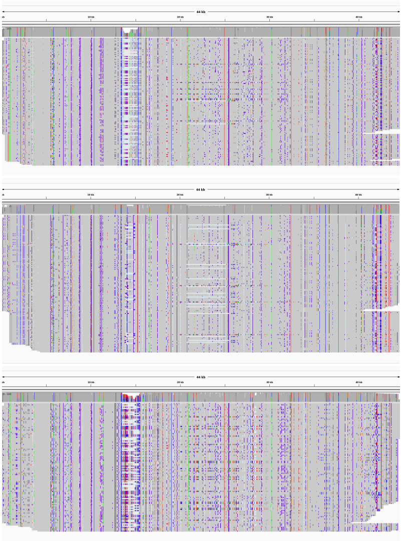
- Assembly of five samples from the Human Pangenome Reference Consortium
- Some preprocessing choices were different compared to CHM13



Sample	$\|R_{\text{Aln}}\|$	k	c_{avg}	Coverage pre-run	Coverage post-run	Explained coverage	ILP runtime	Gap
HG01258	902	157	22	6843835.8	2523671.1	63.1%	7500s	0.02%
HG01361	689	112	27	5798865.7	2479513.9	57.2%	485s	$< 0.01\%$
HG01952	1485	152	27	7601852.0	2801403.8	63.1%	1826s	$< 0.01\%$
HG02257	397	124	21.5	4691861.1	1801916.7	61.6%	55s	$< 0.01\%$
HG03579	811	230	33	15272751.3	6014994.3	60.6%	544s	$< 0.01\%$

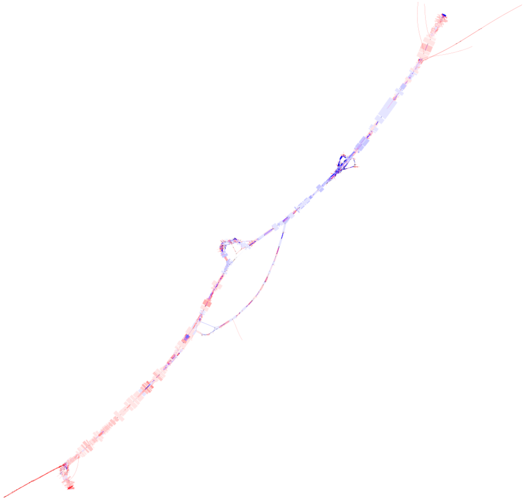
Results: HPRC samples

- Example: HG01258, HG01952, HG02257
- Significant differences in some regions → Junction between coding region and IGS, central part of IGS



Future Challenges

- Room for improvement → Improperly explained coverage



Future Challenges

- Improving the model
- Resolving the order of the copies on the genome
- Haplotyping the copies

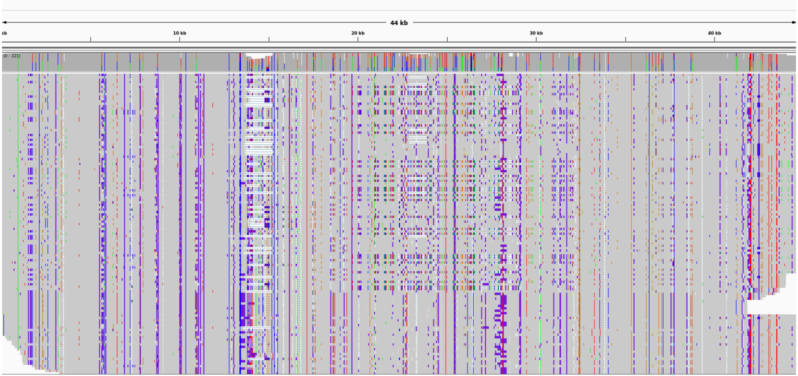
Thank you for your attention!

Literature

- Sergey Nurk et al. *The complete sequence of a human genome.* Science 376 (6588 2022), pp. 44–53. DOI: 10.1126/science.abj6987.
- Mikko Rautiainen and Tobias Marschall. *MBG: Minimizer-based sparse de Bruijn Graph construction.* Bioinformatics 37 (16 2021), pp. 2476–2478. DOI: 10.1093/bioinformatics/btab004.
- Mikko Rautiainen and Tobias Marschall. *GraphAligner: rapid and versatile sequence- to-graph alignment.* Genome Biology 21 (253 2020). DOI: 10.1186/s13059-020-02157-2.
- Jasmijn A. Baaijens et al. *Full-length de novo viral quasispecies assembly through variation graph construction.* In: Bioinformatics 35 (24 2019), pp. 5086–5094. DOI: 1093/bioinformatics/btz443.

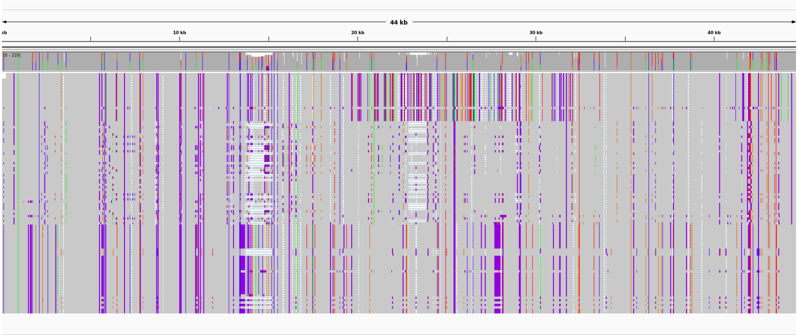
Appendix: Comparison with T2T assembly

- Repeat sequences from our model, aligned against the canonical rDNA unit KY962518.1



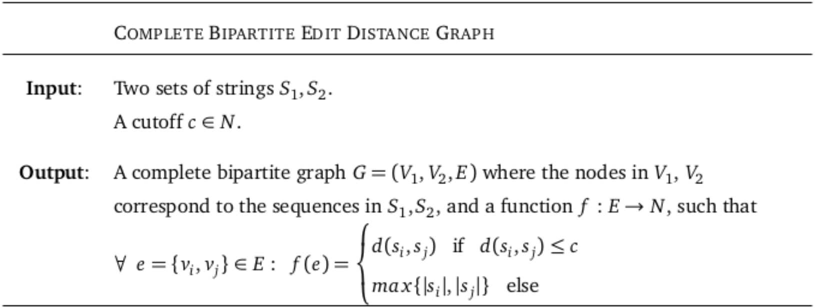
Appendix: Comparison with T2T assembly

- Repeat sequences from the T2T assembly, aligned against the canonical rDNA unit KY962518.1



Appendix: Comparison with T2T assembly

- Idea: Compute edit distance for all pairs of copies from both assemblies
- Find a minimum-weight perfect matching
- Only compute pairs that are similar enough → Cutoff *c*



COMPLETE BIPARTITE EDIT DISTANCE GRAPH

Input: Two sets of strings S_1, S_2.
A cutoff $c \in N$.

Output: A complete bipartite graph $G = (V_1, V_2, E)$ where the nodes in V_1, V_2 correspond to the sequences in S_1, S_2, and a function $f : E \rightarrow N$, such that

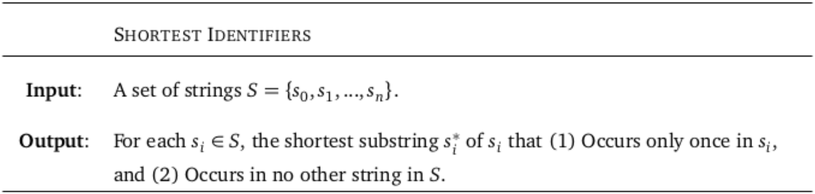
$$\forall\ e = \{v_i, v_j\} \in E:\ \ f(e) = \begin{cases} d(s_i, s_j) & \text{if} \quad d(s_i, s_j) \leq c \\ max\{|s_i|, |s_j|\} & \text{else} \end{cases}$$

Appendix: Comparison with T2T assembly

- The sets of copies differ considerably
- MWPM: Only 53 pairs of repeats with an edit distance $\leqslant 4500$
- For 54 copies from RS model, there is at least one similar copy in the T2T set
- For 215 copies vice versa
- Our copies vary more than the T2T copies → Is this accurate?

Appendix: Comparison with T2T assembly

- Idea: Retrace unique feature of each repeat copy in the HiFi reads → Find SHORTEST IDENTIFIERS
- Homopolymer compression for cleaner results

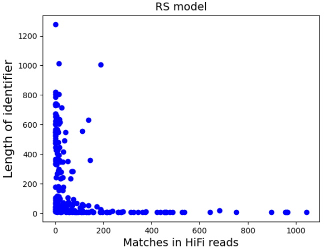


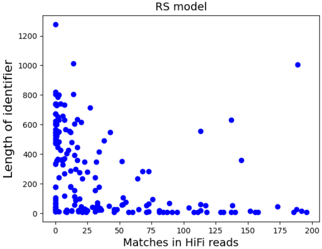
SHORTEST IDENTIFIERS

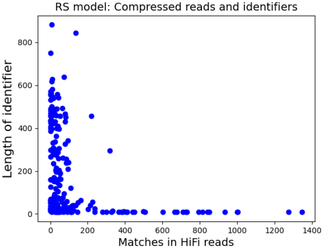
Input: A set of strings $S = \{s_0, s_1, ..., s_n\}$.

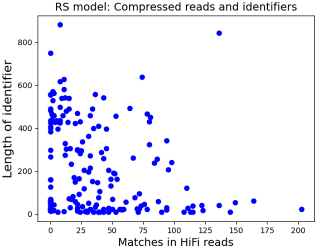
Output: For each $s_i \in S$, the shortest substring s_i^* of s_i that (1) Occurs only once in s_i, and (2) Occurs in no other string in S.

Appendix: Comparison with T2T assembly

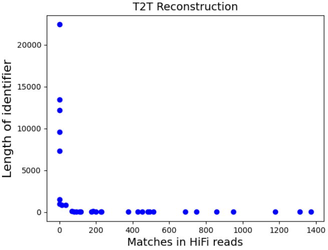

RS model

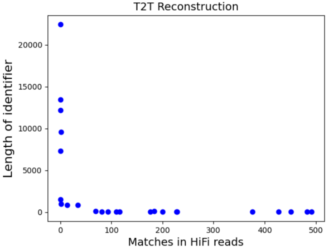

RS model

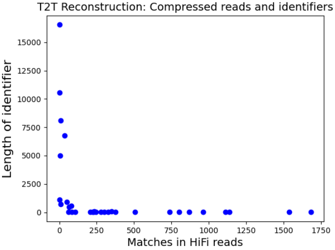

RS model: Compressed reads and identifiers

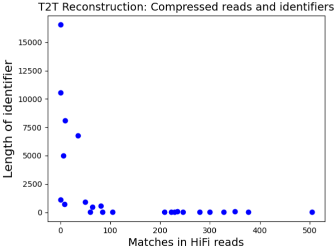

RS model: Compressed reads and identifiers

Appendix: Comparison with T2T assembly


T2T Reconstruction
Length of identifier
Matches in HiFi reads


T2T Reconstruction
Length of identifier
Matches in HiFi reads


T2T Reconstruction: Compressed reads and identifiers
Length of identifier
Matches in HiFi reads


T2T Reconstruction: Compressed reads and identifiers
Length of identifier
Matches in HiFi reads