chromosome-scale scaffolding
human genomes with linked read kmers

René Warren • Lauren Coombe • Jessica Zhang • Ben Vandervalk
Justin Chu • Shaun Jackman • Jeffrey Tse • Inanç Birol

Builds on our QCF scaffold 1

Approach

Linked read mapping
Requires a contig end to match a minimum fraction of the read kmers (parameter \( D \), 0.55, default):

\[
\text{score}(\text{contig, read}) = \frac{\text{kmers(contig)} \cap \text{kmers(read)}}{\text{kmers(contig)}}
\]

Higher specificity: both reads/pair must map same target & kmers with multiple memberships discarded

Gap size estimation

- Train on distances (D) between contig head and tail
- Record \( D \) and Jaccard index (J) for shared barcodes

\[
J(r, p) = \frac{\text{kmers}(r) \cap \text{kmers}(p)}{\text{kmers}(r) \cup \text{kmers}(p)}
\]

- Retrieve intra-contig \( D \) samples with the N closest J
- Output \( \frac{1}{99} \% \) of joined barcode sequence

\[
D_{\text{min}}(r, p) = \text{argmin}_x \sum \| r(x) - p(x) \|
\]

\[
D_{\text{max}}(r, p) = \text{argmax}_x \sum \| r(x) - p(x) \|
\]

Results

Scaffolding NA12878 10xG/Supernova draft

Comparison to human reference

NA12878 human genome ideogram
10xG's Supernova human genome assembly before (blue) and after (green) ARKS scaffolding

Parameters

- \( -z \): ARKS / ARCS pipeline
- \( -c \): ARKS / ARCS pipeline
- \( -e \): ARKS / ARCS pipeline
- \( -m \): ARKS / ARCS pipeline
- \( -l \): ARKS / ARCS pipeline
- \( -a \): ARKS / ARCS pipeline

Performance

Scaffolding NA12878 10xG/Supernova draft

Gap size estimates

References

Software
https://github.com/bcgsc/arks