The Best Finish First: Sequence Finishing with Whole Genome Mapping

Deacon Sweeney, Ph.D.

Bioinformatics Scientist

OpGen Inc.

Outline

Finishing Issues

Whole Genome Mapping

Best-First Sequence Placement

Finishing Issues

"When a species genome is newly assembled, no one knows what's real, what's missing, and what's experimental artifact."

- Monya Baker, Nature Methods: 9:4 (2012)

"While an initial draft […] could be determined in a matter of weeks, the complete sequence required many months or even years of additional experiments"

- Nagarajan et al., BMC Genomics: 11:242 (2010)

Next Gen Sequencing and Finishing

Inexpensive, fast, short reads

Improved base identification accuracy

Issues:

- Duplicate genes and repetitive regions compressed
- Increased mis-assembly rate

Results:

- Confound comparative genomics
- More challenging finishing

Fan, Li. Nature Biotechnology. 30:4 (2012)

The Need for Increased Read Length

"New instruments that generate longer reads [...] will facilitate assembly if the high error rates of these systems can be reduced."“

- Fan & Li, Nature Biotechnology 30:4 (2012) Peking University, Beijing

“It is critical to develop new hybrid sequencing approaches, such as multiplatform strategies including the third-generation long-read technologies“

- Alkan et al. Nature Methods 8 (2011), Howard Hughes Medical Institute

Outline

Finishing Issues

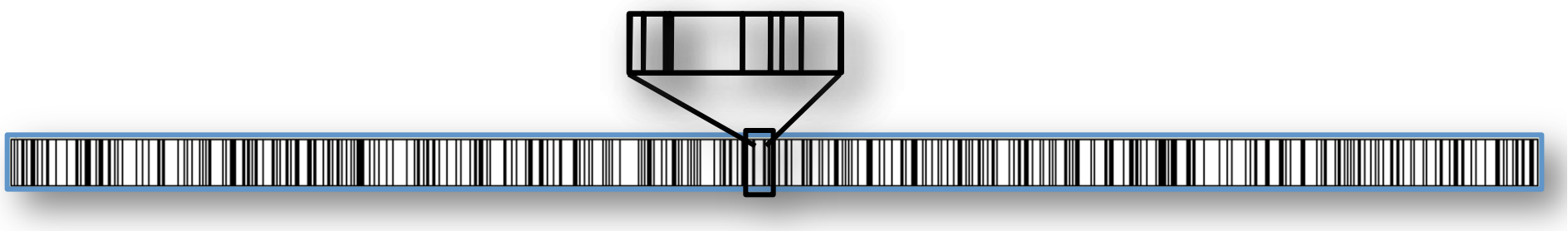
Whole Genome Mapping

Best-First Sequence Placement

Whole Genome Mapping

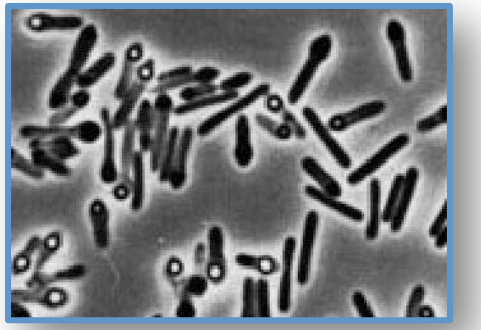
Represents a genome as a set of distances between restriction sites.

acagctctcgaga**ggatcc**tcgtcg**ggatcc**ctcgcgctcgagatcgcgtagcgctagagcgctctagaggctcgcgga
gagctcgcgcgagtgcgtcggggacacattcga**ggatcc**agttagagatcggctcgtgctagaggcctgctcgtagaga
cacagatagacagatagagcggctcgctctcgctgctcggaagtcgctcgcgtaagttcgcgct**ggatcc**cacagctcgc
gctgacacagtcgcgtagagatgcggctgagcgctggcgctgaggctggacagtgctgctgagctcggacagctcgtgt
ggcgc**ggatcc**gtgctcggc**ggatcc**tagggcgtgtcgcgtgctggatgcgctggtgggccccagtttggcggcgctcgc
ggctcggctgctggtcgcctgcttt

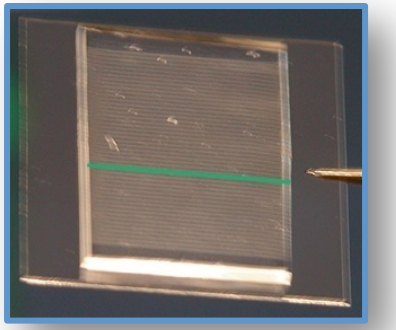


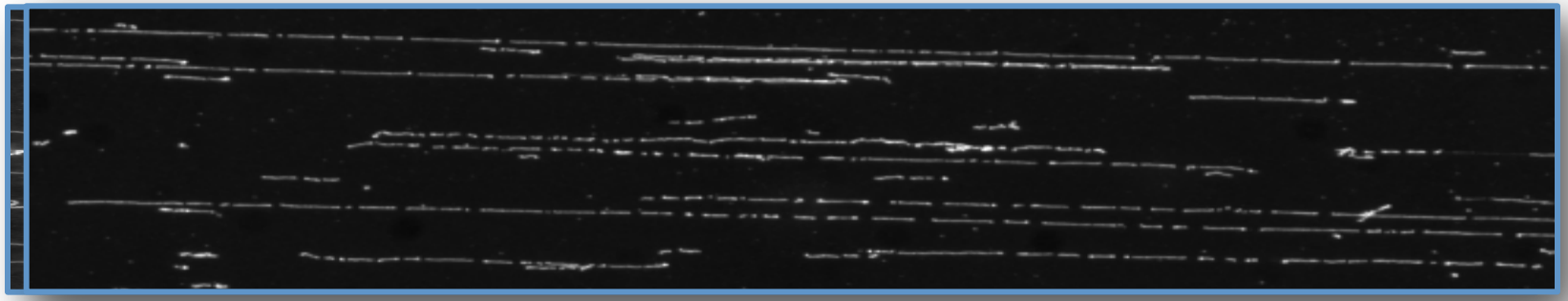
How Whole Genome Maps are Produced

Gently isolate DNA



Pass DNA through channels over a sticky surface

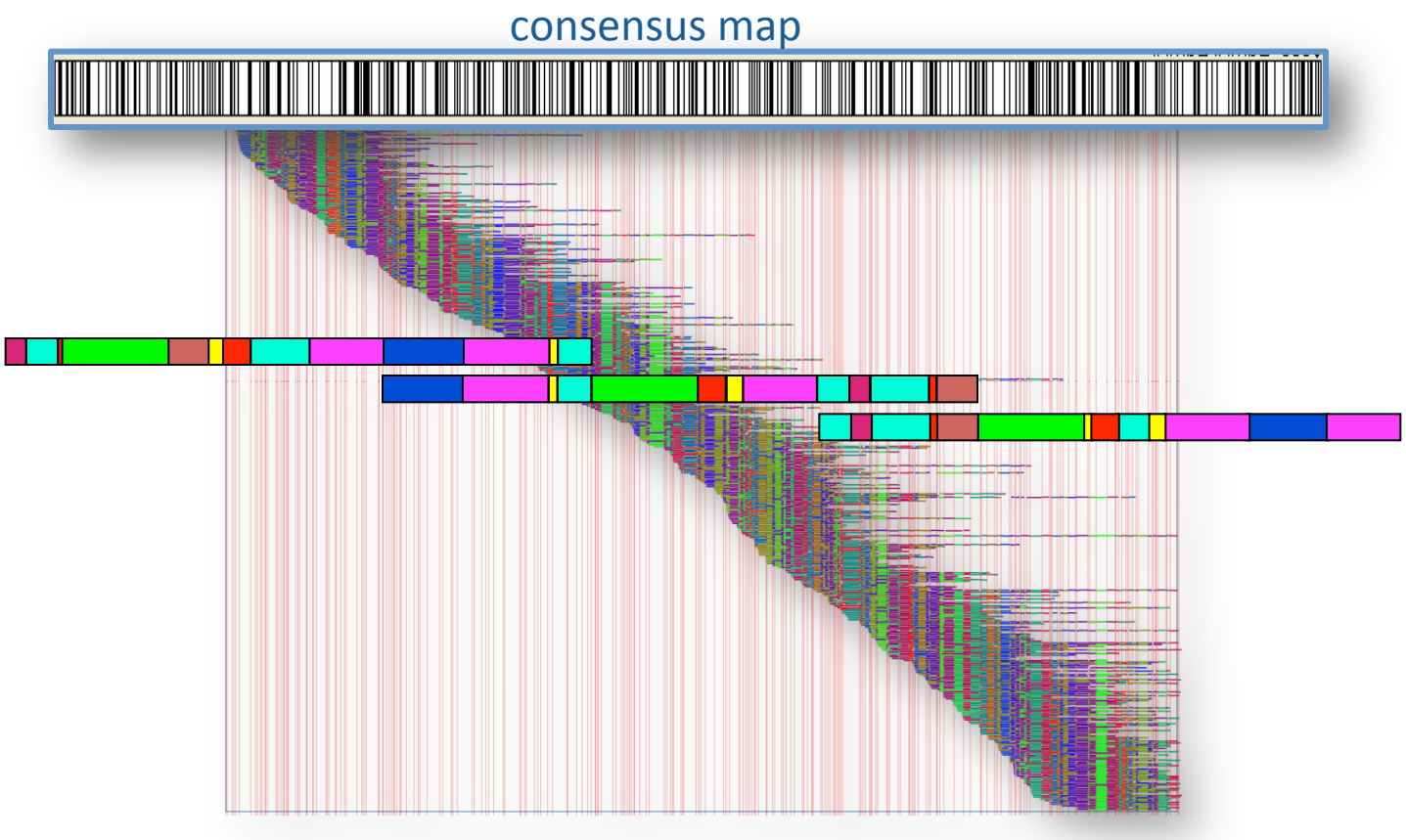




Digest on-surface with restriction enzyme, then image.

Map Assembly

Assemble a chromosomal map by aligning the overlapping patterns.



Outline

Finishing Issues

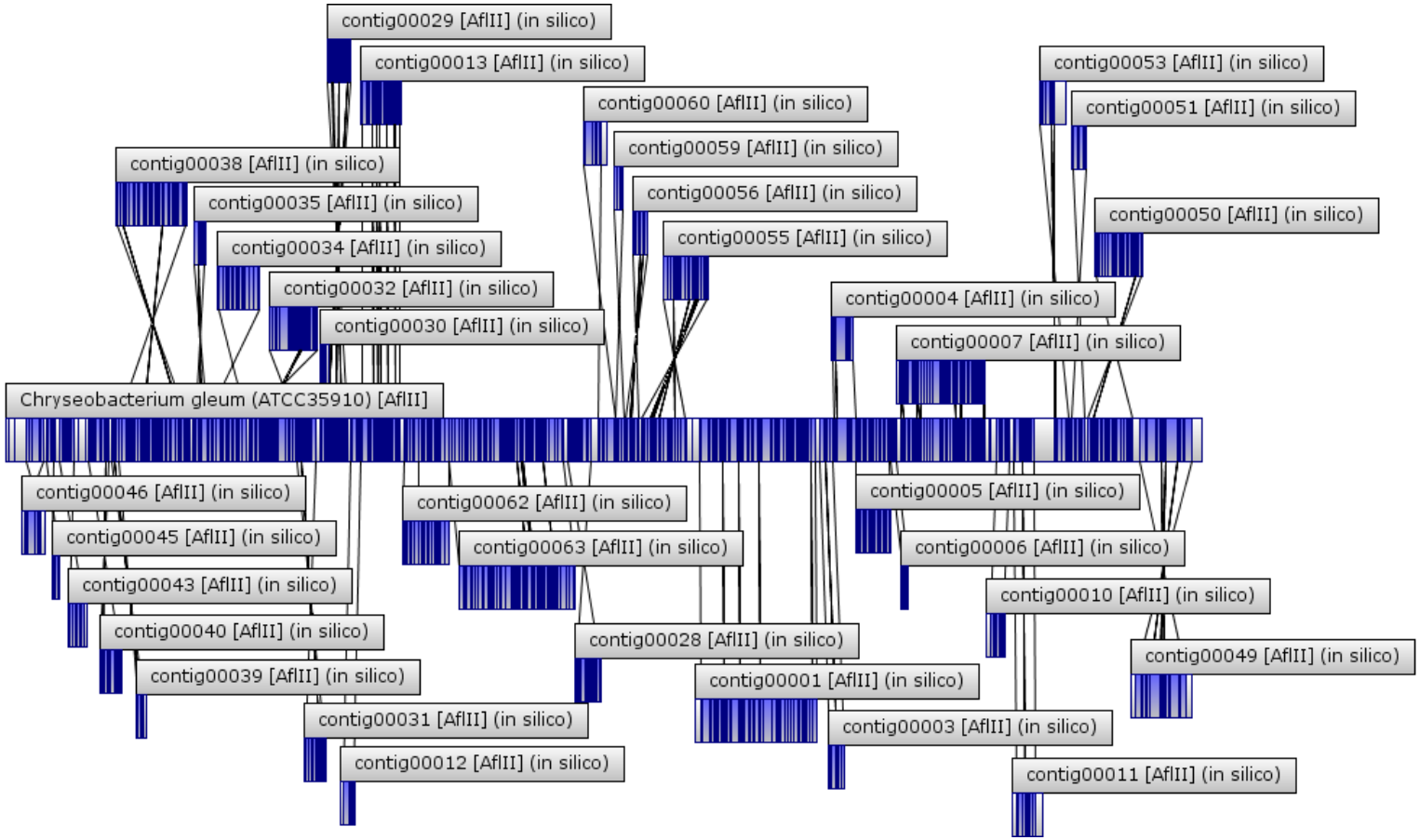
Whole Genome Mapping

Best-First Sequence Placement

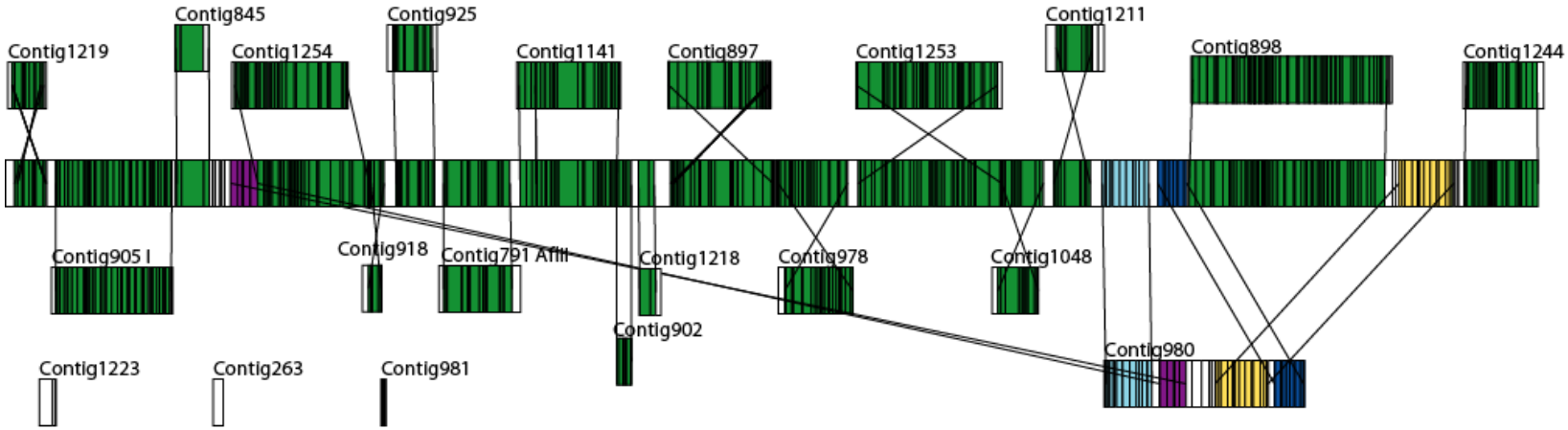
Sequence Placement

SP orders sequence contigs by aligning to WG Maps

Create in silico maps from sequence contigs



Mis-assembled contigs



Best-First Placement

Place in order of decreasing score

Progressively reduce and fragment available regions

Conclusions

Next-gen sequencing brings challenges to sequence accuracy and sequence finishing

Whole Genome Maps verify whole genome sequence assemblies independent of sequencing technology

The Best-First algorithm increases placement rate by progressively reducing background search space