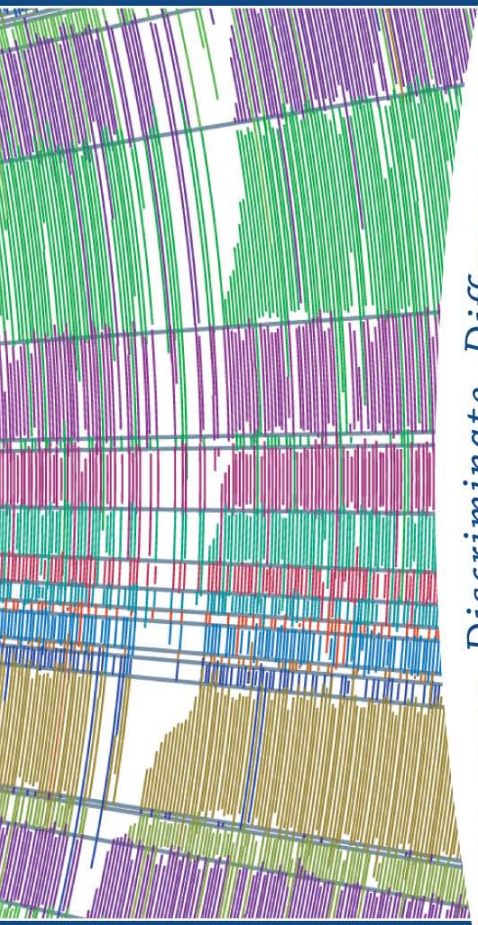


Discover. Discriminate. Differentiate.

A Software System for Validating and Orienting Sequence Contigs Using Optical Maps

Adam M Briska.

Senior Bioinformatics Scientist

OpGen Inc.

Optical Mapping

A system that generates whole-genome, ordered restriction maps.

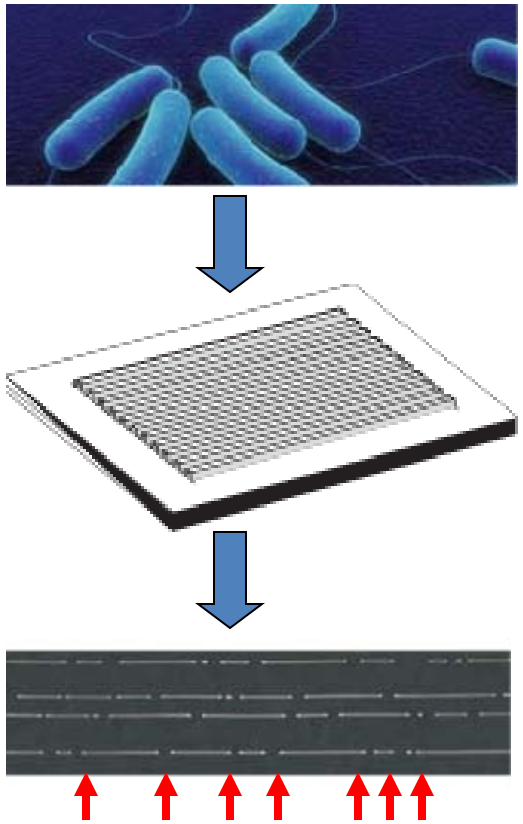
For whole genome analysis of bacteria, yeast or fungi

- high precision
- without the cost of DNA sequencing
- without the inherent bias in microarrays

A *de novo* process...no sequence information required

OpGen

How Optical Mapping Works



Cells gently lysed to release genomic DNA

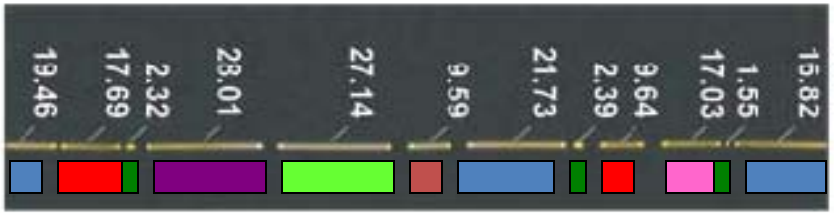
DNA captured in parallel arrays of long single DNA molecules using a microfluidic device

Restriction enzymes cleave the DNA at specific sites

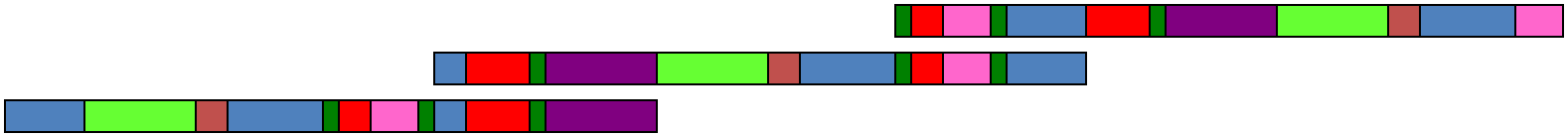
Gaps in the DNA indicate location of cut sites

OpGen

How Optical Mapping Works



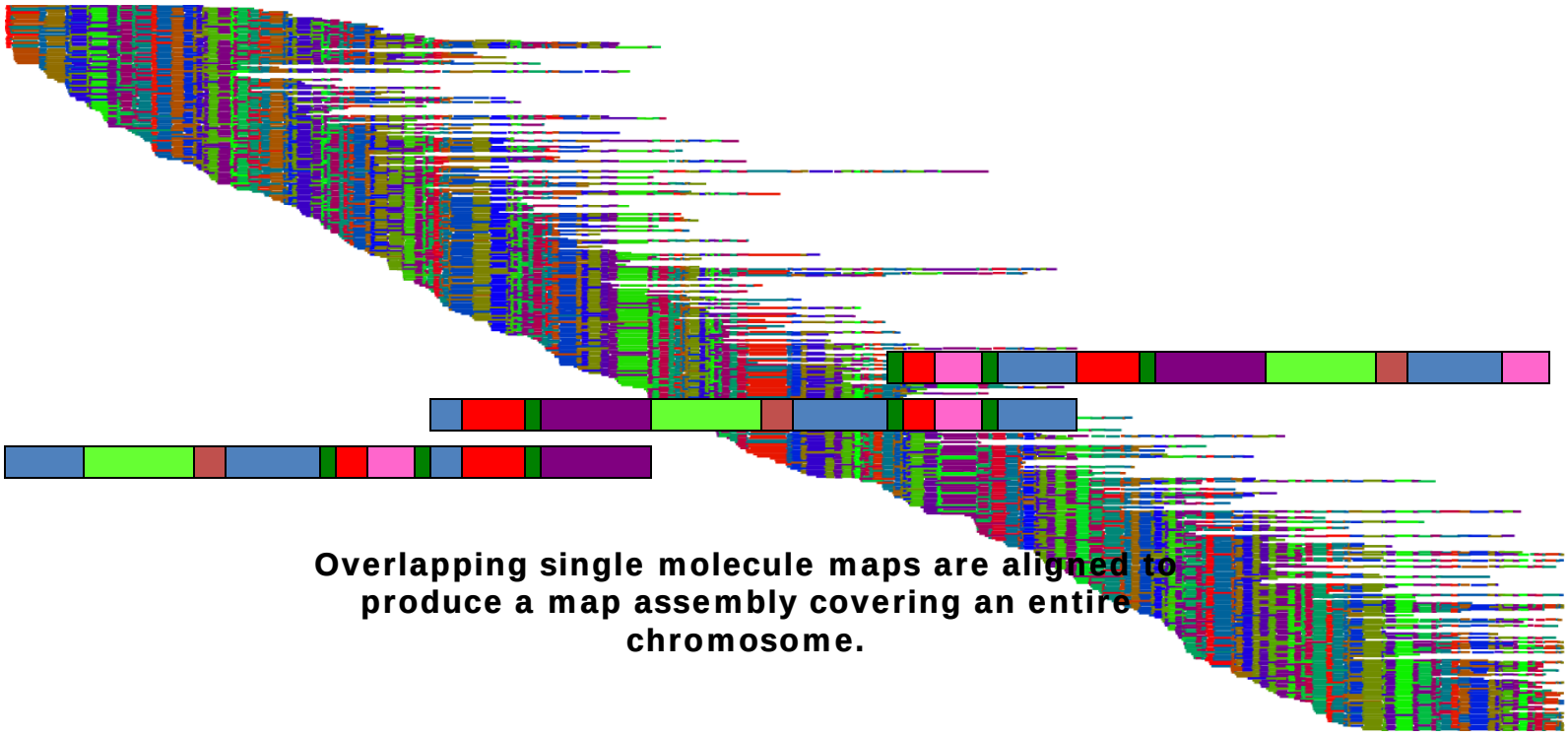
The DNA is stained with fluorescent dye and image analysis measures the size and order of each fragment to create “Single Molecule Maps”.



Overlapping single molecule maps are aligned to produce a map assembly covering an entire chromosome.

OpGen

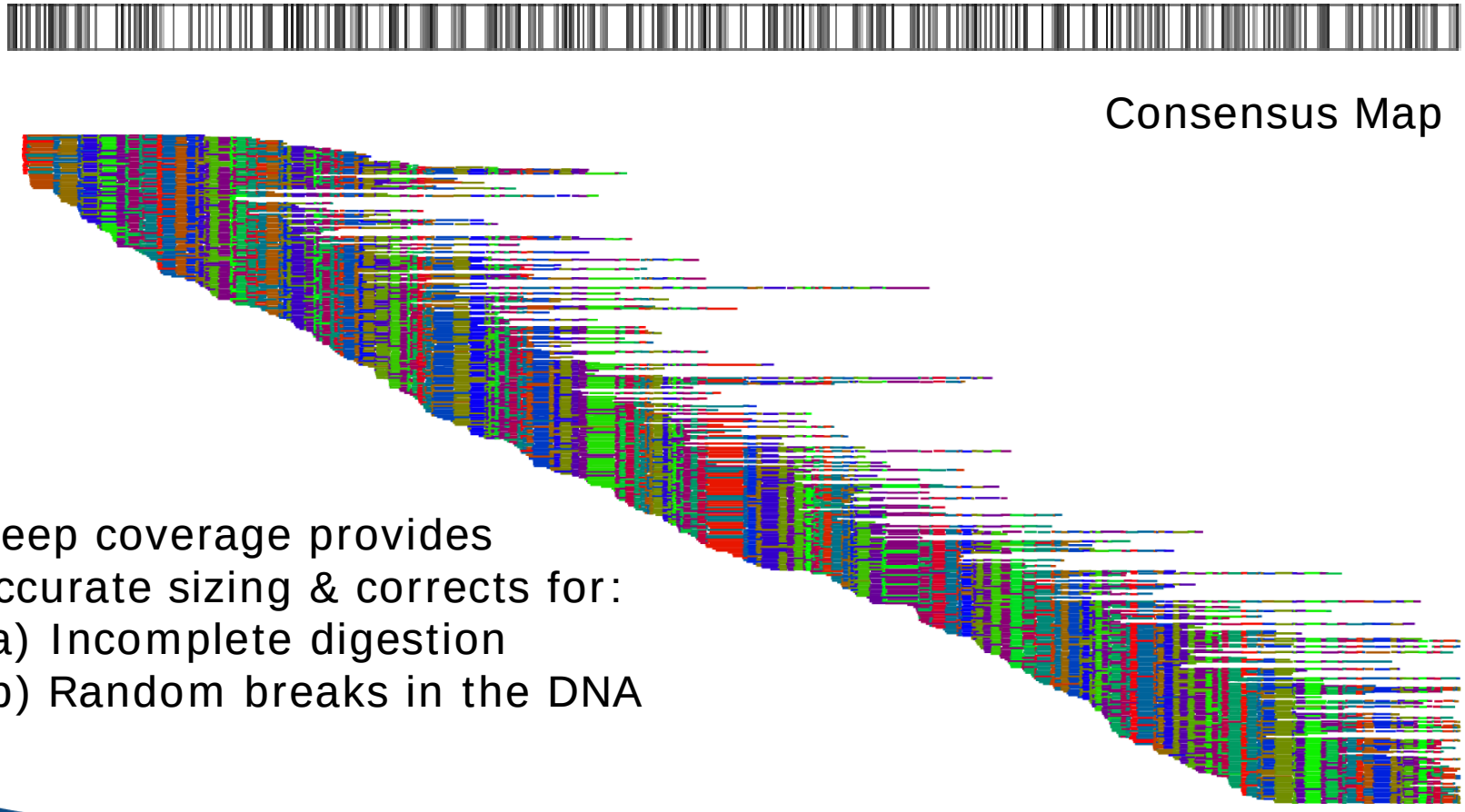
How Optical Mapping Works



Overlapping single molecule maps are aligned to produce a map assembly covering an entire chromosome.

OpGen

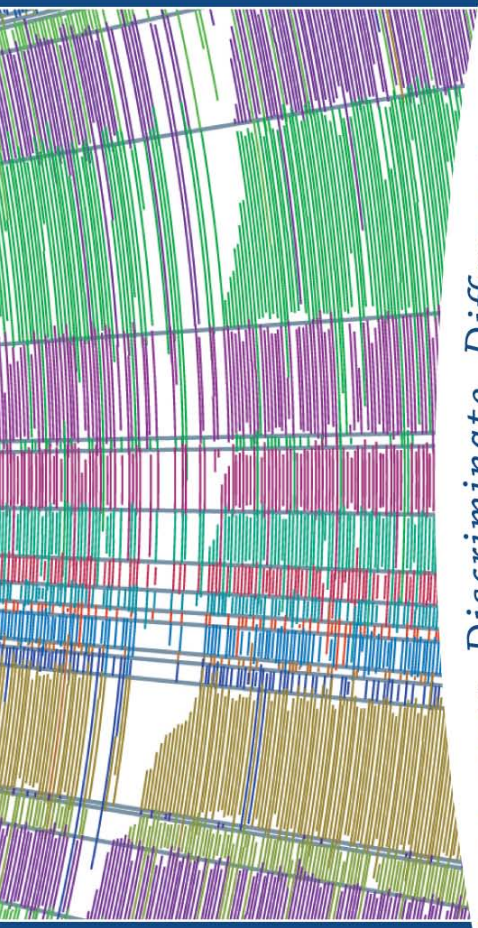
How Optical Mapping Works



Consensus Map

Deep coverage provides
accurate sizing & corrects for:
(a) Incomplete digestion
(b) Random breaks in the DNA

OpGen

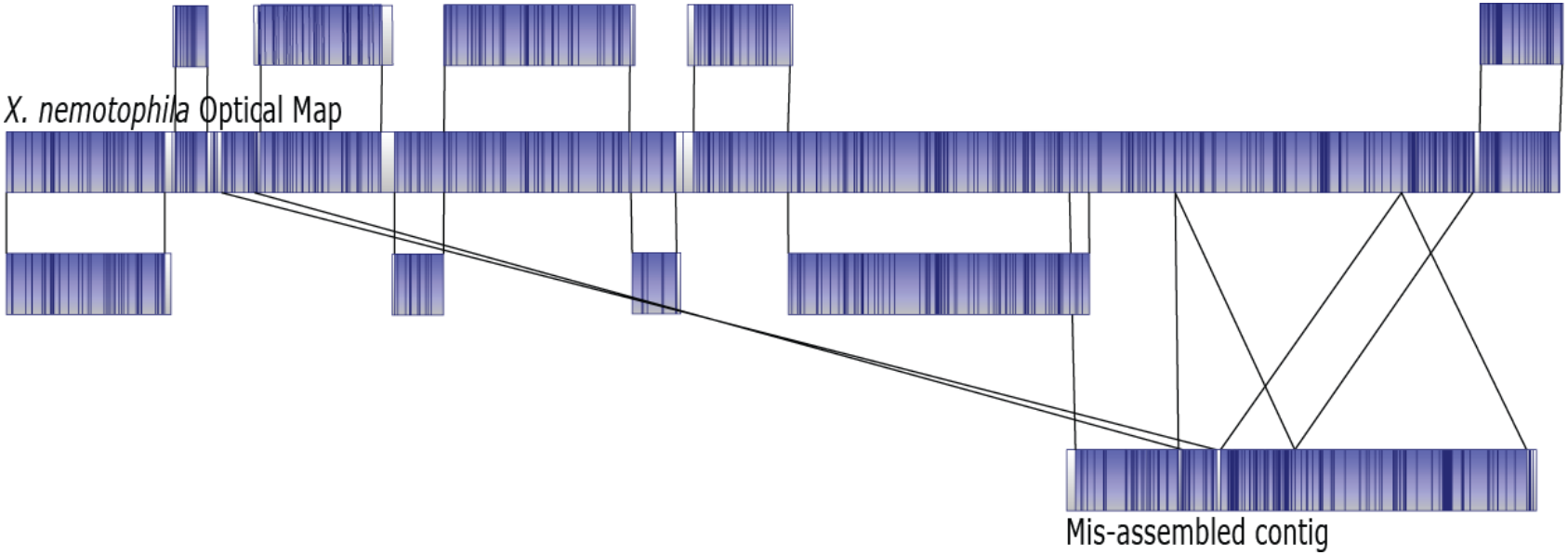


Discover. Discriminate. Differentiate.

Placing Sequence Contigs on an Optical Map

OpGen

Example data



Latreille P, Norton S, Goldman BS, Henkhaus J, Miller N, Barbazuk B, Bode H, Darby G, Du Z, Forst S, Gaudriault S, Goodner B, Goodrich-Blair H, Slater S. “Optical mapping as a routine tool for bacterial genome sequence finishing.” *BMC Genomics* 2007, **8**:321.

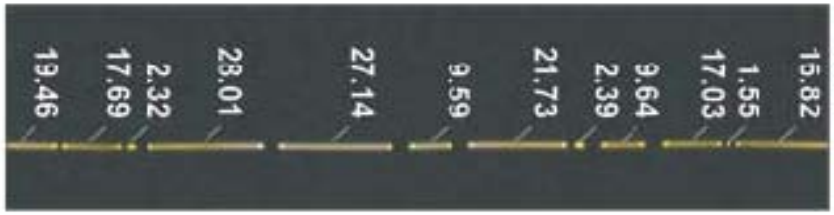
OpGen

Matching Model

Sizing

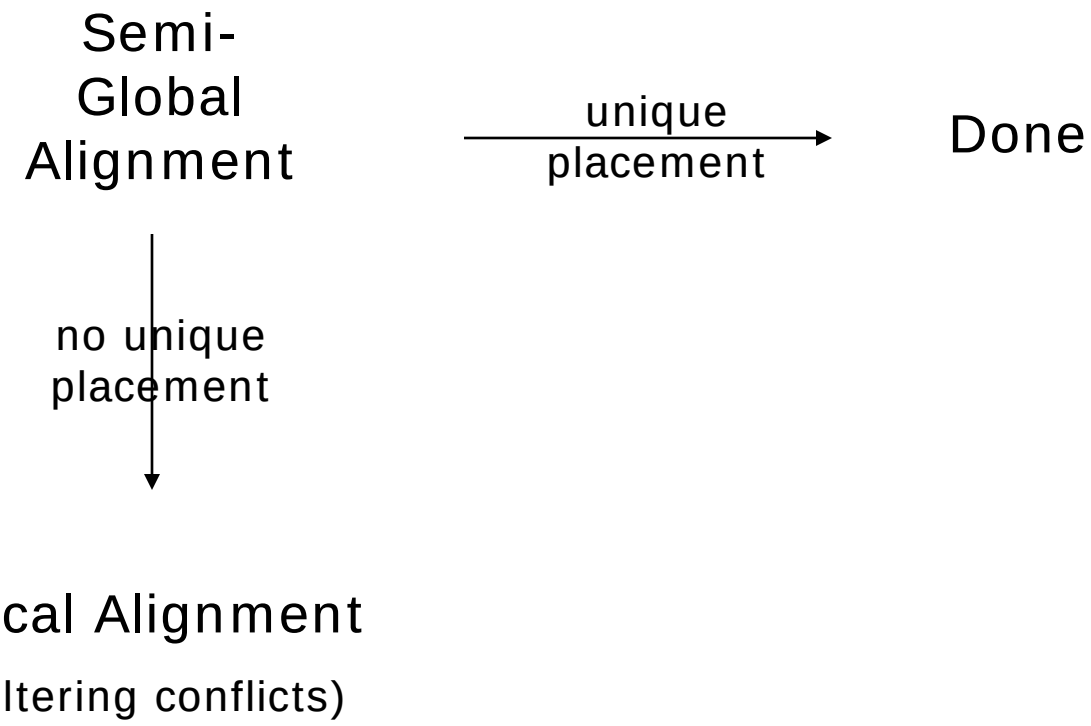
Cut site mismatches

$$\max\begin{cases} 1-\dfrac{|d_M-d_S|}{N\sqrt{d_M V_M}}-C\left(c_M+c_S-\sum_{f\in f_S}e^{-fM}\right) & \text{Normal matches} \\ \max\begin{cases}0\\ d_S\end{cases} & \text{aps} \end{cases}$$



OpGen

Placement logic



Simulation

3 genomes

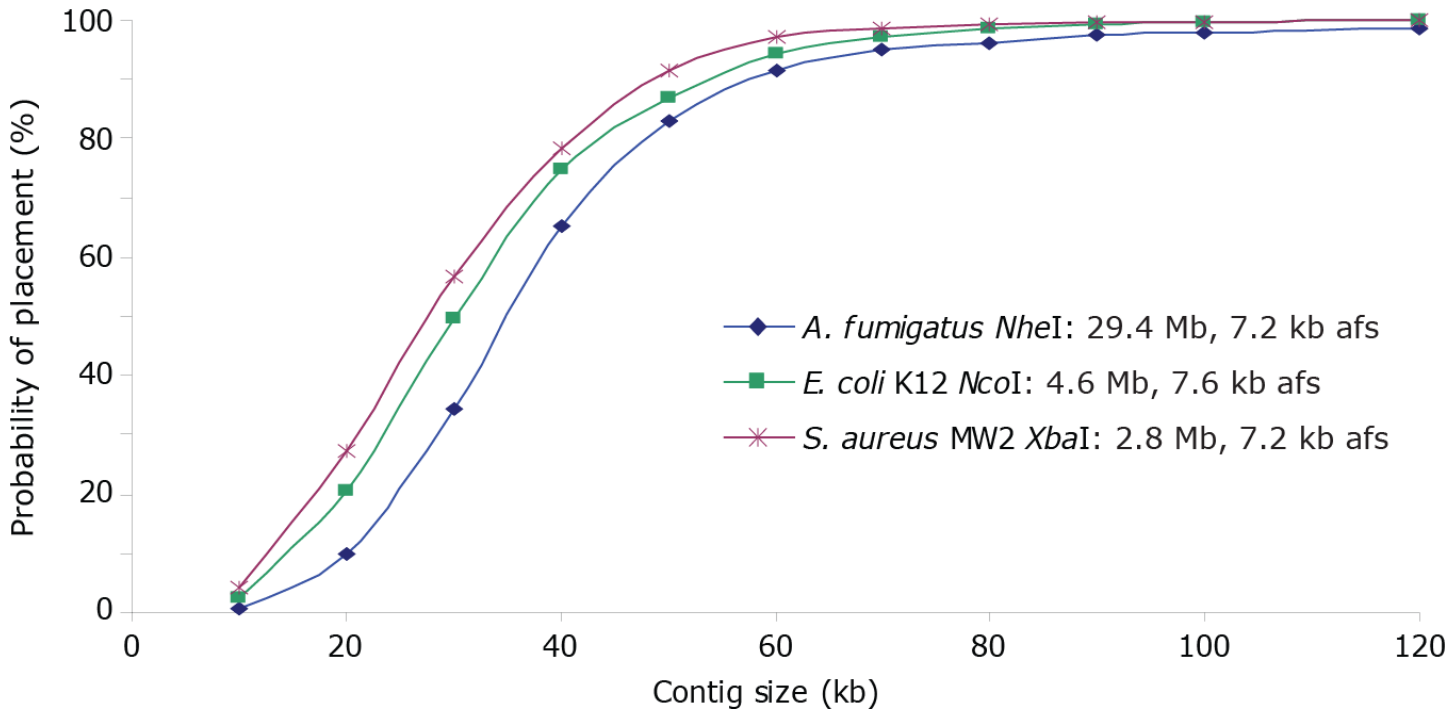
- *A. fumigatus*, *Nhe*I, 29.4 Mb, 7.2 kb afs
- *E. coli* K12, *Nco*I, 4.6 Mb, 7.6 kb afs
- *S. aureus* MW2, *Xba*I, 2.8 Mb, 7.2 kb afs

100 trials for each contig size

100 random contigs for each trial

Placement had to be in the correct location

Simulation results



OpGen

Understanding the placement

How much is this driven by number of cut sites in each contig?

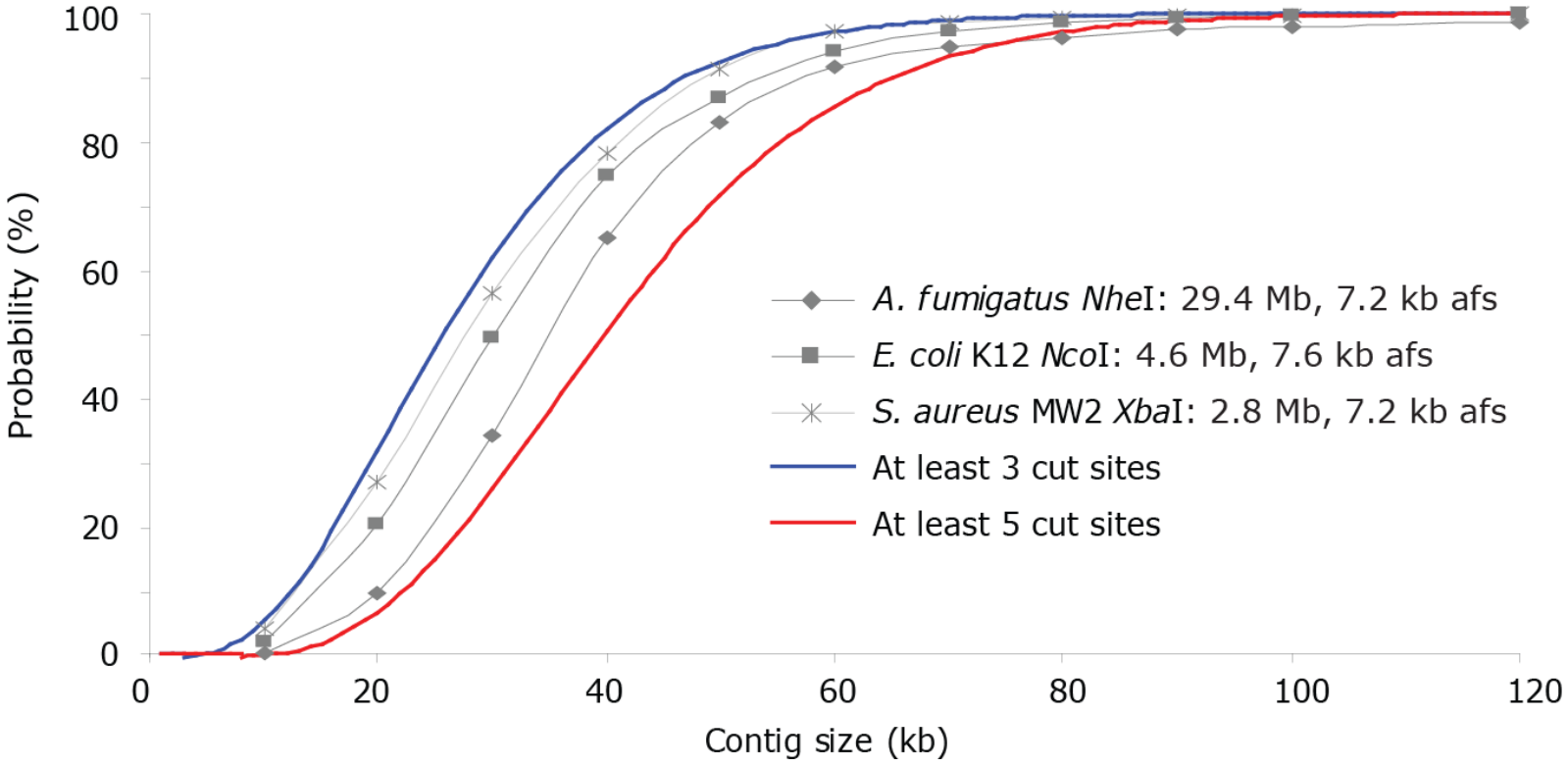
Given the assumption that cut sites are randomly distributed across the genome, the number of cuts per contig is governed by a Poisson distribution:

$$\text{P(at least N cut sites)} = 1 - \sum_{k=0}^{N-1} \frac{e^{-\lambda t}(\lambda t)^k}{k!}$$

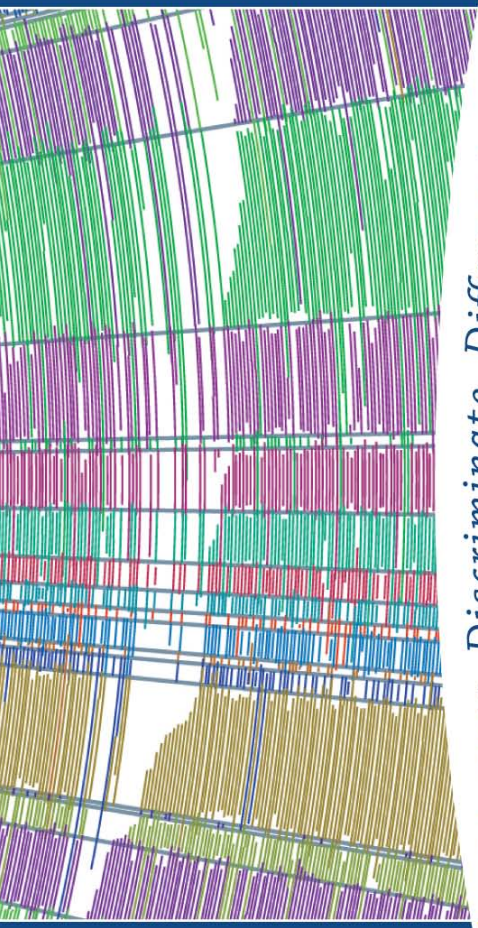
where λ is the enzyme cut rate and t is the size of the contig

OpGen

Understanding the placement



OpGen



Discover. Discriminate. Differentiate.

Questions?

OpGen