PATTERN GENOMICS

An *In Silico* Approach to Designing Toxigenic *Clostridium difficile* Detection Primers

Daniel Fasulo, Ph.D.

daniel.fasulo@patterngenomics.com
(203) 533-9362

Introduction to *In Silico* Assay Design

What is *in silico* assay design?

In silico assay design uses software such as our Daydreamer™ platform to rapidly produce primers and probes that are optimized for a particular diagnostic platform and detection problem.

What are the benefits of *in silico* assay design?

In silico assay design reduces the need for basic research and manual trial-and-error testing in the laboratory. The impact is substantial time and cost savings.

What is the advantage of Daydreamer™?

Many software packages generate primers and probes from a specific sequence, but Daydreamer™ analyzes large collections of draft genomes to find diagnostic patterns and seamlessly expresses them as platform-specific assays. By starting with broad data from many strains, Daydreamer™ finds robust patterns and allows the biologist to focus only on the relevant genomic regions found by the software.

C. difficile Project Overview

Abstract

In this case study, Pattern Genomics' proprietary Daydreamer™ software is utilized to rapidly generate PCR primers specific to toxigenic *Clostridium difficile*.

Relevance

Clostridium difficile is a common hospital-related infection that sickens more than half a million and kills 14,000 people in the US per year. Up to 20% of patients may carry *C. difficile* asymptomatically, and it can be a minor component of normal gut flora. There is a need for rapid diagnostics to detect carriers entering the hospital and provide diagnosis for patients with diarrhea.

Approach

We obtained the complete and draft sequences of 198 *Clostridium* genomes from GenBank derived from 87 toxigenic *C. difficile* strains, 23 non-toxigenic strains, and 88 other *Clostridium* species. We focus on developing PCR primers that are specific to toxigenic *C. difficile*.

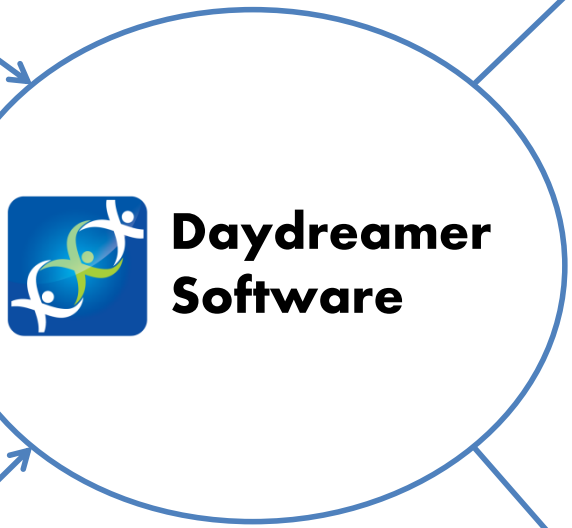
Use of Daydreamer™ Software

Input Run Configuration

198 whole genome sequences

Label toxigenic *C. difficile* as "in" and everything else as "out". Software targets "in" set.

Provide Tm range, amplicon size limits, GC% limits, etc. specific to intended DX platform.



50 minutes on commodity Intel workstation hardware

Outputs

Smallest set of primers pairs that meet design requirements (minimum of 8 candidates).

Comparative genomic output

Results: *In Silico* PCR

Daydreamer™ finds a set of 8 primer pairs that amplify all toxigenic *C. difficile* without cross-reactivity to the other genomes. If no such primers existed, Daydreamer™ would use more pairs to cover subsets of the targeted genomes.

	A	B	C	D	E	F	G	H	I	J
1	GENOME	TYPE	PP0	PP1	PP2	PP3	PP4	PP5	PP6	PP7
71	CD_P25	IN	X	X	X	X	X	X	X	X
72	CD_P32	IN	X	X	X	X	X	X	X	X
73	CD_P33	IN	X	X	X	X	X	X	X	X
74	CD_P36	IN	X	X	X	X	X	X	X	X
75	CD_P46	IN	X	X	X	X	X	X	X	X
76	CD_P59	IN	X	X	X	X	X	X	X	X
77	CD_P64	IN	X	X	X	X	X	X	X	X
78	CD_P69	IN	X	X	X	X	X	X	X	X
79	CD_P71	IN	X	X	X	X	X	X	X	X
80	CD_P72	IN	X	X	X	X	X	X	X	X
81	CD_P74	IN	X	X	X	X	X	X	X	X
82	CD_P78	IN	X	X	X	X	X	X	X	X
83	CD_Y10	IN	X	X	X	X	X	X	X	X
84	CD_Y215	IN	X	X	X	X	X	X	X	X
85	CD_Y231	IN	X	X	X	X	X	X	X	X
86	CD_Y247	IN	X	X	X	X	X	X	X	X
87	CD_Y358	IN	X	X	X	X	X	X	X	X
88	CD_Y41	IN	X	X	X	X	X	X	X	X
89	CD_CD169	OUT	.	.	.	.	.	.	.	.
90	CD_CD178	OUT	.	.	.	.	.	.	.	.
91	CD_CD34	OUT	.	.	.	.	.	.	.	.
92	CD_CD37	OUT	.	.	.	.	.	.	.	.
93	CD_DA00134	OUT	.	.	.	.	.	.	.	.
94	CD_DA00141	OUT	.	.	.	.	.	.	.	.
95	CD_DA00196	OUT	.	.	.	.	.	.	.	.
96	CD_DA00215	OUT	.	.	.	.	.	.	.	.
97	CD_F200	OUT	.	.	.	.	.	.	.	.
98	CD_F480	OUT	.	.	.	.	.	.	.	.

Columns C through J represent 8 primer pairs.

'X' = expected amplification

'.' = no expected amplification

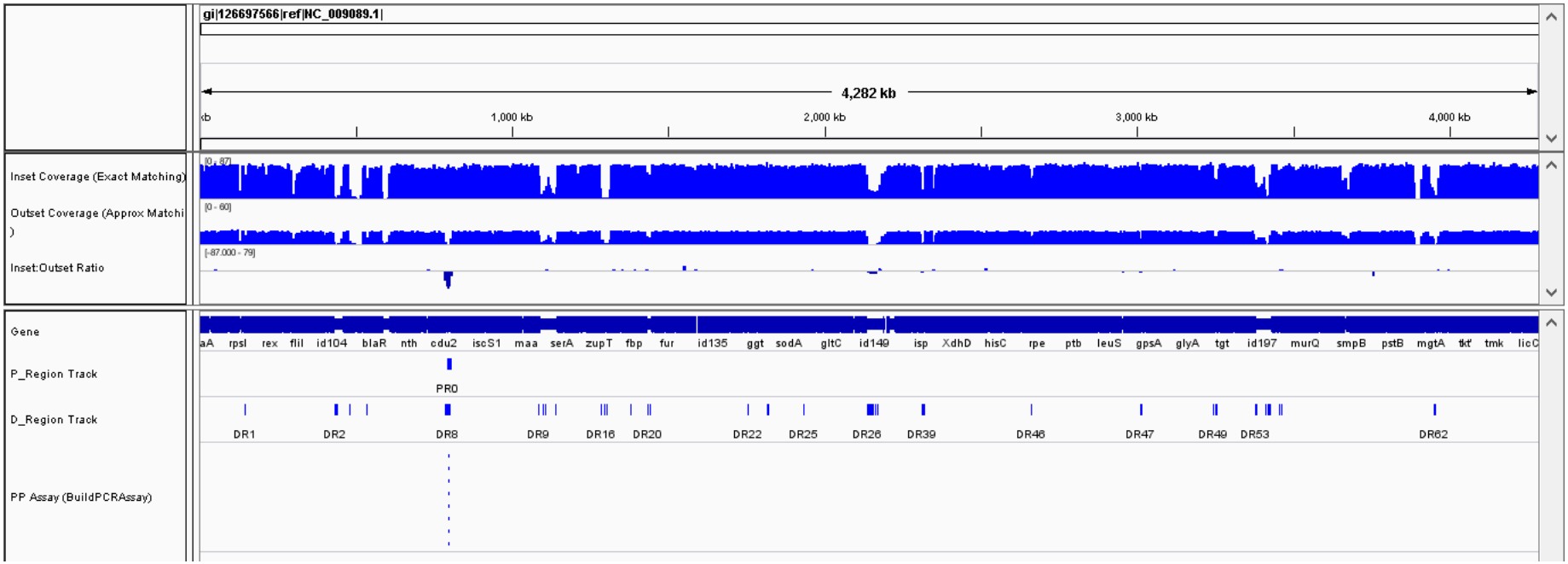
The rows are genomes in the analysis. Only a subset is shown here.

Visualization Overview

Daydreamer™ produces “BED” files that can be loaded into common genome browsers such as the Broad Institute’s IGV or GenBank’s integrated ENSEMBL browser.

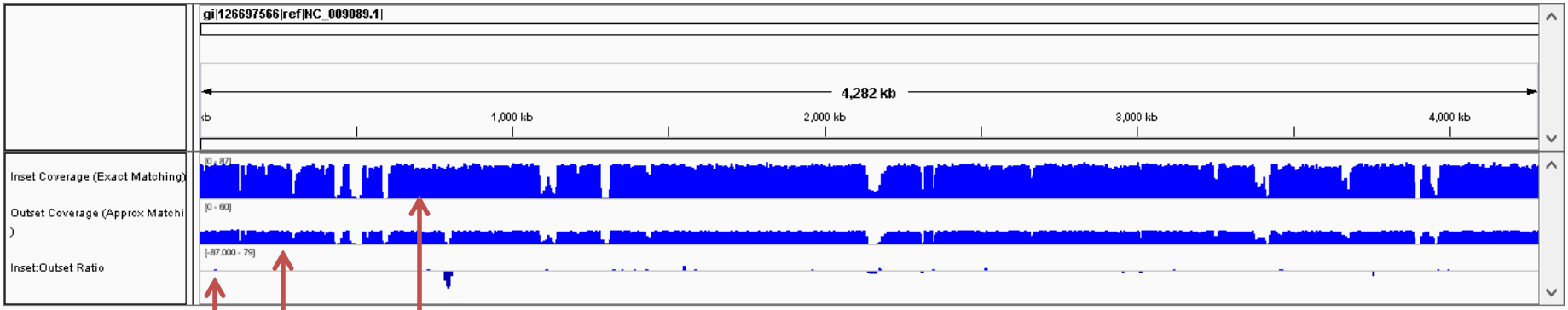
The data can be viewed in the context of any “in” genome in the analysis. Below we are viewing the RefSeq *Clostridium difficile* 630 genome using IGV.

The information being visualized is explained on subsequent slides.



Coverage Histograms

The “coverage histograms” allow one to visually identify regions that are different between the “in” and “out” genome sets, and find areas where other subtypes uncorrelated with “in” and “out” may be present.

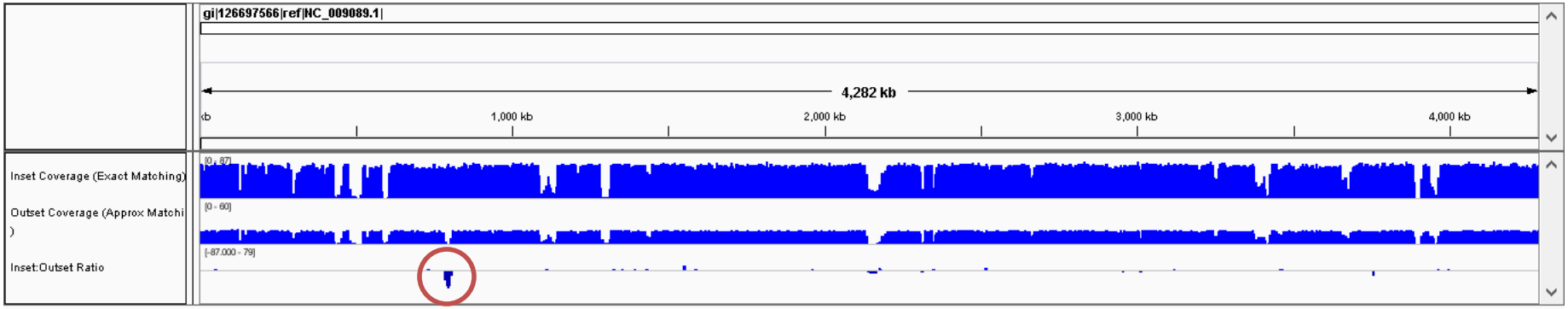


The “inset coverage” shows the number of genomes with the “in” label that are similar to this genome at each position.

The “outset coverage” is similar, except with respect to “out” genomes.

The “ratio histogram” calculates the ratio of “in” to “out” coverage, making differences between the two genome categories more obvious. Note that if the “out” coverage is 0 then a denominator of -1 is used, creating a *downward* spike in the plot corresponding to regions unique to the inset.

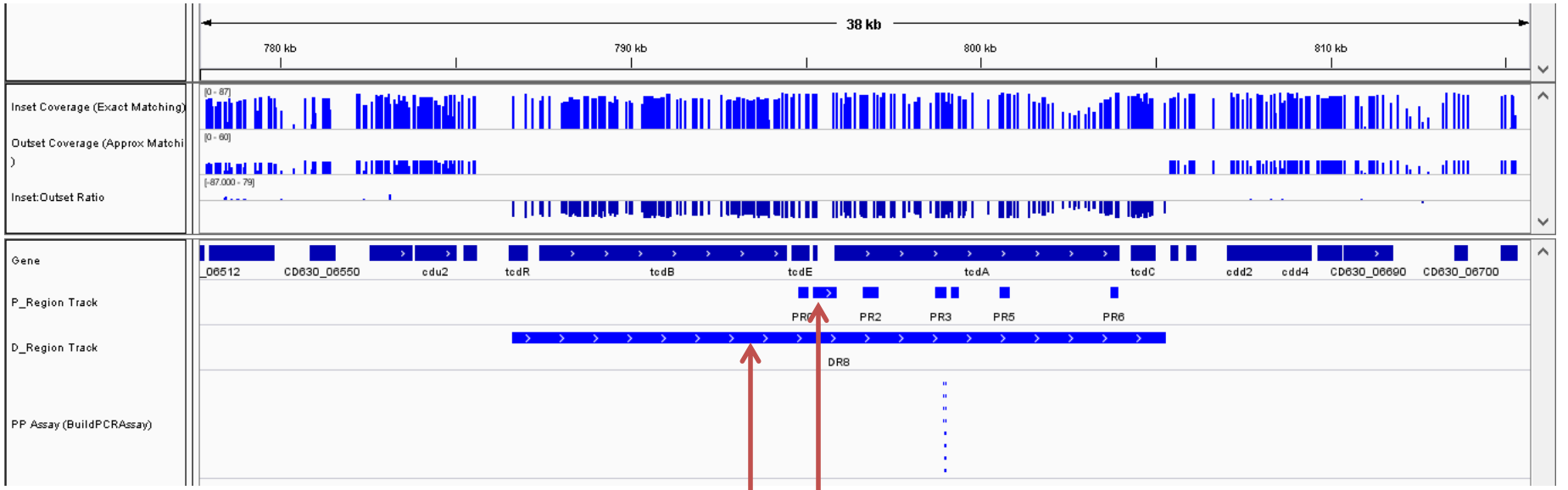
Utilizing Coverage Histograms



There is a strong negative spike here; let's zoom in on this region and explore some of the other visualization data.

Zoom in on Discriminative Region

Daydreamer™ has found the region containing the well-known toxigenicity genes; we see a 38kb portion of the genome containing these genes below.



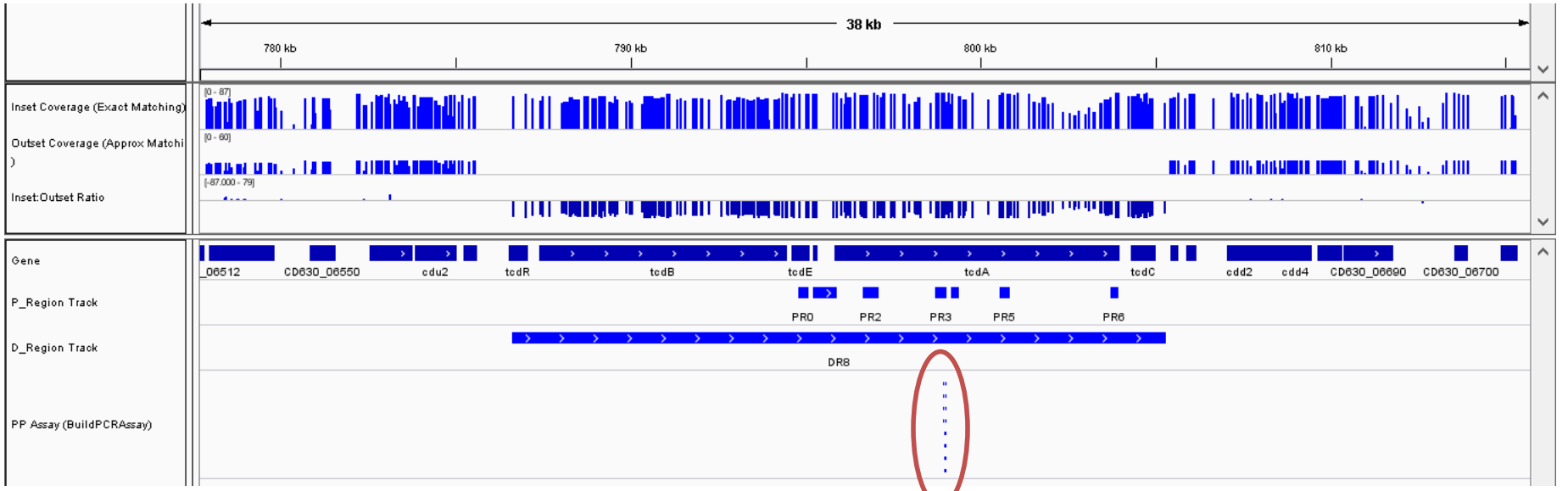
"D_Region" is a *discriminative* region: the candidate primers here are generally found in *some* "in" genomes and *no* "out" genomes.

If no perfect regions are found, then primers must be selected from multiple D_regions.

"P_Region" is a *perfect* region: the candidate primers here are generally found in *every* "in" genome and *no* "out" genomes.

<u>Primer choice matters</u>: in this data set, only 881 potential primers of 99,919, 393 considered are perfect.

Primer Selection



These are the locations of the final primer pairs selected by Daydreamer™ targeting a perfect region of the *tcdA* gene. A variety of heuristics are used to pick the precise positions.

Note that Daydreamer™ does not currently use annotation data, and has no prior knowledge of toxigenicity (or any other) genes. These regions were all discovered by naïve sequence analysis on a large number of genomes.

BLAST Verification

BLAST confirms that the template region for a representative primer set is specific to toxigenic *C. difficile* (shown below); the individual primers also have no matches greater than 85% length / identity to any other target (not shown).

Sequences producing significant alignments:

Select: All None Selected:0

Alignments Download GenBank Graphics Distance tree of results

Description	Max score	Total score	Query cover	E value	Ident	Accession
Clostridium difficile PaLoc containing novel insertion between tcdE and tcdA, strain Ox1485	99.0	99.0	100%	2e-18	100%	HG002394.1
[Clostridium] difficile strain ZP77 toxin A (tcdA) gene, complete cds	99.0	99.0	100%	2e-18	100%	KC292125.1
[Clostridium] difficile strain VPI10463 toxin A (tcdA) gene, complete cds	99.0	99.0	100%	2e-18	100%	KC292122.1
[Clostridium] difficile strain UK1 toxin A (tcdA) gene, complete cds	99.0	99.0	100%	2e-18	100%	KC292117.1
[Clostridium] difficile strain ZP50 toxin A (tcdA) gene, complete cds	99.0	99.0	100%	2e-18	100%	KC292116.1
[Clostridium] difficile strain ZP17 toxin A (tcdA) gene, complete cds	99.0	99.0	100%	2e-18	100%	KC292104.1
[Clostridium] difficile strain ZP80 toxin A (tcdA) gene, complete cds	99.0	99.0	100%	2e-18	100%	KC292092.1
[Clostridium] difficile strain ZP74 toxin A (tcdA) gene, complete cds >gb|KC292084.1| [Clostridium] difficile strain ZP30 toxin A (tcdA) gene, complete cds >gb|KC2	99.0	99.0	100%	2e-18	100%	KC292072.1
Clostridium difficile ATCC 9689 toxin A (tcdA) gene, complete cds >gb|KC292071.1| [Clostridium] difficile strain JN031 toxin A (tcdA) gene, complete cds >gb|KC2	99.0	99.0	100%	2e-18	100%	KC292070.1
[Clostridium] difficile strain ZP81 toxin A (tcdA) gene, complete cds >gb|KC292099.1| [Clostridium] difficile strain ZP75 toxin A (tcdA) gene, complete cds >gb|KC2	99.0	99.0	100%	2e-18	100%	KC292062.1
[Clostridium] difficile strain GZ1 toxin A (tcdA) gene, complete cds >gb|KC292065.1| [Clostridium] difficile strain SH12 toxin A (tcdA) gene, complete cds >gb|KC29	99.0	99.0	100%	2e-18	100%	KC292061.1
[Clostridium] difficile strain ZP48 toxin A (tcdA) gene, complete cds >gb|KC292080.1| [Clostridium] difficile strain ZP5 toxin A (tcdA) gene, complete cds >gb|KC29	99.0	99.0	100%	2e-18	100%	KC292059.1
[Clostridium] difficile strain GZ5 toxin A (tcdA) gene, complete cds >gb|KC292066.1| [Clostridium] difficile strain GZ7 toxin A (tcdA) gene, complete cds	99.0	99.0	100%	2e-18	100%	KC292058.1
[Clostridium] difficile strain SH8 toxin A (tcdA) gene, complete cds >gb|KC292069.1| [Clostridium] difficile strain SH5 toxin A (tcdA) gene, complete cds >gb|KC29	99.0	99.0	100%	2e-18	100%	KC292057.1
[Clostridium] difficile strain GZ14 truncated toxin A (tcdA) gene, complete cds >gb|KC292060.1| [Clostridium] difficile strain ZP47 truncated toxin A (tcdA) gene, co	99.0	99.0	100%	2e-18	100%	KC292056.1
[Clostridium] difficile strain JF09 TcdA (tcdA) gene, complete cds	99.0	99.0	100%	2e-18	100%	JQ809335.1
Clostridium difficile BI9 chromosome	99.0	99.0	100%	2e-18	100%	FN668944.1
Clostridium difficile BI1 chromosome, complete sequence	99.0	99.0	100%	2e-18	100%	FN668941.1
Clostridium difficile M68, complete genome	99.0	99.0	100%	2e-18	100%	FN668375.1
Clostridium difficile complete genome, strain 2007855	99.0	99.0	100%	2e-18	100%	FN665654.1
Clostridium difficile complete genome, strain M120	99.0	99.0	100%	2e-18	100%	FN665653.1
Clostridium difficile complete genome, strain CF5	99.0	99.0	100%	2e-18	100%	FN665652.1
Clostridium difficile R20291 complete genome	99.0	99.0	100%	2e-18	100%	FN545816.1
Clostridium difficile CD196 complete genome, strain CD196	99.0	99.0	100%	2e-18	100%	FN538970.1
Clostridium difficile 630 complete genome	99.0	99.0	100%	2e-18	100%	AM180355.1
C.difficile cdu2, cdu1, tcdD, tcdB, tcdE, tcdA, tcdC, cdd1, cdd2, cdd3, and cdd4 genes	99.0	99.0	100%	2e-18	100%	X92982.1
C.difficile toxin A gene, complete cds	99.0	99.0	100%	2e-18	100%	M30307.1
Clostridium difficile toxA gene for Toxin A and flanks containing an unidentified reading frame	99.0	99.0	100%	2e-18	100%	X51797.1

Conclusion

- We have shown that Daydreamer™ is capable of designing relevant strain typing PCR primer sets from large scale sequencing data in a matter of minutes to hours.

- Detection of toxigenic *C. difficile* is predicted to be accomplished with only a single PCR reaction using carefully selected primers. In other cases, combinations of markers may be needed. See for example our published, lab-tested results on methicillin-resistant *Staphylococcus aureus*.

- Once raw assembly data is available, a combination of powerful, efficient algorithms and useful visualization output enable a single user to create a computationally optimized design that is ready to be tested in the lab.

- Although not shown in this presentation, the software also has the capability to generate probe sequences and help the user identify mis-labeled or otherwise atypical samples.

PATTERN
GENOMICS