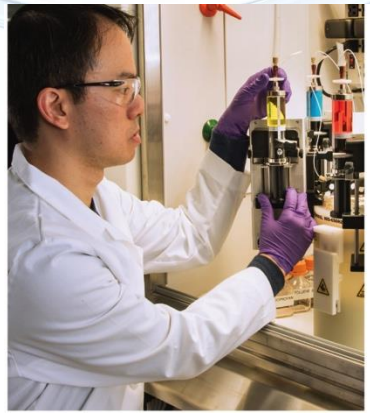
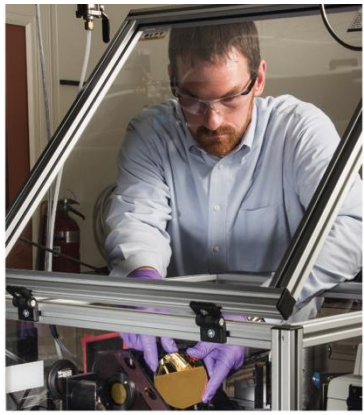
Standards for NGS-Based Microbial Strain Identification





Scott Jackson
Leader, Complex Microbial Systems Group
Biosystems and Biomaterials Division
MML

MML Microbial Metrology Mission Statement

"To develop advanced measurements that will permit the exploitation of microbes to promote human health, precision medicine and advanced manufacturing"

NIST
National Institute of Standards and Technology
U.S. Department of Commerce

MATERIAL MEASUREMENT LABORATORY

WHOLE GENOME SEQUENCING AS AN INDUSTRY STANDARD FOR DEMONSTRATING IDENTITY AND UNIQUENESS OF PROBIOTIC STRAINS

Microbial Genomes 101

- Most microbial genomes are a single chromosome, circular
  - Some microbes have 2 (or more) chromosomes, but this is rare
  - Plasmids, usually under 100kb, are common in bacteria

- The average size of a microbial genome is ~5 million nucleotides; but this can range from ~2 million to ~7 million

- Strain-level genomic diversity varies depending on the species
  - In *E. coli*, two isolates (strains) may differ by 50% of their genome
  - In *B. anthracis*, two isolates (strains) that are temporally and geographically distinct may differ by only a few nucleotides (0.00001% different)

MICROBIAL GENOMIC REFERENCE MATERIALS

Nate Olson

STRAIN SELECTION

Strain	Reasoning		Size (bp)[1]	GC%[1]
Salmonella enterica LT2[2]	Common foodborne pathogen	Chromosome	4.8 Mb	52
		Plasmid	94 kb	53
Staphylococcus aureus	Ubiquitous opportunistic pathogen	Chromosome	2.8 Mb	33
	Clinical Isolate from CNH[3]	Plasmid	25 kb	29
Pseudomonas aeruginosa	High GC content	Chromosome	6.3 Mb	67
	Clinical Isolate from CNH[3]			
Clostridium sporogenes[4]	Low GC content	Chromosome	4.1 Mb	28

[1] Genome size and GC content from `http://www.ncbi.nlm.nih.gov/genome`

[2] Full Name *Salmonella enterica* subspecies enterica serovar Typhimurium LT2

[3] Children's National Hospital

[4] Information based on draft assembly

PRODUCTION

Produced by local vendor

For each strain

- pure culture
- single batch of DNA
- ~ 1500 vials
- 3µg per vial

CHARACTERIZED PROPERTIES

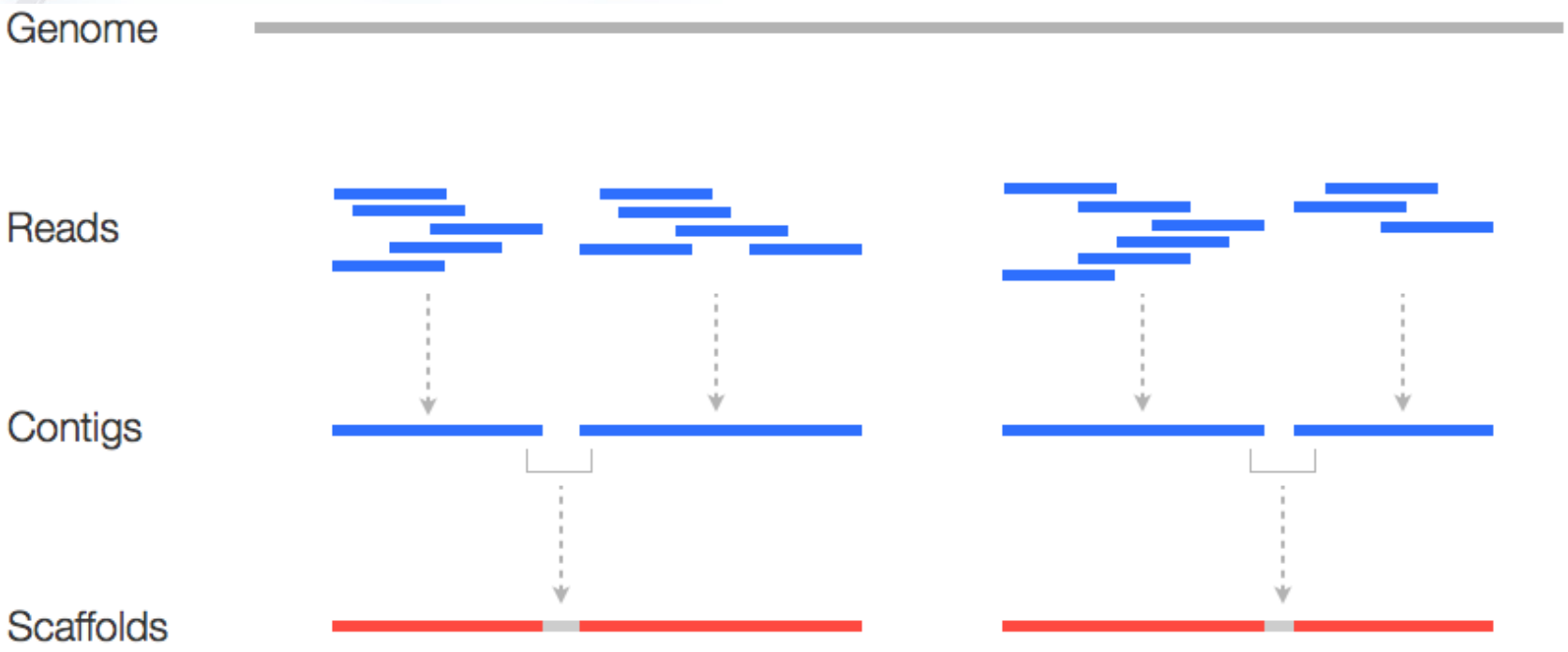
1. Genome Assembly
2. Base Level Purity
3. Genomic Contaminants
4. DNA Stability

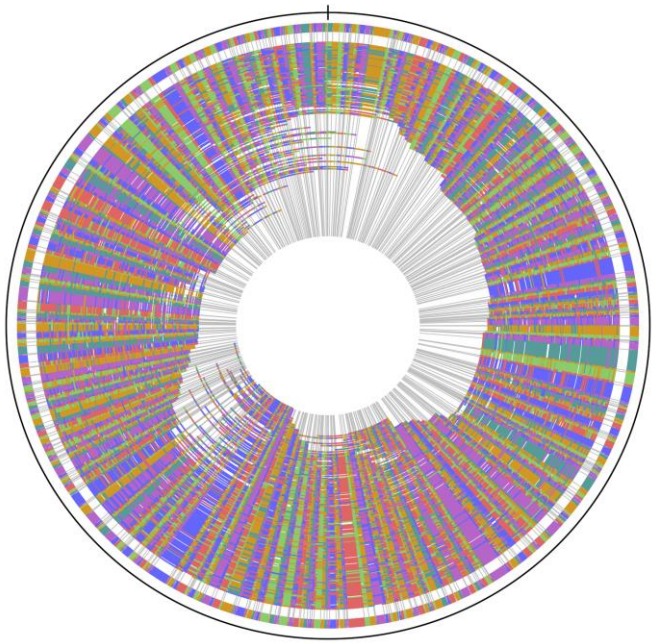
NEXT-GENERATION SEQUENCING

Seq Platform	Vials	Libraries	Read Length	Targeted Coverage Library	Total
Pac Bio RSII	1	1	8 kb		200
Illumina MiSeq	8[a]	2	2 X 300 bp	175	2800
Ion Torrent PGM	8[a]	1	400 bp	37.5	600
				Total Coverage:	**3600**

[a] The same vials were sequenced with both platforms

HIGH QUALITY GENOME ASSEMBLY





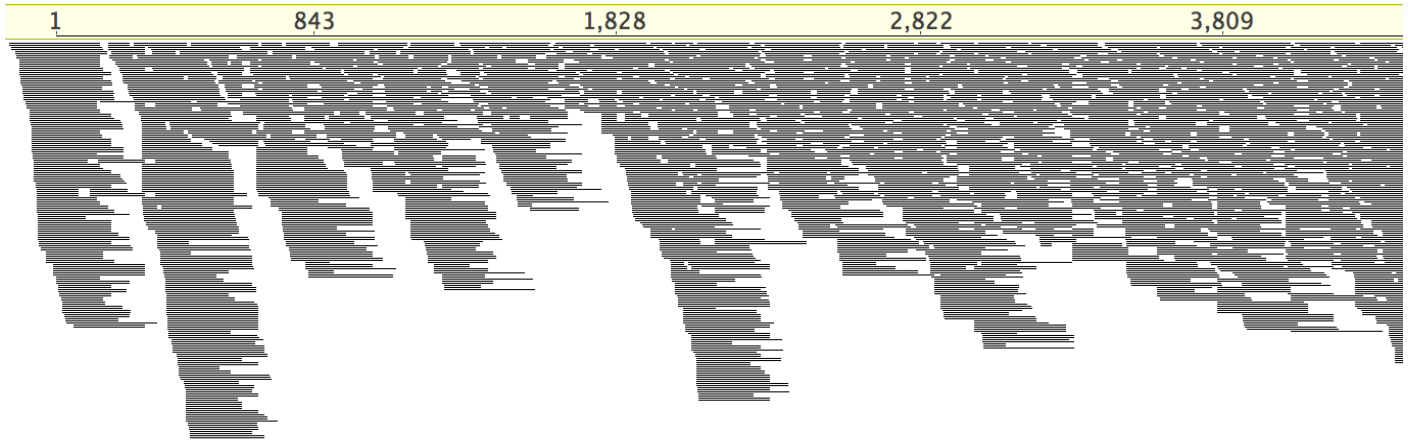
CHARACTERIZED PROPERTIES

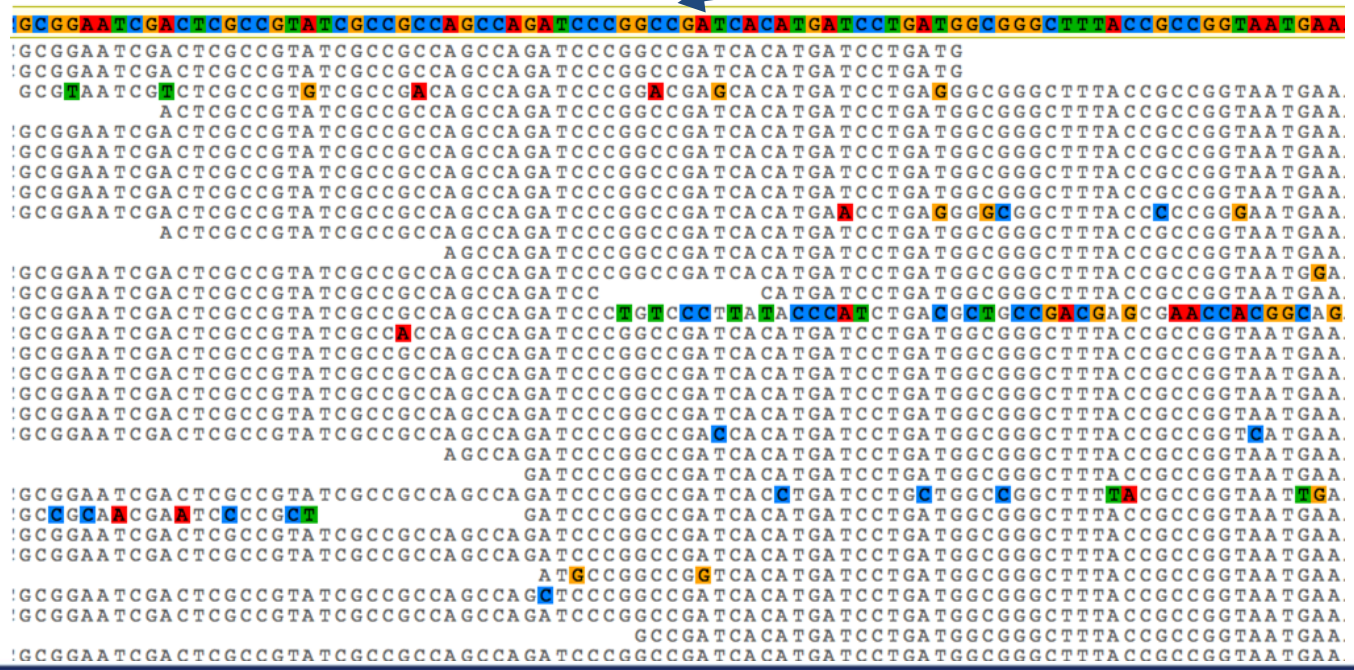
1. Genome Assembly
2. Base Level Purity
3. Genomic Contaminants
4. DNA Stability

Calculate Base-Level Purity

Sequencing Reads

Align to Reference

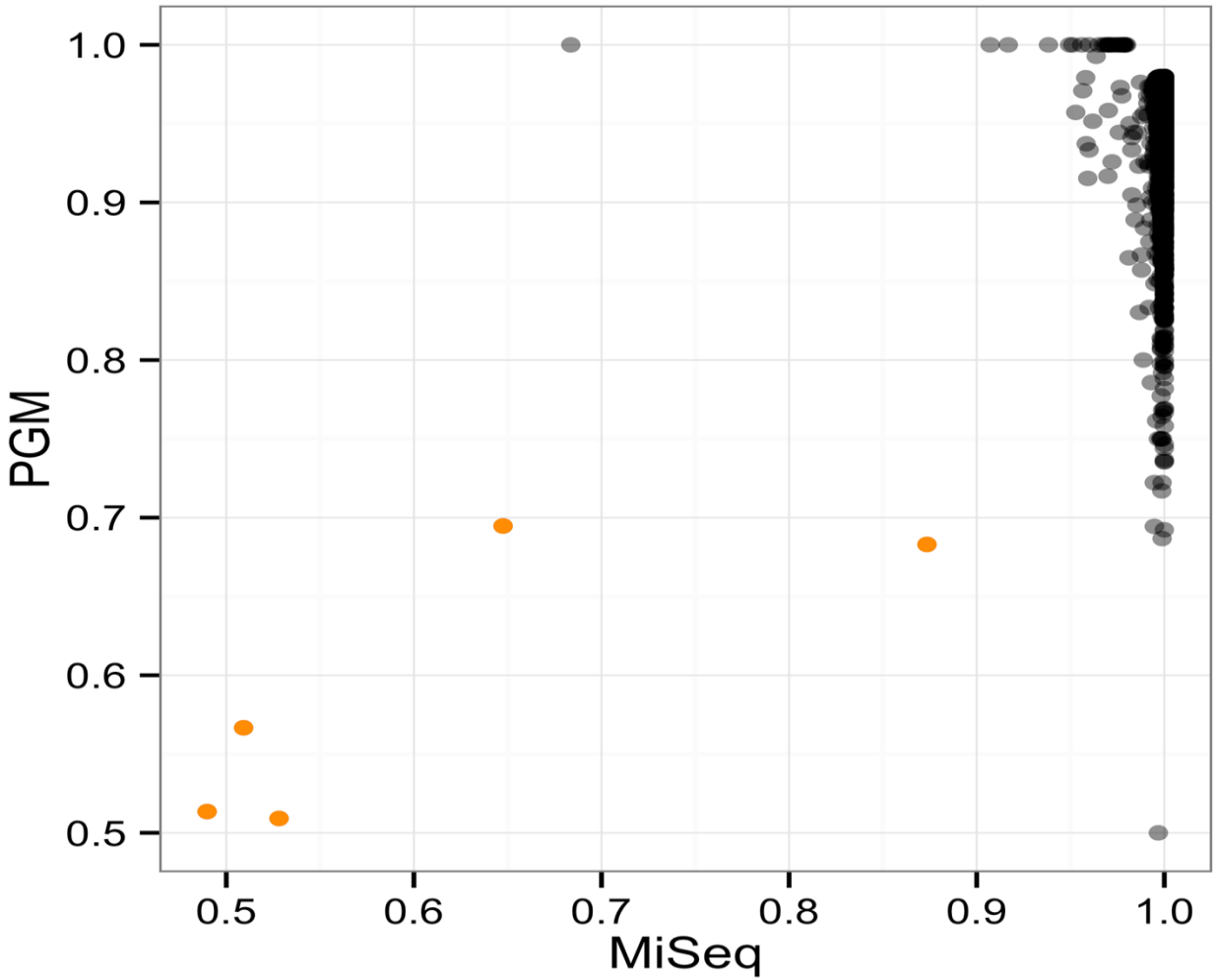




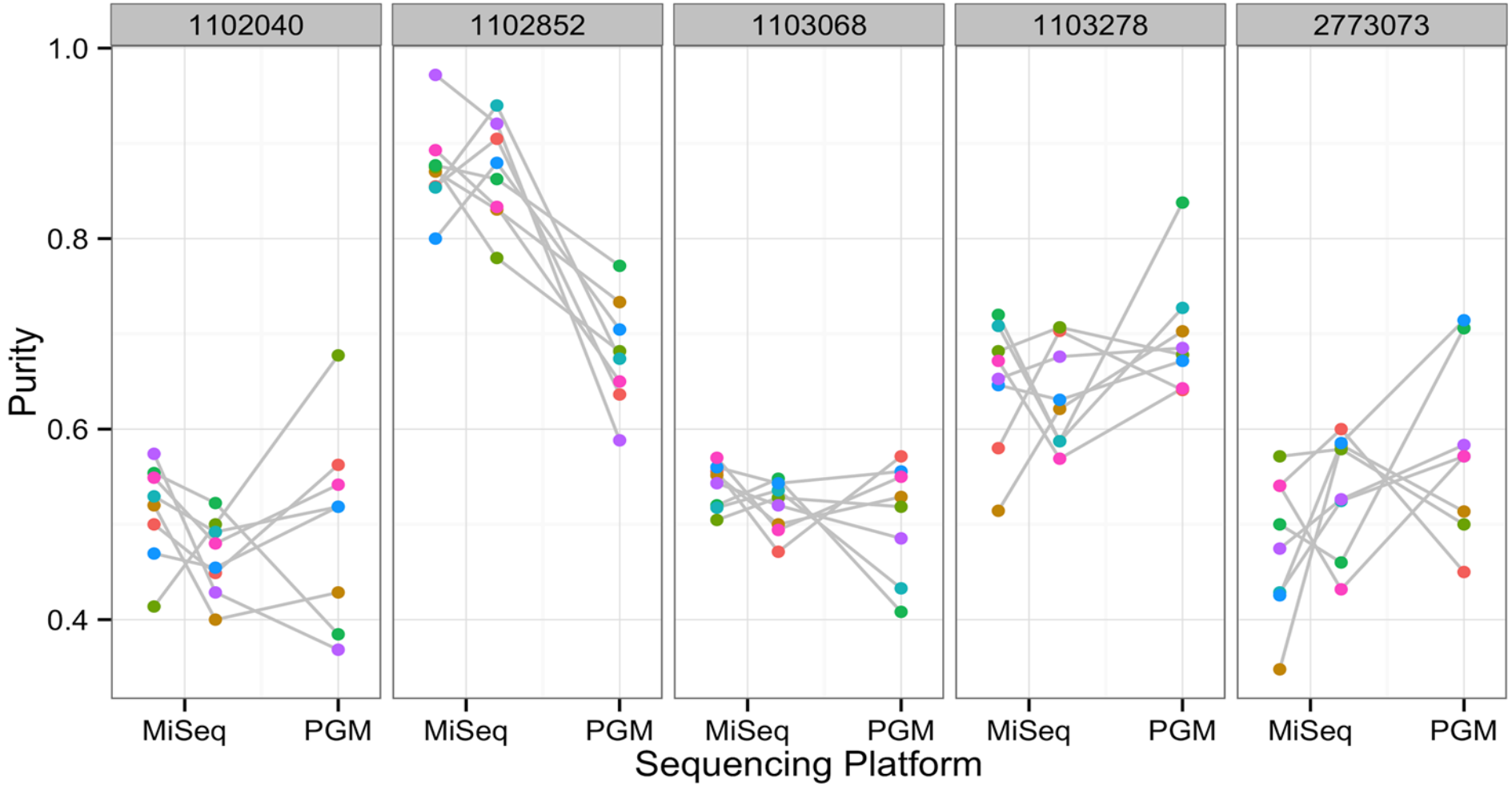
BASE LEVEL PURITY: RESULTS MG001

- 19 out of 4.8 Mb have purity values less than 0.98 for both platforms

- 5 positions with purity less than 0.95



BASE LEVEL PURITY: RESULTS MG001



There’s no such thing as a pure culture

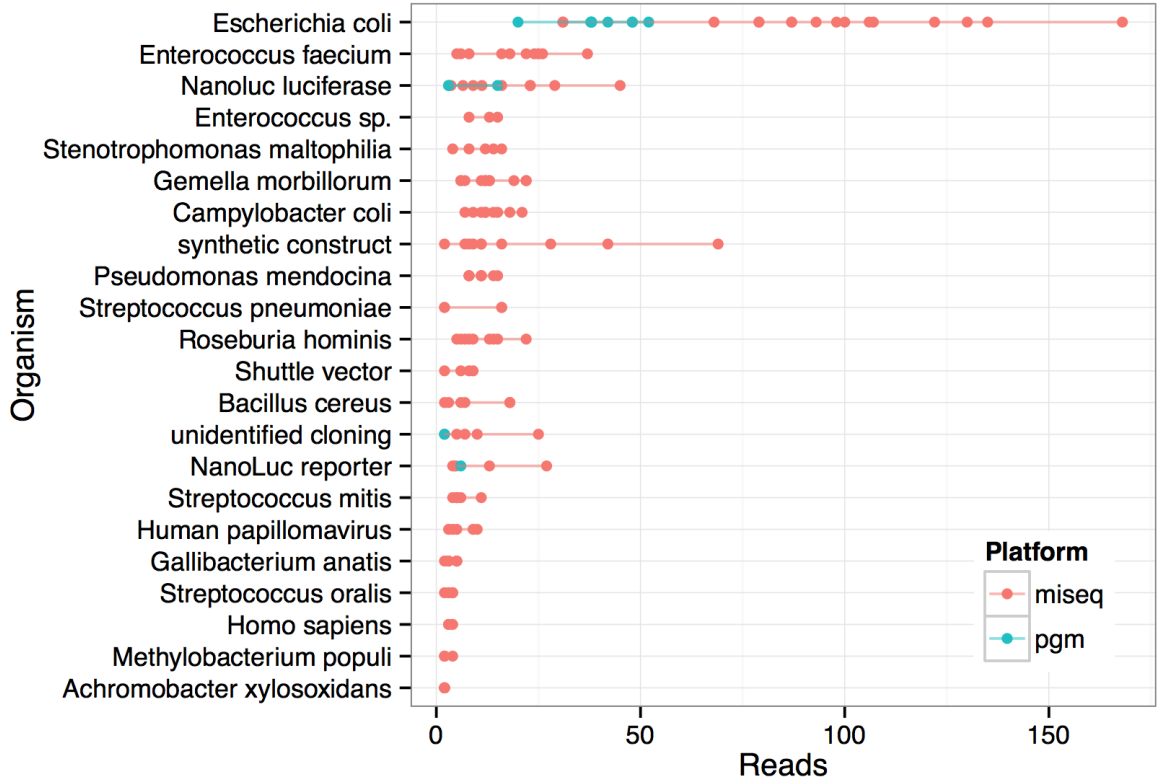
Someone, Sometime, Somewhere

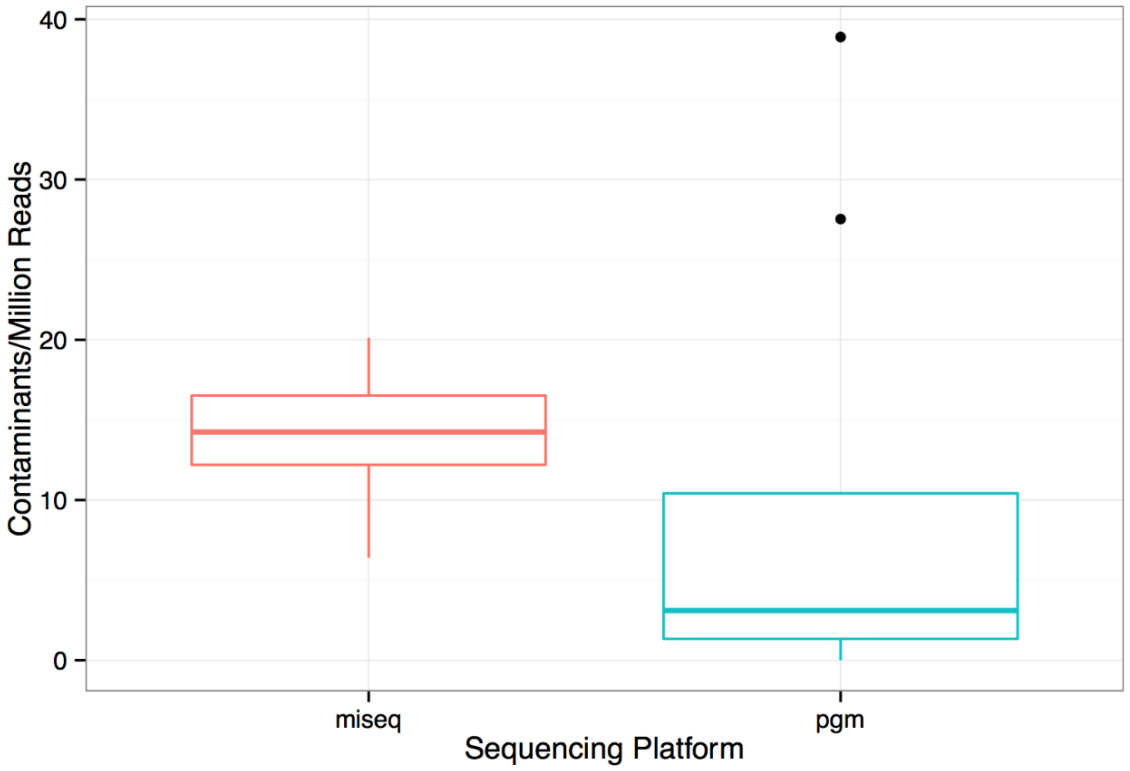
CHARACTERIZED PROPERTIES

1. Genome Assembly
2. Base Level Purity
3. Genomic Contaminants
4. DNA Stability

ORGANISM-LEVEL PURITY ANALYSIS OF "PURE" SALMONELLA GENOMIC DNA

Name	Strain	Biosample	Size	%GC
MG001	*Salmonella enterica* LT2	SAMN02854572	4.8 Mb	52





Analysis done via Pathoscope

GENOMIC CONTAMINANTS: CONCLUSIONS

Likely contaminant sources

- Sequencing reagents
- Bioinformatic errors

Fit for purpose

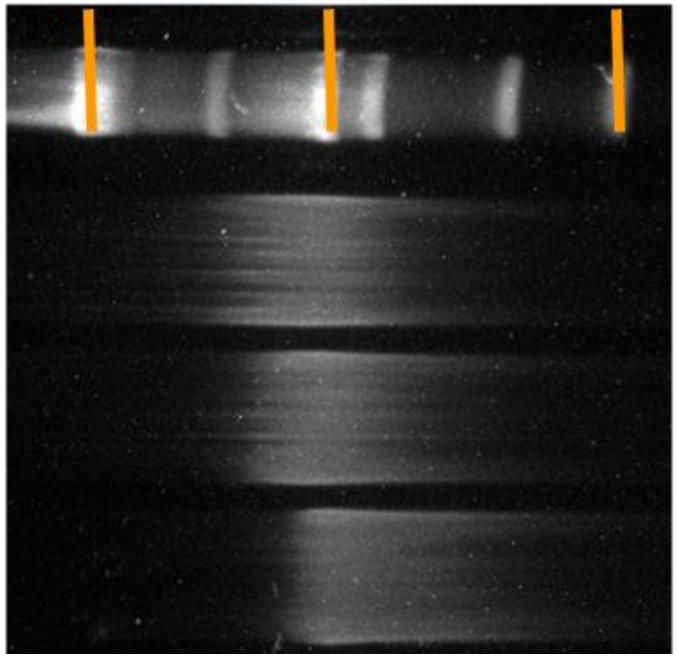
- 99.995% minimum genomic purity

CHARACTERIZED PROPERTIES

1. Genome Assembly
2. Base Level Purity
3. Genomic Contaminants
4. DNA Stability

DNA Stability: Methods

37°C Treatment



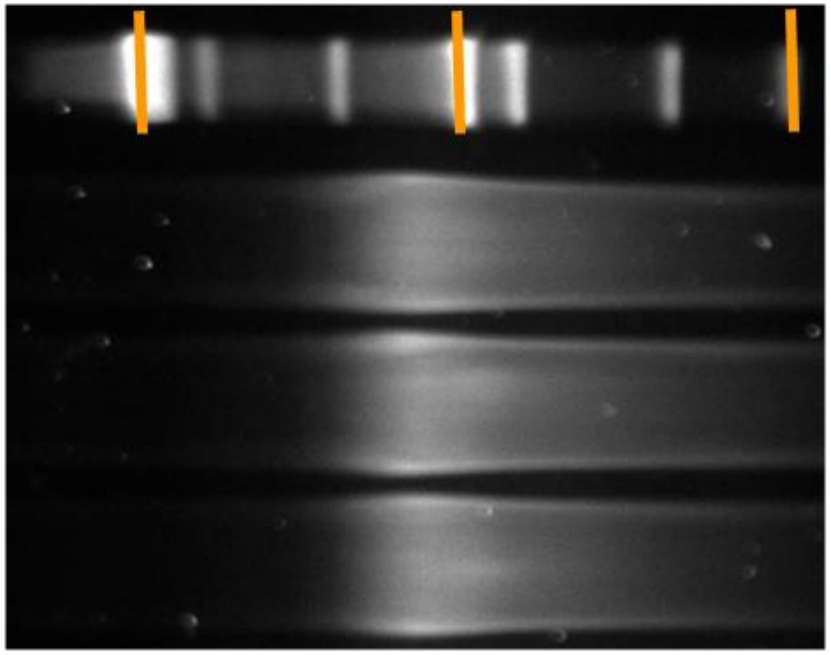
Ladder

Control

2 weeks

8 weeks

4°C Treatment



Ladder

Control

2 weeks

8 weeks

COMPUTATIONAL REPRODUCIBILITY

Data

NCBI

Code

GitHub

Environment

docker

PROPOSED STANDARD FOR THE PROBIOTIC INDUSTRY

- A High Quality “Finished” Genome Sequence Must be Provided for Each Strain
- This only need be done once, on the master cell stock
- Currently costs $1000-$5000 per strain
- Each of these reference genomes be deposited to a database that’s accessible to other manufacturers

RELEASE ASSAYS

- PCR assays can be designed based on the whole genome sequence data and can be used for release assays

- But consider, a quick and dirty NGS run of the release product can also demonstrate identity (99.9% confidence in identify) relative to the reference genome, for $100.

NIST-NRC POSTDOCTORAL POSITIONS AVAILABLE

National Institute of Standards and Technology
U.S. Department of Commerce

National Research Council (NRC)

- US Citizens ONLY
- Within 5 years of PhD
- 2 years @ ~$70k/y

Inquire for details:

Scott Jackson scott.jackson@nist.gov

QUESTIONS?

Scott Jackson scott.jackson@nist.gov

#NISTPathogen #NISTMicrobiome