Characterizing the performance of the MinION for real-time detection

SAND2018-2680C

Raga Krishnakumar

Michael Bartsch

Sara Bird

Steven Branda

Harrison Edwards

Harikrishnan Jayamohan

Kamlesh Patel

Joe Schoeniger

Anupama Sinha

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC., a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

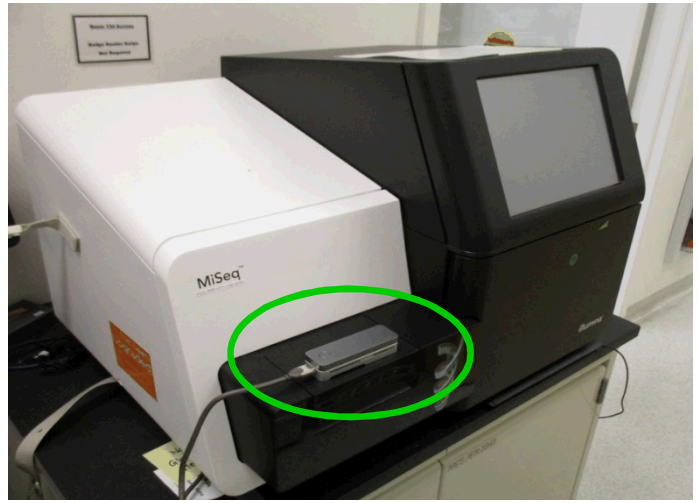
BIOSCIENCE RESEARCH FOUNDATION

Sandia National Laboratories

Nanopore technology – Fieldable 3rd generation sequencing

- Oxford Nanopore Techonolgy (ONT)'s MinION uses electrical impedance to "read" DNA as it transits through arrayed nanopores on disposable chip

- Major advantages
  - Long reads
  - Real-time data
  - Interactive sequencing (ReadUntil)
  - Field-ready, small footprint



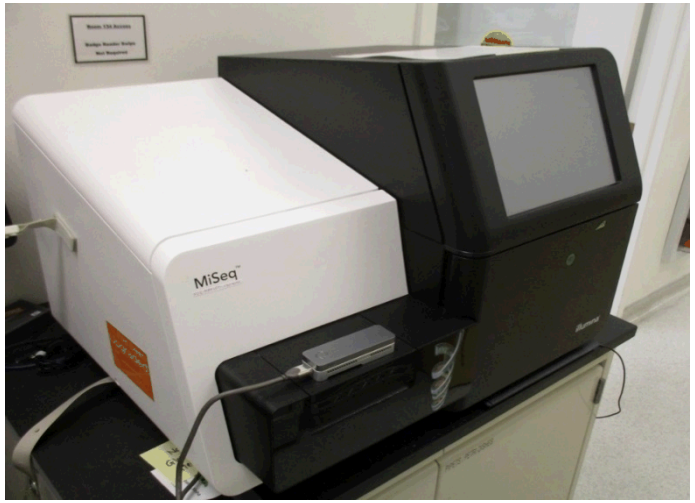
A. Sinha, H. Jayamohan

nanoporetech.com

Nanopore technology – Fieldable 3rd generation sequencing

- What can this sequencer address in the field of real-time detection?
  - Co-occurring genes (eg resistance, toxins), unique splice variants etc
  - Needle in a haystack – small amounts of pathogen DNA in large background samples

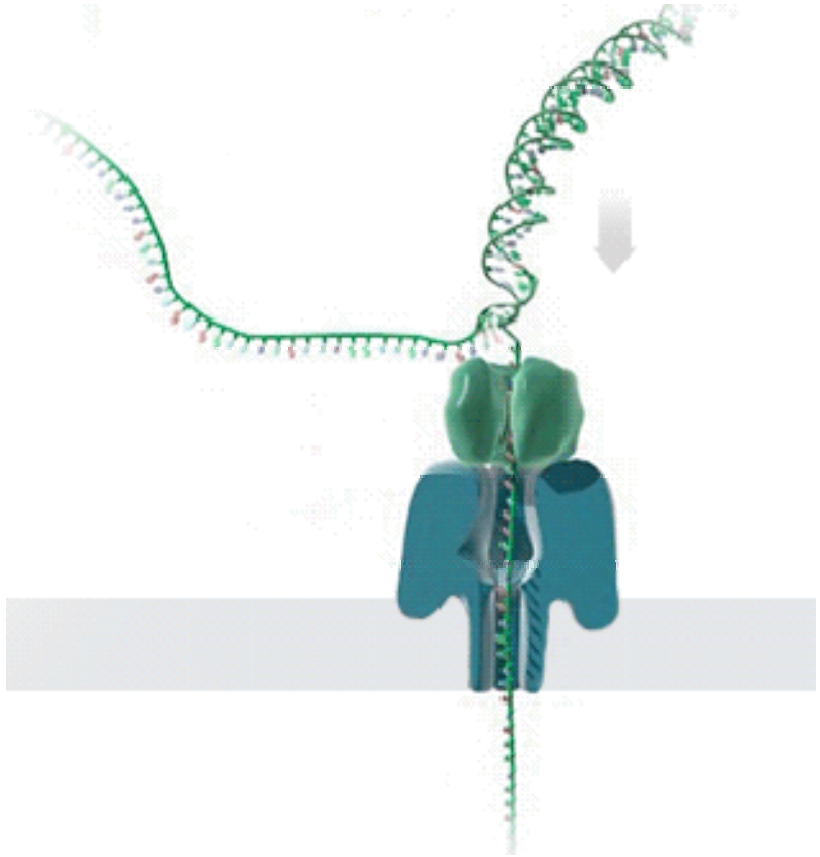
- What needs to be done?
  - Understanding how error rate affects ability to detect mutations (statistical confidence of calls)
  - Benchmark error (stochastic vs deterministic, factors affecting)
  - Develop experimental and bioinformatic pipelines that allow us to generate large amounts of accurate data

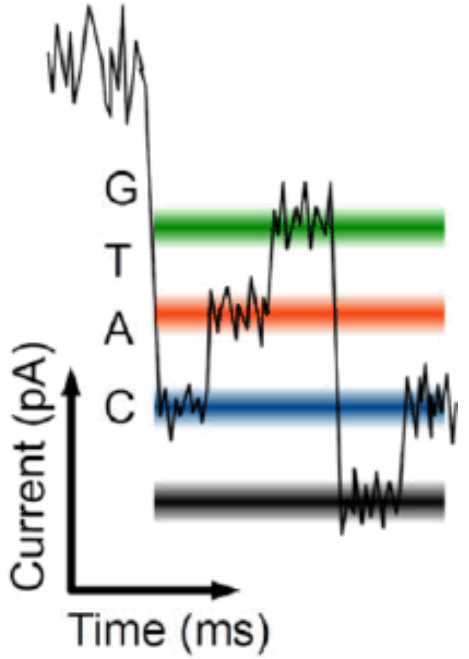


A. Sinha, H. Jayamohan

nanoporetech.com

Nanopore sequencing technology

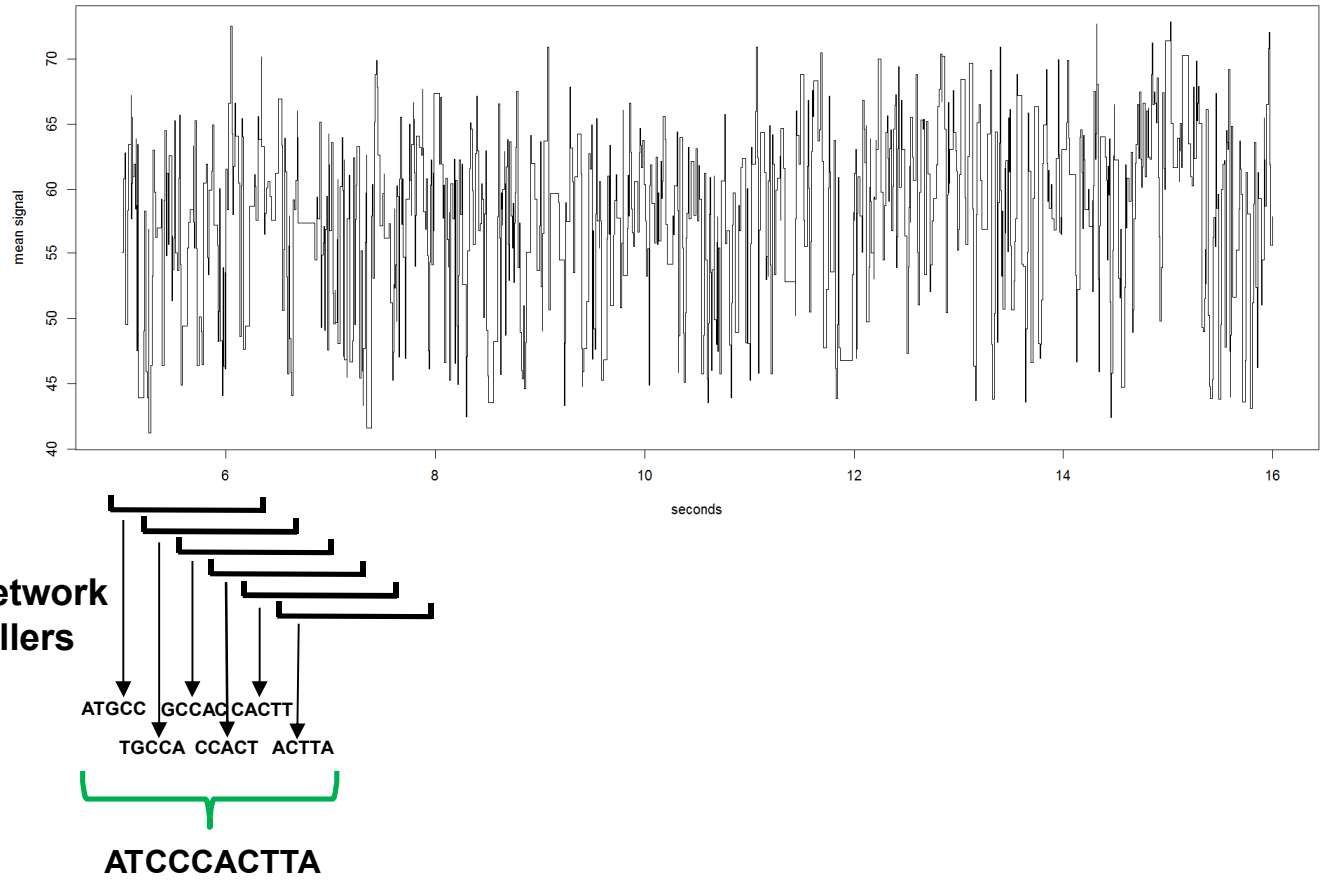




Steinbock and Radenovic 2016

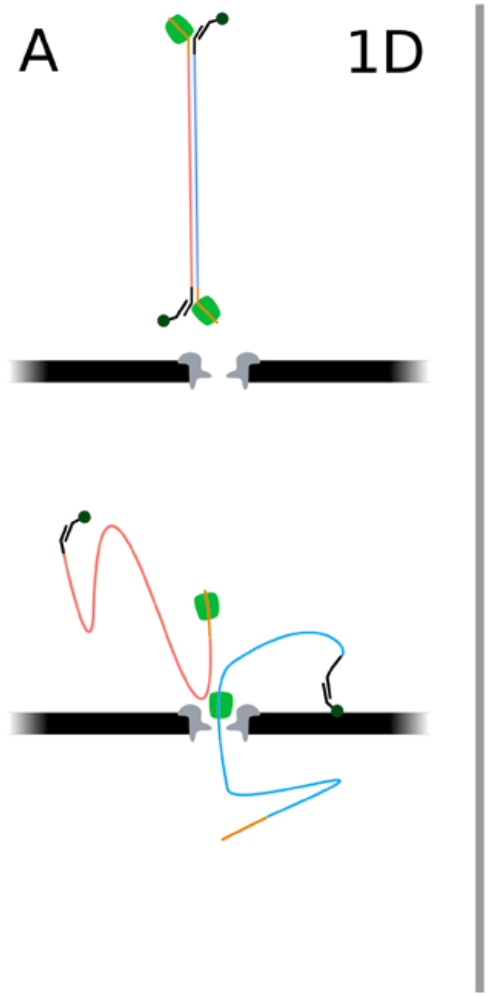
MinION data analysis

Raw and/or event-based “squiggle” plot generated in real time during sequencing



BIOSCIENCE RESEARCH FOUNDATION
Sandia National Laboratories

1D vs 2D vs $1D^2$

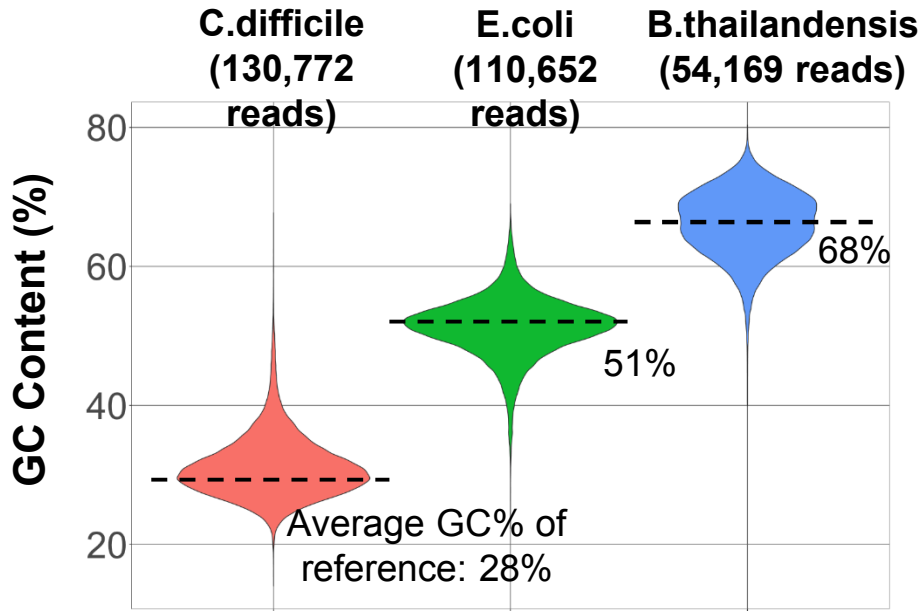


de Lannoy et al (F1000 Research 2017)

BIOSCIENCE RESEARCH FOUNDATION
Sandia National Laboratories

Understanding error sources in nanopore sequencing

GC content of underlying sample is overall reproducible

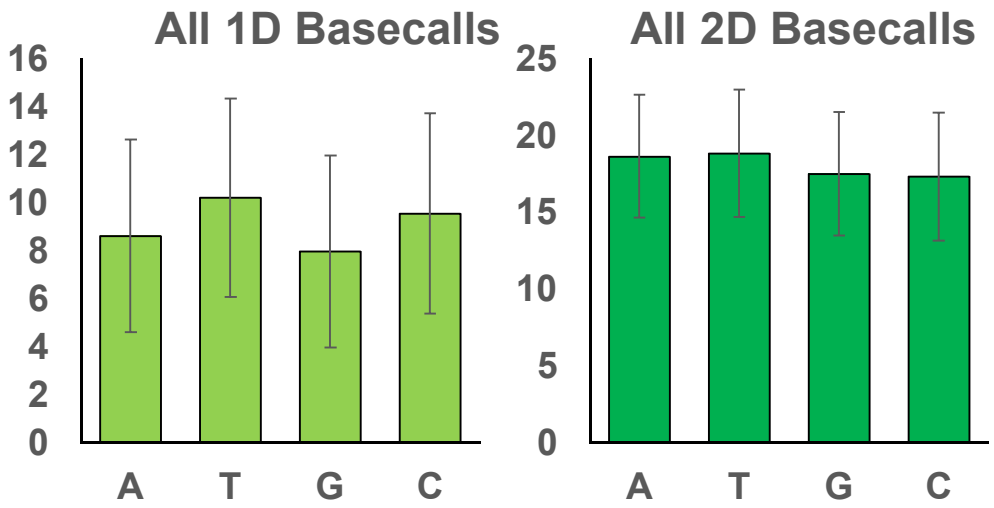


Krishnakumar et al (Scientific Reports 2018)

BIOSCIENCE RESEARCH FOUNDATION
Sandia National Laboratories

Understanding error sources in nanopore sequencing

There are systematic base-specific biases



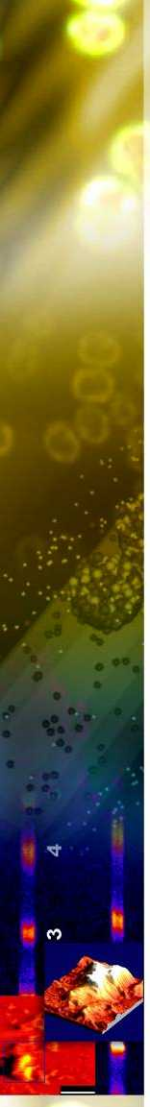
Understanding error sources in nanopore sequencing

There is strong context-specific dependence of base quality

Q score of next base call will be

E. coli (2D)	1	2	3	4	5	6	7	8	9	10
Lowest - 1	0.46	0.19	0.19	0.07	0.04	0.02	0.01	0.01	0.01	0.00
2	0.25	0.22	0.27	0.11	0.06	0.04	0.02	0.01	0.01	0.00
3	0.13	0.15	0.31	0.16	0.11	0.06	0.04	0.03	0.01	0.01
4	0.07	0.08	0.22	0.21	0.17	0.11	0.06	0.05	0.03	0.01
5	0.04	0.04	0.14	0.16	0.21	0.16	0.10	0.08	0.04	0.02
6	0.02	0.03	0.08	0.11	0.17	0.21	0.16	0.13	0.07	0.03
7	0.01	0.02	0.05	0.07	0.11	0.17	0.20	0.20	0.12	0.04
8	0.01	0.01	0.03	0.04	0.07	0.11	0.16	0.27	0.21	0.08
9	0.00	0.00	0.02	0.02	0.04	0.06	0.09	0.20	0.35	0.20
Highest - 10	0.00	0.00	0.01	0.01	0.02	0.03	0.04	0.09	0.24	0.57

Given Current Base Call Quality Score Decile



0 — 1

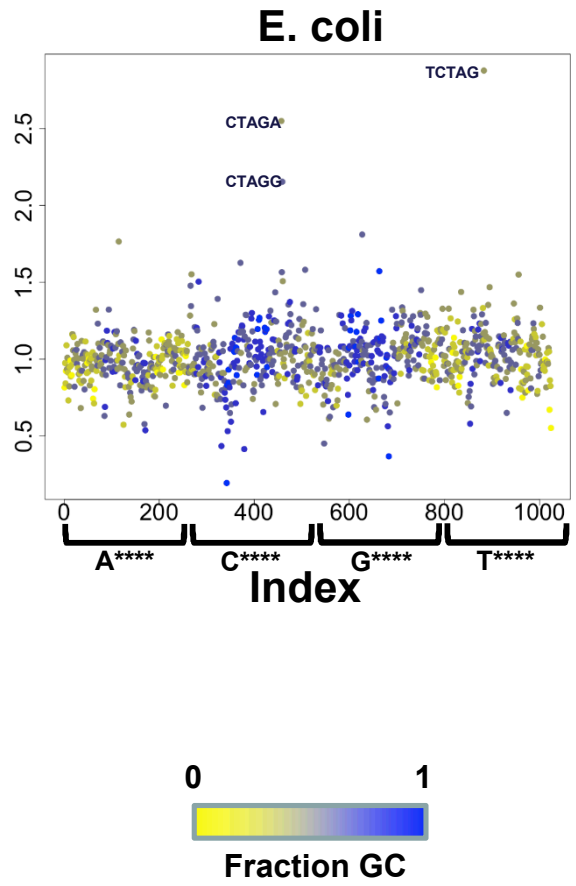
Probability

BIOSCIENCE RESEARCH FOUNDATION

Sandia National Laboratories

Understanding error sources in nanopore sequencing

There are context-specific k-mer biases

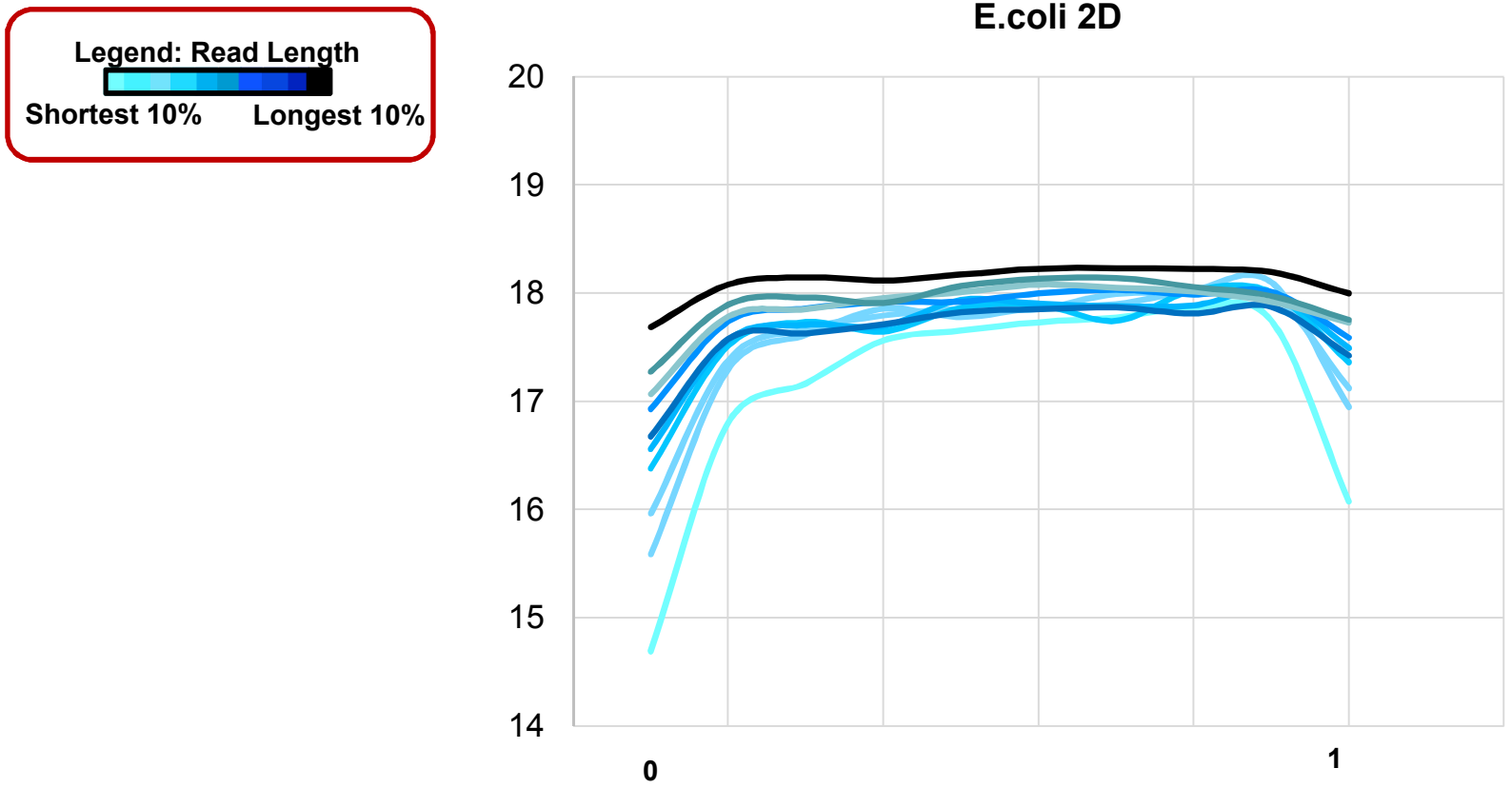


BIOSCIENCE RESEARCH FOUNDATION

Sandia National Laboratories

Understanding error sources in nanopore sequencing

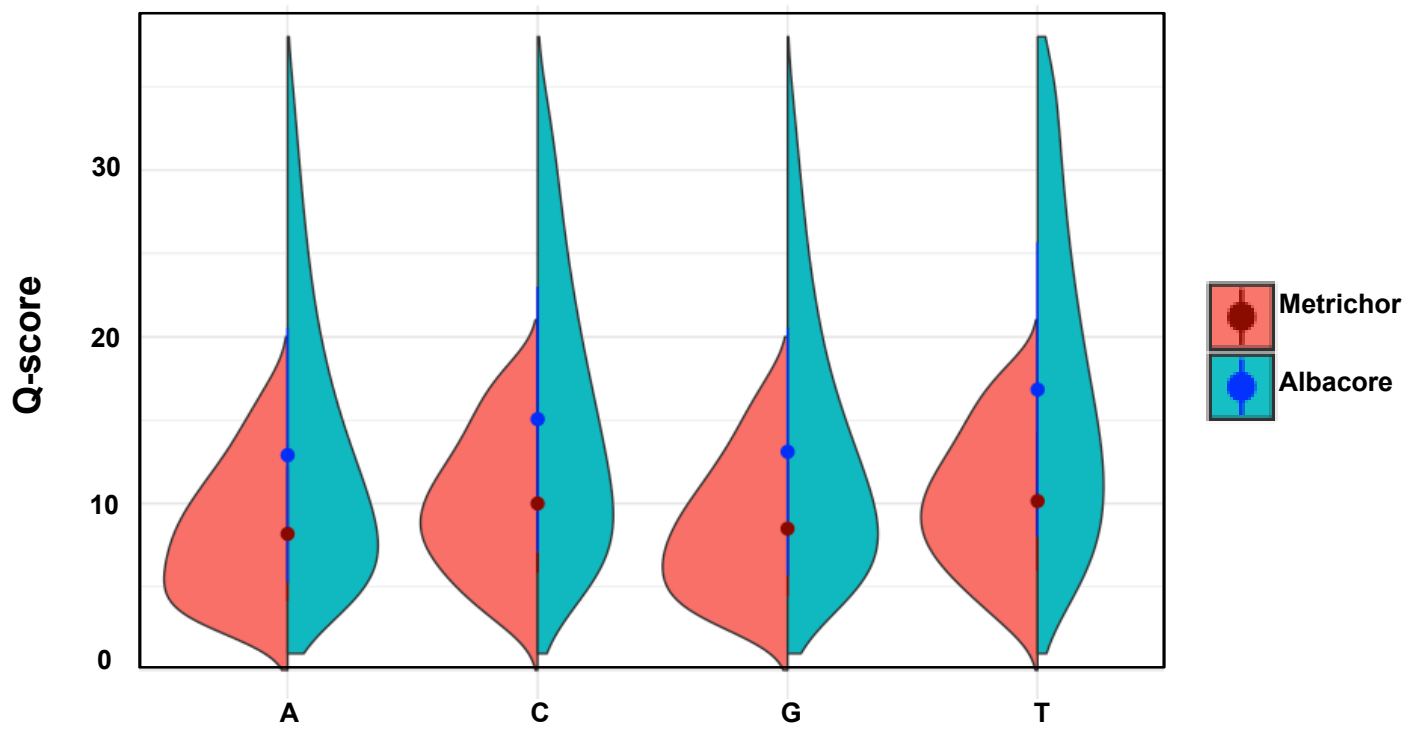
There are positional biases, length-dependent biases along the read



BIOSCIENCE RESEARCH FOUNDATION

Sandia National Laboratories

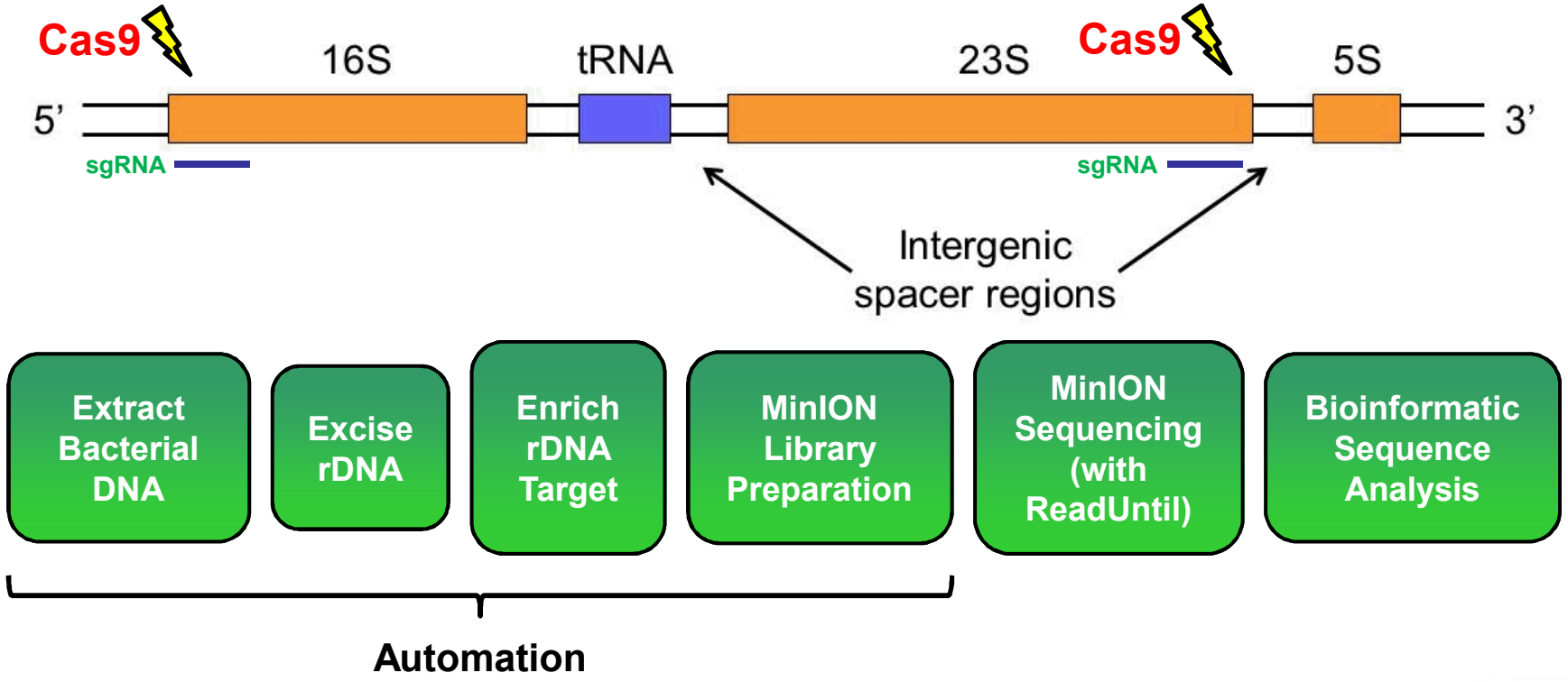
Evolving basecallers
Metrichor vs Albacore



We utilize this information to identify genomic regions/sequences that are optimal for high-confidence detection

Real-time automated pathogen identification by enhanced ribotyping (RAPIER)

- Goal: Rapid, universal, amplification-free, pathogen ID & characterization suitable for clinical, fieldable, & low-resource settings
- Universally conserved bacterial rDNA locus for species typing
- Long reads span the entire locus

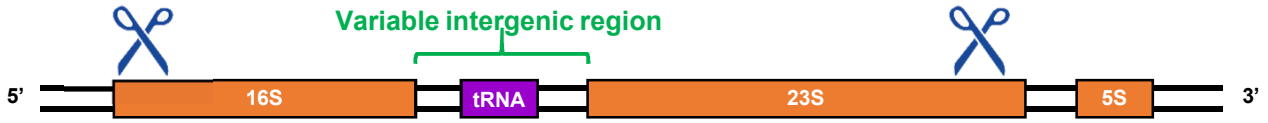


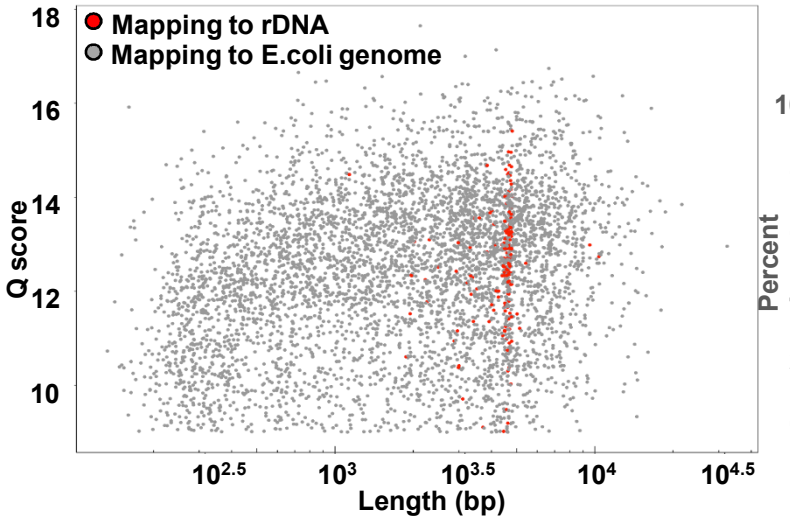
M. Bartsch and S. Bird

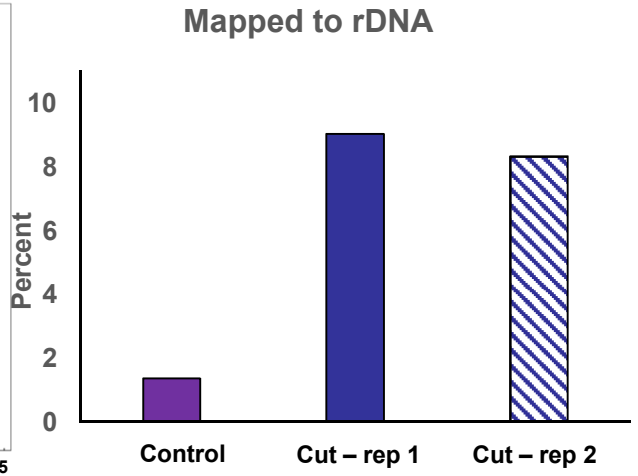
BIOSCIENCE RESEARCH FOUNDATION

Sandia National Laboratories

Cas9-based selection of rDNA





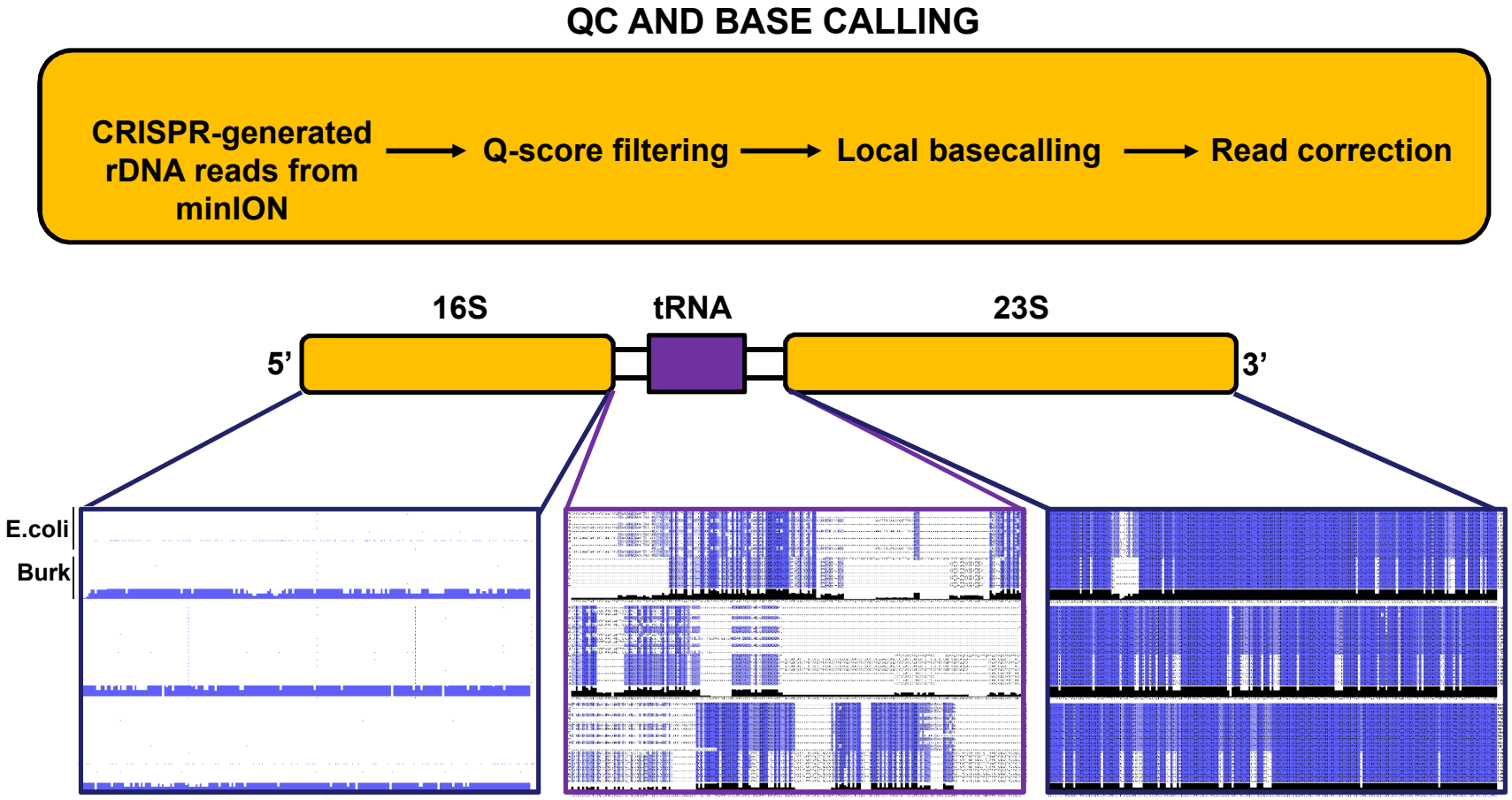


Cutting out the rDNA locus enriches its sequencing

A. Sinha, S. Bird and R. Krishnakumar

BIOSCIENCE RESEARCH FOUNDATION
Sandia National Laboratories

Sorting bacterial species from error-prone rDNA reads - pipeline



BIOSCIENCE RESEARCH FOUNDATION

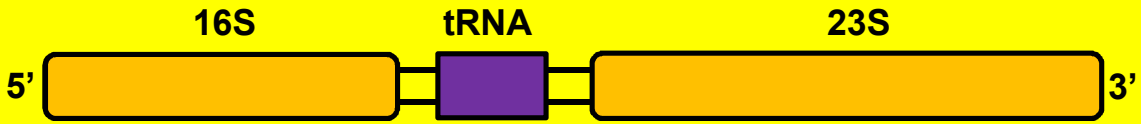
Sandia National Laboratories

Sorting bacterial species from error-prone rDNA reads - pipeline

QC AND BASE CALLING

CRISPR-generated rDNA reads from minION → Q-score filtering → Local basecalling → Read correction

ALIGNMENT 1 – 16S AND 23S FAMILY SORTING



ALIGNMENT 2 – INTERGENIC AND tRNA SUBGROUP SORTING

Future Perspectives

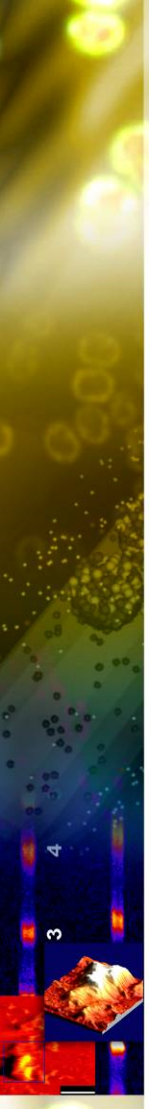
Knowledge of error behavior allows for smart selection of target

Experimental enrichment targets (eg. Cas9) can enhance probability of detection

Developing models to make rapid diagnostic calls is made possible by

1) **the groundwork laid by the above**
2) **implementing real-time selective sequencing (Read Until)**
3) **empirical testing of real-world samples**

Mike Bartsch: Real-time selection can search for ‘needles in a haystack’



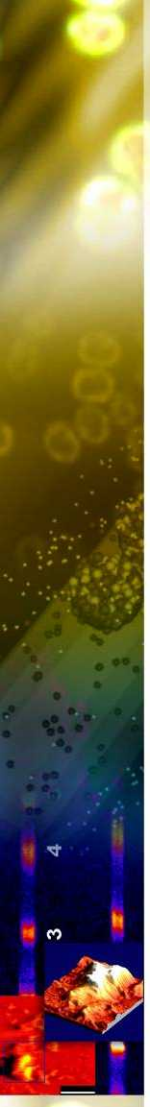
BIOSCIENCE RESEARCH FOUNDATION

Sandia National Laboratories

Acknowledgements

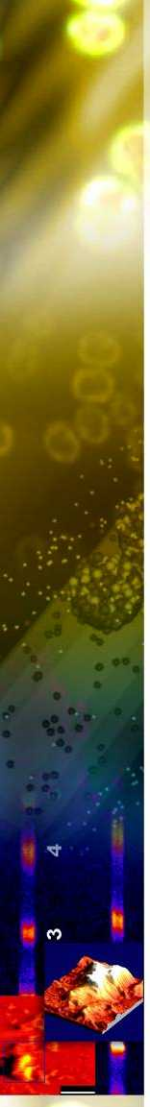
Mike Bartsch (PI)
Sara Bird
Steve Branda
Harrison Edwards
Hari Jayamohan
Ken Patel
Joe Schoeniger
Anupama Sinha

Funding – Sandia LDRD



BIOSCIENCE RESEARCH FOUNDATION

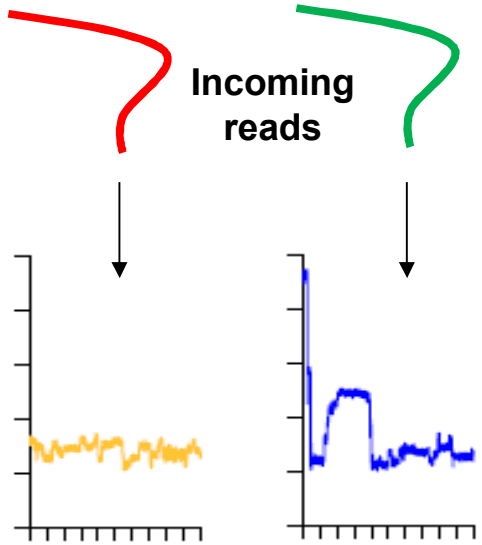
Sandia National Laboratories



BIOSCIENCE RESEARCH FOUNDATION

Sandia National Laboratories

RUBRIC – Read Until with Basecall- and Reference-Informed Criteria



Real time
basecall

TTCATAG

TTAGATC

REJECT

SEQUENCE

Real time
alignment

TTAATAAGGGTTAGATCG
CTTCTTTTCTAAGGCTAG
CTATTTGTTTCTTTCTTTC

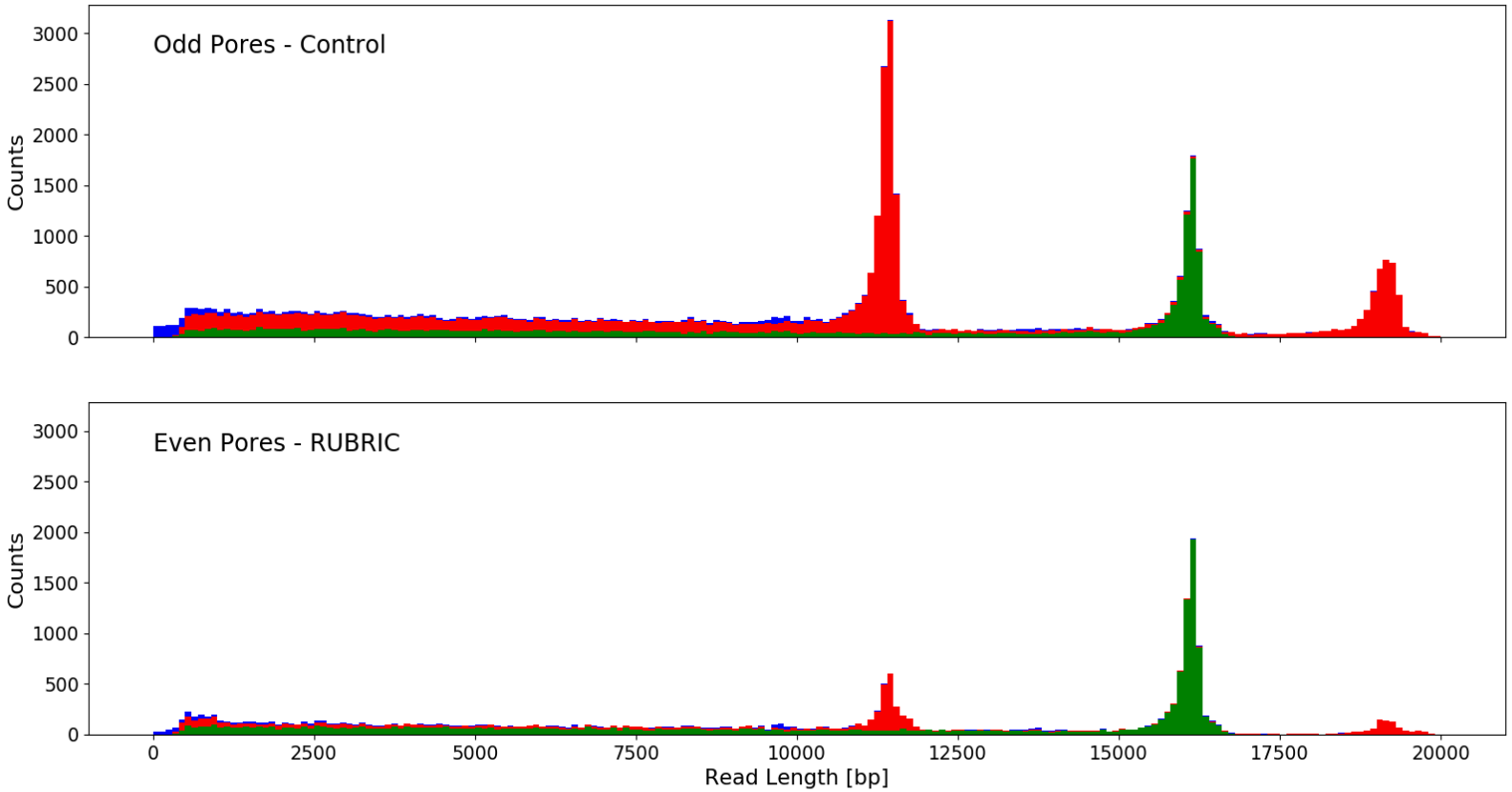
Sandia RUBRIC

BIOSCIENCE RESEARCH FOUNDATION

Sandia National Laboratories

RUBRIC – Read Until with Basecall- and Reference-Informed Criteria

Bacteriophage Lambda

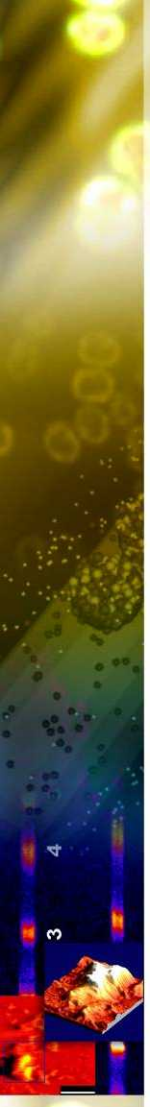


BIOSCIENCE RESEARCH FOUNDATION

Sandia National Laboratories

Long read sequencing technology

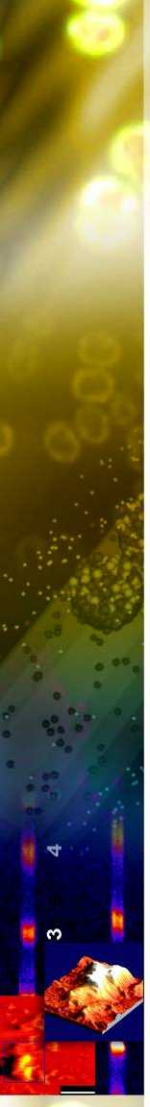
	Long read (eg. MinION)	Short read (eg. MiSeq)
LENGTH	**Upto 100s of kb** • Assembly of structurally challenging or repetitive regions • Splice variants • 3D strucutre	**Upto 600 bp**
AMPLIFICATION REQUIRED?	**No**	**Yes**
REAL TIME OUTPUT	**Yes**	**No**
SELECTIVE SEQUENCING	**Yes**	**No**
THROUGHPUT	**0.88 Gb/hr**	**0.25 Gb/hr**
COST	**$17/Gb**	**$53/Gb**
ACCURACY	**~85 – 95%**	**~99.99%**



BIOSCIENCE RESEARCH FOUNDATION

Sandia National Laboratories

Figure 3



Q score of next base call will be

E. coli (2D)	1	2	3	4	5	6	7	8	9	10
Lowest - 1	0.46	0.19	0.19	0.07	0.04	0.02	0.01	0.01	0.01	0.00
2	0.25	0.22	0.27	0.11	0.06	0.04	0.02	0.01	0.01	0.00
3	0.13	0.15	0.31	0.16	0.11	0.06	0.04	0.03	0.01	0.01
4	0.07	0.08	0.22	0.21	0.17	0.11	0.06	0.05	0.03	0.01
5	0.04	0.04	0.14	0.16	0.21	0.16	0.10	0.08	0.04	0.02
6	0.02	0.03	0.08	0.11	0.17	0.21	0.16	0.13	0.07	0.03
7	0.01	0.02	0.05	0.07	0.11	0.17	0.20	0.20	0.12	0.04
8	0.01	0.01	0.03	0.04	0.07	0.11	0.16	0.27	0.21	0.08
9	0.00	0.00	0.02	0.02	0.04	0.06	0.09	0.20	0.35	0.20
Highest - 10	0.00	0.00	0.01	0.01	0.02	0.03	0.04	0.09	0.24	0.57

Given Current Base Call Quality Score Decile

C. diff (2D)	1	2	3	4	5	6	7	8	9	10
Lowest - 1	0.44	0.25	0.15	0.06	0.04	0.02	0.02	0.01	0.01	0.00
2	0.21	0.29	0.24	0.10	0.06	0.04	0.03	0.01	0.01	0.00
3	0.10	0.18	0.28	0.16	0.11	0.07	0.06	0.03	0.02	0.01
4	0.05	0.10	0.21	0.20	0.16	0.11	0.09	0.04	0.03	0.01
5	0.03	0.06	0.13	0.15	0.19	0.16	0.14	0.07	0.04	0.02
6	0.02	0.04	0.09	0.11	0.16	0.19	0.20	0.10	0.06	0.03
7	0.01	0.02	0.05	0.07	0.10	0.15	0.27	0.17	0.11	0.05
8	0.01	0.01	0.03	0.04	0.06	0.10	0.21	0.24	0.20	0.10
9	0.00	0.01	0.02	0.02	0.04	0.06	0.13	0.19	0.31	0.22
Highest - 10	0.00	0.00	0.01	0.01	0.02	0.03	0.06	0.09	0.22	0.56

Given Current Base Call Quality Score Decile

B. thail (2D)	1	2	3	4	5	6	7	8	9	10
Lowest - 1	0.46	0.20	0.14	0.09	0.05	0.03	0.02	0.01	0.01	0.00
2	0.25	0.23	0.21	0.14	0.08	0.05	0.02	0.01	0.01	0.00
3	0.14	0.17	0.23	0.19	0.12	0.07	0.04	0.02	0.02	0.01
4	0.08	0.10	0.17	0.22	0.18	0.11	0.06	0.03	0.03	0.01
5	0.05	0.06	0.11	0.17	0.22	0.17	0.10	0.06	0.05	0.02
6	0.03	0.03	0.06	0.11	0.18	0.21	0.16	0.09	0.09	0.03
7	0.02	0.02	0.04	0.07	0.12	0.18	0.21	0.15	0.15	0.05
8	0.01	0.01	0.02	0.05	0.08	0.13	0.18	0.20	0.24	0.08
9	0.01	0.01	0.01	0.03	0.04	0.07	0.11	0.15	0.37	0.19
Highest - 10	0.00	0.00	0.01	0.01	0.02	0.03	0.05	0.07	0.25	0.57

Given Current Base Call Quality Score Decile

0 — 1

Probabilit
y

BIOSCIENCE RESEARCH FOUNDATION

Sandia National Laboratories

Figure 4

E.coli	1D Run				2D Run															
					2D template				2D complement				1D reads				2D reads			
	A	C	T	G	A	C	T	G	A	C	T	G	A	C	T	G	A	C	T	G
Lowest - 1	0.37	0.16	0.15	0.33	0.37	0.16	0.15	0.32	0.34	0.15	0.15	0.36	0.36	0.17	0.17	0.33	0.28	0.24	0.22	0.27
2	0.32	0.19	0.16	0.33	0.31	0.19	0.17	0.33	0.32	0.18	0.15	0.35	0.31	0.20	0.20	0.33	0.25	0.26	0.23	0.26
3	0.27	0.22	0.20	0.31	0.28	0.22	0.19	0.32	0.28	0.21	0.18	0.33	0.28	0.22	0.22	0.31	0.23	0.27	0.24	0.25
4	0.24	0.25	0.23	0.28	0.25	0.24	0.21	0.30					0.26	0.23	0.23	0.29	0.23	0.27	0.25	0.25
5	0.23	0.26	0.25	0.26	0.23	0.26	0.24	0.27	0.26	0.23	0.21	0.30	0.24	0.25	0.25	0.27	0.23	0.27	0.25	0.25
6	0.22	0.27	0.26	0.25	0.22	0.28	0.26	0.25	0.24	0.25	0.24	0.27	0.22	0.27	0.27	0.25	0.23	0.26	0.25	0.25
7	0.21	0.28	0.27	0.24	0.21	0.28	0.27	0.24	0.23	0.27	0.26	0.24	0.21	0.28	0.28	0.23	0.24	0.25	0.25	0.25
8	0.20	0.28	0.29	0.23	0.20	0.29	0.28	0.23	0.21	0.29	0.29	0.20	0.20	0.29	0.29	0.22	0.24	0.25	0.25	0.25
9	0.19	0.29	0.30	0.22	0.19	0.30	0.29	0.22	0.20	0.31	0.31	0.17	0.19	0.30	0.30	0.21	0.25	0.24	0.26	0.25
Highest - 10	0.15	0.33	0.33	0.19	0.15	0.34	0.32	0.19	0.18	0.33	0.37	0.12	0.15	0.34	0.34	0.18	0.25	0.24	0.25	0.25

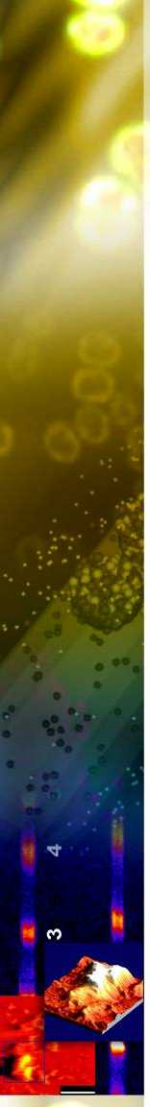
Quality Score Decile

C.diff	2D template				2D complement				2D reads			
	A	C	T	G	A	C	T	G	A	C	T	G
Lowest - 1	0.34	0.19	0.14	0.33	0.29	0.16	0.15	0.40	0.28	0.24	0.23	0.25
2	0.36	0.18	0.19	0.27	0.35	0.16	0.17	0.33	0.31	0.19	0.31	0.19
3	0.35	0.17	0.27	0.22	0.36	0.16	0.22	0.27	0.32	0.17	0.35	0.16
4	0.34	0.16	0.33	0.18					0.33	0.16	0.36	0.15
5	0.33	0.16	0.36	0.15	0.36	0.16	0.27	0.21	0.34	0.15	0.37	0.14
6	0.33	0.16	0.38	0.14	0.35	0.16	0.32	0.17	0.35	0.14	0.38	0.13
7	0.32	0.15	0.41	0.12	0.34	0.15	0.37	0.13	0.36	0.13	0.38	0.13
8	0.32	0.15	0.43	0.10	0.33	0.16	0.42	0.10	0.37	0.12	0.39	0.12
9	0.32	0.15	0.45	0.09	0.32	0.16	0.47	0.06	0.38	0.12	0.39	0.12
Highest - 10	0.31	0.13	0.51	0.05	0.27	0.14	0.57	0.02	0.38	0.11	0.39	0.12

Quality Score Decile

B.thail	2D template				2D complement				2D reads			
	A	C	T	G	A	C	T	G	A	C	T	G
Lowest - 1	0.39	0.14	0.15	0.32	0.35	0.15	0.13	0.37	0.26	0.25	0.21	0.29
2	0.27	0.20	0.14	0.39					0.19	0.32	0.18	0.31
3	0.19	0.26	0.16	0.39	0.29	0.20	0.12	0.39	0.16	0.35	0.17	0.32
4	0.15	0.31	0.17	0.37	0.24	0.25	0.13	0.38	0.14	0.36	0.17	0.33
5	0.13	0.34	0.17	0.36	0.20	0.30	0.16	0.34	0.14	0.36	0.17	0.34
6	0.12	0.37	0.17	0.35	0.17	0.35	0.18	0.30	0.14	0.35	0.16	0.35
7	0.11	0.38	0.17	0.34					0.14	0.35	0.16	0.35
8	0.09	0.40	0.17	0.33	0.14	0.40	0.20	0.26	0.14	0.34	0.16	0.36
9	0.08	0.42	0.17	0.32	0.12	0.44	0.22	0.22	0.14	0.34	0.16	0.36
Highest - 10	0.05	0.49	0.17	0.29	0.10	0.47	0.26	0.17	0.15	0.33	0.15	0.37

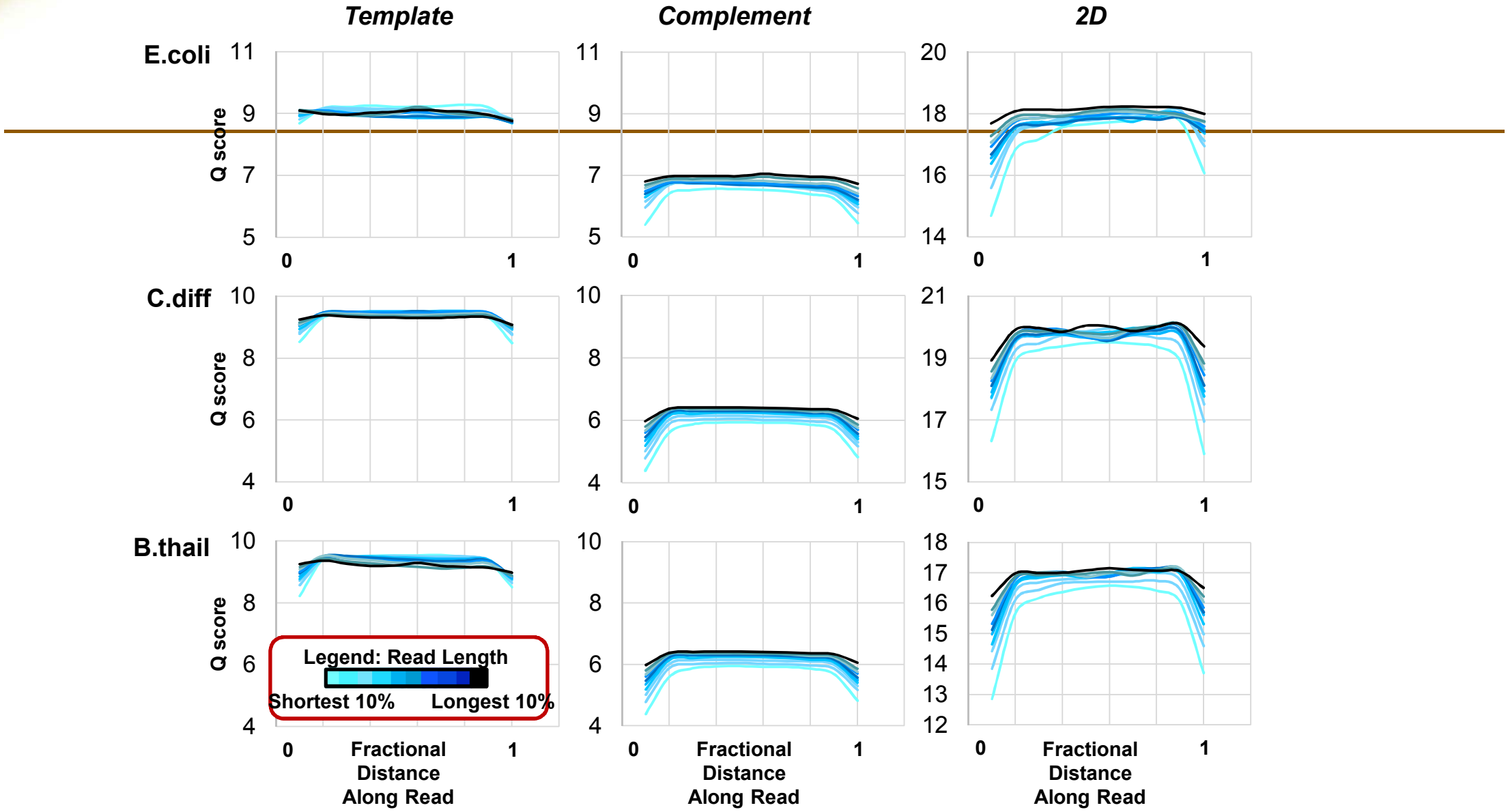
Quality Score Decile



BIOSCIENCE RESEARCH FOUNDATION

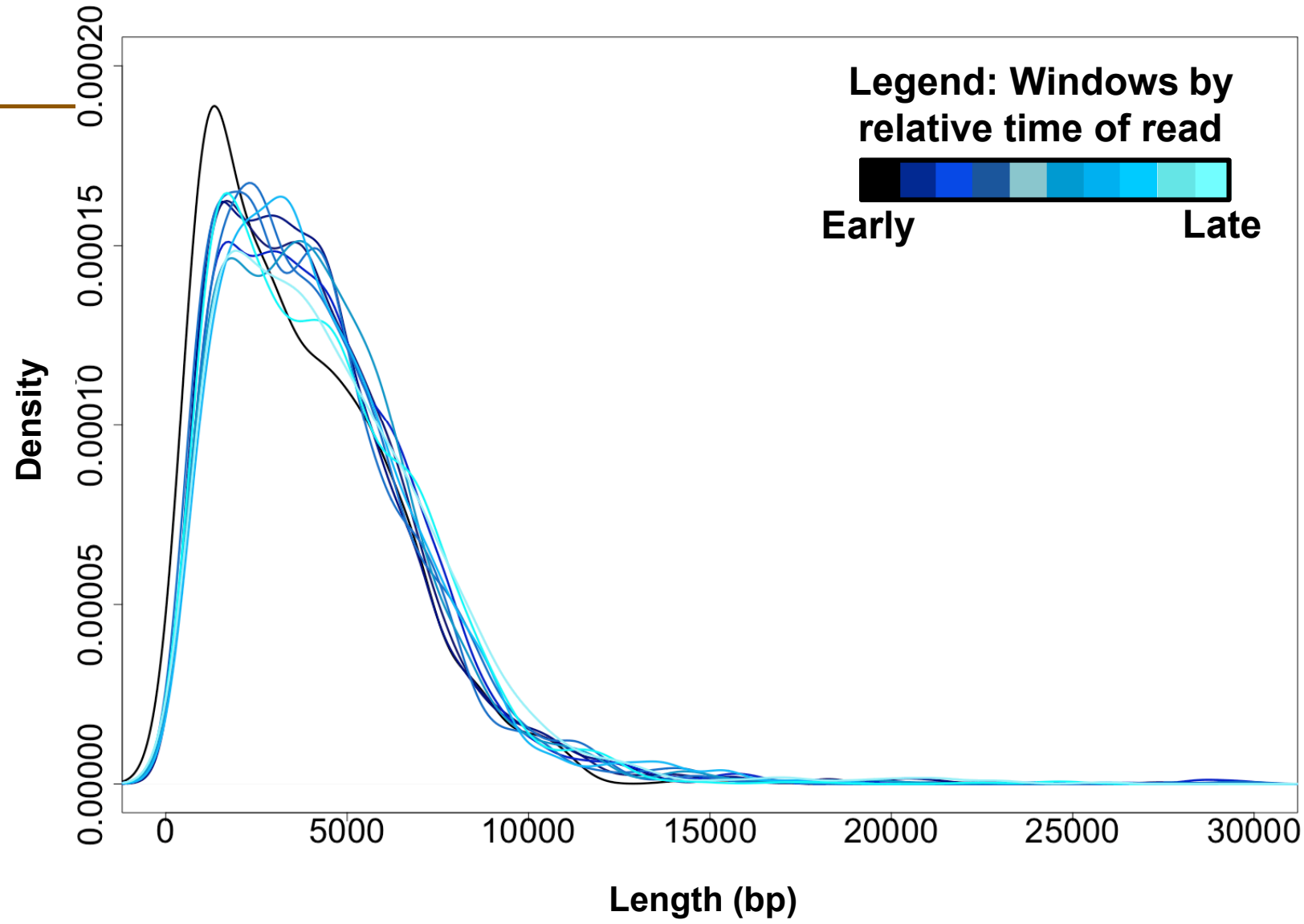
Sandia National Laboratories

Figure 5



BIOSCIENCE RESEARCH FOUNDATION
Sandia National Laboratories

Figure 7



BIOSCIENCE RESEARCH FOUNDATION

Sandia National Laboratories