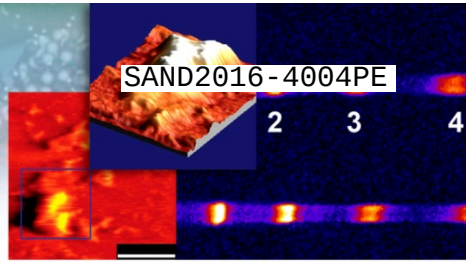
SANDIA BIOSCIENCE

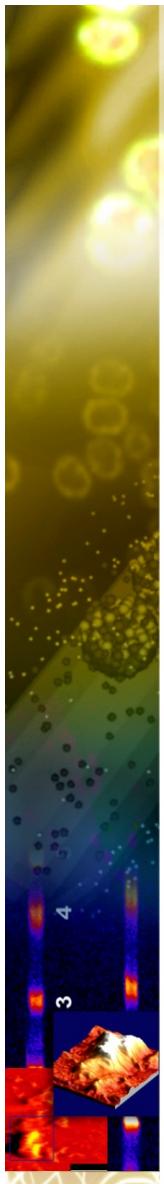


Bioinformatics and Computational Biology at a National Laboratory

Corey Hudson
MUII Conference
April 5, 2016

Mission of biology at a National Lab

- The mission of the national labs:
  - **Biodefense**: Tools to predict and detect biological attacks
  - **Bioenergy**: As a Department of Energy facility, national labs are heavily involved in the US government energy mission - including bioenergy
    - **Historically, this meant developing developing next generation biofuels**
    - **More recently, biological production of high value petroleum derivable chemicals, e.g., *adipic acid* which is petro-derived and produces huge amounts of N_2O.**

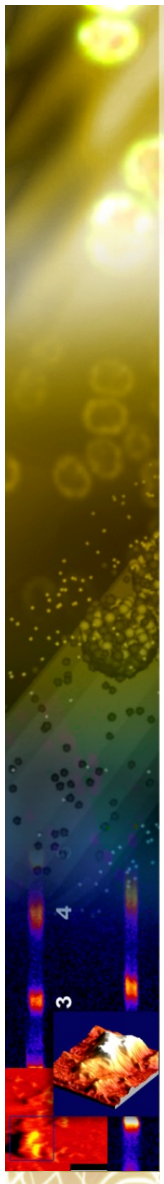


SANDIA BIOSCIENCE

My postdoctoral work at Sandia

- Started **2012** as a **postdoctoral researcher**
- Projects:
  - Characterize genome and mobilome of highly antibiotic resistant bacteria: *Klebsiella pneumoniae* BAA-2146

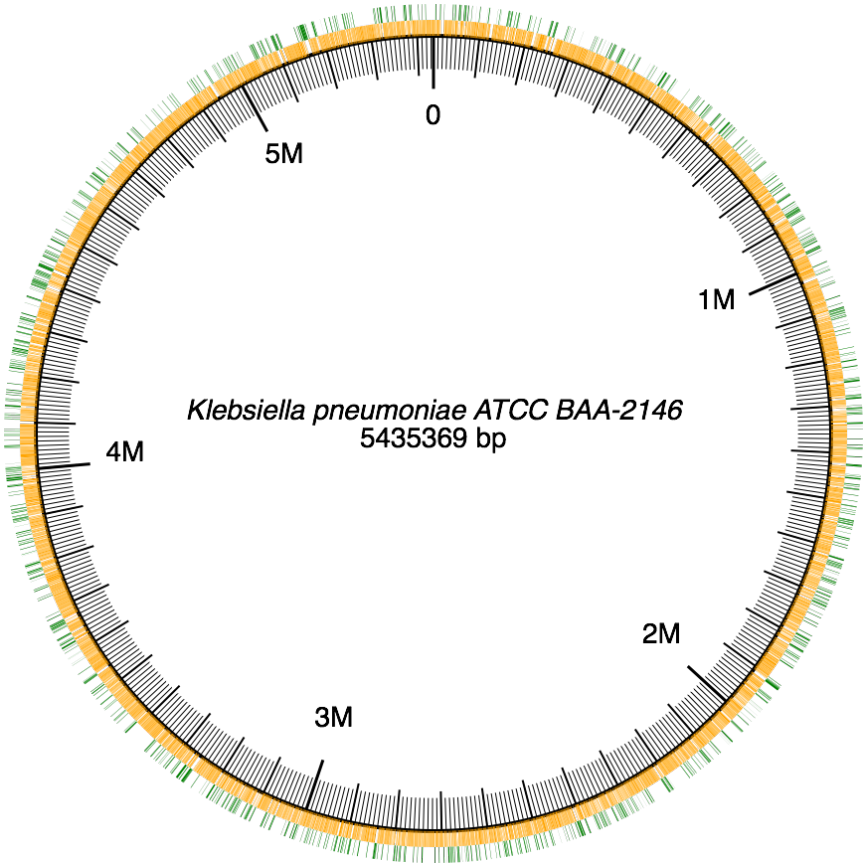
    Responsible for nosocomial infections in India - known as the New Delhi *Klebsiella*. First US isolate.
  - Predict lignolytic enzymes in Sevilleta Arid Lands in New Mexico microbiome.
  - Build database of high quality mobile genomic islands in bacteria.



SANDIA BIOSCIENCE

Identifying resistance elements in multidrug resistant genomes

- *Klebsiella pneumoniae* BAA-2146
- Strain sequenced from 2010 isolate
- Carries NDM-1 plasmid
- Resistance tested for 34 antimicrobial and antimicrobial/inhibitor combinations tested

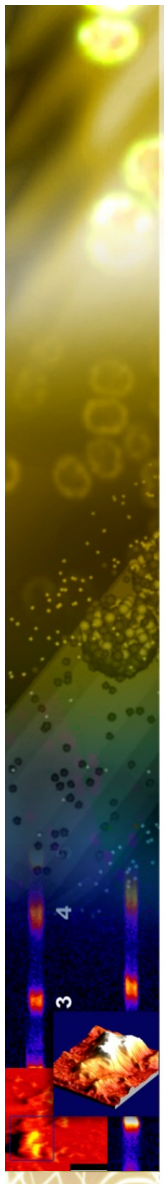


Hudson, Bent, Meagher, Williams. (2014) Resistance Determinants and Mobile Genetic Elements of an NDM-1-Encoding *Klebsiella pneumoniae* Strain. *PLoS One* 9(6): e99209

SANDIA BIOSCIENCE

Mobile genomic elements as a source of pathogenicity and resistance

- Plasmids (inter-bacterial)
  - Primary source of resistance in *Klebsiella pneumoniae*
  - Can also include independently mobile resistance integrons
- Genomic islands (inter-bacterial)
  - Common source of pathogenicity in bacteria
- Chromosomal integrons (intra-bacterial)
  - Common source of resistance in bacteria
- Insertion sequences (intra-bacterial)
  - Often capable of affecting virulence through the activation and repression of neighboring genes

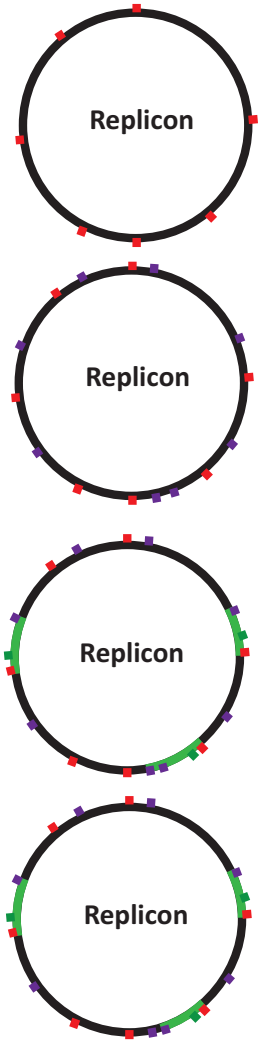


SANDIA BIOSCIENCE

Features of genomic islands

- Genomic islands are large insertions into the genome by means of an integrase.
- These genomic regions can carry large numbers of genes.
  - Allows them to rework bacterial metabolism (e.g., iron uptake)
  - Many carry bacterial toxins
  - Secretion systems and effector proteins
  - Capsule synthesis proteins

Islander: An algorithm for discovering high-quality islands

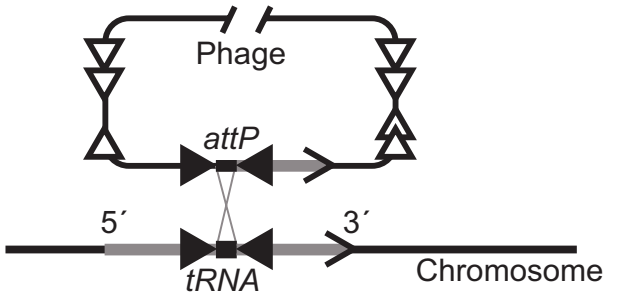


Place tRNA/tmRNA on replicon

BLAST tRNA/tmRNA against replicon to find fragments

Identify integrases occurring between tRNA/tmRNA and their associated fragments

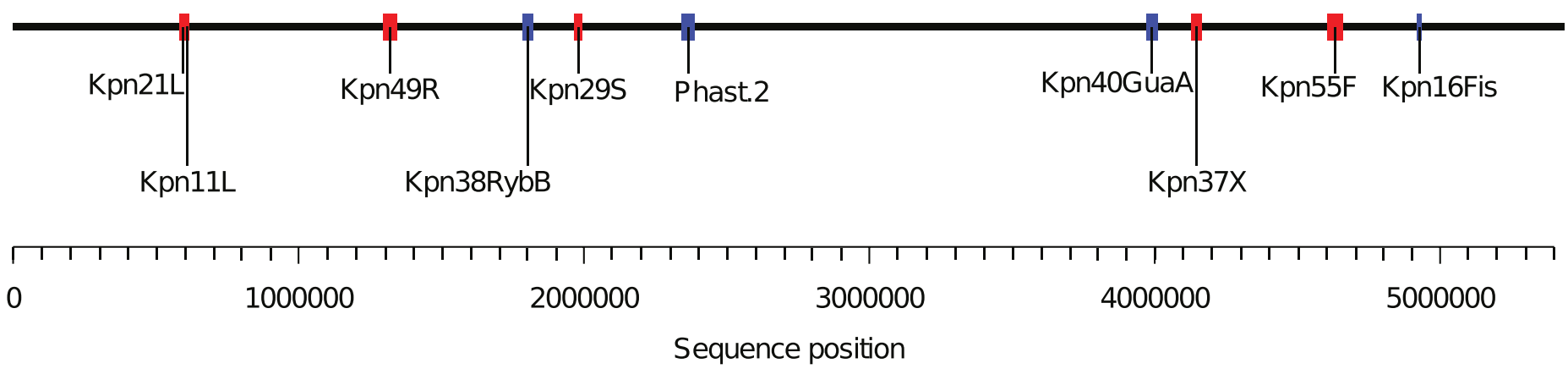
Deduplicate overlapping regions



SANDIA BIOSCIENCE

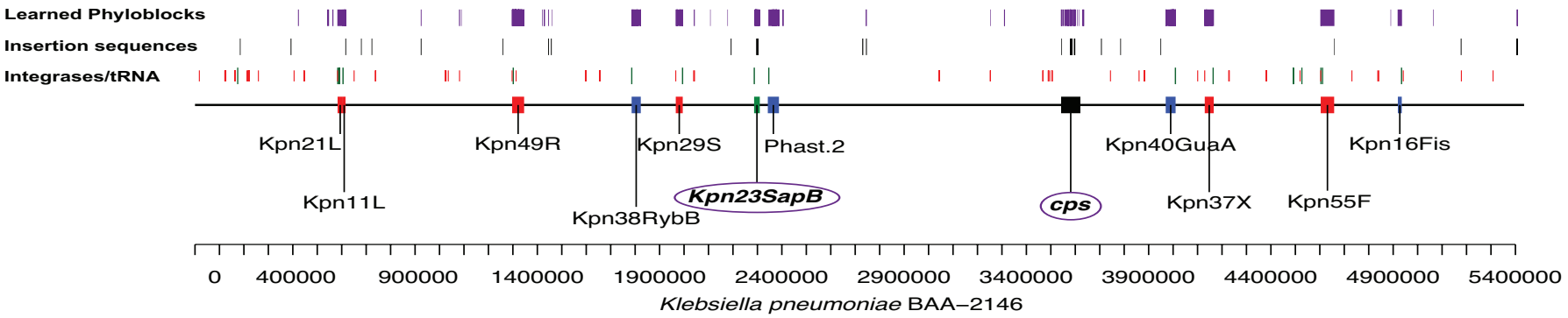
Sandia National Laboratories

Annotating genomic islands in *Klebsiella pneumoniae* BAA-2146



- Annotated 6 genomic islands using this algorithm (Islander)
- Included 4 additional islands using PHAge Search Tool (PHAST)
- Amounts to 342 kb out of 5429 kb

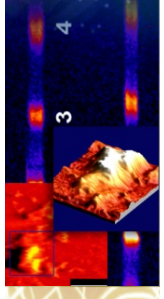
Using comparative genomics to ID islands and and insertion sequences



- Points to four insertion sequences
  - 2 IS*Ecp1*
    - 1 carrying β-lactamase CTX-M-15 that was recently transferred to the chromosome (a first in any complete genome) from one of its plasmids and in plasmids, predicted to confer resistance to β-lactam
  - 2 IS*Kpn18*
    - 1 disrupting RamR, a negative regulator of RamA, predicted to confer tigecycline resistance

Switching gears: Breaking down lignin

- Lignin, an irregular, complex aromatic molecule in higher plants and some algae, is the second most abundant biopolymer on the planet, making up between 10% and 25% of plant biomass.
- Limits the ability to break-down plant materials into biofuels.
- Question: Can high-lignin turnover environments be prospected to identify new classes of lignin breakdown enzymes?

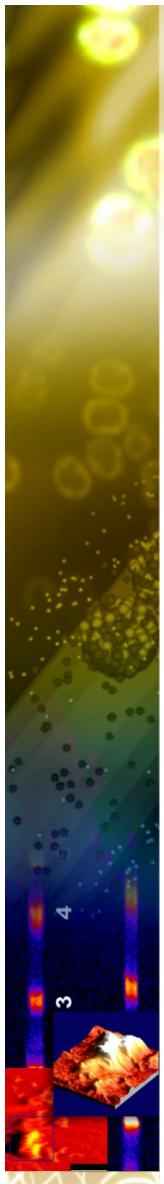


SANDIA BIOSCIENCE

Hudson, et al (2015) Lignin-Modifying Processes in the Rhizosphere of Aridland Grasses. *Environmental Microbiology* 17(12)

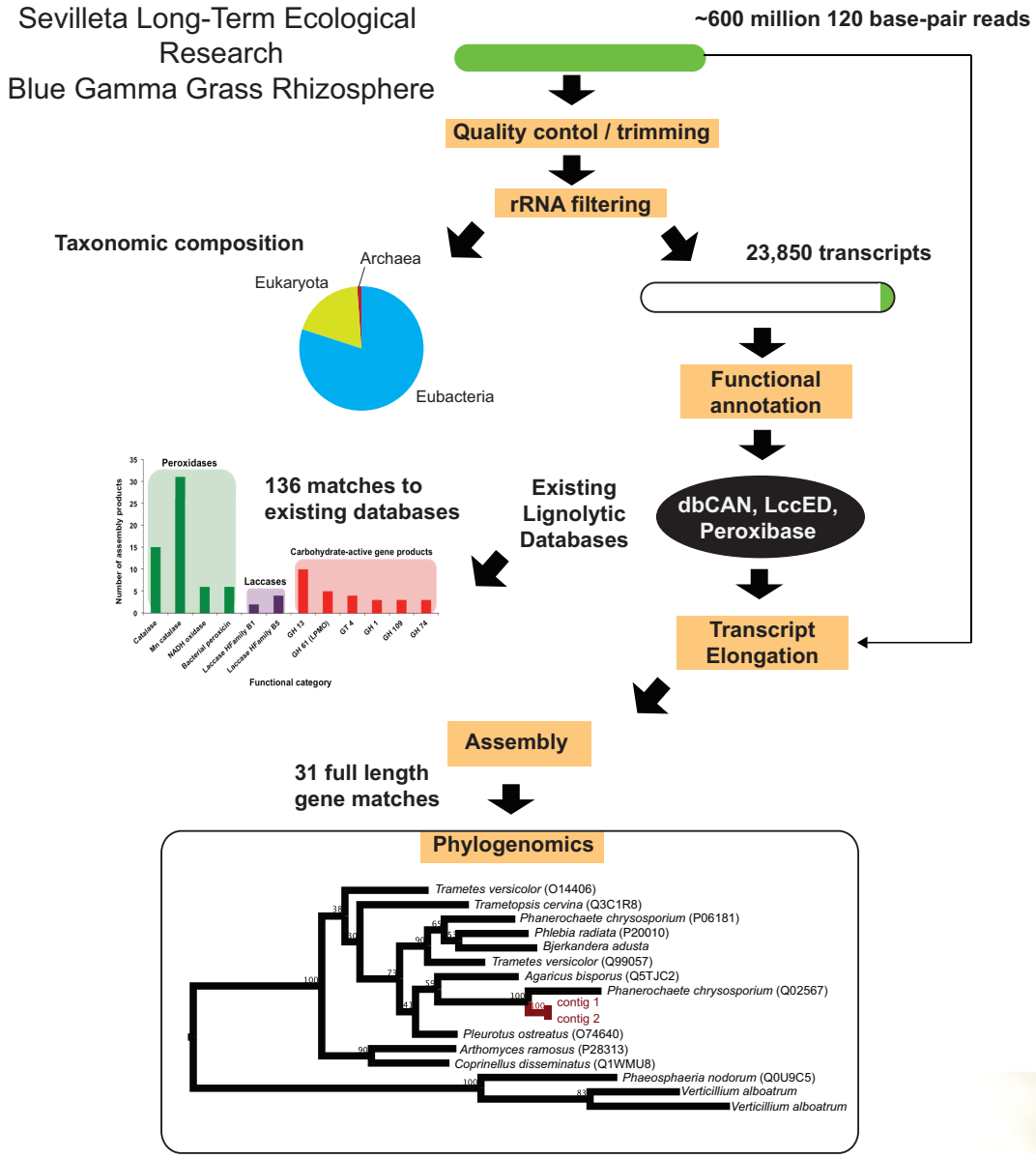
Tools for engineering microbial lignolytic breakdown

- Most microbiome annotators work at the broadest of functional categories (e.g., *secondary metabolism* or *virulence, disease* and *defense*)
  - Not useful discovering novel genes in a metagenome
  - Inefficiently annotate entire metagenome
- Need for tools that are targeted
  - Include expert knowledge of gene feature
  - Target annotation
  - Use rule-in / rule-out gene families with careful thresholding to remove spurious reads



SANDIA BIOSCIENCE

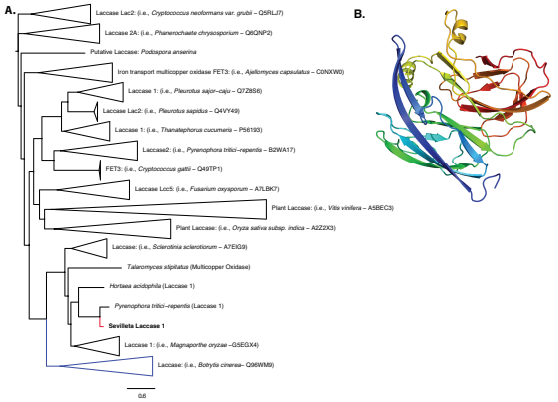
Pipeline for Lignin-active Enzymes

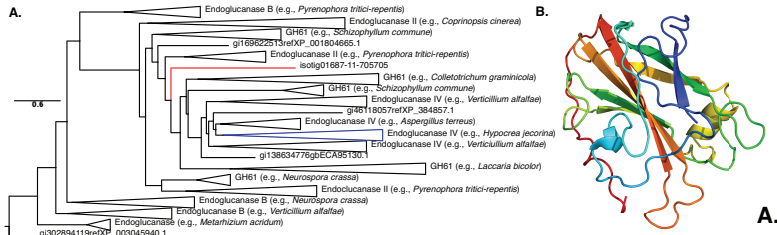


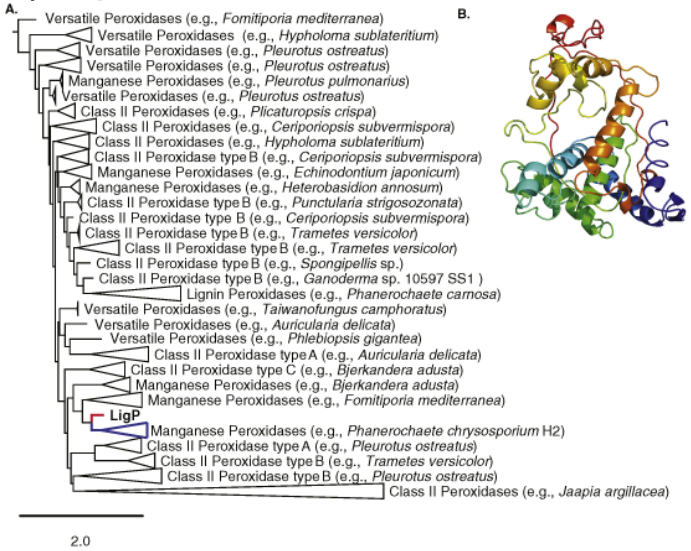
SANDIA BIOSCIENCE

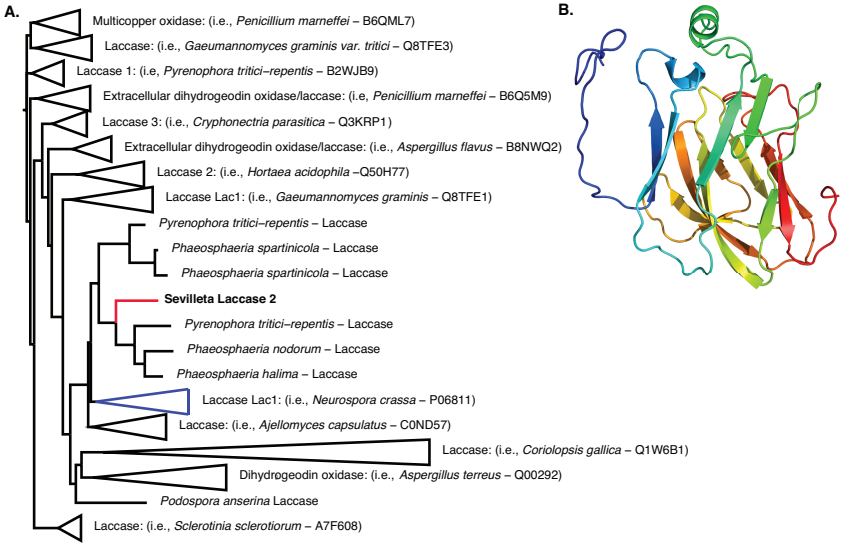
Sandia National Laboratories

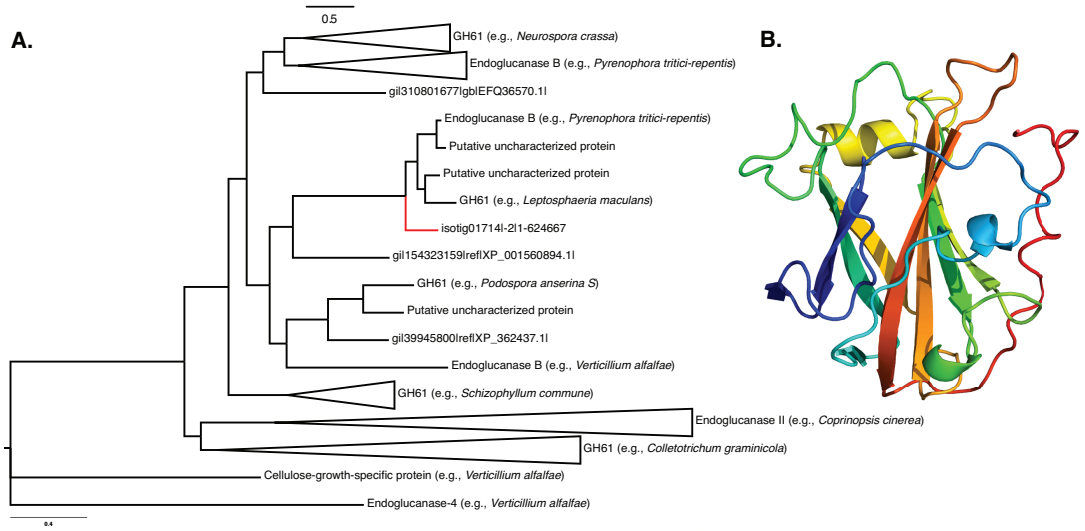
Five new families of enzymes









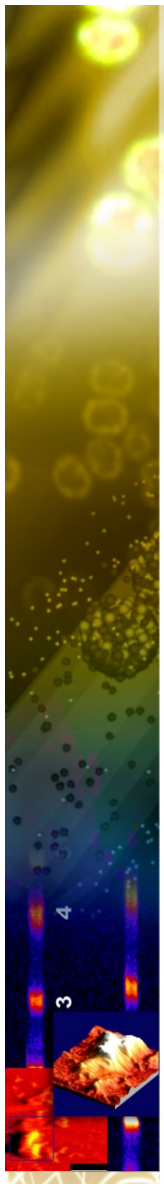


SANDIA BIOSCIENCE

Sandia National Laboratories

Update on project, since publication

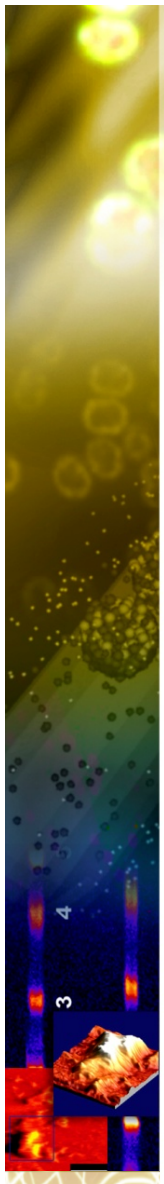
- Enzymes are currently being synthesized to allow *in situ* testing of the their efficacy in lignin breakdown.
- Early days for this work, and involves considerable work in optimizing production and activity.
- *Remains to be seen how much lignin breakdown will stay a focus of the DOE.*



SANDIA BIOSCIENCE

My staff work at Sandia

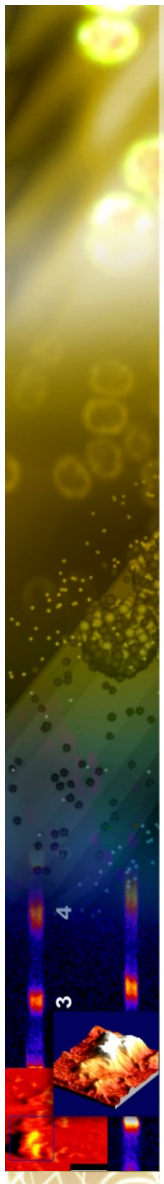
- Started **2014** as a **Senior Member of the Technical Staff**
- Projects:
  - Experimental identification of mobile genetic elements
  - Lab coordinator for DTRA Cooperative Biological Engagement Program at Lakka Beach in Freetown, Sierra Leone
  - Triage of biologically produced chemicals of high value as drop-in fuel blendstocks
  - Identification of engineered biological organisms
  - User-in-the-loop active machine learning in cybersecurity (not part of today's talk)



SANDIA BIOSCIENCE

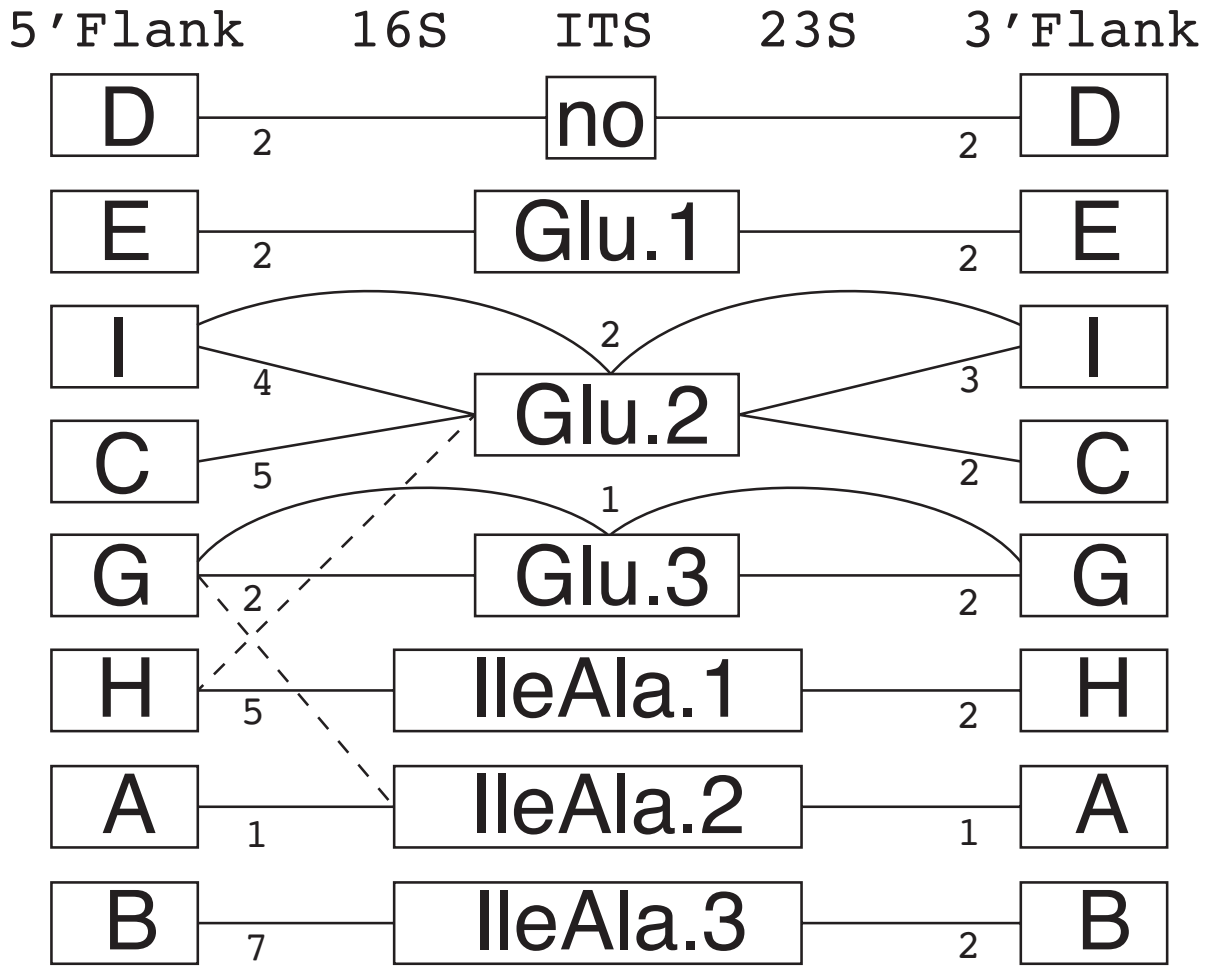
Experimental characterization of mobilome

- Bacteria release mobile elements
  - Under stress
  - In stationary phase
- Research question: **Can we supplement and improve bioinformatic tools for mobile element detection using experimental means**?

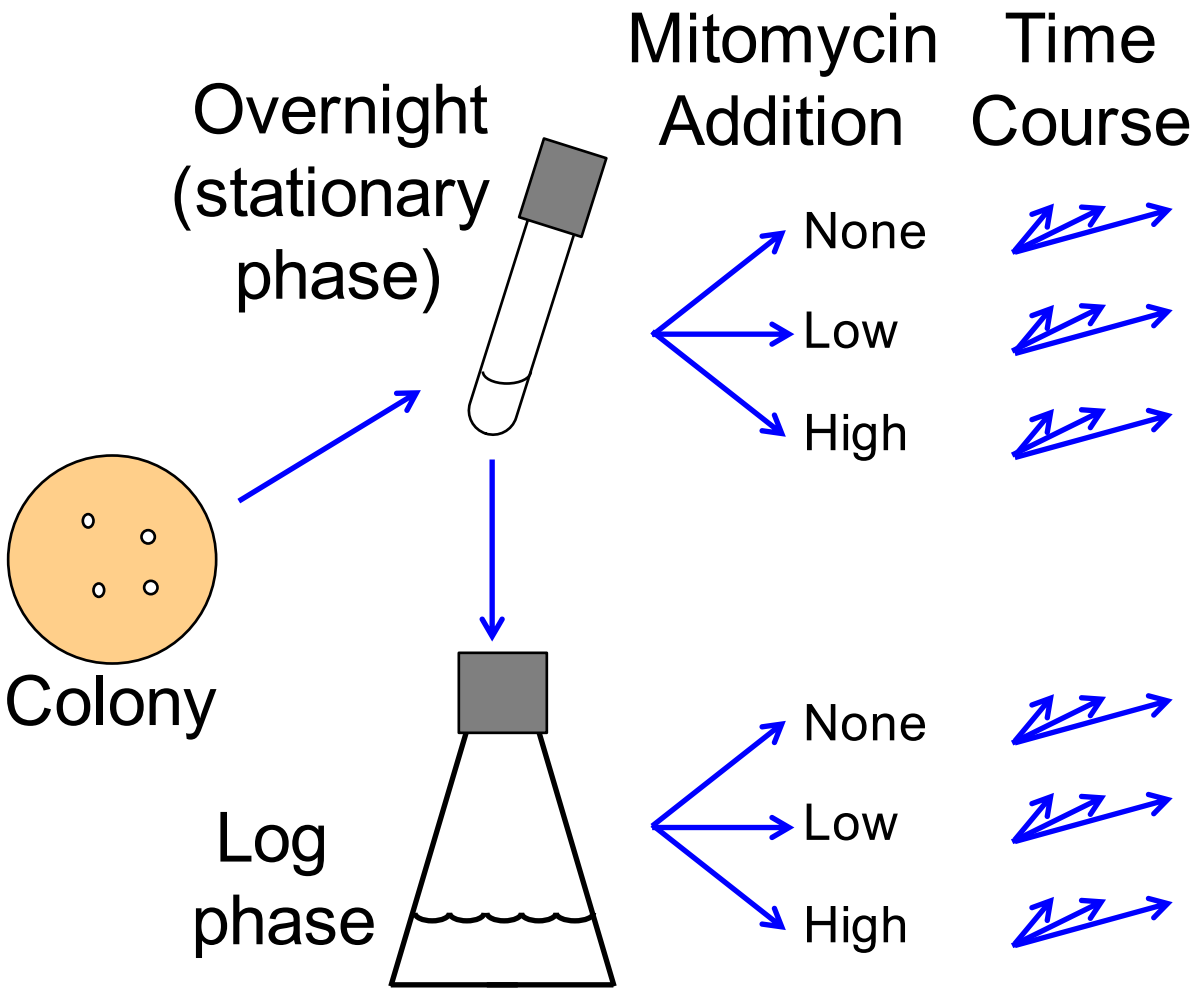


SANDIA BIOSCIENCE

Discovery while assembling genome: PacBio identification of recombinant subpopulation



Protocol

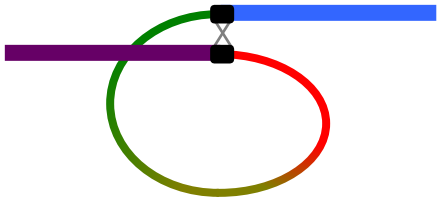


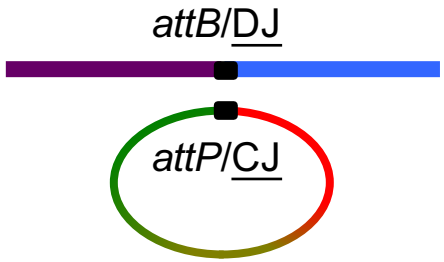
Per sample:

- Treat with or without DNase overnight
- Nextera
- SE Illumina NextSeq
- Juxtaposer software

Detecting Excision and Circularization

Mobility event



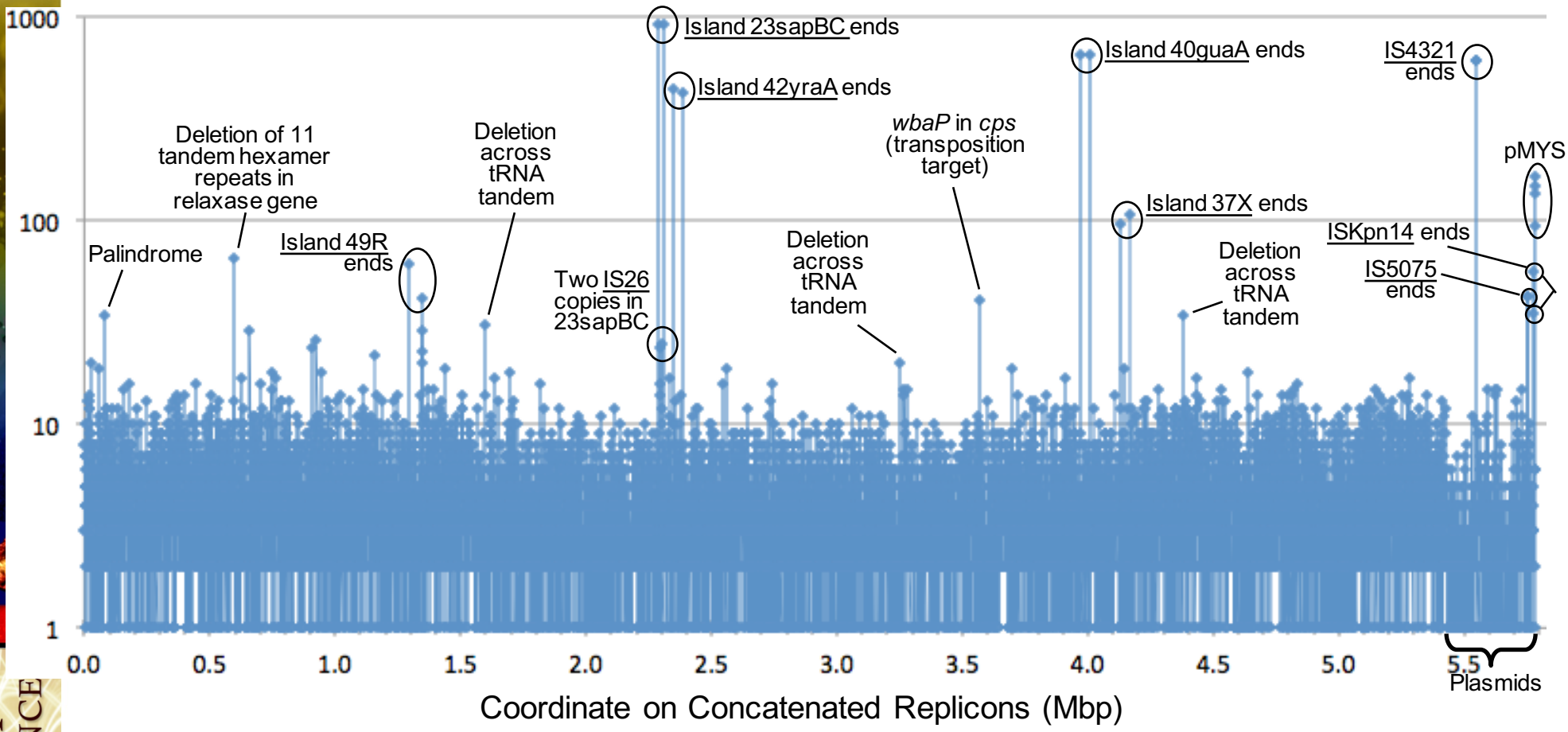


NGS reads

Standard reads

Recombinant reads

Juxtapositions in the genome

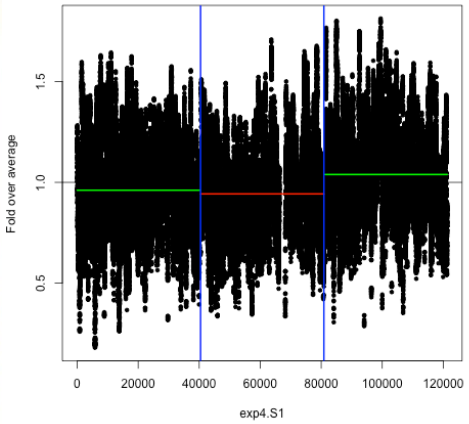


SANDIA BIOSCIENCE

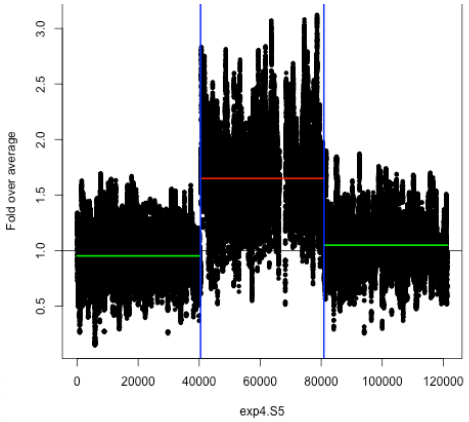
Sandia National Laboratories

Mitomycin and Exonuclease Treatment

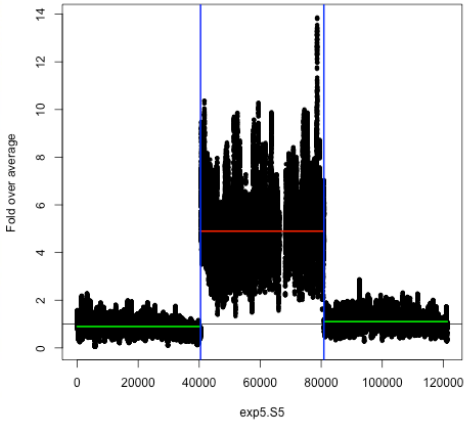
Island Kpn40guaA



No MMC
(no Exo)

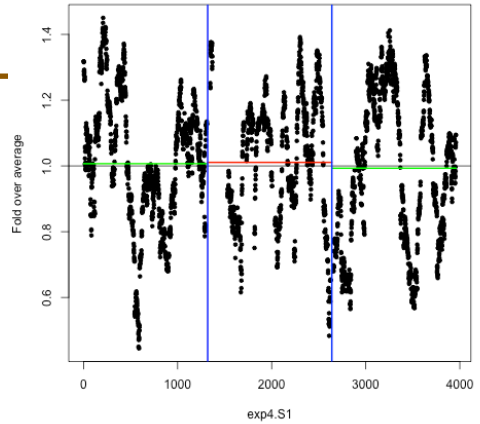


High
Mitomycin 2h
(no Exo)

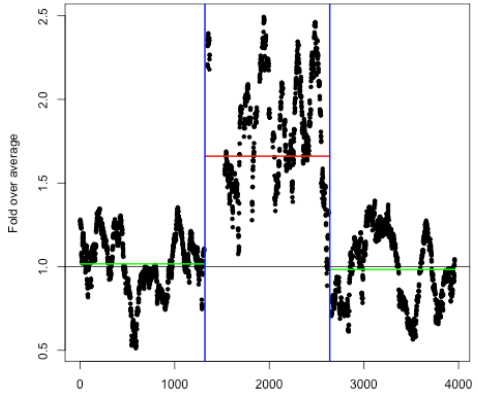


High MMC 2h,
Exo-treated

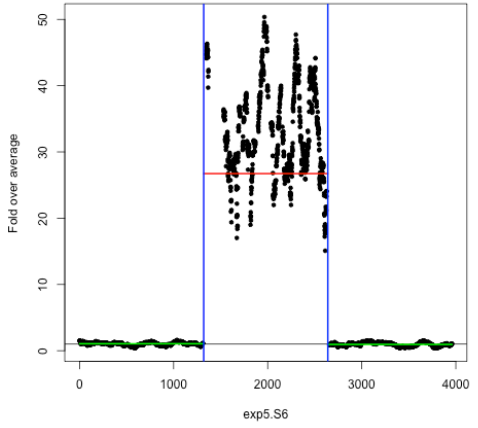
Insertion seq. IS4321



Log Phase
(no Exo)



Overnight
(no Exo)



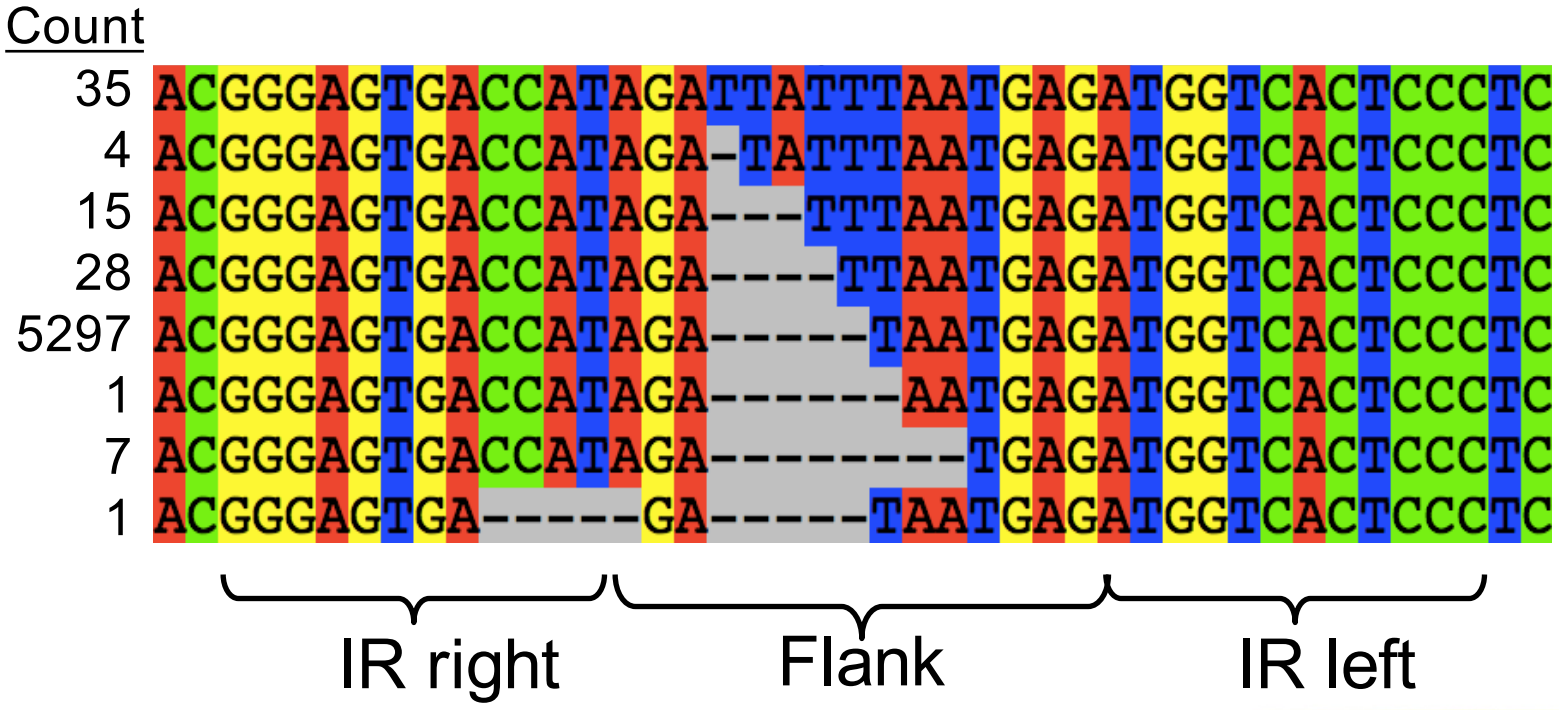
Overnight,
Exo-treated

SANDIA BIOSCIENCE

Sandia National Laboratories

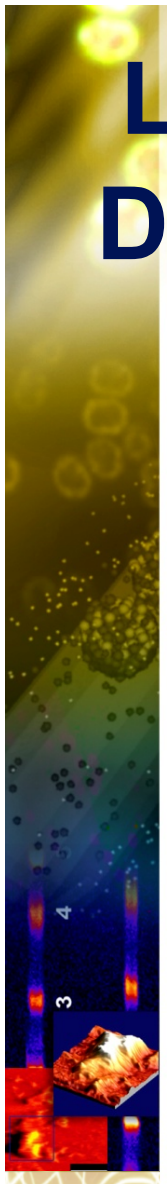
IS4321 and IS5075 precision

- Very similar, each at one copy at left end of Tn21
- Typically reach beyond their Inverted Repeats, i.e., they are site-specific
- Varying extent of flanking sequence captured in Circular Junctions



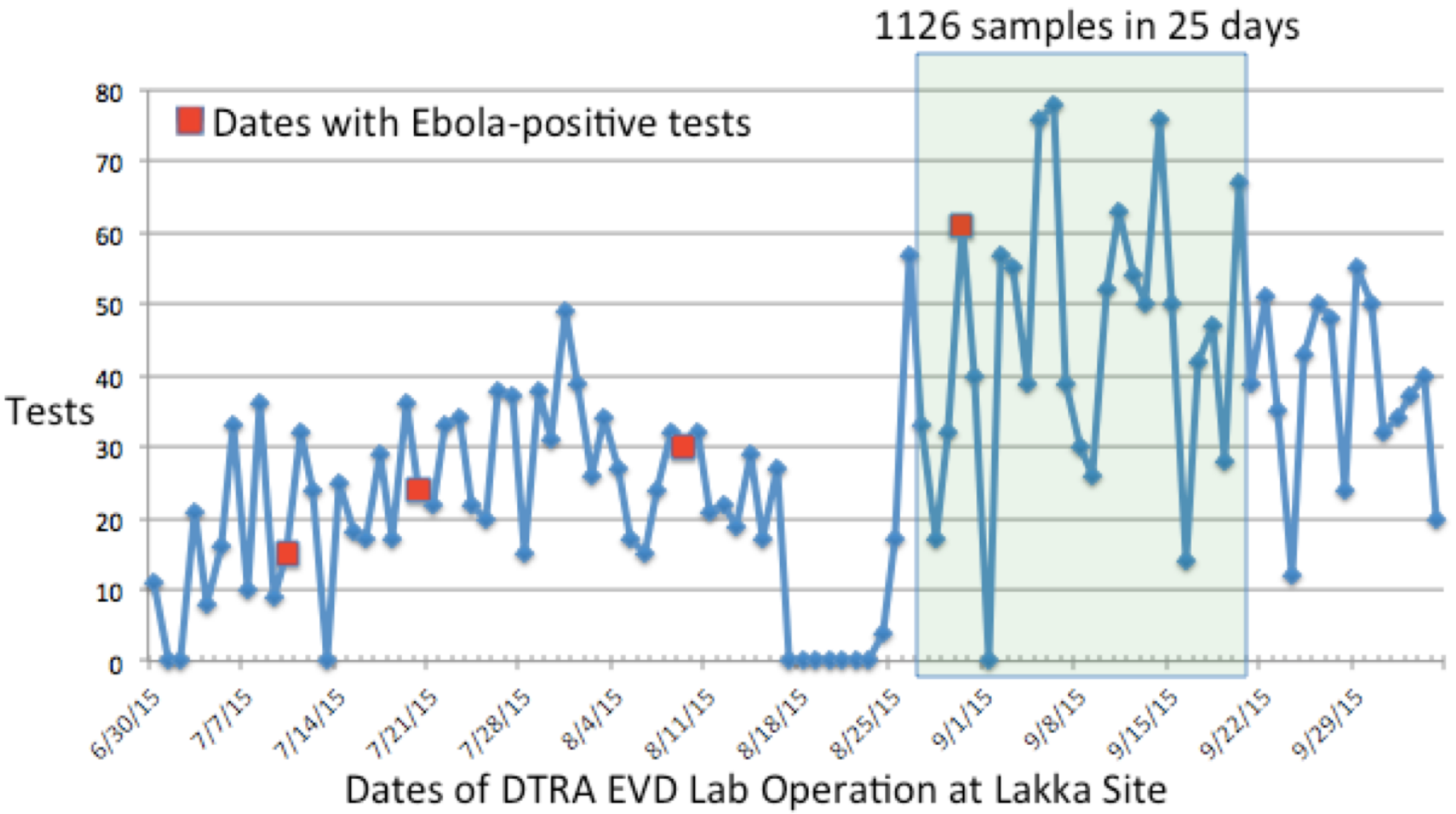
Lab Coordinator at Lakka Beach Ebola Diagnostic Lab: Freetown, Sierra Leone

- DTRA Cooperative Biological Engagement Program (CBEP) – Mobile Ebola Virus Disease (EVD) diagnostic laboratory
- Built and staffed by contractor MRI Global
- Sandia: Lab Coordinator and oversight role
  - Assist with daily operations
  - Interpret/verify test results and report out to local Ministry of Health and Sanitation (MoHS)
  - Lab liasion to MoHS, CDC, WHO, and NGOs
  - Attend (weekly) MoHS meetings at Former UN Special Court complex

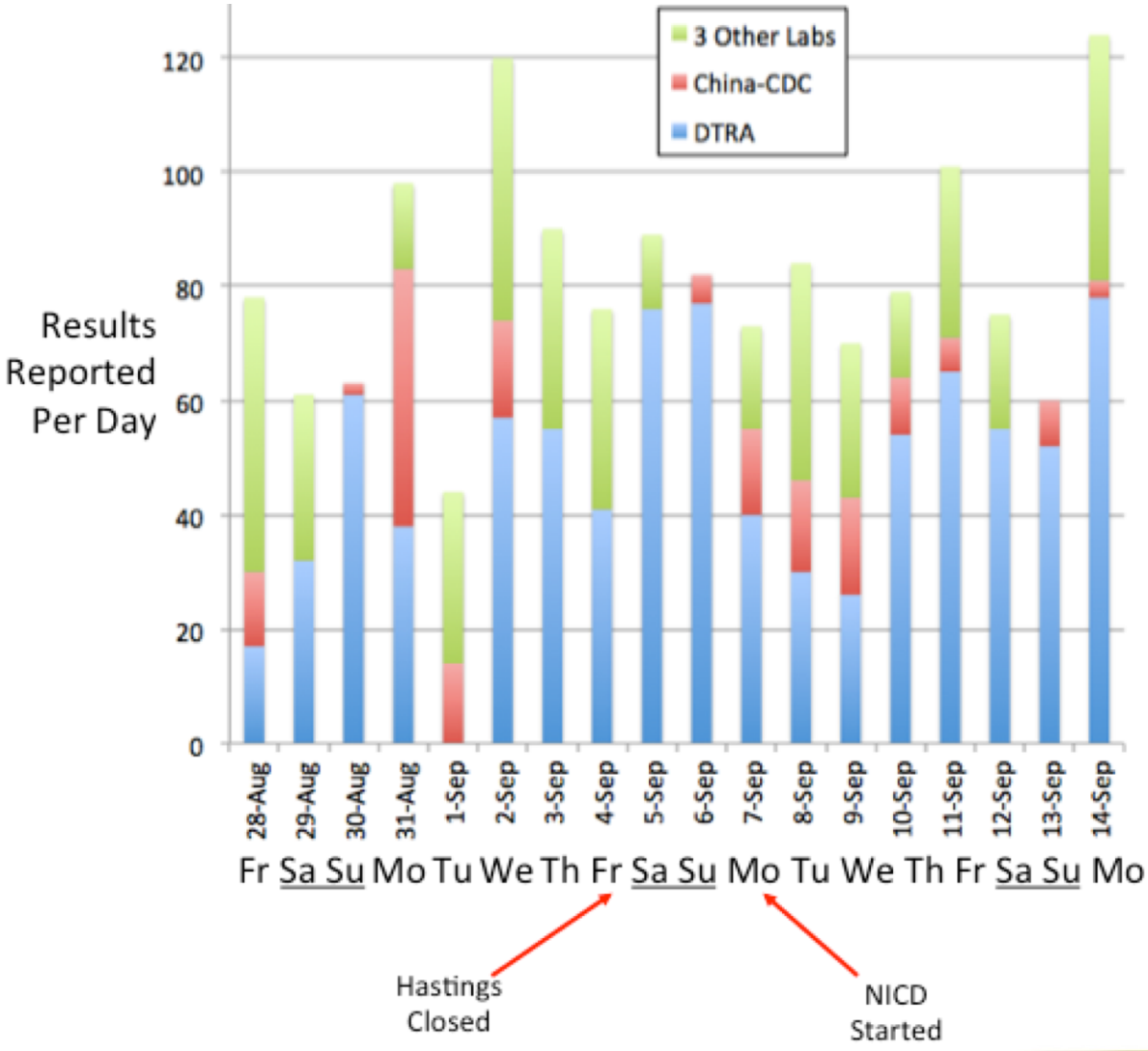


SANDIA BIOSCIENCE

Sample processing



Relative volume of lab



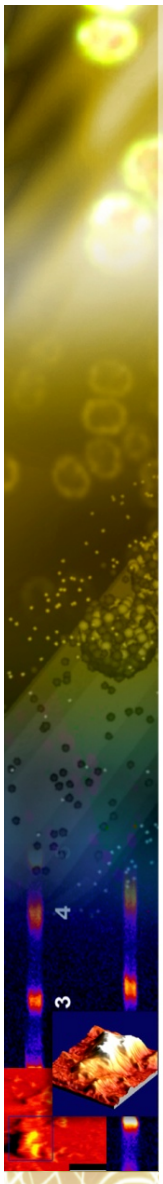
Total Tests
DTRA: 854
Others: 613

SANDIA BIOSCIENCE

Sandia National Laboratories

Screening chemicals for biofuel blendstocks

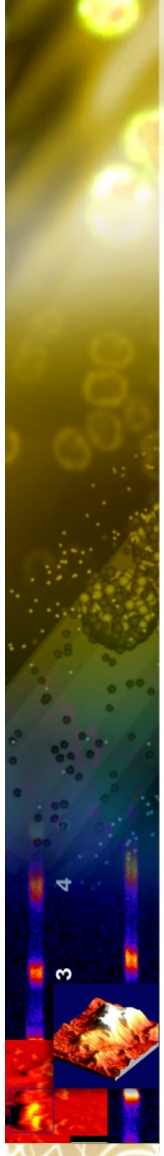
- Goal: Offset 30% of petroleum usage with Low Greenhouse Gas chemicals, without a dip in engine performance by 2025.
- Assumptions:
  - 1) Compounds with high Research Octane Numbers (RON: Translation-they burn well) will be able to substantially improve the quality of consumer grade fuel. (Caveats: A bunch of caveats, a whole bunch of caveats)
  - 2) We may not need to know the exact RON values.
  - 3) We have limited time and resources to critically evaluate all chemical properties of all potential chemicals.
  - 4) There are numerous chemicals.



SANDIA BIOSCIENCE

Enter Machine Learning

- Data collection of measured RON from 159 compounds
- Simple RON prediction using machine learning (i.e., Random Forest Classifier)
  - Model provides a very fast and course grained treatment for the assignment of chemicals into high/low RON class (e.g., > RON 85) and for triaging chemicals with no previous combustion measurements.
  - Mines data from public sources, including ChemSpider and PubChem.
  - ChemSpider properties primarily include chemical features (e.g., Boiling Point, Vapor Enthalpy, etc.)
  - PubChem includes 881 substructural fingerprint features: including hierarchical element counts, chemical ring systems, simple atom pairs, simple atom nearest neighbors, atomic neighborhoods, simple and complex smarts patterns.



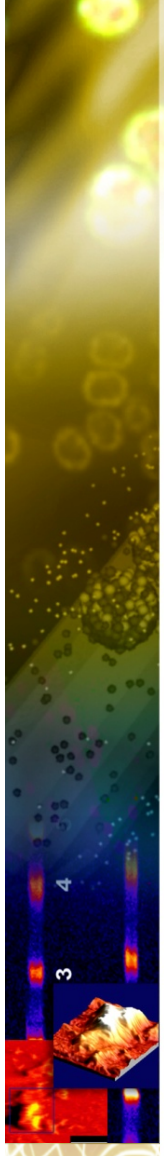
SANDIA BIOSCIENCE

Quality of Predictor

Performance of classifier:

- 100 sub-sampled cross-validations (with 50% leave out)

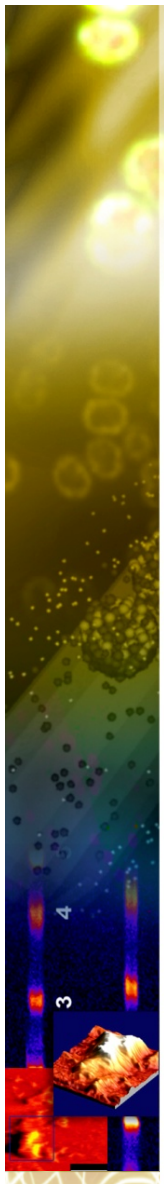
Metric	Mean value	Std. dev
Accuracy	0.80	0.09
Precision	0.82	0.12
Sensitivity	0.80	0.16
Receiver Operator Characteristic (AUC)	0.88	0.07



SANDIA BIOSCIENCE

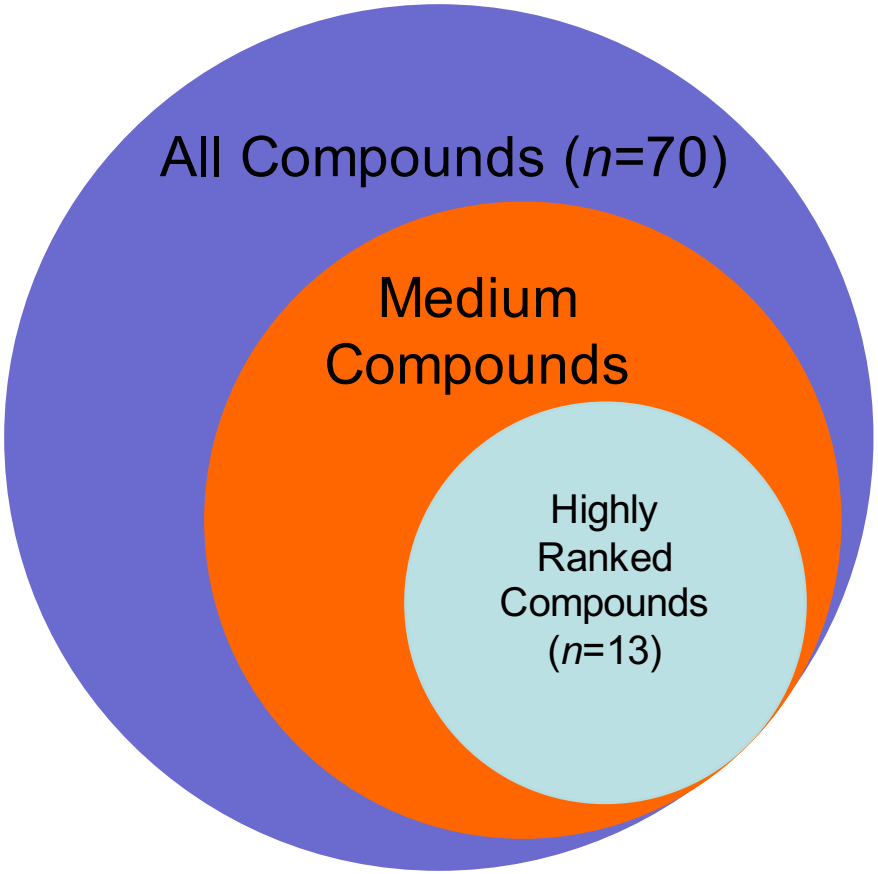
Chemoinformatics Data Mining

Features	Weight	Type
OH Rate Constant	0.0382	Physical
Surface Tension	0.0368	Physical
SMARTS Pattern: C-C-C-C-C-C	0.0349	Structural
KOC (pH 5.5)	0.0323	Physical
Vapor Pressure (mmHg at 25° C)	0.0313	Physical
SMARTS Pattern: C-C-C-C-C	0.0313	Structural
Octanol-Water Distribution Constant (pH 7.4)	0.0302	Physical



SANDIA BIOSCIENCE

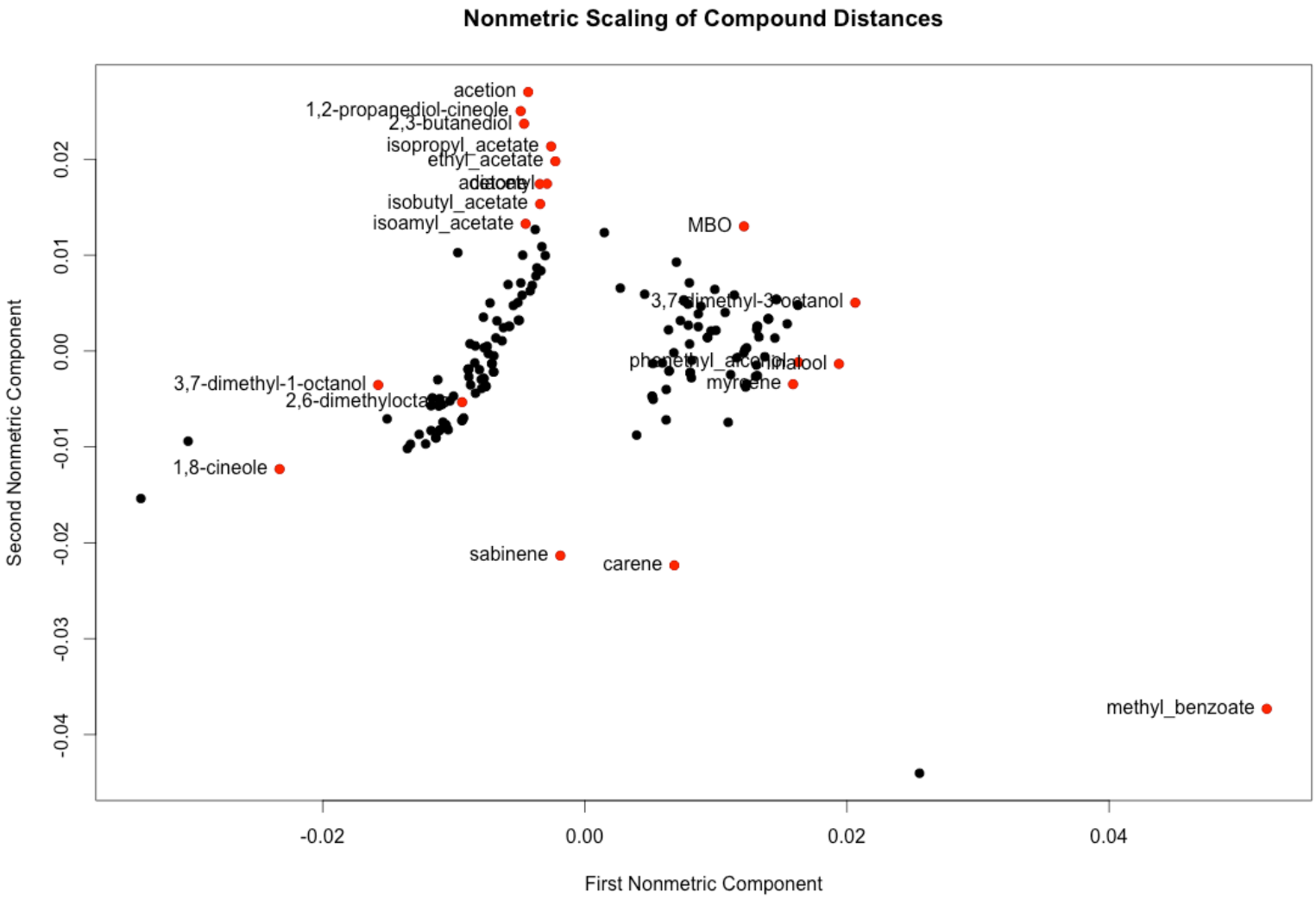
Initial Ranking for 70 Compounds of Interest



Compounds predicted to have RON > 85 in order of probability in class	
1. Isooctane	8. Isoprenol
2. Methylcyclopentane	9. Isobutanol
3. Ethanol	10. 3 methyl 1 butanol
4. Methyl butyrate	11. Butyl acetate
5. Ethyl isobutyrate	12. Toluene
6. Methyl 2-methylbutyrate	13. Isoamyl acetate
7. 2-methyl 2 butanol	

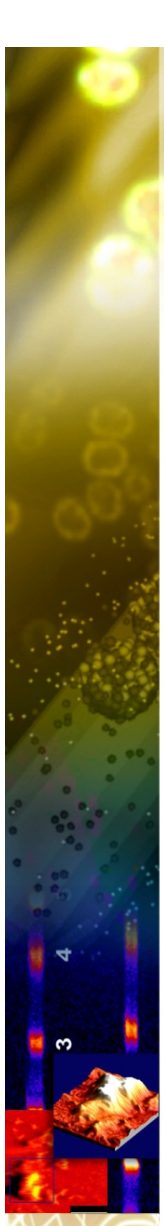
SANDIA BIOSCIENCE

Testing 20 new compounds and the limits of the model



Wrap-Up

- Career Track in a National Lab
  - Postdoc
  - Senior Member of Technical Staff
- Biodefense: Basic and applied work in emerging infectious diseases
- Bioenergy: Work in the production of high-value chemicals
- Work on mission areas



SANDIA BIOSCIENCE

Acknowledgements

Collaborators

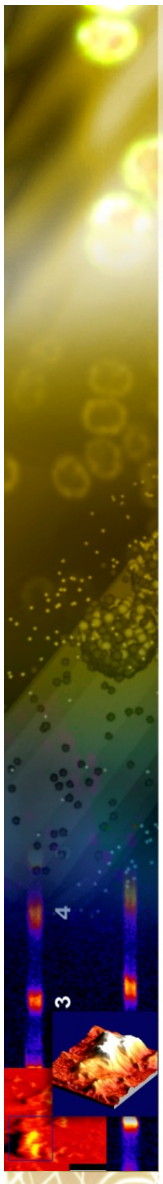
Kelly Williams	Gavin Conant	Leanne Whitmore
J. Chris Peers	Joe Schoeniger	Kunal Poorey
Seema Singh	Amy Powell	Debjit Ray
Zach Bent	Robert Meagher	Devin Petersohn
Britney Lau	Julian Wagner	

Funding

LDRD (Sandia) DOE (BETO, AOP, CSP)

DHS DOS DTRA

Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.



SANDIA BIOSCIENCE