NGS biases, systematic sequencing errors, and accuracy

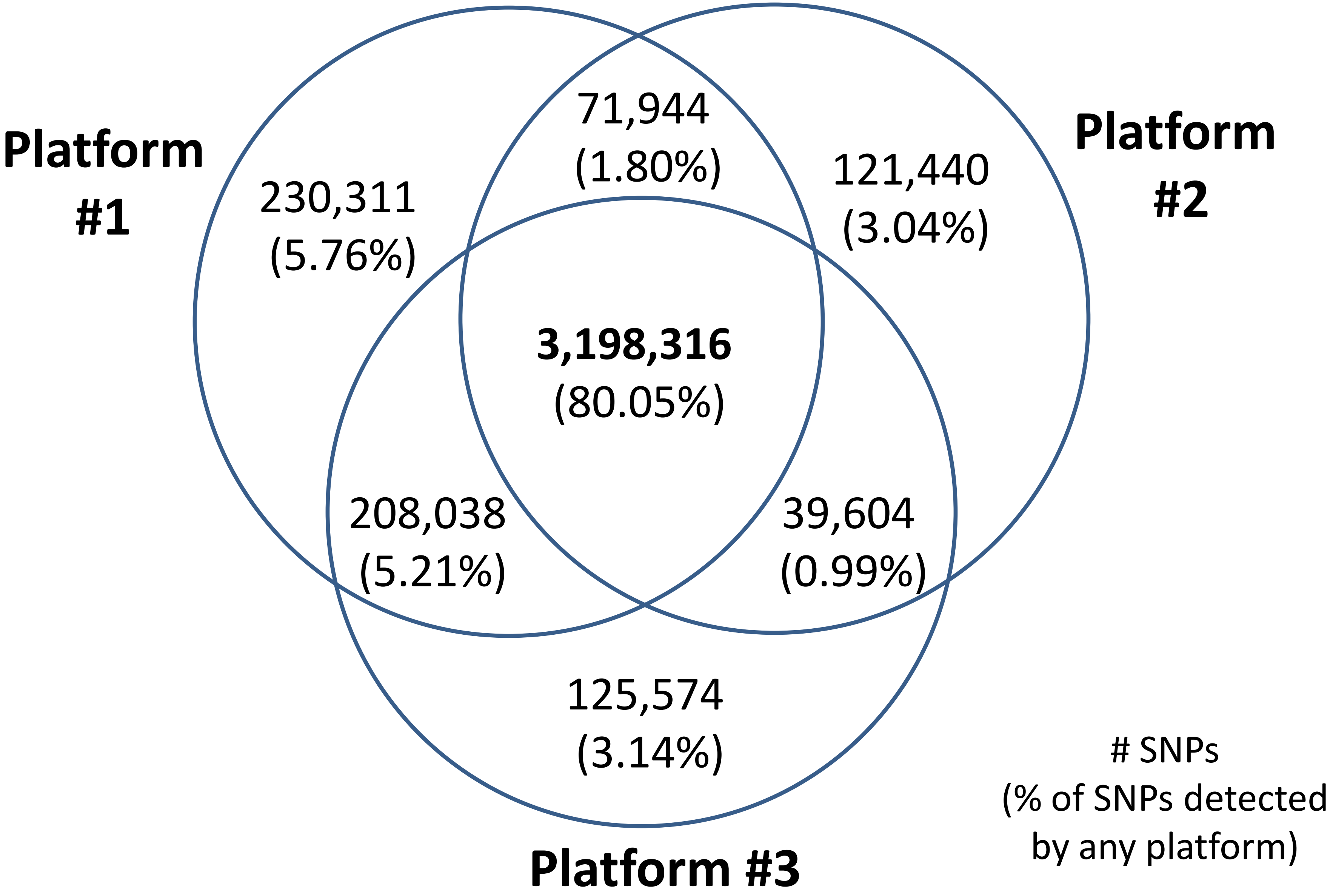
Justin Zook and Marc Salit

January 31, 2012

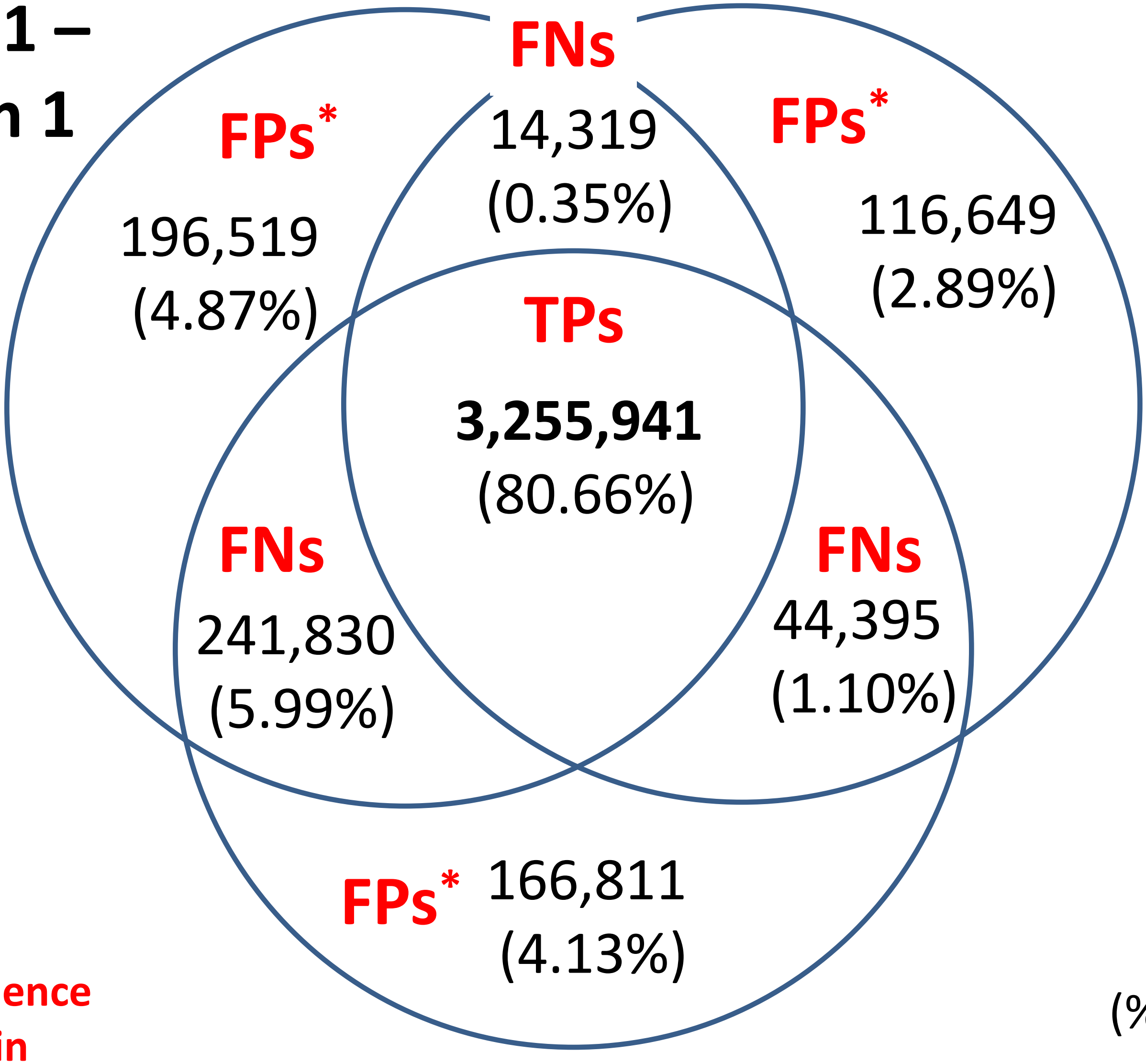
Reference Materials and NGS

- Comparison of SNPs on same sample
  - Different sequencing platforms
  - Different algorithms
  - Prospective Reference Material
- What causes these differences?
- Systematic errors and biases
- Utility of Reference Materials

Whole genome sequencing technologies disagree about 100,000's of SNPs



Different bioinformatics algorithms also disagree about 100,000's of SNPs



Platform 1 – Algorithm 1

Platform 2

Platform 1 – Algorithm 2

FPs* 196,519 (4.87%)

FNs 14,319 (0.35%)

FPs* 116,649 (2.89%)

TPs **3,255,941** (80.66%)

FNs 241,830 (5.99%)

FNs 44,395 (1.10%)

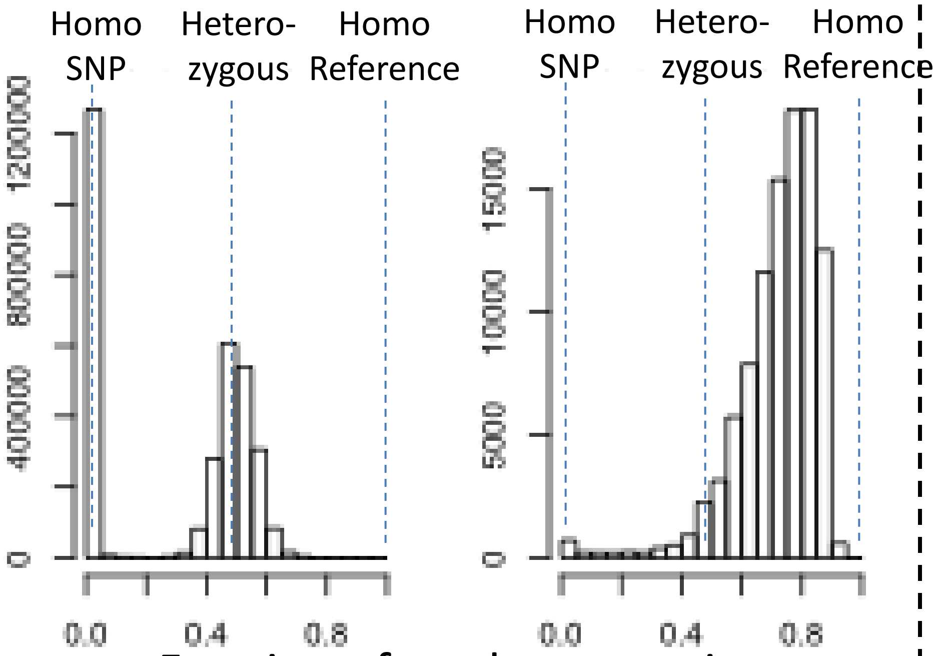
FPs* 166,811 (4.13%)

# SNPs
(% of SNPs detected by any method)

* If high-confidence reference call in another platform

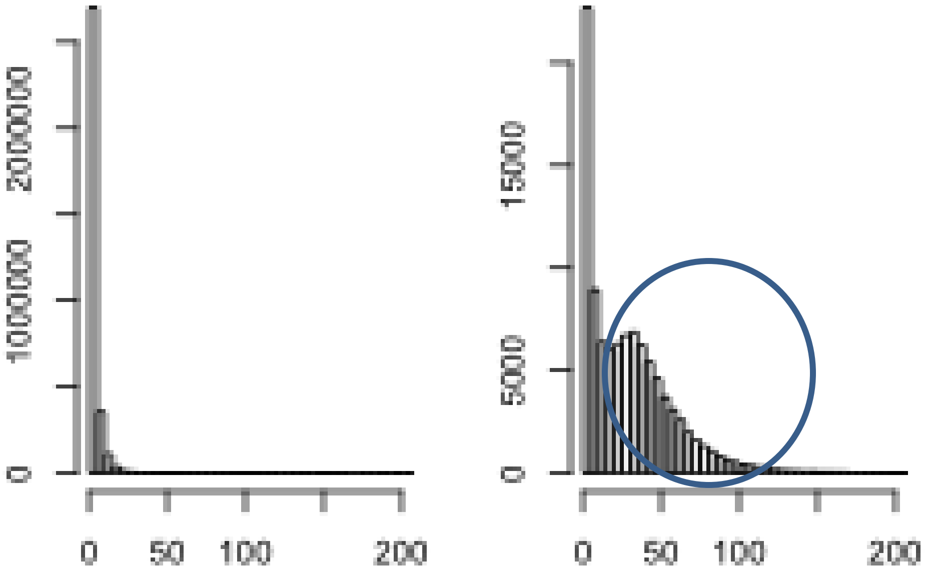
Some false positives have distinctive characteristics

True Positives **False Positives**



Fraction of reads supporting reference base

True Positives **False Positives**



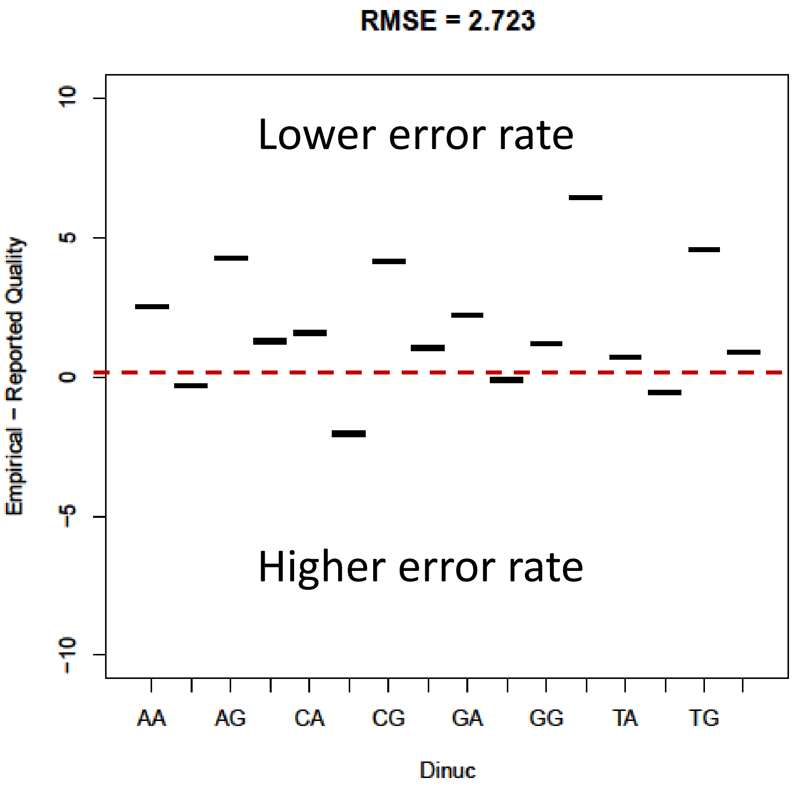
Strand bias test

Systematic errors cause more problems than random errors

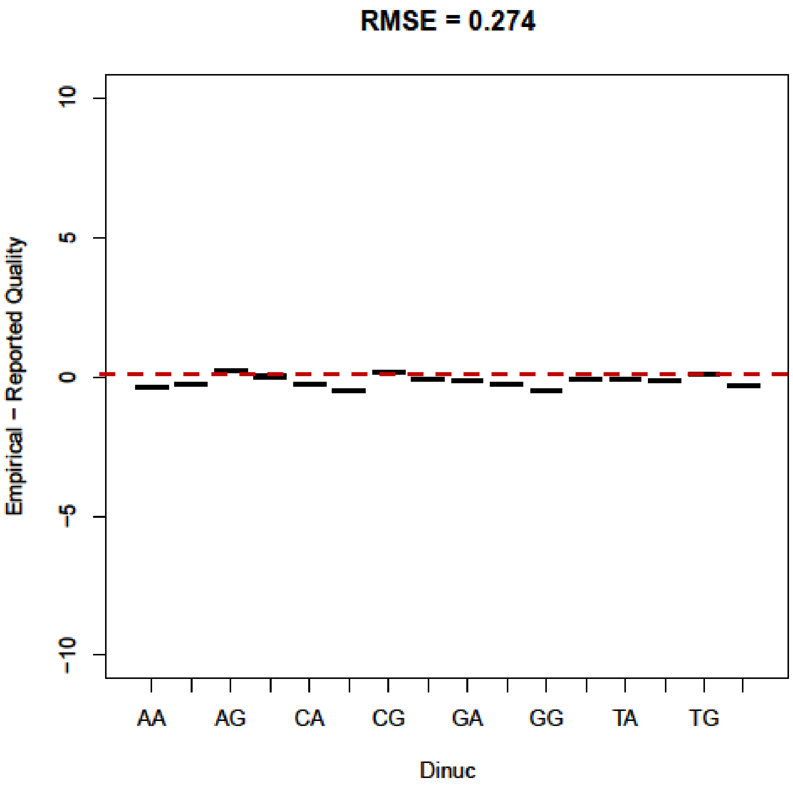
- Random errors can be modeled statistically
- High coverage sequencing minimizes importance of random errors
- Systematic errors remain at high coverage
- Many systematic errors are platform- or run-specific

Reference Materials can be used to detect and correct some systematic errors

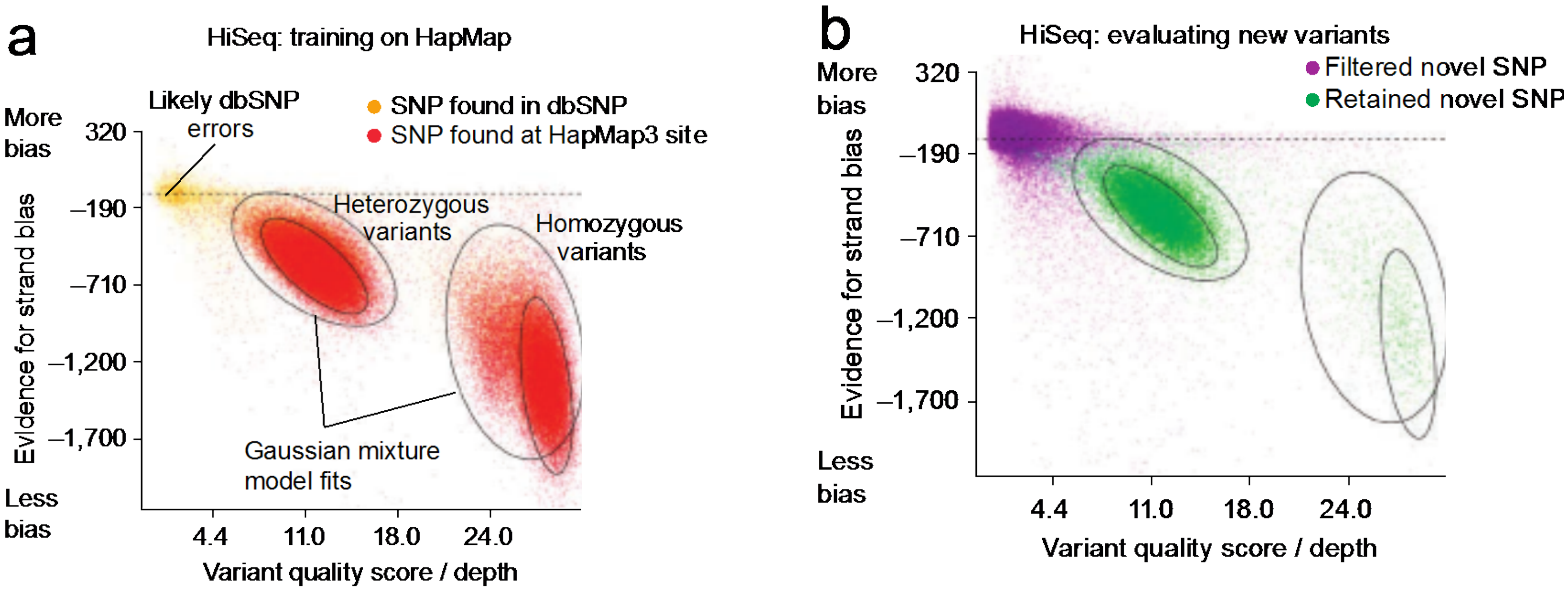
Before Recalibration



After Recalibration



Other algorithms recalibrate variant quality scores using known variants



DePristo, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data, *Nature Genetics*, **2011**, 43, 491.

Other systematic errors are more difficult to detect or correct

- Mapping/alignment ambiguities
  - Homopolymers and tandem repeats (STRs)
  - Complex variants
  - Multiallelic variants

- PCR problems
  - Homopolymers and tandem repeats (STRs)
  - GC-bias

Other types of variants are more difficult than SNPs

- Indels (scale – 1-10s of bases)
- Large insertions and deletions (>10s of bases)
- Copy number variants (CNVs)
- Inversions
- Complex structural rearrangements

Technologies are improving

- Higher coverage
- Less GC bias (e.g., new Illumina chemistry)
- Lower error rates (e.g., SOLiD ECC chemistry)
- Fewer systematic sequencing errors
- Longer reads -> fewer mapping ambiguities

Bioinformatics algorithms are improving

- Better base calling
- Faster and more accurate mapping
- Hybrid *de novo* assembly algorithms
- Methods to account for systematic errors and biases
- Algorithms to detect more complex variants

Reference Materials (RMs) can help

- Accuracy is very important for forensic applications
- Synthetic DNA RMs
  - Can be spiked-in to any sample
  - Can be used to detect and correct some SSEs
  - Can test detectability of specific types of variants
- Whole genome RMs
  - Characterized by multiple technologies
  - Will help improve technologies and algorithms
  - Provide constant benchmarks for rapidly changing technologies and algorithms

Questions?

- Contact information
  - Marc Salit: salit@nist.gov
  - Justin Zook: jzook@nist.gov