DATA ANALYSIS AND VALIDATION ON ONCOKB, GATEWAYSEQ, AND MYELOSEQ-HD

Justin Caringal & Ajay Khanna, Dr. Eric Duncavage Lab
Washington University of Saint Louis
Department of Pathology & Immunology

ANALYSIS OF ONCOKB API CALLS IN GATEWAYSEQ

STARTING POINT AND GOALS

- GatewaySeq: A tumor-only, high coverage targeted next generation sequencing assay for the identification of gene mutations, copy number alterations, microsatellite instability, tumor mutational burden, and gene fusions

- Evaluate three different methods for looking up variants in OncoKB from MSKCC via web API for interpretation provided to physicians
  - byGenomicChange, byProteinChange, byHGVSg

BRAF → 7,140453136,140453136,A,T p.V600E 7:g.140453136A>T

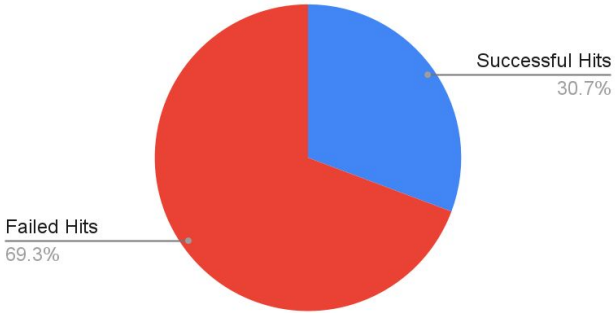
https://pathologyservices.wustl.edu/items/gatewayseq-ngs-panel-with-interpretation/

TABLE ANNOTATIONS (142 SAMPLES)

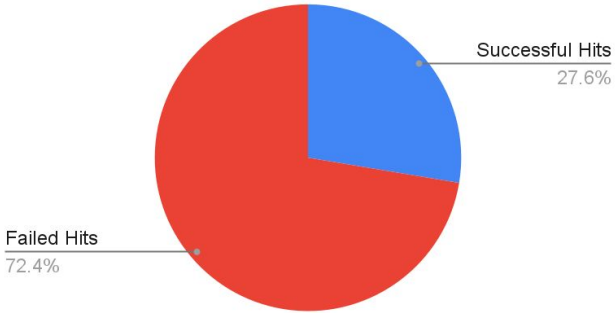
Comparisons	Differences
Genomic vs. Protein	+60
Genomic vs. HGVSg	-2
Protein vs. HGVSg	-60

	Successful Hits	Failed Hits
Genomic	538	1214
Protein	484	1268
HGVSg	540	1212
Total Searches	1752	

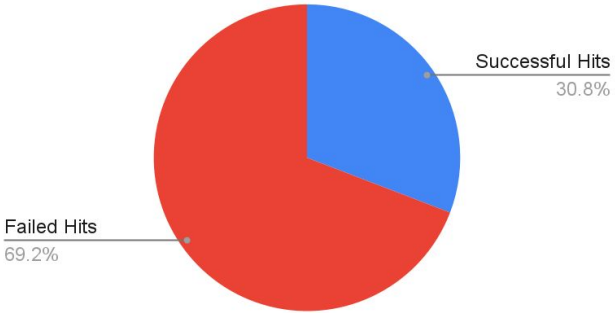
byGenomicChange Calls



byProteinChange Calls



byHGVSg Calls



DISCREPANCIES BETWEEN GATEWAYSEQ AND ONCOKB

- AKT1
- ATM
- B2M
- CD79B
- CHEK1
- DICER1
- FGFR1
- HRAS
- ITPKB
- MYD88
- NF1
- NFKBIE
- NTRK3
- PAX8
- PTPRD
- RAD51B
- RSPO3
- SGK1
- SMARCA4
- SMARCB1
- TCF3
- TFEB
- TNFAIP3

Gene/Transcript	1752
Correct ID	1489
Incorrect ID	263
Discrepancy Frequency	0.15

- OncoKB and GatewaySeq differed in choice of gene transcript
- GatewaySeq uses the Ensembl Canonical transcript
- AA coordinates may differ, could affect lookups

GATEWAYSEQ CONCLUSION

- Overall, small differences between API calls
- byProteinChange lagged behind both byGenomicChange and byHGVSg
- In cases of successful hit discrepancies
  - byProteinChange failure: No p.syntax available (e.g. splice variant)
  - byGenomicChange failure: Complex variants may not be found
- HGVSg is marginally better than Genomic
  - Example: ARID1A, ENST00000324856 (PASS) TAG→AA, complex variant
- Future Directions: Might be better to use byHGVSg in the pipeline

COVERAGE COMPARISONS IN MYELOSEQ-HD

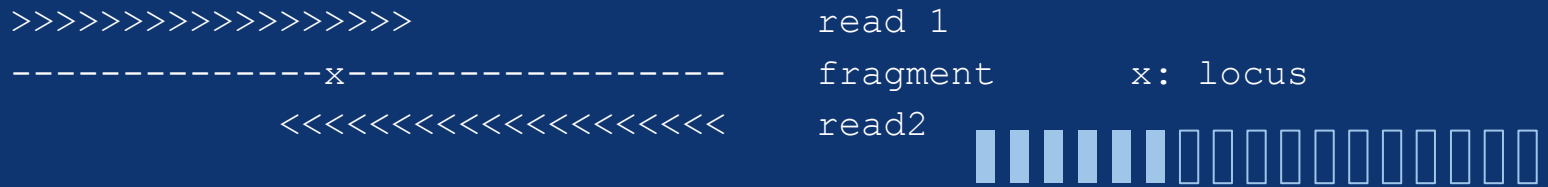
MYELOSEQ-HD

- Targeted sequencing assay for 49 genes and gene hotspots that are recurrently mutated in myeloid neoplasms, such as MDS and AML
- Uses a high coverage UMIs-based error corrected sequencing approach to achieve >95% sensitivity for previously identified mutations with VAFs ≥0.25%

https://pathologyservices.wustl.edu/items/myeloseq/

READ COVERAGE

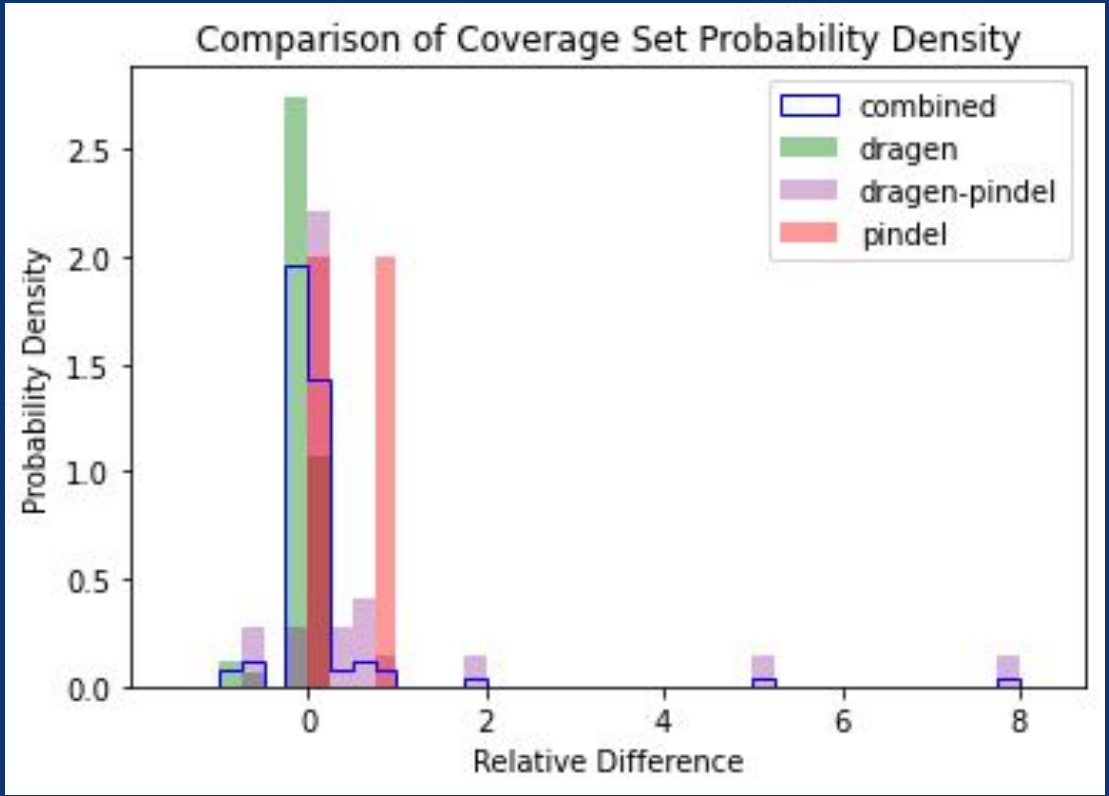
- New MyeloSeq-HD feature:
  - For previously identified variants, limit of detection (LOD) depends on sampling error
  - Sampling error depends on coverage
- Paired-End Sequencing reads can overlap with small enough fragment sizes
  - What is the coverage of loci in overlaps, 1 (collapsed) or 2?

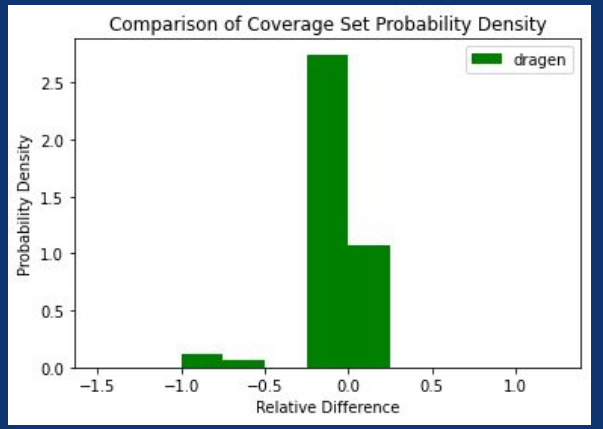


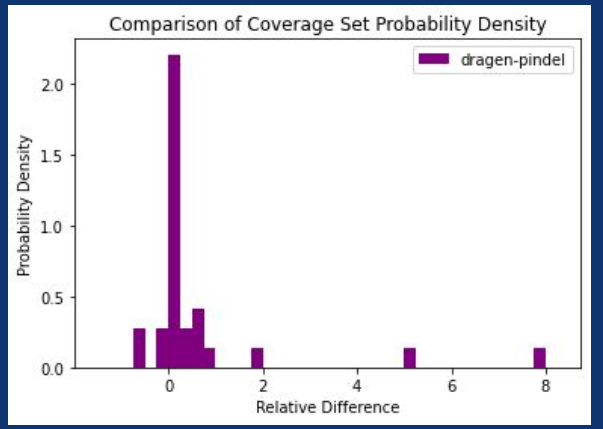
GENERAL PIPELINE

- MyeloSeq-HD coverage:
  - Dragen aligns FASTQ files, paired-end sequencing (150 base pairs), reports region coverage → BED
  - Call specific variants (Dragen, Pindel, combination) → VCF/JSON
    - Pindel & Dragen-Pindel uses custom Python script
- Goal: Compare output VCF and BED coverage to measure discrepancies

COMPARISONS (TIER 1-3 VARIANTS)







Pindel not shown (2 cases)

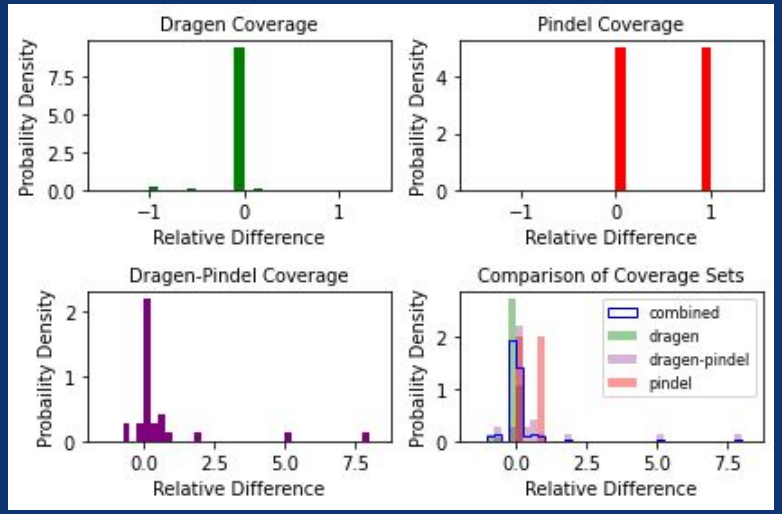
COMPARISONS (TIER 1-3 VARIANTS)

	dragen	pindel	dragen-pindel
Total extractions	67	2	29
Average Relative Difference	-0.03362	0.47623	0.599
Average Difference in Coverage	-85.46269	249.5	292.06897
Absolute differences greater than 50	5 (7.46%)	1 (50%)*	16 (55.17%)

*sample too small to make any meaningful conclusions

MYELOSEQ-HD CONCLUSION

- Dragen is the preferred method
  - <8% over 50 (| JSON – BED |)
- Data on Pindel is too small to draw meaningful conclusions
  - Only two Pindel TIER 1-3 variants in Batch 247
- Dragen-Pindel more prone to wild swings
- Future Directions: Integrating more Dragen, more data on Pindel



DRAGEN (67 VARIANTS)

JSON-BED
2496
2496
486
-9 to 0 (57)
-10
-11
-20
-78
-1664
-1778
-3865

- Coverage: order of thousands
- Most (57/67) were near-identical (-9 to 0)
- Examined NRAS variant alignments

CHROM	POS	REF	ALT	GENE	JSON BED	JSON-BED
chr1	114716127	C	T	NRAS	5664	7442	-1778

  - With Samtools mpileup (-Q10 -q20), able to reproduce JSON (variant caller) coverage of 5664
- May want to generate own coverage, or look at options for Dragen BED (region) coverage

JSON-BED
4629
2683
1734
1683
1624
650
615
578
473
445
432
286
174
98
11
0 to 5 (11)
-33
-3056
-4187

DRAGEN-PINDEL (29 VARIANTS)

- Many more large discrepancies
  - Tends to be higher in JSON (variant caller)
  - Possibly due to filtering script
- Did not have time to investigate further
  - We are thinking about removing filtering script

Pindel-only contains 2 cases

THANK YOU!

Special thanks to the Spencer Lab!

Questions?


slidesgo

Remove: slide 4

GOALS

- Implement similar structures using OncoKB API calls to byProteinChange and byHGVSg
- Explore JSON-to-table generation and annotation
- Compare API calls
  - Elapsed time
  - Successful hits

Remove: slide 5

EXAMPLE:

TWJV-GATEWAY-SEQ-S16-474-LIB2 REPORT JSON

	byGenomicChange	byProteinChange	byHGVSg
Total PASS	7	7	7
Skipped PASS	0	0	0
Annotated PASS	2	2	2
Total Filtered	26	26	26
Skipped Filtered	0	0	0
Annotated Filtered	3	3	3
Total Annotated	5	5	5
Elapsed (secs)	37.475	38.758	34.927

Implemented Elapsed time feature to be cross-compatible between machines

Remove: slide 6

JSON-TO-TABLE METHOD

- JSON → annotated JSON comparison inefficient
  - Unique variant duplicates increases query time
  - Queries for each specific tumor type
- JSON → table → annotated table more efficient
  - Removes duplicates
  - General query for information
  - Easier to analyze

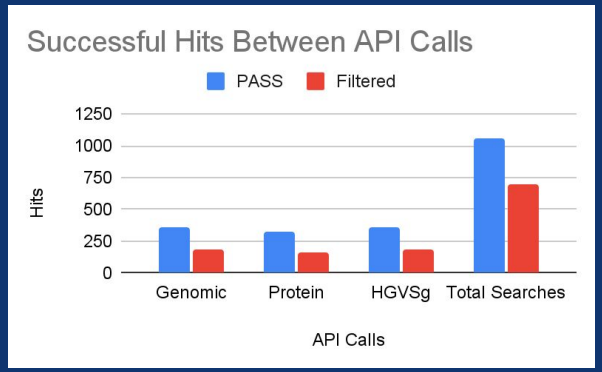
Remove: slide 8

ELAPSED TIME

	byGenomicChange	byProteinChange	byHGVSg
Total (secs)	390.918	299.029	318.2
Mean (secs)	0.223	0.171	0.182
Variance (secs²)	0.04379	0.03104	0.02863
Standard Deviation (secs)	0.209	0.176	0.169
Approx. Total Elapsed Time (secs)			1008.147
Total Elapsed, Standard Formatting		16 min, 48.147 secs	

Remove: slide 9

PASS VS. FILTERED VARIANTS



	PASS	Filtered	**PASS (%)**	**Filtered (%)**
Genomic	356	182	**33.62**	**26.26**
Protein	325	159	**30.69**	**22.94**
HGVSg	358	182	**33.81**	**26.26**
Total Searches	1059	693