Analysis of OncoKB API Calls

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

Starting Point (apIdx)

- Program takes input JSON file (unannotated)
- Program outputs annotated JSON file
  - Processed further down the pipeline (PDF generation)
- Also outputs .err file
  - Outputs statistics comparing PASS and Filtered Variants
  - Total, Skipped, Annotated
- All calls using byGenomicChange

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

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

JSON-to-Table Method

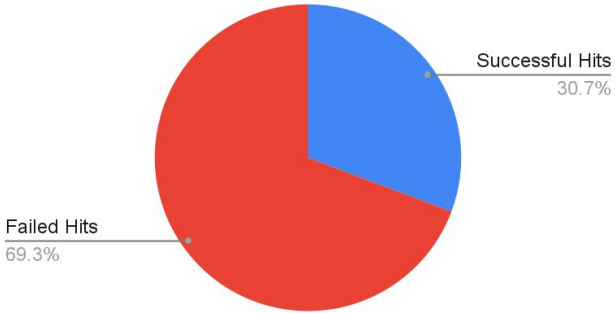
- JSON → annotated JSON comparison inefficient
  - Unique variant duplicates increases query time
  - Queries for each specific tumor type
- JSON → table → annotated table more efficient
  - Removes duplicates
  - General query for information
  - Easier to analyze

Table Annotations (142 JSON files)

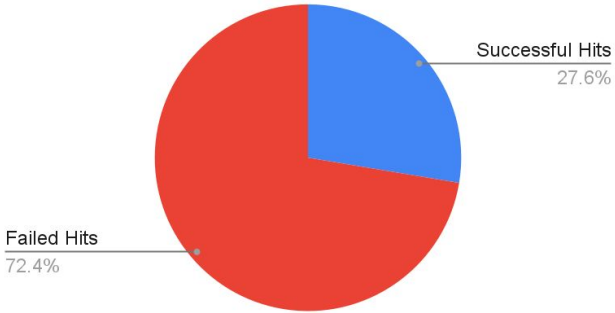
Comparisons	Differences
Genomic vs. Protein	+60
Genomic vs. HGVSg	-2
Protein vs. HGVSg	-60

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

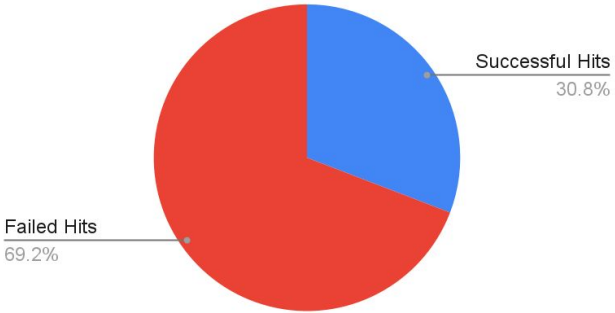
byGenomicChange Calls



byProteinChange Calls



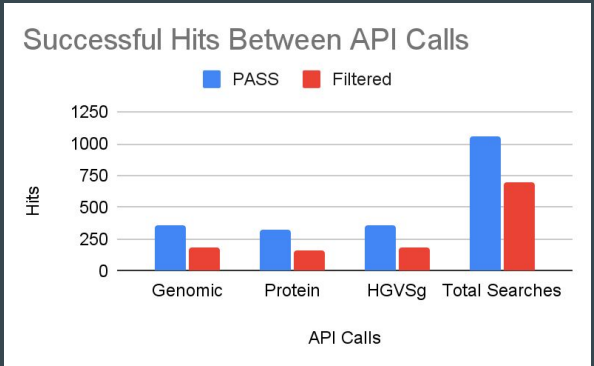
byHGVSg Calls



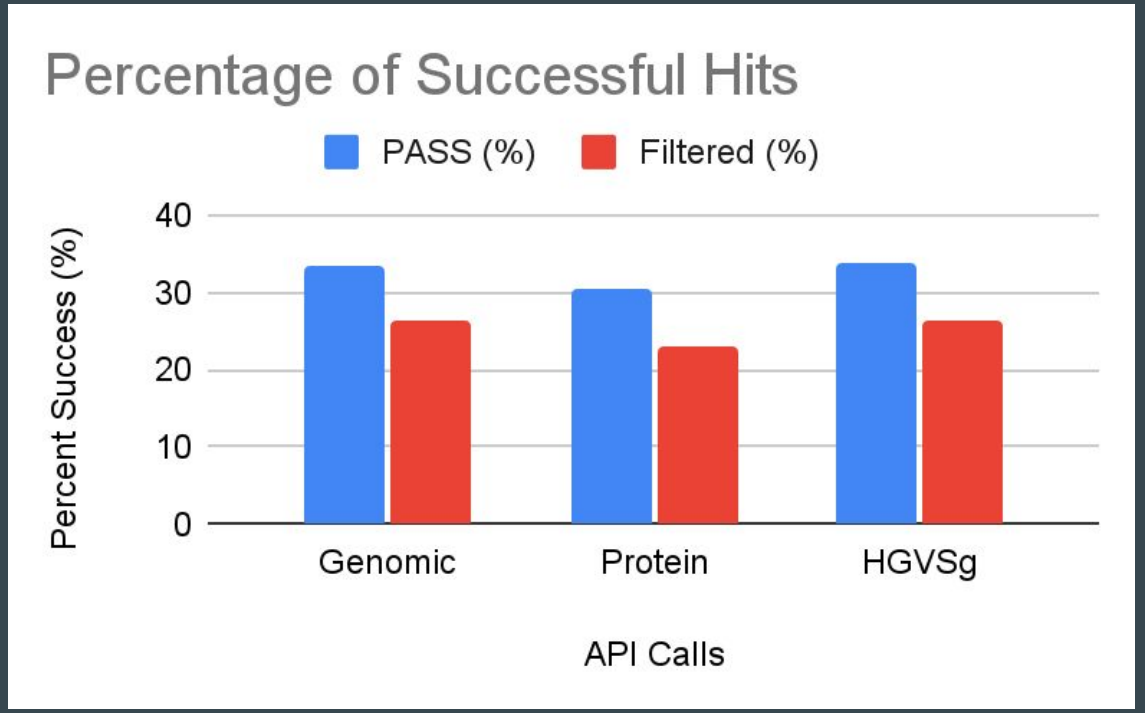
Elapsed Time

	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	

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		



Discrepancies between the University 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 WashU differed in their gene-transcript pairs
- Comparing with Ensembl, WashU is using the ‘canonical’ transcript (ensembl annotation canonical)

Conclusion

- Overall, small differences between API calls
- Protein lagged behind both Genomic and HGVSg
- In cases of successful hit discrepancies (XOR)
  - c.syntax was used in place of p.syntax for all (60)
  - Complex variants may not be found using p.syntax
- Other possible reasons
  - Transcript ID mismatch
  - Further investigation possible

Conclusion (cont'd)

- HGVSg is marginally better than Genomic
  - Possibly down to conversion of MAF and HGVSg strings
  - Genes:
    - MET, ENST00000397752 (PASS)
      - SNV: T→C, unknown, possible genomicLocation syntax error
    - ARID1A, ENST00000324856 (PASS)
      - INDEL: TAG→AA, complex variant
- May be possible to keep Genomic in the pipeline