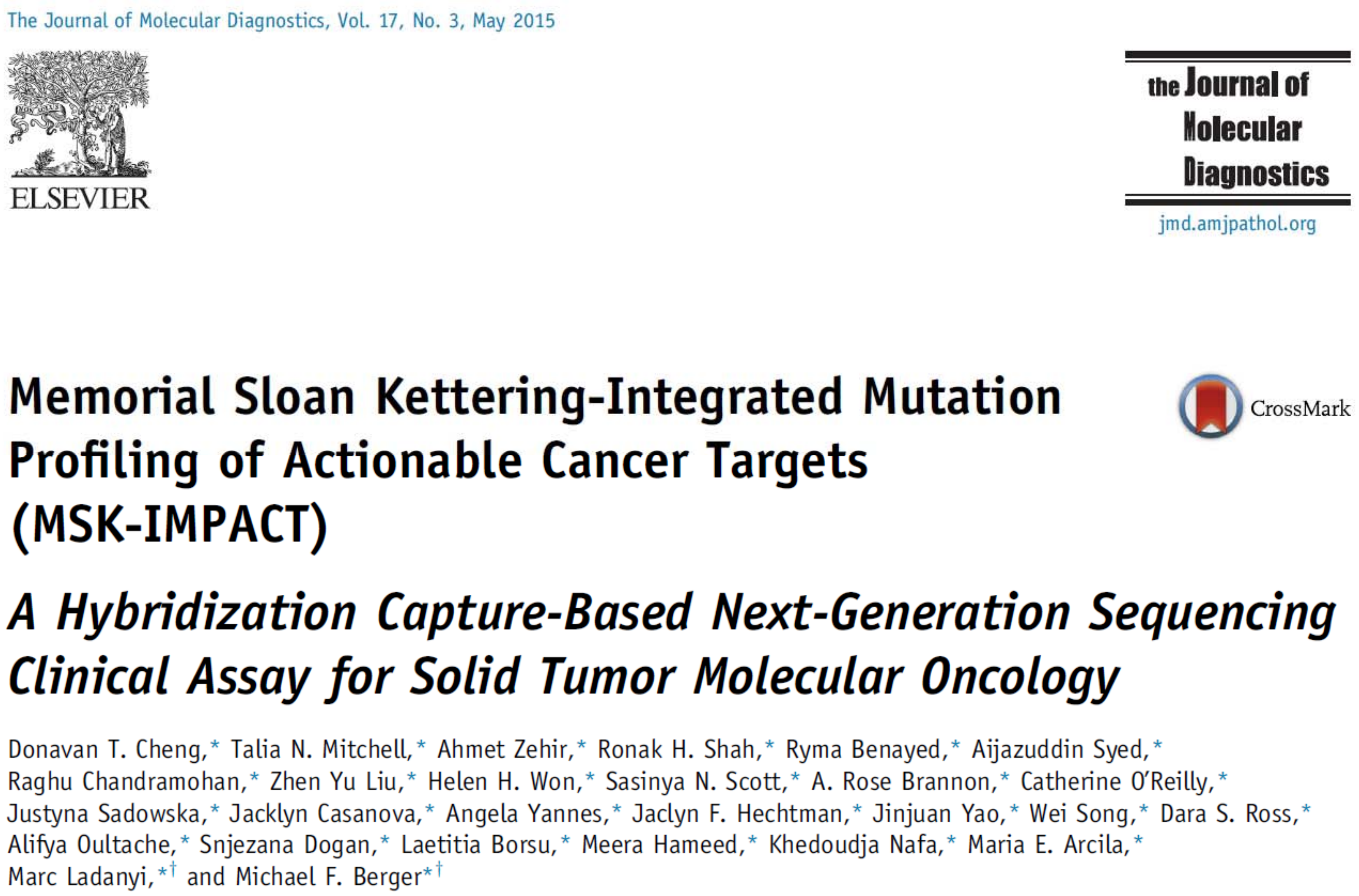
HOKKAIDO UNIVERSITY

Tumor profiling testの実践 MSK-IMPACT

北海道大学病院　臨床研究開発センター
特任助教　天野虎次

第25回抗悪性腫瘍薬開発フォーラム『ゲノム医療実用化の最前線』＠2018/06/30

MSK-IMPACT



The Journal of Molecular Diagnostics, Vol. 17, No. 3, May 2015

ELSEVIER

the Journal of Molecular Diagnostics

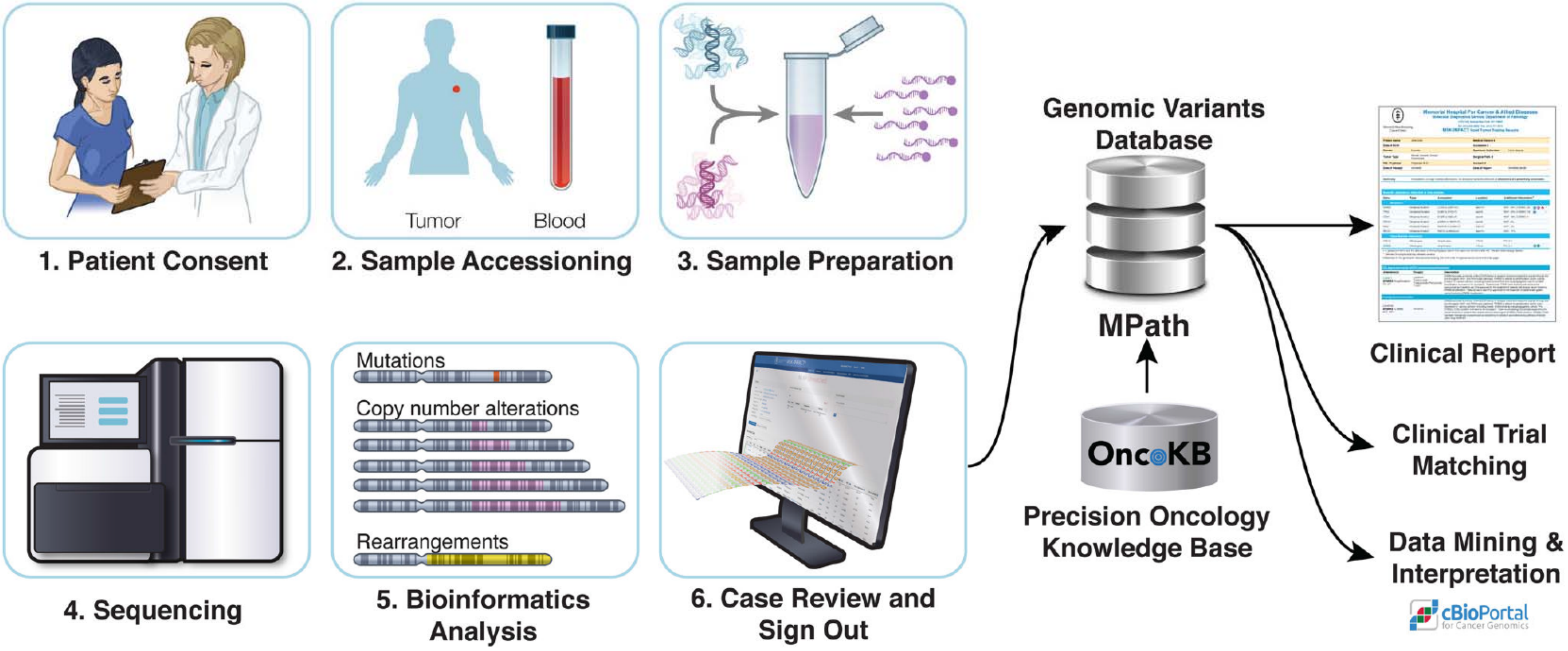
jmd.amjpathol.org

Memorial Sloan Kettering-Integrated Mutation Profiling of Actionable Cancer Targets (MSK-IMPACT)

A Hybridization Capture-Based Next-Generation Sequencing Clinical Assay for Solid Tumor Molecular Oncology

Donavan T. Cheng,* Talia N. Mitchell,* Ahmet Zehir,* Ronak H. Shah,* Ryma Benayed,* Aijazuddin Syed,* Raghu Chandramohan,* Zhen Yu Liu,* Helen H. Won,* Sasinya N. Scott,* A. Rose Brannon,* Catherine O'Reilly,* Justyna Sadowska,* Jacklyn Casanova,* Angela Yannes,* Jaclyn F. Hechtman,* Jinjuan Yao,* Wei Song,* Dara S. Ross,* Alifya Oultache,* Snjezana Dogan,* Laetitia Borsu,* Meera Hameed,* Khedoudja Nafa,* Maria E. Arcila,* Marc Ladanyi,*† and Michael F. Berger*†

MSK-IMPACT Clinical Workflow



Zehir et al., Nat Med. 2017 June ; 23(6): 703–713

MSK-IMPACT Bioinformatics Team

＜ Pathology : Diagnostic Molecular Pathology Service＞

Dr. Ladanyi (PI)　Lab

- 主なStuff : 8人
- Researcher : 10+人
- Technician: 4+人（MSK-IMPACT専任）

＜ CMO (Center for Molecular Oncology) ＞

Dr. Berger (PI)　Lab

- Researcher : 6+人
- MSK-IMPACT Bioinformatics team : 8人

cBioPortal Bioinformatics team : 20人　+　SE(10+人)

実際の運用規模 (2016年)

- IMPACT update（2016/06/13）; 4 batch 12pool (n=12142)

		Mon	Tue	Wed	Thr	Fri
6/13	New batch	229,230,231	232,233,234	235,236,237	238,239,240	-
	Sequence	222,223,224, 225,226	-	227,228	-	229,230,231

- **100-140 patients/week (1 pool = 10~12 case)**
- **Turn Around Time : 14+日（accession ~ sign out）**
- **QC meeting：毎週月曜:60分程度**
- **sign-out　：Pathologistと日程調整して適宜 (20 case / mtg)**

MSK-IMPACT-pipeline

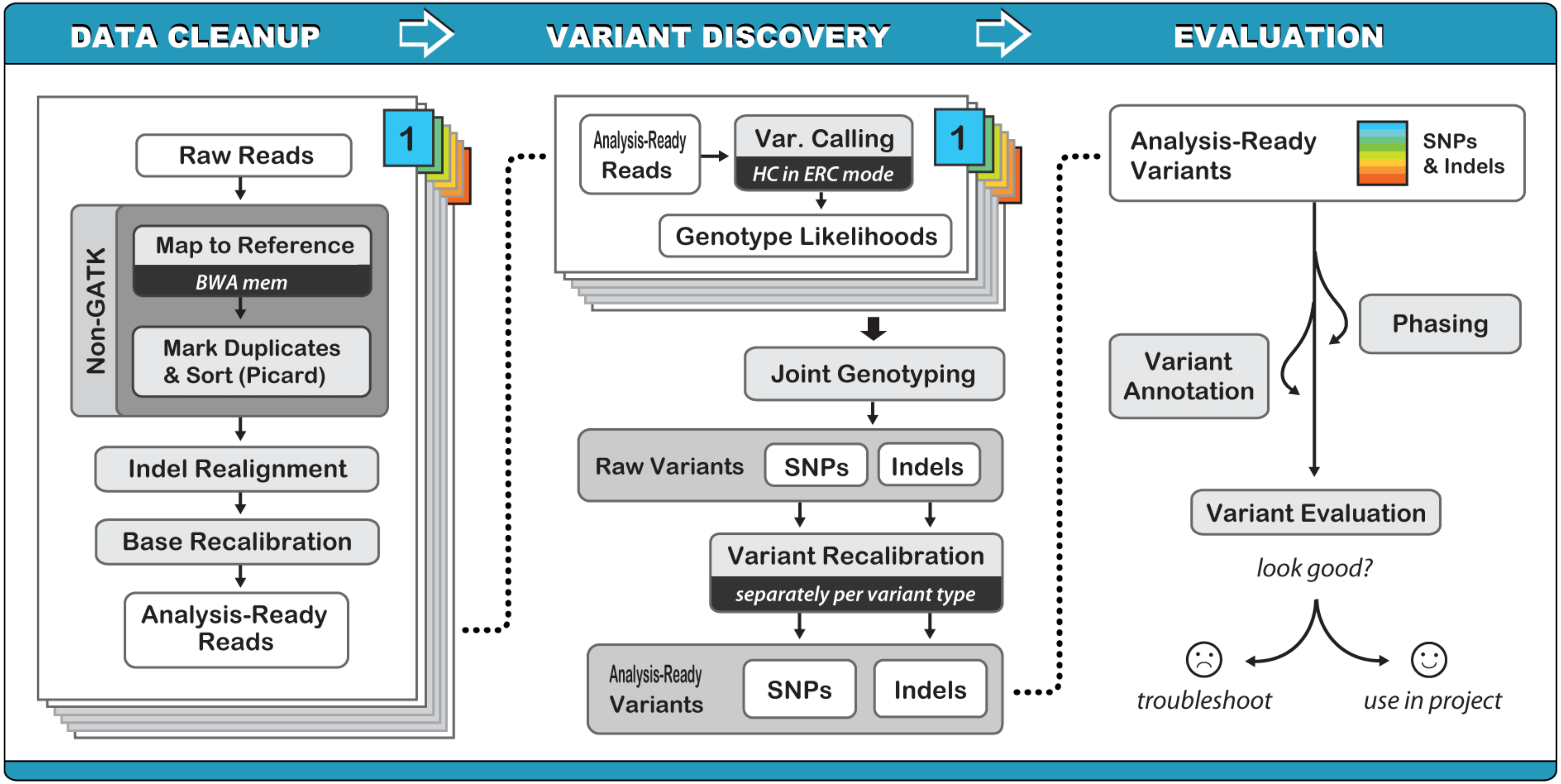
● 使用している解析ツールは一般的なもの。

GATK, Picard, samtools 等々。

● 使用するpackageの選択は、比較検討を行って確認して決定している。(ABRA、Delly など）

● Filteringおよび一部の解析に自作プログラムを使用

Analysis (GATK Best Practices)



DATA CLEANUP ⇨ VARIANT DISCOVERY ⇨ EVALUATION

Raw Reads

Non-GATK

Map to Reference

BWA mem

Mark Duplicates & Sort (Picard)

Indel Realignment

Base Recalibration

Analysis-Ready Reads

1

Analysis-Ready Reads

Var. Calling

HC in ERC mode

Genotype Likelihoods

1

Joint Genotyping

Raw Variants SNPs Indels

Variant Recalibration

separately per variant type

Analysis-Ready Variants SNPs Indels

Analysis-Ready Variants

SNPs & Indels

Phasing

Variant Annotation

Variant Evaluation

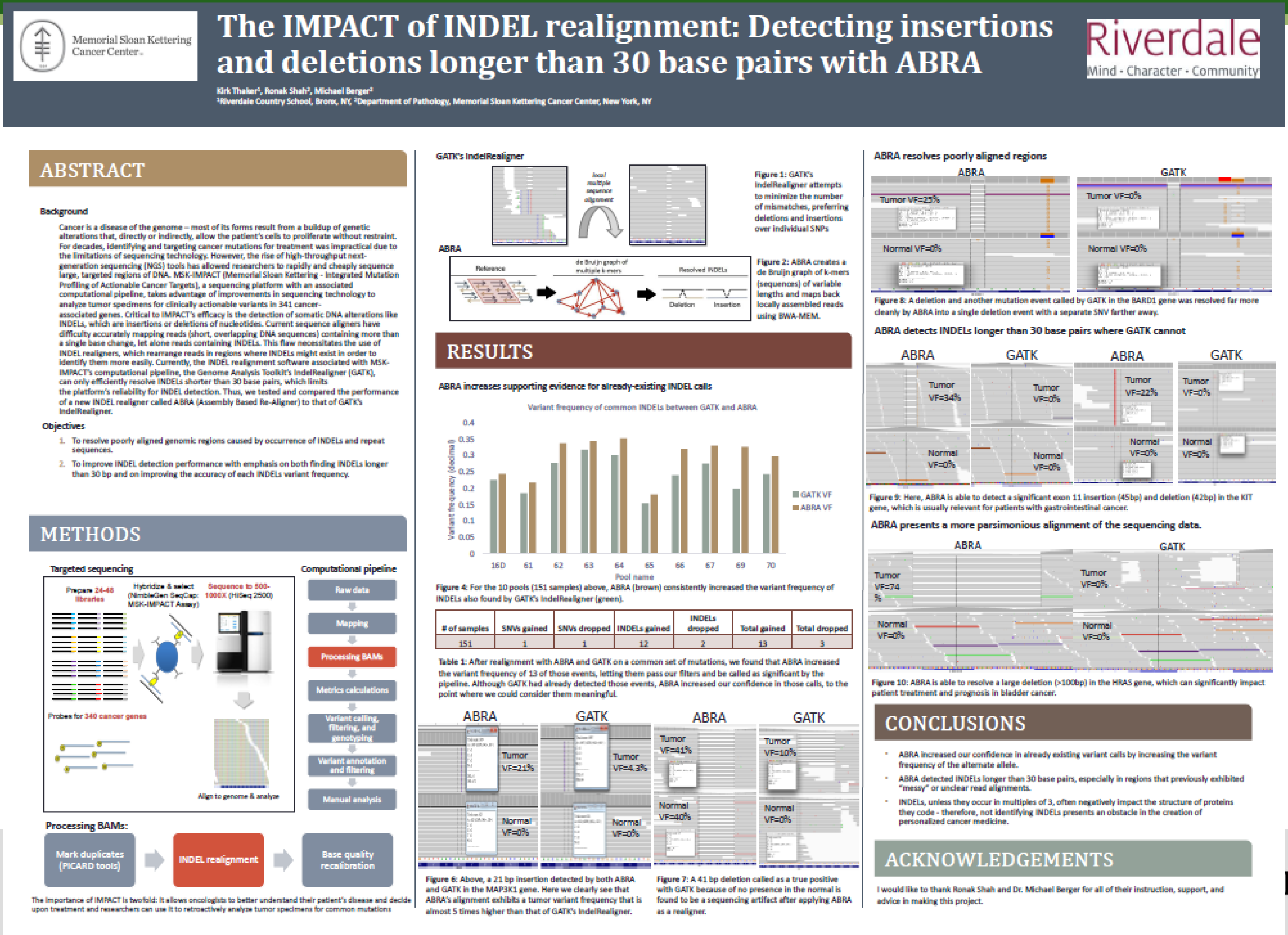
look good?

☹ *troubleshoot*

☺ *use in project*

https://software.broadinstitute.org/gatk/

Example of the Package Comparison



The IMPACT of INDEL realignment: Detecting insertions and deletions longer than 30 base pairs with ABRA

Kirk Thaker[1], Ronak Shah[2], Michael Berger[2]
[1]Riverdale Country School, Bronx, NY, [2]Department of Pathology, Memorial Sloan Kettering Cancer Center, New York, NY

ABSTRACT

Background

Cancer is a disease of the genome – most of its forms result from a buildup of genetic alterations that, directly or indirectly, allow the patient's cells to proliferate without restraint. For decades, identifying and targeting cancer mutations for treatment was impractical due to the limitations of sequencing technology. However, the rise of high-throughput next-generation sequencing (NGS) tools has allowed researchers to rapidly and cheaply sequence large, targeted regions of DNA. MSK-IMPACT (Memorial Sloan Kettering - Integrated Mutation Profiling of Actionable Cancer Targets), a sequencing platform with an associated computational pipeline, takes advantage of improvements in sequencing technology to analyze tumor specimens for clinically actionable variants in 341 cancer-associated genes. Critical to IMPACT's efficacy is the detection of somatic DNA alterations like INDELs, which are insertions or deletions of nucleotides. Current sequence aligners have difficulty accurately mapping reads (short, overlapping DNA sequences) containing more than a single base change, let alone reads containing INDELs. This flaw necessitates the use of INDEL realigners, which rearrange reads in regions where INDELs might exist in order to identify them more easily. Currently, the INDEL realignment software associated with MSK-IMPACT's computational pipeline, the Genome Analysis Toolkit's IndelRealigner (GATK), can only efficiently resolve INDELs shorter than 30 base pairs, which limits the platform's reliability for INDEL detection. Thus, we tested and compared the performance of a new INDEL realigner called ABRA (Assembly Based Re-Aligner) to that of GATK's IndelRealigner.

Objectives

1. To resolve poorly aligned genomic regions caused by occurrence of INDELs and repeat sequences.
2. To improve INDEL detection performance with emphasis on both finding INDELs longer than 30 bp and on improving the accuracy of each INDELs variant frequency.

METHODS

Targeted sequencing

Prepare 24-48 libraries
Hybridize & select (NimbleGen SeqCap: MSK-IMPACT Assay)
Sequence to 500-1000X (HiSeq 2500)
Probes for 340 cancer genes
Align to genome & analyze

Computational pipeline

Raw data
Mapping
Processing BAMs
Metrics calculations
Variant calling, filtering, and genotyping
Variant annotation and filtering
Manual analysis

Processing BAMs:

The importance of IMPACT is twofold: it allows oncologists to better understand their patient's disease and decide upon treatment and researchers can use it to retroactively analyze tumor specimens for common mutations

GATK's IndelRealigner

Figure 1: GATK's IndelRealigner attempts to minimize the number of mismatches, preferring deletions and insertions over individual SNPs

ABRA

Figure 2: ABRA creates a de Bruijn graph of k-mers (sequences) of variable lengths and maps back locally assembled reads using BWA-MEM.

RESULTS

ABRA increases supporting evidence for already-existing INDEL calls

Variant frequency of common INDELs between GATK and ABRA
Variant frequency (decimal)
0
0.05
0.1
0.15
0.2
0.25
0.3
0.35
0.4
16D
61
62
63
64
65
66
67
69
70
Pool name
GATK VF
ABRA VF

Figure 4: For the 10 pools (151 samples) above, ABRA (brown) consistently increased the variant frequency of INDELs also found by GATK's IndelRealigner (green).

# of samples	SNVs gained	SNVs dropped	INDELs gained	INDELs dropped	Total gained	Total dropped
151	1	1	12	2	13	3

Table 1: After realignment with ABRA and GATK on a common set of mutations, we found that ABRA increased the variant frequency of 13 of those events, letting them pass our filters and be called as significant by the pipeline. Although GATK had already detected those events, ABRA increased our confidence in those calls, to the point where we could consider them meaningful.

Figure 6: Above, a 21 bp insertion detected by both ABRA and GATK in the MAP3K1 gene. Here we clearly see that ABRA's alignment exhibits a tumor variant frequency that is almost 5 times higher than that of GATK's IndelRealigner.

Figure 7: A 41 bp deletion called as a true positive with GATK because of no presence in the normal is found to be a sequencing artifact after applying ABRA as a realigner.

ABRA resolves poorly aligned regions

Figure 8: A deletion and another mutation event called by GATK in the BARD1 gene was resolved far more cleanly by ABRA into a single deletion event with a separate SNV farther away.

ABRA detects INDELs longer than 30 base pairs where GATK cannot

Figure 9: Here, ABRA is able to detect a significant exon 11 insertion (45bp) and deletion (42bp) in the KIT gene, which is usually relevant for patients with gastrointestinal cancer.

ABRA presents a more parsimonious alignment of the sequencing data.

Figure 10: ABRA is able to resolve a large deletion (>100bp) in the HRAS gene, which can significantly impact patient treatment and prognosis in bladder cancer.

CONCLUSIONS

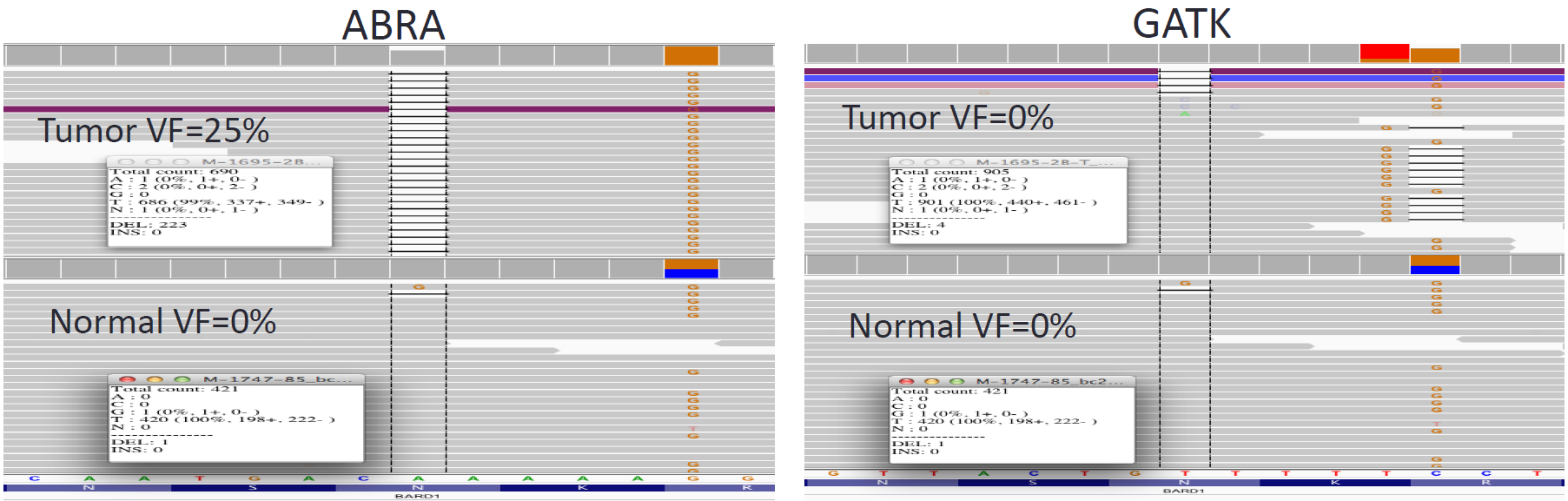
- ABRA increased our confidence in already existing variant calls by increasing the variant frequency of the alternate allele.
- ABRA detected INDELs longer than 30 base pairs, especially in regions that previously exhibited "messy" or unclear read alignments.
- INDELs, unless they occur in multiples of 3, often negatively impact the structure of proteins they code - therefore, not identifying INDELs presents an obstacle in the creation of personalized cancer medicine.

ACKNOWLEDGEMENTS

I would like to thank Ronak Shah and Dr. Michael Berger for all of their instruction, support, and advice in making this project.

Assembly-base and Mapping-base

ABRA resolves poorly aligned regions



```
M-1695-2B...
Total count: 690
A : 1 (0%, 1+, 0- )
C : 2 (0%, 0+, 2- )
G : 0
T : 686 (99%, 337+, 349- )
N : 1 (0%, 0+, 1- )
-----------------
DEL: 223
INS: 0
```

```
M-1747-85_bc...
Total count: 421
A : 0
C : 0
G : 1 (0%, 1+, 0- )
T : 420 (100%, 198+, 222- )
N : 0
-----------------
DEL: 1
INS: 0
```

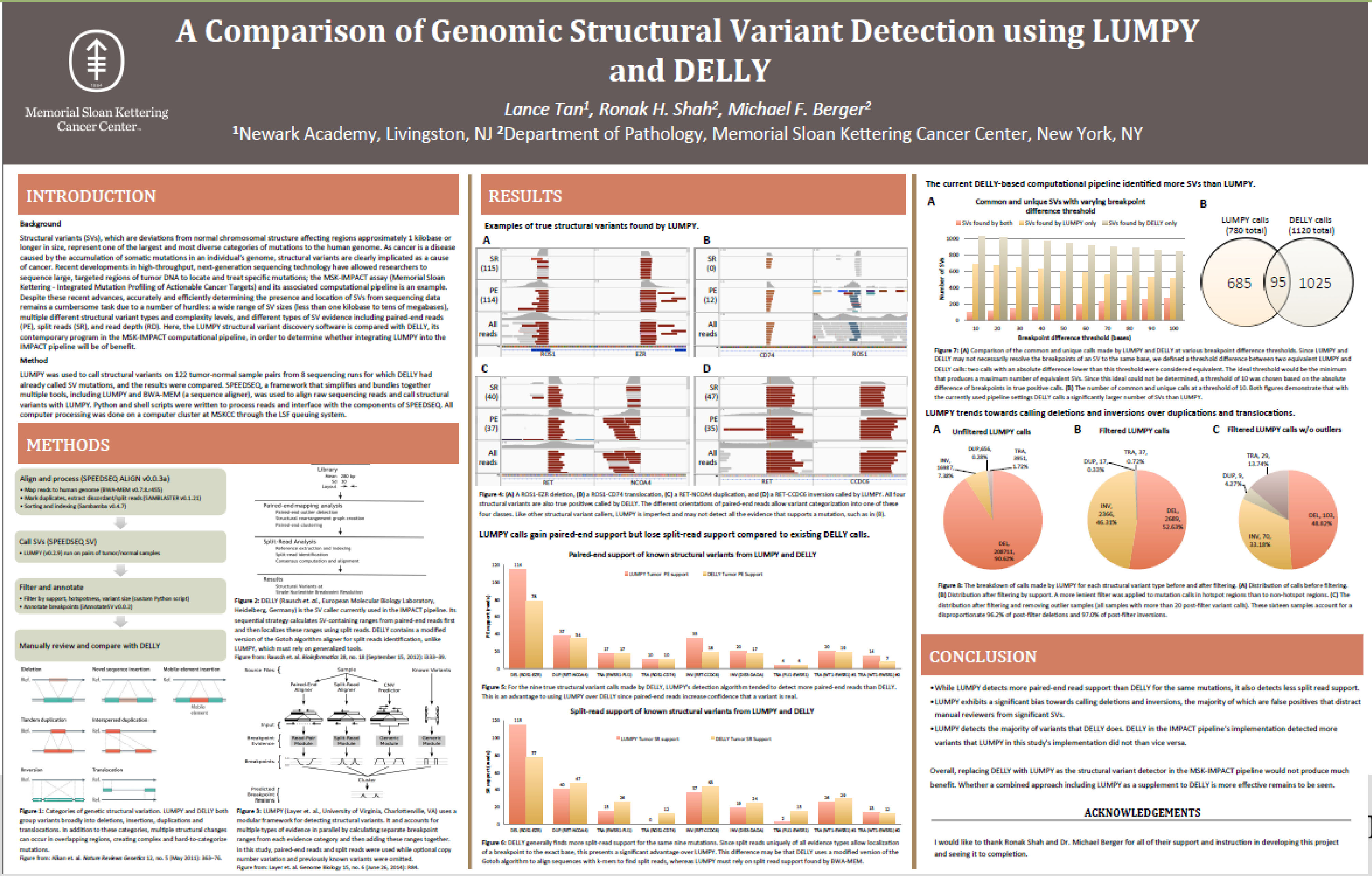
```
M-1695-2B-T_...
Total count: 905
A : 1 (0%, 1+, 0- )
C : 2 (0%, 0+, 2- )
G : 0
T : 901 (100%, 440+, 461- )
N : 1 (0%, 0+, 1- )
-----------------
DEL: 4
INS: 0
```

```
M-1747-85_bc2...
Total count: 421
A : 0
C : 0
G : 1 (0%, 1+, 0- )
T : 420 (100%, 198+, 222- )
N : 0
-----------------
DEL: 1
INS: 0
```

Figure 8: A deletion and another mutation event called by GATK in the BARD1 gene was resolved far more cleanly by ABRA into a single deletion event with a separate SNV farther away.

https://www.slideshare.net/rshah7/comparison-of-lumpy-vs-delly-for-structural-variant-detection

Example of the Package Comparison



A Comparison of Genomic Structural Variant Detection using LUMPY and DELLY

Memorial Sloan Kettering Cancer Center

Lance Tan[1], Ronak H. Shah[2], Michael F. Berger[2]

[1]Newark Academy, Livingston, NJ [2]Department of Pathology, Memorial Sloan Kettering Cancer Center, New York, NY

INTRODUCTION

Background

Structural variants (SVs), which are deviations from normal chromosomal structure affecting regions approximately 1 kilobase or longer in size, represent one of the largest and most diverse categories of mutations to the human genome. As cancer is a disease caused by the accumulation of somatic mutations in an individual's genome, structural variants are clearly implicated as a cause of cancer. Recent developments in high-throughput, next-generation sequencing technology have allowed researchers to sequence large, targeted regions of tumor DNA to locate and treat specific mutations; the MSK-IMPACT assay (Memorial Sloan Kettering - Integrated Mutation Profiling of Actionable Cancer Targets) and its associated computational pipeline is an example. Despite these recent advances, accurately and efficiently determining the presence and location of SVs from sequencing data remains a cumbersome task due to a number of hurdles: a wide range of SV sizes (less than one kilobase to tens of megabases), multiple different structural variant types and complexity levels, and different types of SV evidence including paired-end reads (PE), split reads (SR), and read depth (RD). Here, the LUMPY structural variant discovery software is compared with DELLY, its contemporary program in the MSK-IMPACT computational pipeline, in order to determine whether integrating LUMPY into the IMPACT pipeline will be of benefit.

Method

LUMPY was used to call structural variants on 122 tumor-normal sample pairs from 8 sequencing runs for which DELLY had already called SV mutations, and the results were compared. SPEEDSEQ, a framework that simplifies and bundles together multiple tools, including LUMPY and BWA-MEM (a sequence aligner), was used to align raw sequencing reads and call structural variants with LUMPY. Python and shell scripts were written to process reads and interface with the components of SPEEDSEQ. All computer processing was done on a computer cluster at MSKCC through the LSF queuing system.

METHODS

Align and process (SPEEDSEQ ALIGN v0.0.3a)
- Map reads to human genome (BWA-MEM v0.7.8-r455)
- Mark duplicates, extract discordant/split reads (SAMBLASTER v0.1.21)
- Sorting and indexing (Sambamba v0.4.7)

Call SVs (SPEEDSEQ SV)
- LUMPY (v0.2.9) run on pairs of tumor/normal samples

Filter and annotate
- Filter by support, hotspotness, variant size (custom Python script)
- Annotate breakpoints (AnnotateSV v0.0.2)

Manually review and compare with DELLY

Library
Mean: 280 bp
SD: 30
Layout
Paired-end mapping analysis
Paired-end outlier detection
Structural rearrangement graph creation
Paired-end clustering
Split-Read Analysis
Reference extraction and indexing
Split-read identification
Consensus computation and alignment
Results
Structural Variants at
Single Nucleotide Breakpoint Resolution

Figure 2: DELLY (Rausch et. al., European Molecular Biology Laboratory, Heidelberg, Germany) is the SV caller currently used in the IMPACT pipeline. Its sequential strategy calculates SV-containing ranges from paired-end reads first and then localizes these ranges using split reads. DELLY contains a modified version of the Gotoh algorithm aligner for split reads identification, unlike LUMPY, which must rely on generalized tools.
Figure from: Rausch et. al. *Bioinformatics* 28, no. 18 (September 15, 2012): i333–39.

Figure 1: Categories of genetic structural variation. LUMPY and DELLY both group variants broadly into deletions, insertions, duplications and translocations. In addition to these categories, multiple structural changes can occur in overlapping regions, creating complex and hard-to-categorize mutations.
Figure from: Alkan et. al. *Nature Reviews Genetics* 12, no. 5 (May 2011): 363–76.

Figure 3: LUMPY (Layer et. al., University of Virginia, Charlottesville, VA) uses a modular framework for detecting structural variants. It and accounts for multiple types of evidence in parallel by calculating separate breakpoint ranges from each evidence category and then adding these ranges together. In this study, paired-end reads and split reads were used while optional copy number variation and previously known variants were omitted.
Figure from: Layer et. al. Genome Biology 15, no. 6 (June 26, 2014): R84.

RESULTS

Examples of true structural variants found by LUMPY.

Figure 4: (A) A ROS1-EZR deletion, (B) a ROS1-CD74 translocation, (C) a RET-NCOA4 duplication, and (D) a RET-CCDC6 inversion called by LUMPY. All four structural variants are also true positives called by DELLY. The different orientations of paired-end reads allow variant categorization into one of these four classes. Like other structural variant callers, LUMPY is imperfect and may not detect all the evidence that supports a mutation, such as in (B).

LUMPY calls gain paired-end support but lose split-read support compared to existing DELLY calls.

Figure 5: For the nine true structural variant calls made by DELLY, LUMPY's detection algorithm tended to detect more paired-end reads than DELLY. This is an advantage to using LUMPY over DELLY since paired-end reads increase confidence that a variant is real.

Figure 6: DELLY generally finds more split-read support for the same nine mutations. Since split reads uniquely of all evidence types allow localization of a breakpoint to the exact base, this presents a significant advantage over LUMPY. This difference may be that DELLY uses a modified version of the Gotoh algorithm to align sequences with k-mers to find split reads, whereas LUMPY must rely on split read support found by BWA-MEM.

The current DELLY-based computational pipeline identified more SVs than LUMPY.

Figure 7: (A) Comparison of the common and unique calls made by LUMPY and DELLY at various breakpoint difference thresholds. Since LUMPY and DELLY may not necessarily resolve the breakpoints of an SV to the same base, we defined a threshold difference between two equivalent LUMPY and DELLY calls: two calls with an absolute difference lower than this threshold were considered equivalent. The ideal threshold would be the minimum that produces a maximum number of equivalent SVs. Since this ideal could not be determined, a threshold of 10 was chosen based on the absolute difference of breakpoints in true positive calls. (B) The number of common and unique calls at a threshold of 10. Both figures demonstrate that with the currently used pipeline settings DELLY calls a significantly larger number of SVs than LUMPY.

LUMPY trends towards calling deletions and inversions over duplications and translocations.

Figure 8: The breakdown of calls made by LUMPY for each structural variant type before and after filtering. (A) Distribution of calls before filtering. (B) Distribution after filtering by support. A more lenient filter was applied to mutation calls in hotspot regions than to non-hotspot regions. (C) The distribution after filtering and removing outlier samples (all samples with more than 20 post-filter variant calls). These sixteen samples account for a disproportionate 96.2% of post-filter deletions and 97.0% of post-filter inversions.

CONCLUSION

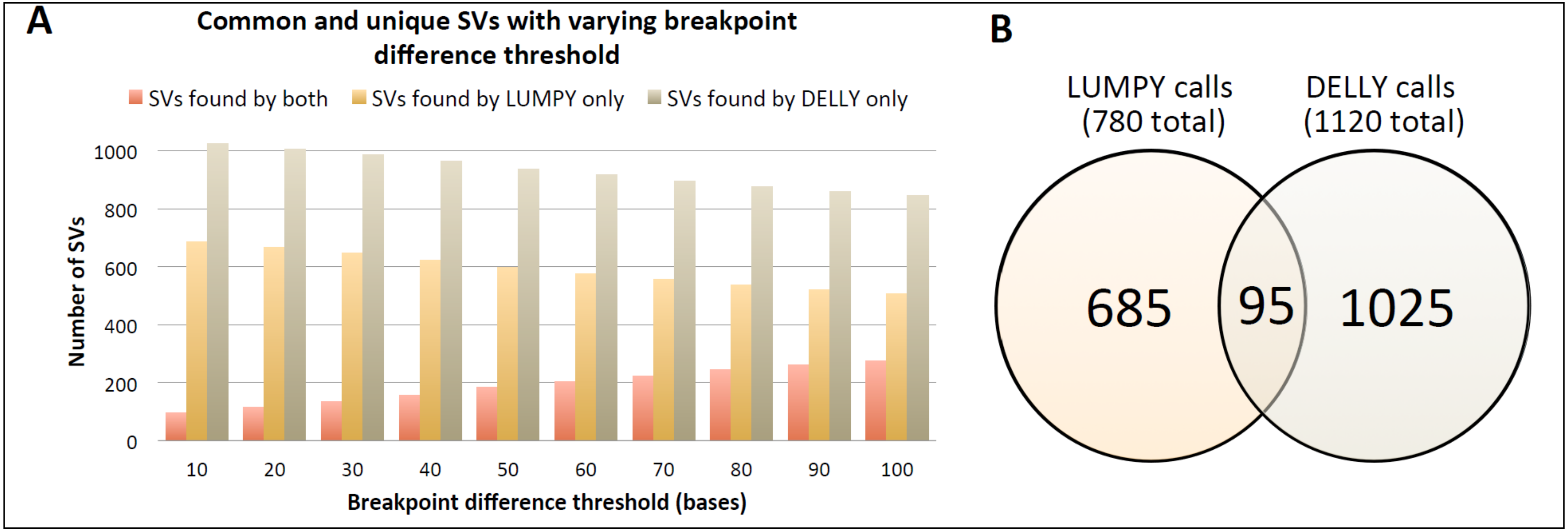
- While LUMPY detects more paired-end read support than DELLY for the same mutations, it also detects less split read support.
- LUMPY exhibits a significant bias towards calling deletions and inversions, the majority of which are false positives that distract manual reviewers from significant SVs.
- LUMPY detects the majority of variants that DELLY does. DELLY in the IMPACT pipeline's implementation detected more variants that LUMPY in this study's implementation did not than vice versa.

Overall, replacing DELLY with LUMPY as the structural variant detector in the MSK-IMPACT pipeline would not produce much benefit. Whether a combined approach including LUMPY as a supplement to DELLY is more effective remains to be seen.

ACKNOWLEDGEMENTS

I would like to thank Ronak Shah and Dr. Michael Berger for all of their support and instruction in developing this project and seeing it to completion.

Comparing Tools with Different Algorithm



https://www.slideshare.net/rshah7/comparison-of-lumpy-vs-delly-for-structural-variant-detection

Variant Caller Comparison in WGS & WES

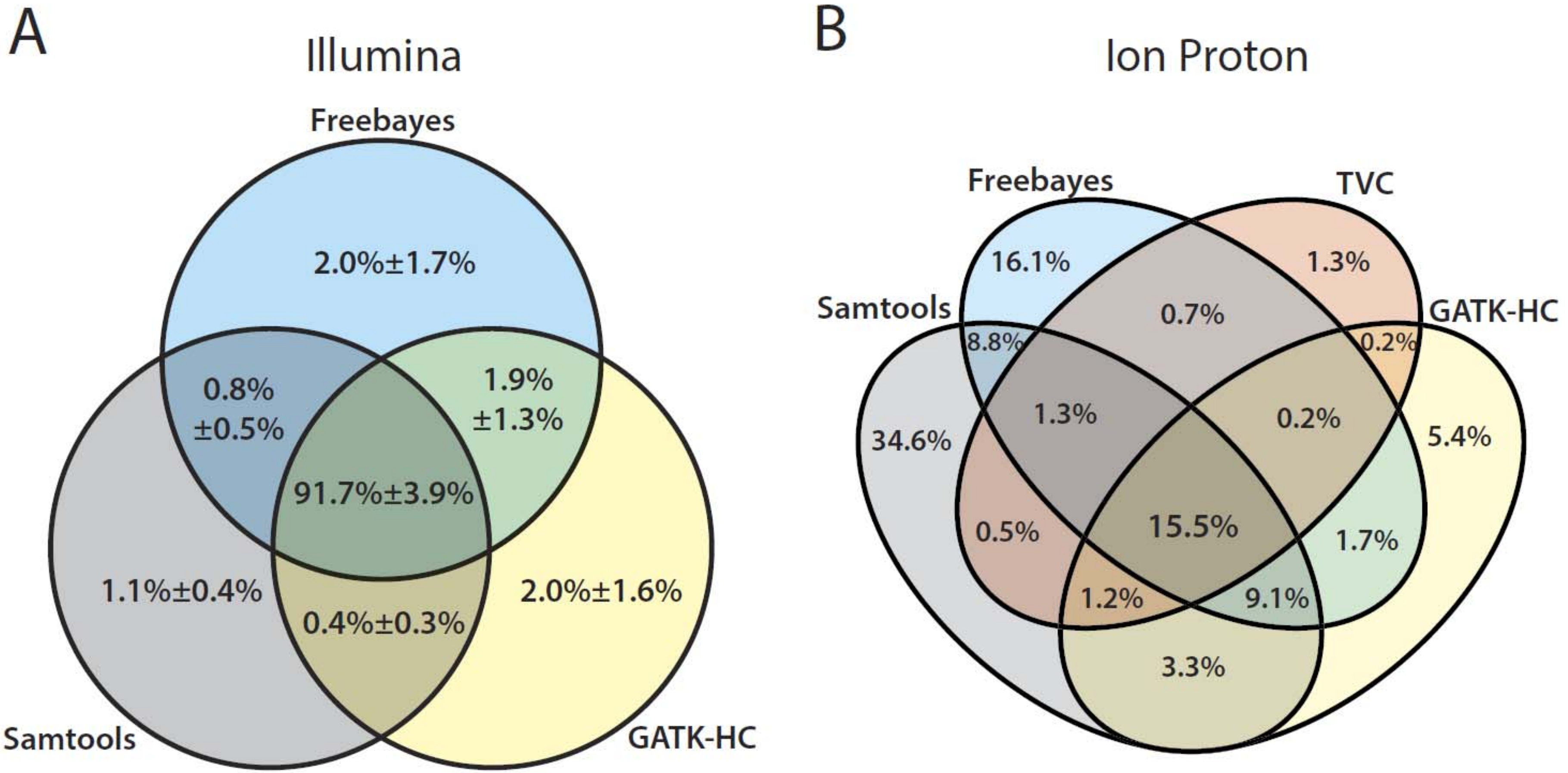
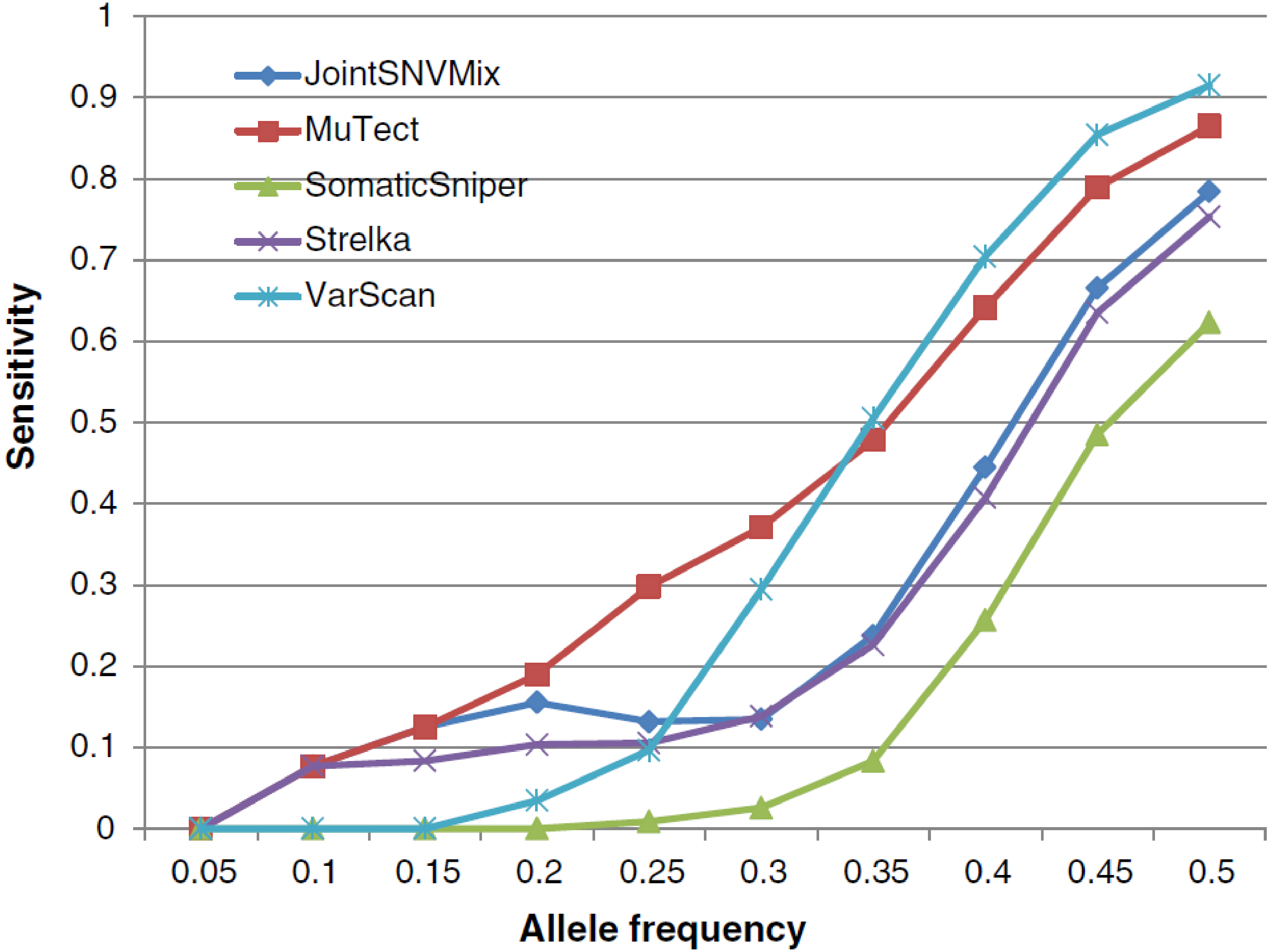


Figure 3. Venn diagrams summarizing called variants by different callers. The mean percentage with standard deviation of confidence variant calls with equal to or higher than the quality score threshold of 20 are represented for (**A**) Illumina data sets and (**B**) Ion Proton data set.

Sohyun Hwang, Scientific Reports volume5, 17875 (2015)

Variant Caller Comparison in WES



Qingguo Wang, Genome Med. 2013; 5(10): 91.

Variant Caller Comparison in WES & Target Seq

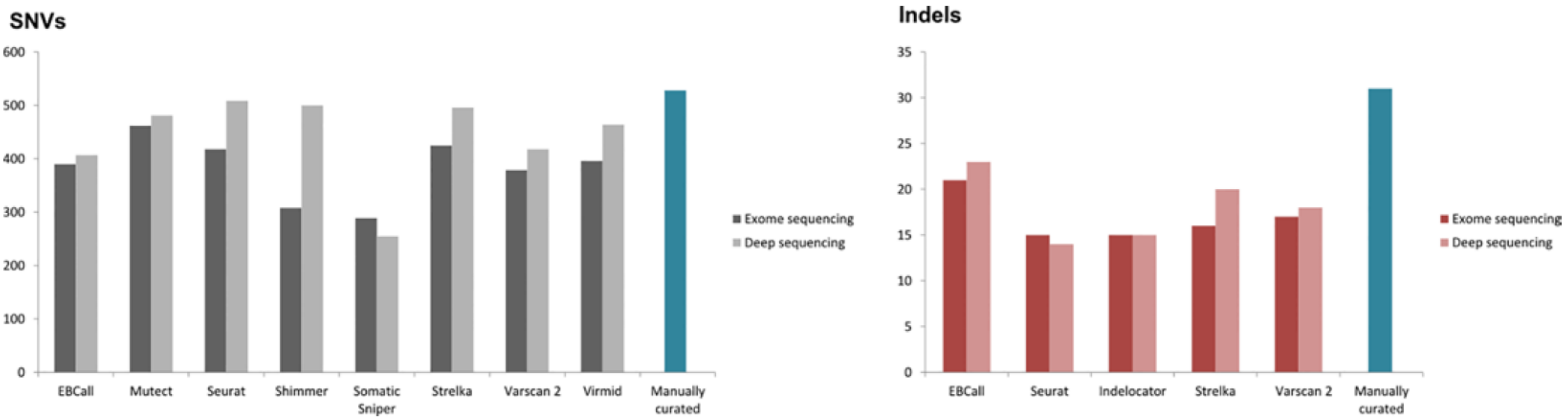
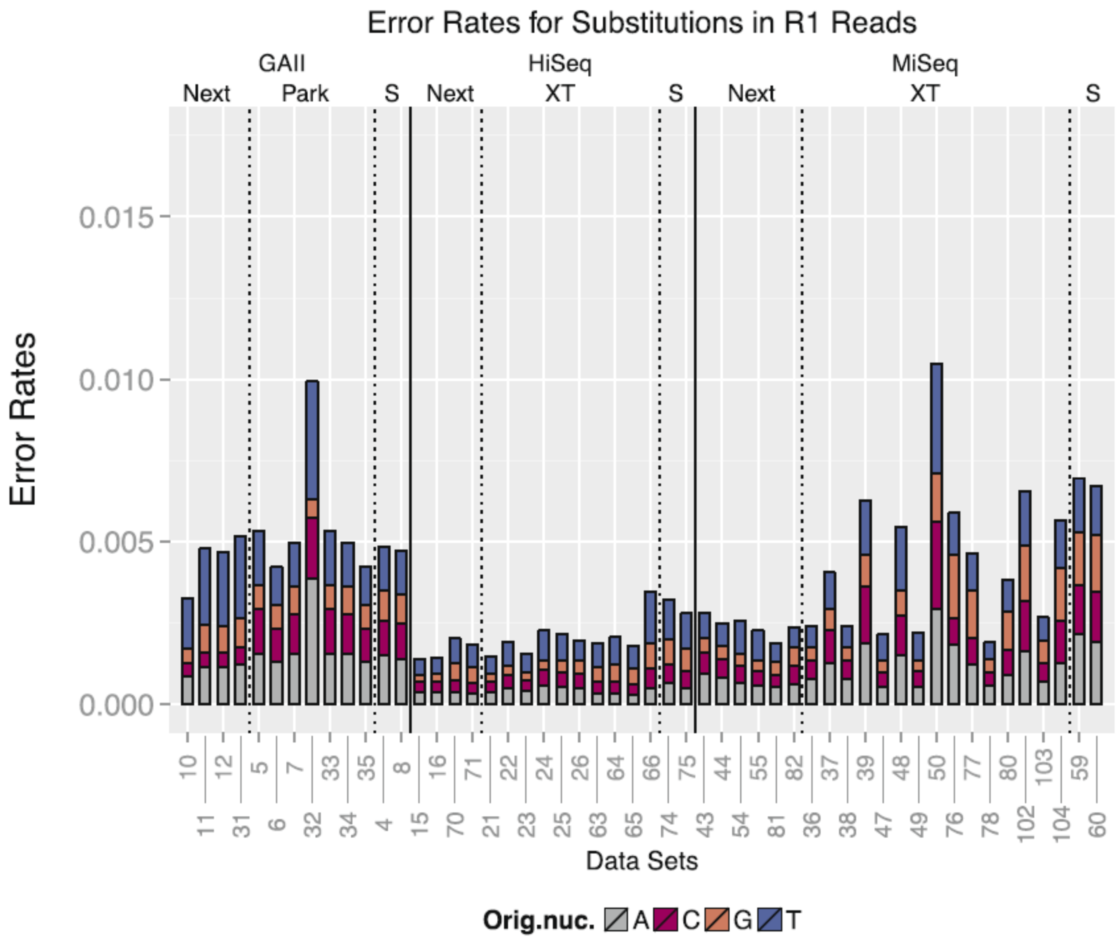


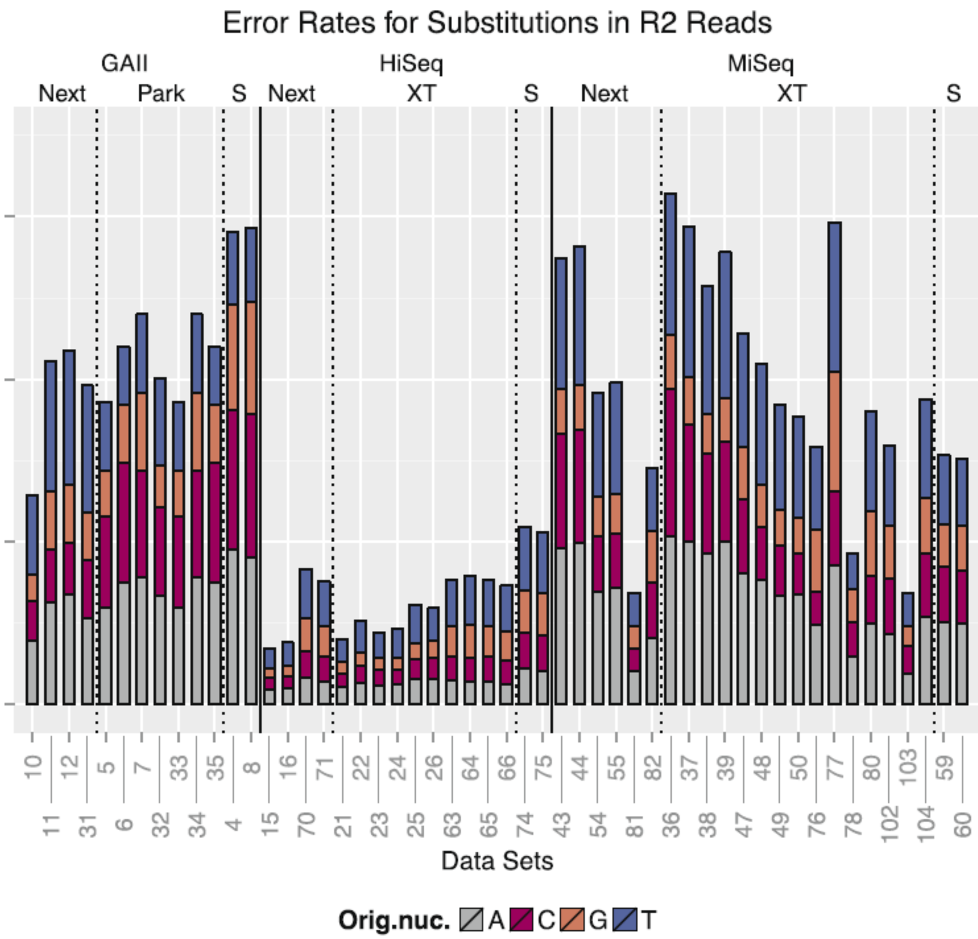
Fig 6. Variant caller sensitivity. Variant caller sensitivity for detecting the manually curated mutations for SNVs and indels are shown in left and right panels, respectively. The y-axis depicts the number of variant calls. The dark and light grey bars represent calls in the exome and targeted deep sequencing data, respectively.

AB Krøigård et al., PLoS One. 2016; 11(3)

Error Rate Comparison among Platforms



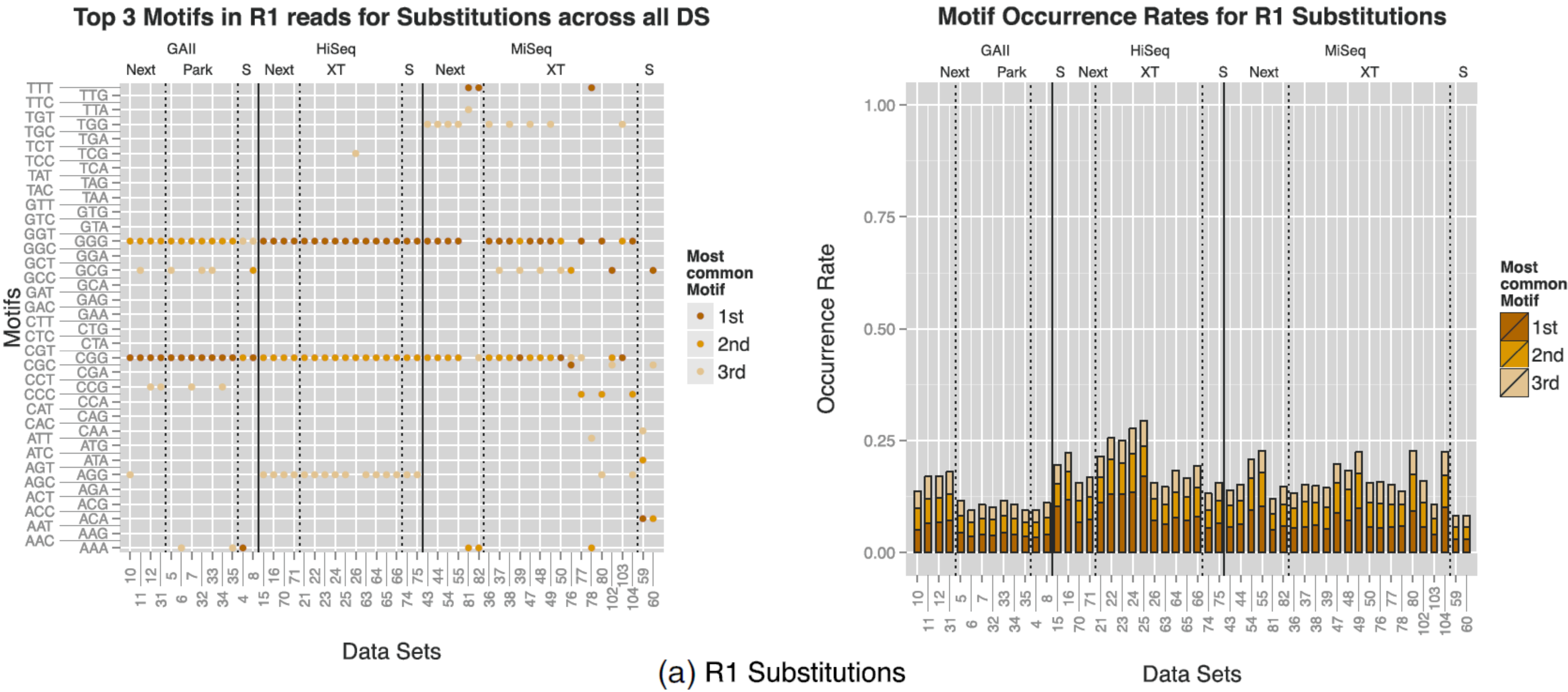
Error Rates for Substitutions in R1 Reads



Error Rates for Substitutions in R2 Reads

Melanie Schirmer, BMC Bioinformatics, 2016(17):125

Error Related Motif (3mers preceding errors)


Top 3 Motifs in R1 reads for Substitutions across all DS

Motif Occurrence Rates for R1 Substitutions

(a) R1 Substitutions

Melanie Schirmer, BMC Bioinformatics, 2016(17):125

HOKKAIDO UNIVERSITY

Mutect Defaults Filter Settings

Filters used in high-confidence mode

1. Proximal Gap
2. Poor Mapping
3. Strand Bias
4. Clustered Position
5. Observed in Control

Filters applied in all MuTect modes

1. Tumor and normal LOD scores
2. Possible contamination
3. Normal LOD score and dbsnp status
4. Triallelic Site Filter

https://gatkforums.broadinstitute.org/gatk/discussion/4464/how-mutect-filters-candidate-mutations

Empirical Variant Filter for Somatic Mutation

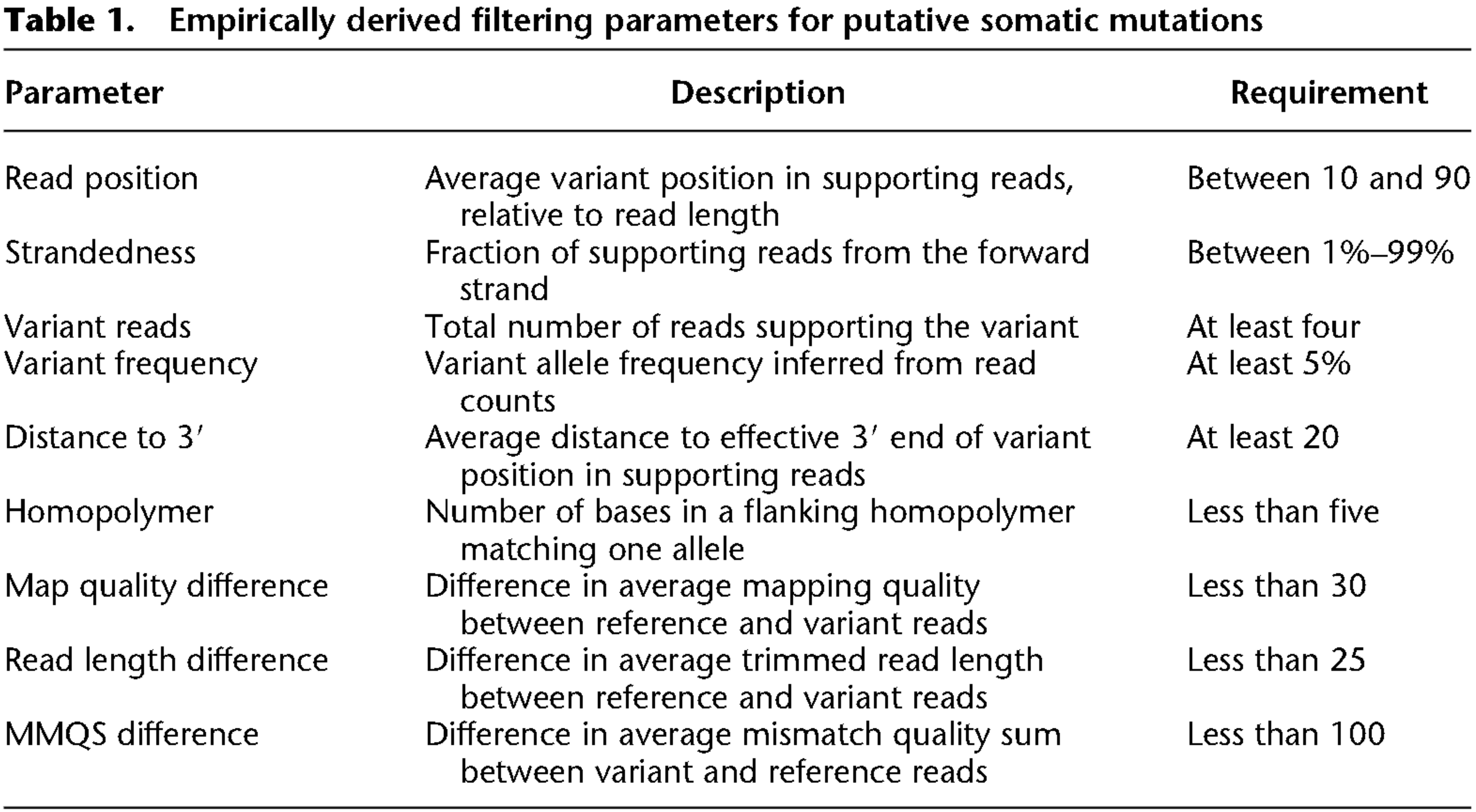
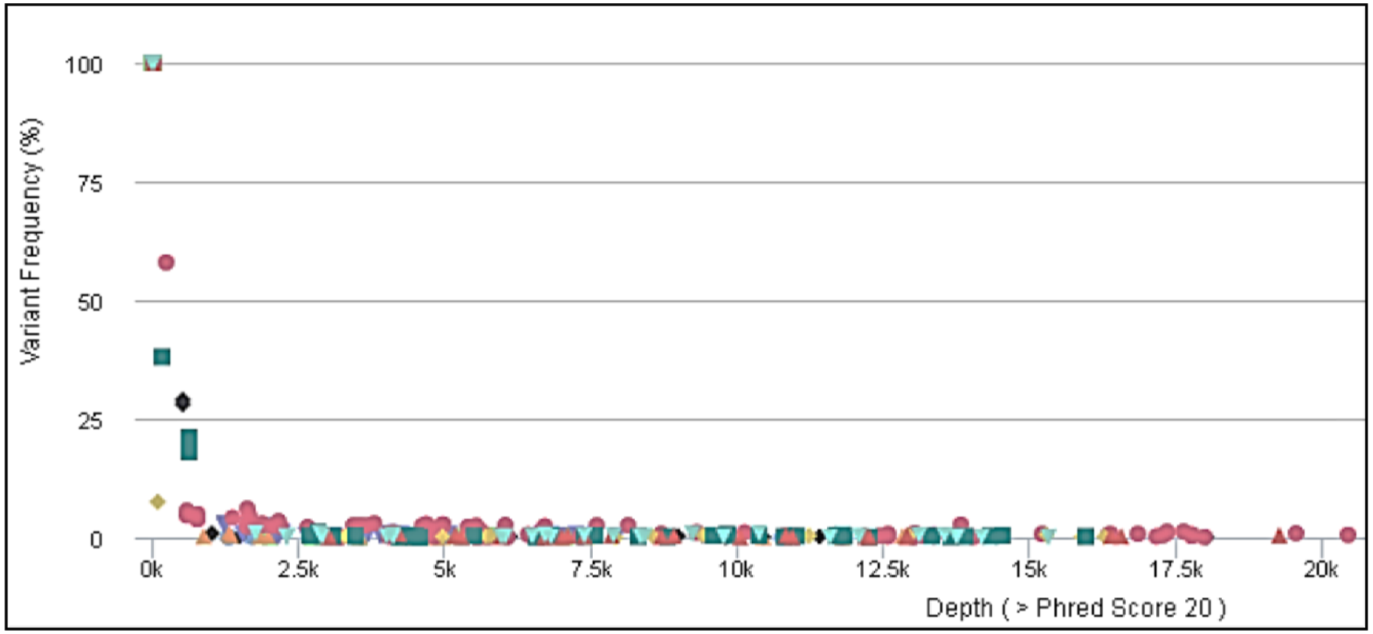


Table 1. Empirically derived filtering parameters for putative somatic mutations

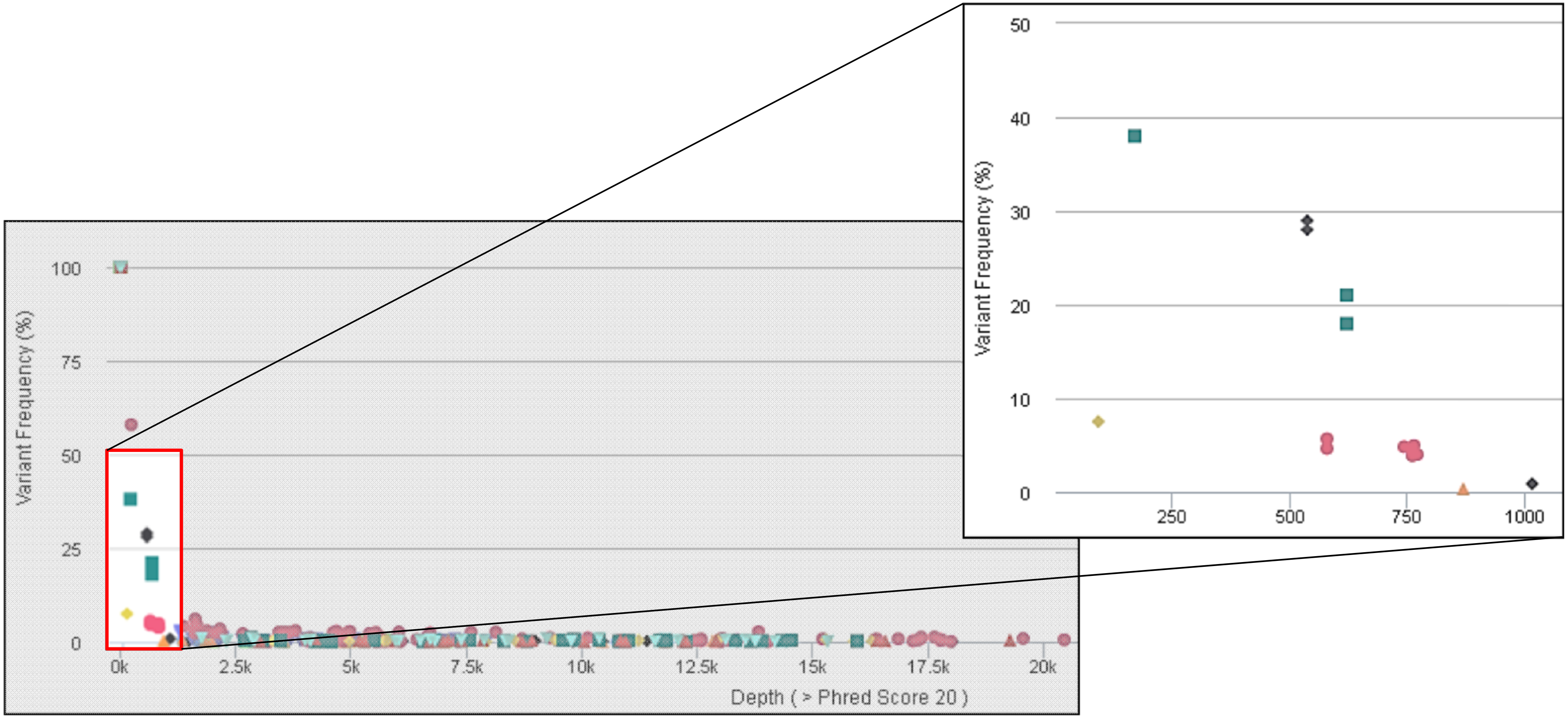
Parameter	Description	Requirement
Read position	Average variant position in supporting reads, relative to read length	Between 10 and 90
Strandedness	Fraction of supporting reads from the forward strand	Between 1%–99%
Variant reads	Total number of reads supporting the variant	At least four
Variant frequency	Variant allele frequency inferred from read counts	At least 5%
Distance to 3′	Average distance to effective 3′ end of variant position in supporting reads	At least 20
Homopolymer	Number of bases in a flanking homopolymer matching one allele	Less than five
Map quality difference	Difference in average mapping quality between reference and variant reads	Less than 30
Read length difference	Difference in average trimmed read length between reference and variant reads	Less than 25
MMQS difference	Difference in average mismatch quality sum between variant and reference reads	Less than 100

Koboldt DC, Genome Res. 2012 Mar;22(3):568-76.

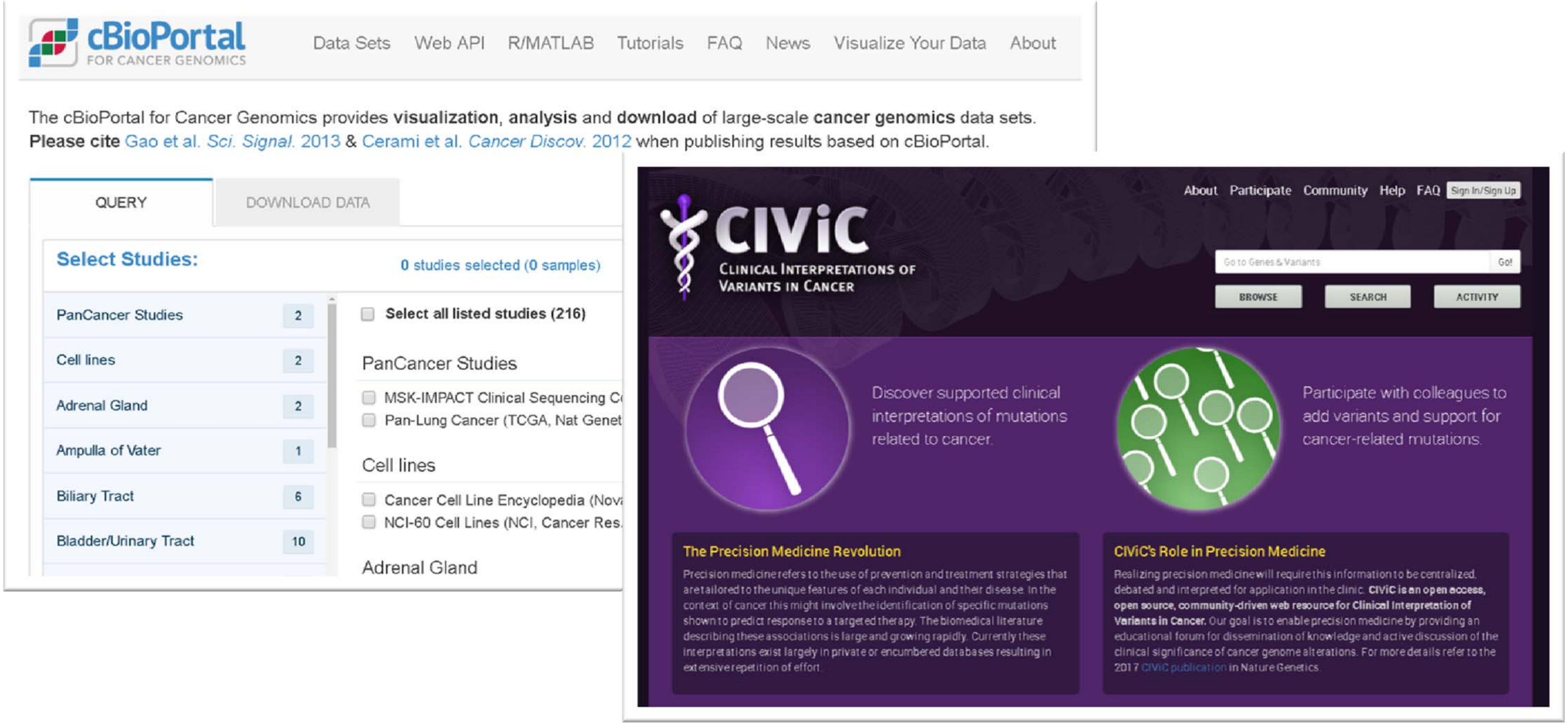
Variant Filtering in Another Cases



Variant Filtering in Another Cases



Mannual Review



The cBioPortal for Cancer Genomics provides **visualization**, **analysis** and **download** of large-scale **cancer genomics** data sets. **Please cite** Gao et al. *Sci. Signal.* 2013 & Cerami et al. *Cancer Discov.* 2012 when publishing results based on cBioPortal.

Select Studies:

0 studies selected (0 samples)

□ **Select all listed studies (216)**

PanCancer Studies

□ MSK-IMPACT Clinical Sequencing C

□ Pan-Lung Cancer (TCGA, Nat Genet

Cell lines

□ Cancer Cell Line Encyclopedia (Nova

□ NCI-60 Cell Lines (NCI, Cancer Res.

Adrenal Gland

Discover supported clinical interpretations of mutations related to cancer.

Participate with colleagues to add variants and support for cancer-related mutations.

The Precision Medicine Revolution

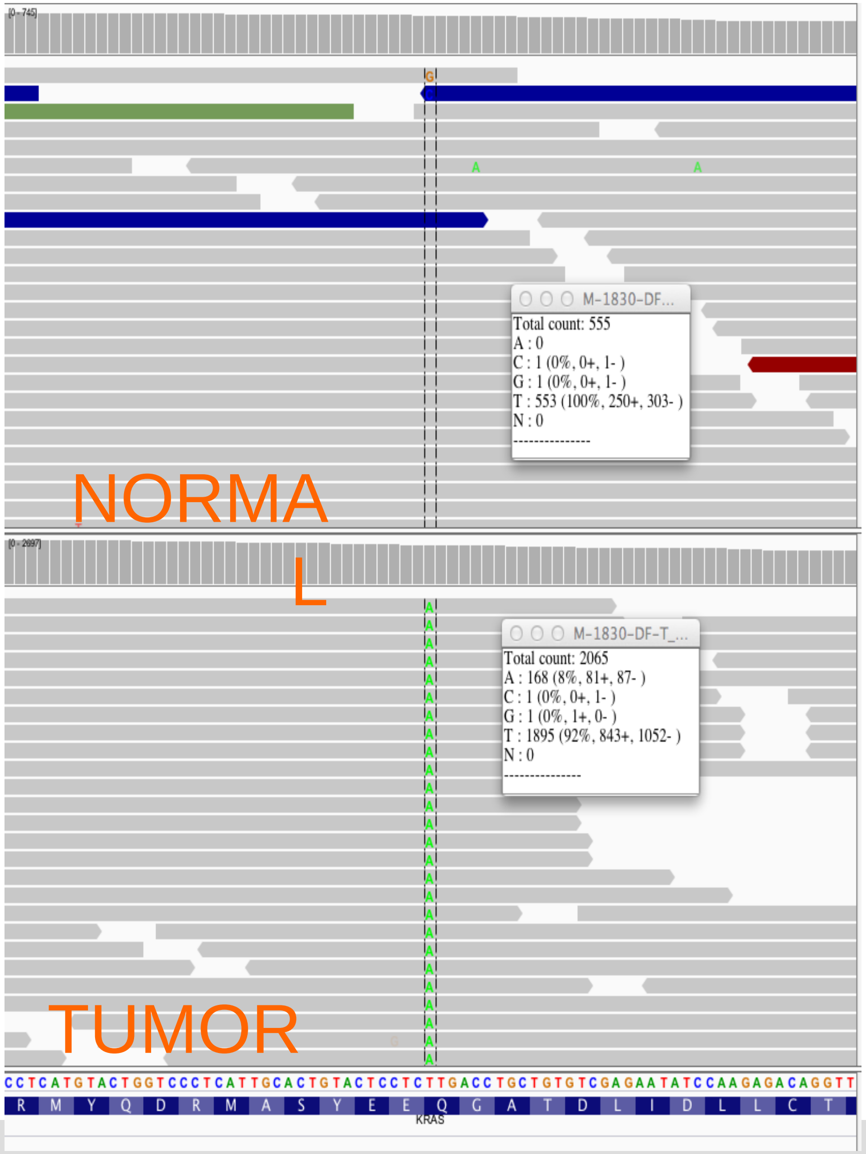
Precision medicine refers to the use of prevention and treatment strategies that are tailored to the unique features of each individual and their disease. In the context of cancer this might involve the identification of specific mutations shown to predict response to a targeted therapy. The biomedical literature describing these associations is large and growing rapidly. Currently these interpretations exist largely in private or encumbered databases resulting in extensive repetition of effort.

CIViC's Role in Precision Medicine

Realizing precision medicine will require this information to be centralized, debated and interpreted for application in the clinic. **CIViC is an open access, open source, community-driven web resource for Clinical Interpretation of Variants in Cancer.** Our goal is to enable precision medicine by providing an educational forum for dissemination of knowledge and active discussion of the clinical significance of cancer genome alterations. For more details refer to the 2017 CIViC publication in Nature Genetics.

http://www.cbioportal.org/
https://civic.genome.wustl.edu/home

Mannual Review

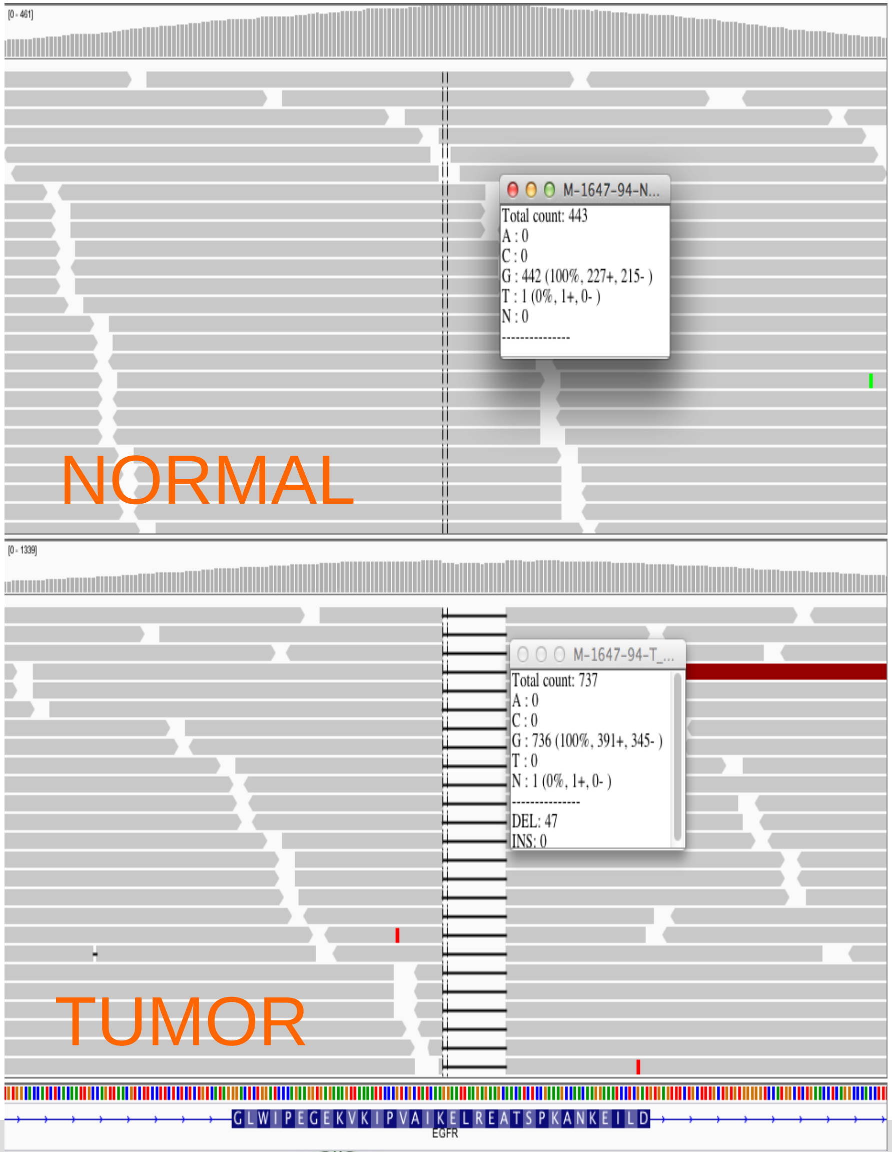

[0 - 745]
NORMA
L
[0 - 2697]
TUMOR
CCTCATGTACTGGTCCCTCATTGCACTGTACTCCTCTTGACCTGCTGTGTCGAGAATATCCAAGAGACAGGTT
R M Y Q D R M A S Y E E Q G A T D L I D L L C T
KRAS

M-1830-DF...

```
Total count: 555
A : 0
C : 1 (0%, 0+, 1- )
G : 1 (0%, 0+, 1- )
T : 553 (100%, 250+, 303- )
N : 0
---------------
```

M-1830-DF-T_...

```
Total count: 2065
A : 168 (8%, 81+, 87- )
C : 1 (0%, 0+, 1- )
G : 1 (0%, 1+, 0- )
T : 1895 (92%, 843+, 1052- )
N : 0
---------------
```


[0 - 461]
NORMAL
[0 - 1339]
TUMOR
G L W I P E G E K V K I P V A I K E L R E A T S P K A N K E I L D
EGFR

M-1647-94-N...

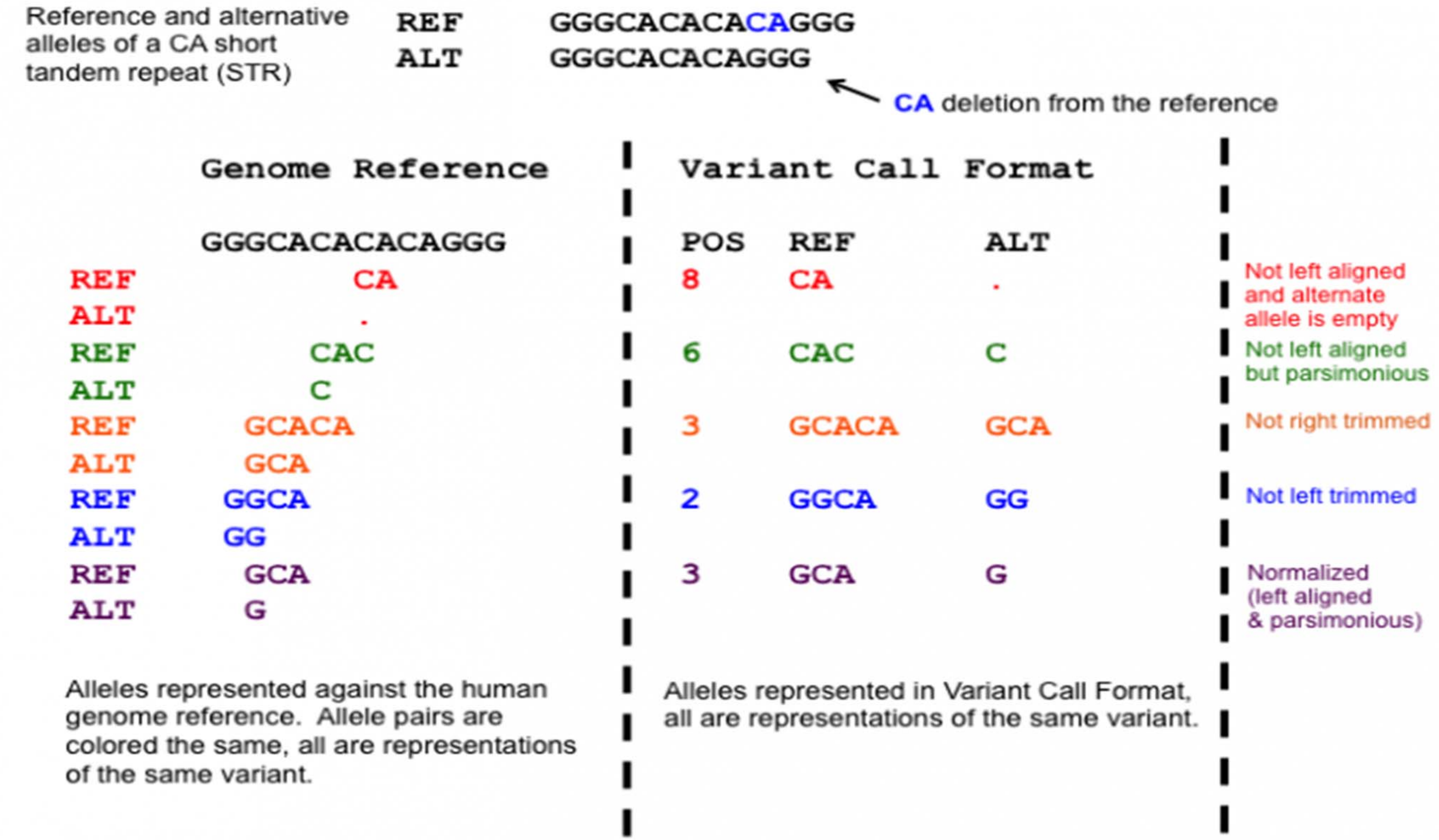
```
Total count: 443
A : 0
C : 0
G : 442 (100%, 227+, 215- )
T : 1 (0%, 1+, 0- )
N : 0
---------------
```

M-1647-94-T_...

```
Total count: 737
A : 0
C : 0
G : 736 (100%, 391+, 345- )
T : 0
N : 1 (0%, 1+, 0- )
---------------
DEL: 47
INS: 0
```

※) Cited from GATK Workshop slides

Variant Normalization (Left-Alignment)



Reference and alternative alleles of a CA short tandem repeat (STR)

REF GGGCACACACAGGG

ALT GGGCACACAGGG

CA deletion from the reference

	Genome Reference	POS	REF	ALT	
	GGGCACACACAGGG				
REF	CA	8	CA	.	Not left aligned and alternate allele is empty
ALT	.				
REF	CAC	6	CAC	C	Not left aligned but parsimonious
ALT	C				
REF	GCACA	3	GCACA	GCA	Not right trimmed
ALT	GCA				
REF	GGCA	2	GGCA	GG	Not left trimmed
ALT	GG				
REF	GCA	3	GCA	G	Normalized (left aligned & parsimonious)
ALT	G				

Variant Call Format

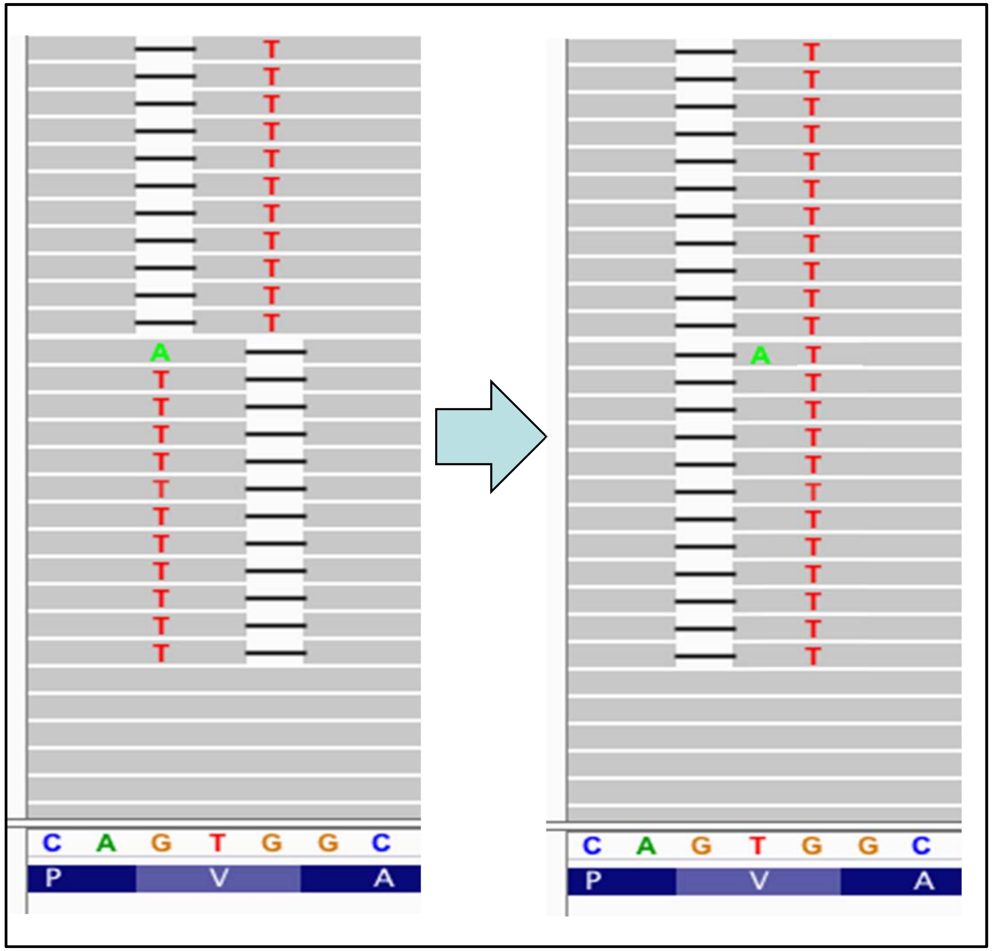
Alleles represented against the human genome reference. Allele pairs are colored the same, all are representations of the same variant.

Alleles represented in Variant Call Format, all are representations of the same variant.

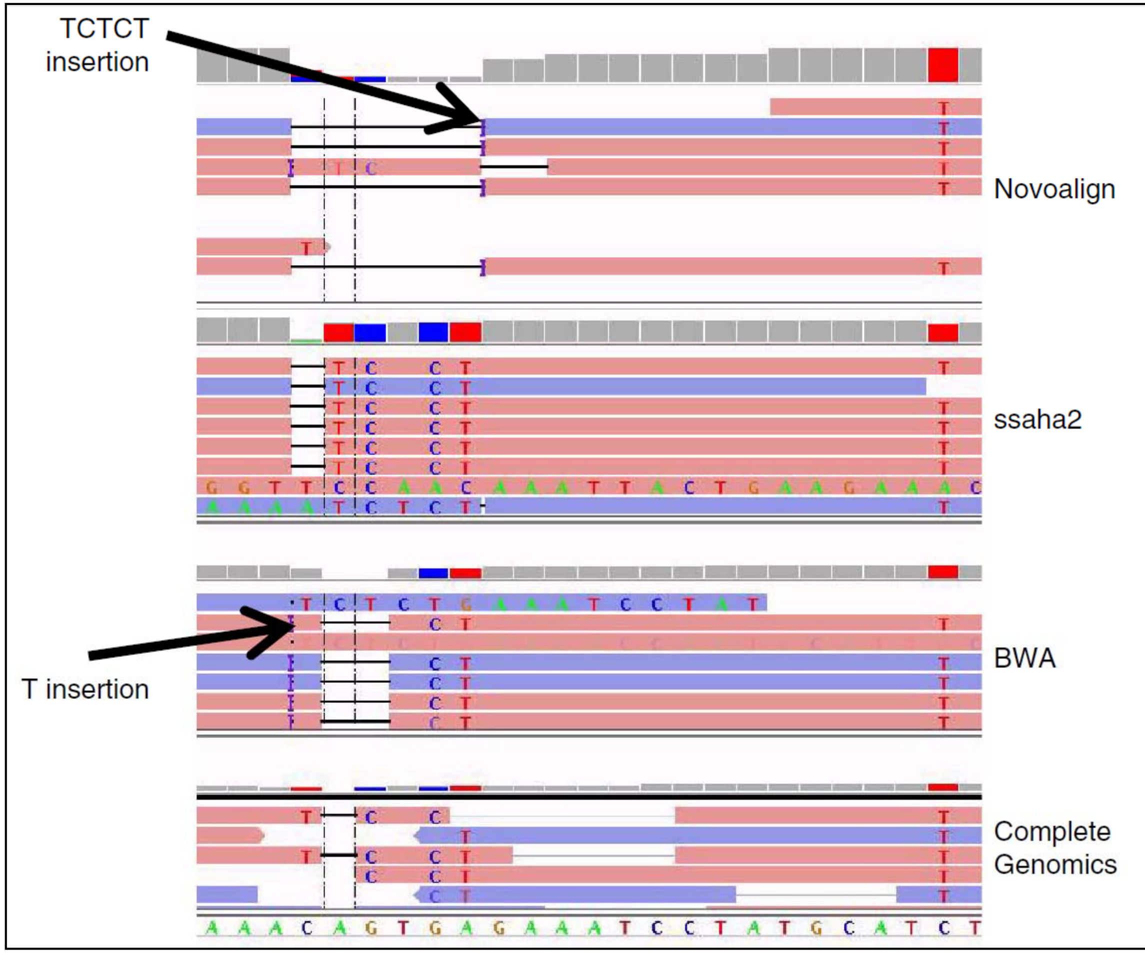
https://genome.sph.umich.edu/wiki/Variant_Normalization

Normalization and Complex Variant

単純な例



複雑な例



Zook JM, Nat Biotechnol. 2014 Mar;32(3):246-51