A method for analysis to detect somatic mutations precisely using Japanese reference genome sequence, JG2.0

○Nozomi Kasazumi, Shuji Kaieda, Kouichiro Aya, Satoshi Kira, Akiyuki Sato and Masanari Kitagawa（TAKARA BIO INC.）

Clontech TaKaRa cellartis

Introduction

GRCh38/hg38 is the international human reference genome sequence and the most used for NGS analysis such as detection of variants. However, because it was constructed from genomic information from donors of European or African ancestry, detection of variants in Japanese samples using GRCh38/hg38 has the risk to detect false variants due to differences in genomic sequences among ethnic groups. JG2.0 is the Japanese reference genome sequence constructed in Tohoku Medical Megabank Organization (ToMMo) to address this problem. It was constructed by performing de novo assembly of three Japanese male genome and integrating them. JG2.0 is expected to exhibit advantages in NGS analysis of Japanese samples by using as a reference sequence instead of GRCh38/hg38. In this study, we developed a highly accurate mutation analysis method using JG2.0, especially for somatic mutation analysis.

Accuracy of somatic variant calling with JG2.0

We evaluated whether JG2.0 is a suitable reference genome for somatic variant calling. The results showed that it is difficult to state that JG2.0 improved the accuracy of detection of mutations in Japanese samples compared to GRCh38.

0.7 % more reads were aligned in target region of JG2.0 compared to GRCh38 (Figure 1). Somatic mutations detected using JG2.0 was on average 8.4% less than those using GRCh38 (Figure 2). A number of variants detected only when using JG2.0 were suspected to be false positives. Specifically, the positions of these variants on JG2.0 corresponded to multiple positions on GRCh38 encoding paralogous genes, and the reference sequences of these positions were different (Figure 3). In addition, the reads with variants and reads with the same sequence as the reference sequence were aligned to other regions when aligned to GRCh38, respectively. Therefore, we suspected that these variants were false positives due to a missing part of segmental duplications in JG2.0.

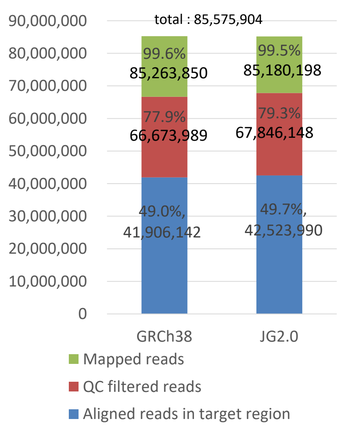

Figure 1. # of aligned reads.

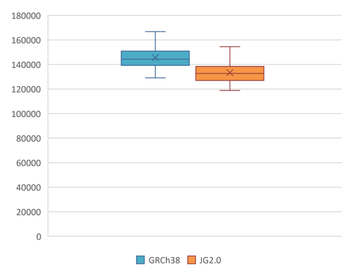

Figure 2. # of detected SNPs.

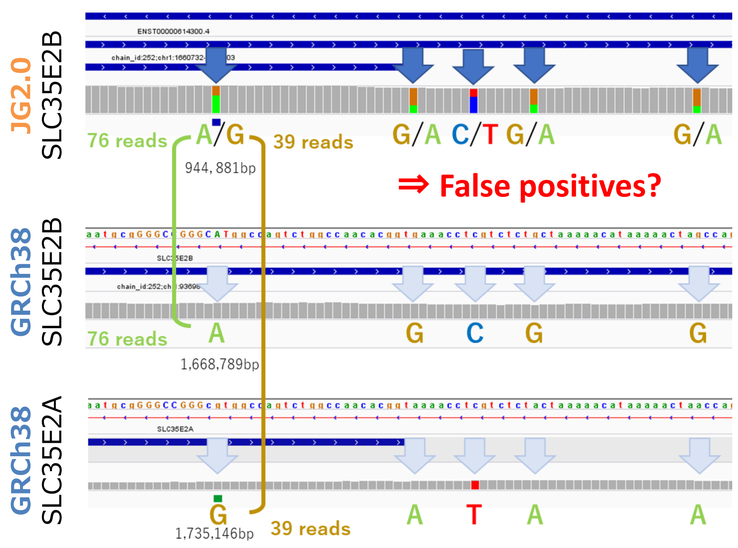

Figure 3. Example of suspected false positive SNPs.

※ We performed somatic mutation analysis of 31 Japanese gastric cancer patients using each of JG2.0beta (JG2.0) and GRCh38.p13 (GRCh38). We downloaded fastq files of Whole-exome sequencing 150-bp paired-end reads of tumor samples. QC filtered reads were mapped to GRCh38/JG2.0 and putative somatic variants were called by DRAGEN software(version 07.021.595.3.7.5). Default parameters were used for each analysis.

Differences between JG2.0 and hg38 potentially cause false detection

We estimated the amount of false detections due to differences in genome structure among ethnic populations and difficult-to-read regions.

The results suggest that the use of JG2.0 results in false positives and false negatives due to some missing reference genome sequences, and the use of GRCh38 results in false positives due to differences between ethnic populations.

216 Mbp (7.0%) of GRCh38 (primary assembly) and 261 Mbp (8.5%) of JG2.0 had no corresponding region on JG2.0/ GRCh38, which are unique to GRCh38/JG2.0 (Figure 4). In addition, 16 Mbp (0.5%) of GRCh38 corresponded to two or more regions in JG2.0, indicating that there are multiple regions in JG2.0 with sequences similar to those regions. Compared to them, the region of JG2.0 corresponding to two or more regions of GRCh38 was 37 Mbp (1.2%), more than twice as long (Figure 4).

We also detected 4,571,012 SNPs and 984,911 short indels between GRCh38 and JG2.0 sequences. They are likely to be miss-detected as SNPs/short indels that the samples have when the genome of Japanese samples are analyzed using GRCh38/hg38.

※ we analyzed the chain file (JG2.0.0beta_to_GRCh38.p13.genome.chain) using transanno (version 2.4.0) to compare JG2.0 with GRCh38 sequences. A chain file is a correspondence table of the reference positions of two reference genome sequences and is created based on the result of alignment of one reference genome to the other.

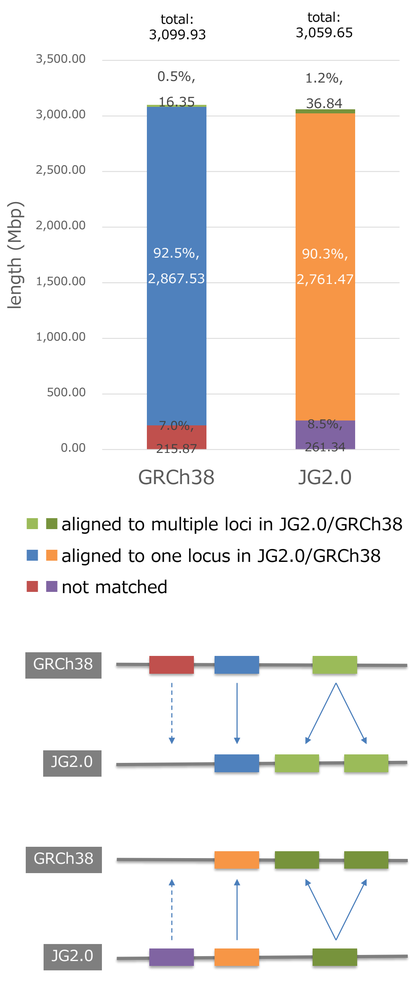

Figure 4. Comparison between JG2.0 and GRCh38.

Somatic mutation analysis workflow using both GRCh38 and JG2.0

The results of variant calling using JG2.0 and comparison between JG2.0 and GRCh38 suggested that there are some problems with detecting variants using JG2.0/GRCh38, respectively. JG2.0 and GRCh38 have specific segmental duplication, long insertion/deletion and missing genomic region, which would result in detection of false positives. JG2.0 may have more such false positives because JG2.0 have more region which corresponds to two or more regions in GRCh38. On the other hand, use of GRCh38 results in false positives due to genomic differences between European/African and Japanese.

In order to reduce false positives, we developed a somatic mutation analysis workflow using both GRCh38 and JG2.0 reference (Figure 5). At first, in this workflow, Japanese whole genome or whole exome sequencing data is aligned to GRCh38 and JG2.0 separately, and variant calling using each references is conducted independently. Subsequently, at the comparison step, JG2.0 reference information and called variants on JG2.0 compared to those on GRCh38. According to each situation, we determine whether each variants on GRCh38 is TRUE or FALSE by considering the six false positive filtering tags (Figure 6, Table 1).

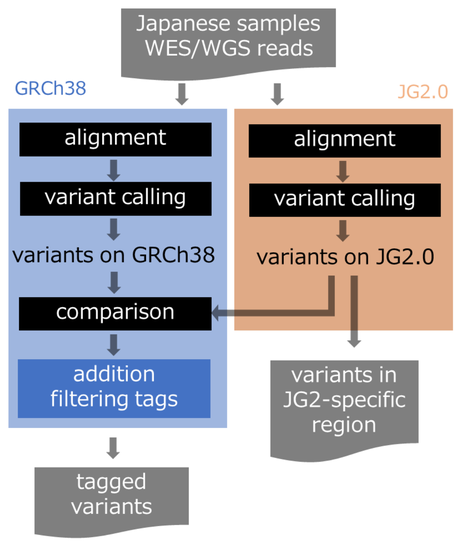

Figure 5. schematic diagram of workflow.

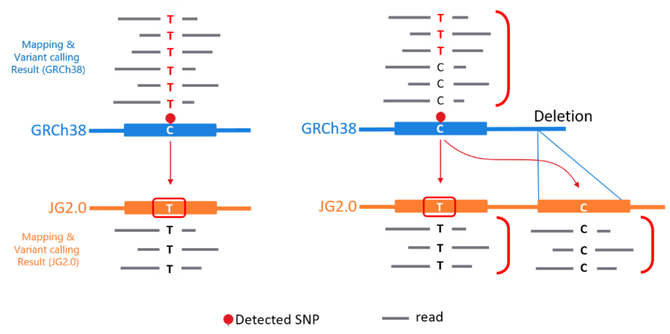

Figure 6. Examples of tagged SNPs.

Table 1. Description of tags

Tag name	False positive level	description
ALT_is_JG_ref	moderate	”Variant detected on GRCh38” matches “reference sequence of JG2.0”
not_detected_in_JG	moderate	No variants detected on JG2.0
JG_multi_loci	low	Variant detected position on GRCh38 corresponds to multiple coordinates on JG2.0
ALT_is_matched_JG_ref	high	“Detected variant on GRCh38” matches “reference sequence of corresponding multiple coordinates of JG2.0”
suspicion_multi_map_in_JG	high	“The total read depth on corresponding multiple coordinates on JG2” and “the depth on GRCh38” are close.
high_depth_compare_JG	low	The read depth on GRCh38 is very high compared to the corresponding depth of JG2.0

Verification of the workflow

Method 1

First, we tested the defined tags for healthy Japanese and European samples to confirm the effectiveness of filtering the false positive variants. Because the workflow is to tag false positives that are more likely to be detected in Japanese samples, it is expected that a greater proportion of variants detected in Japanese samples will be tagged than those detected in European samples.

Table 2. # of tags on average.

Ethnicity of samples	European		Japanese	
Tags	#	%	#	%
ALT_is_JG_ref	54274.7	43.9%	58948.7	50.4%
not_detected_in_JG	39583.7	32.0%	45806.3	39.1%
JG_multi_loci	531.0	0.4%	484.3	0.4%
ALT_is_matched_JG_ref	140.3	0.1%	125.3	0.1%
suspicion_multi_map_in_JG	34.3	0.0%	27.0	0.0%
high_depth_compare_JG	1239.0	1.0%	1222.7	1.0%
Total tagged SNPs	**57655.3**	**46.6%**	**62251.7**	**53.2%**

Result 1

As hypothesized, 46.6% of variants of European samples were tagged, whereas 53.2% of those of Japanese samples (Table 2, Figure 7). This suggests that the addition of JG2.0 information suppressed false positives, which are more likely to occur in Japanese samples.

Method 2

Second, to show that this workflow would not over-filter SNPs, we compare the proportion of tagged variants among known gastric cancer mutations and among all variants.

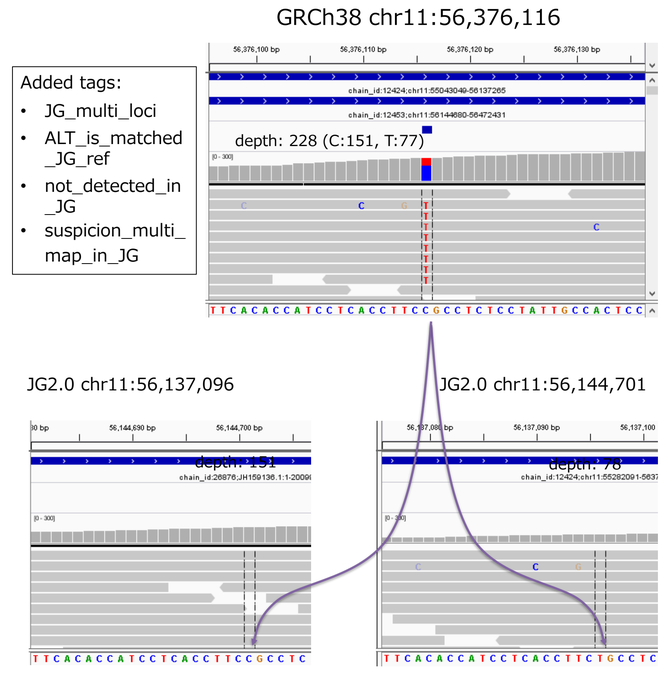

Figure 7. Example of tagged SNP of Japanese sample.

Result 2

54.2% of variants detected using GRCh38 are tagged on average. But when limited to mutations listed in COSMIC catalog as SNPs related to gastric cancer, 22.3% fewer variants were filtered out (Table 3).

Table 3. # of tagged SNPs

Sample	#1	#2	#3	average
All SNPs	55.4%	55.5%	51.6%	54.2%
SNPs related to gastric cancer	33.7%	32.7%	29.3%	31.9%

Conclusion

By annotating the mutations detected using GRCh38 with the results of mutation analysis using JG2.0, we have developed an accurate mutation analysis that takes into account differences in genome sequences among ethnic groups. This method enables more efficient search for mutations by suppressing false positives and narrowing down disease-causing mutation candidates, even in mutation analysis using only somatic cells, which are particularly prone to false positive detection.