nature portfolio

https://doi.org/10.1038/s41587-021-00994-5

Supplementary information

Toward best practice in cancer mutation detection with whole-genome and whole-exome sequencing

In the format provided by the authors and unedited

Towards best practice in cancer mutation detection with whole-genome and whole-exome sequencing

Materials and methods

Analysis of inter-/intra-center variations for WES and WGS

We performed the following analysis to further investigate inter- and intra-center reproducibility based on concordance of SNV detection in any pair of NGS runs, in association of SNVs defined in our call set, as described in the section for "Creating a high-confidence call truth set". We used the Jaccard index[1] to measure the concordance of SNVs from any "non-self" pair of runs:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}$$

Four comparison groups were included in this analysis: a) WES (12 repeats/3 centers), b) WGS (limited to exome regions, 12 repeats/3 centers), c) WGS (12 repeats/3 centers), and d) WES vs WGS (limited to exome regions, 24 repeats/3 centers). We did each of the four comparison groups with "all reads" (original coverage) and "downsampled reads" (fixed coverage for each NGS run, 50X for WGS and 150X for WES).

In addition, to investigate reproducibility represented by overall SNVs called in a given pair of NGS runs, we also broke down SNVs into three subgroups: a) Repeatable: SNVs defined in "HighConf" and "MedConf" categories in the call set; b) Gray zone: SNVs defined in "LowConf" and "Unclassified" categories in the call set; and c) Non-Repeatable: SNVs not defined in the call set. Jaccard scores for any pair of NGS runs were calculated as depicted in **Supplementary Fig. 2**, with two exemplar WGS runs from EA_1 and FD_1, and formulas based on: 1) overall SNV; 2) Repeatable; 3) Gray zone; and 4) Non-Repeatable.

The average value of Jaccard scores from all possible pairs of NGS runs in each of the four comparison groups was aggregated at groups for "Overall pairs", "intra-center", or "inter-center".

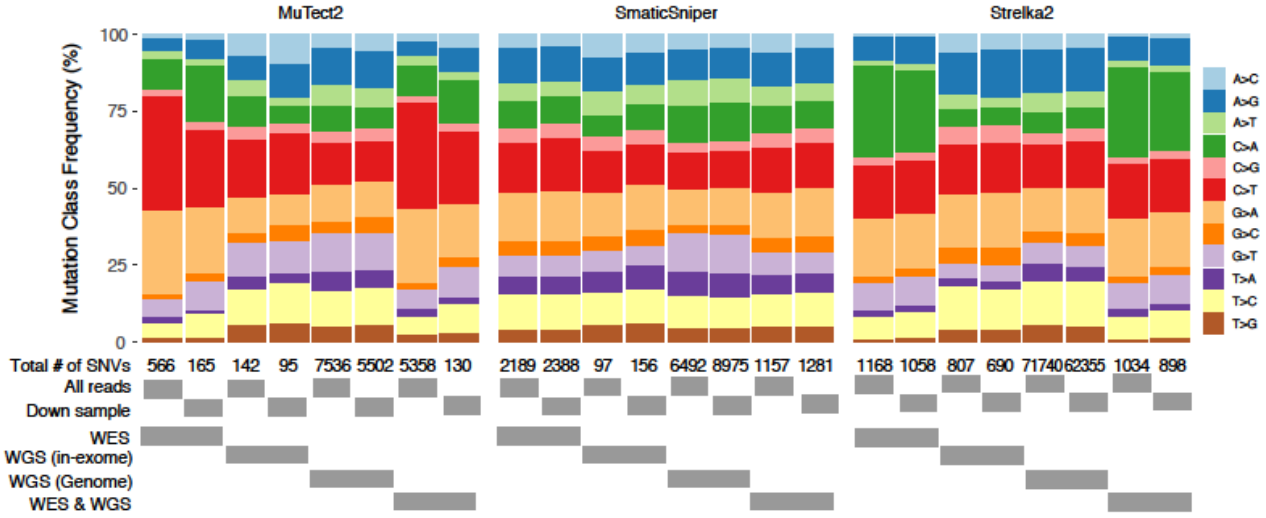
$$J(\textit{Repeatable}) = \frac{(d+e) \cap (e+f)}{(d+e) \cup (e+f)} = \frac{e}{(d+e+f)}$$

$$J(\textit{Gray zone}) = \frac{(h+i) \cap (i+j)}{(h+i) \cup (i+j)} = \frac{i}{(h+i+j)}$$

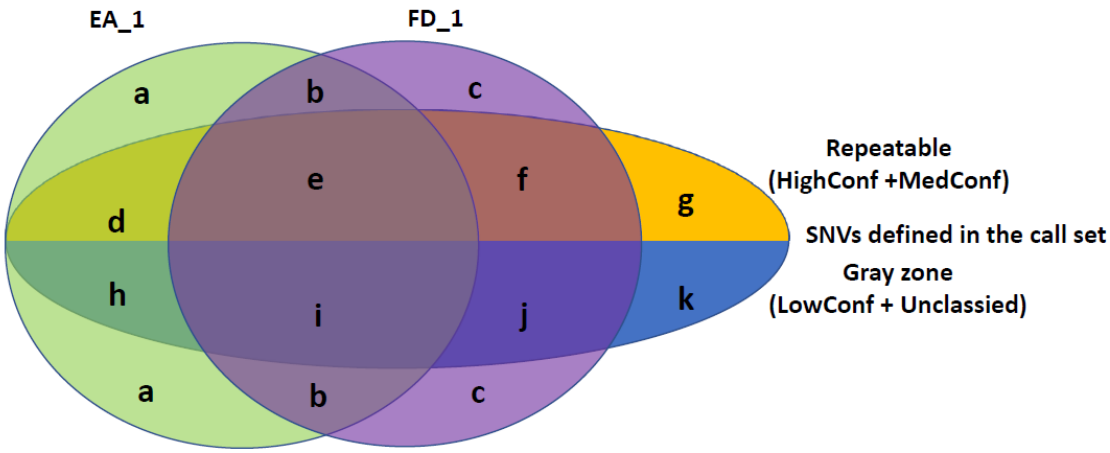
$$J(\textit{Non-Repeatable}) = \frac{(a+b) \cap (b+c)}{(a+b) \cup (b+c)} = \frac{b}{(a+b+c)}$$

Multi-variant analysis with 3-way interactions was conducted to explore the source of variation in Jaccard index (JMP Genomics 9.0). Five factors, caller (Strelka2, SomaticSniper, and MuTect2), type (all reads vs. downsampled), SNV_subset (overall, Repeatable, Gray zone, and Non-Repeatable), pair_group (inter-center vs. intra-center), platform (WES vs. WGS), as well as their interaction terms were included in the model. A coefficient of determination (R^2) of 0.98 was achieved in the model fitting. F statistics and the corresponding P-values were calculated for all the factors. We also performed Student's t-test to evaluate Jaccard index change between WES and WGS, within and across pair_groups.

Supplementary Figures



Supplementary Fig. 1. Distribution of mutation type for Non-Repeatable SNVs called by three callers, MuTect2, Strelka2, and SomaticSniper. Overall numbers of Non-Repeatable SNV in each of platform comparison group are shown on the top of figures. Percentage of mutation type is presented by the color scheme shown at the left. "all reads": all raw reads from each NGS run were used for the mutation calling; "down sample": reads from each NGS run were downsampled to a fixed coverage, 50X for WGS and 150X for WES.



Supplementary Fig. 2. Venn diagram of overlapping calls from two repeated runs in comparison to the reference call set. Two exemplar WGS runs from EA_1 and FD_1, and the reference call set which was divided into "Repeatable" (includes HighConf and MedConf calls) and "Gray zone" (LowConf and Unclassified calls), were shown in three call sets. Number of SNVs in each unique section of the Venn diagram was labelled as "a-k".

References

1. Jaccard, P. The Distribution of the Flora in the Alpine Zone. *New Phytologist* **11**, 37–50 (1912).