SV evaluation report sample FR05812606

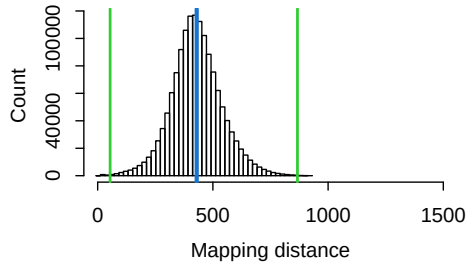
February 10, 2019

μ mean in control, σ stdev in control, z z-score. OK if $|z| \leq 2$, else !!!

1 QC from bam alignment file

Input bam file: FR05812606.bam
Percent reads pairs not mapping as proper pair: 1.3 [μ 2.1, σ 1.5, z -0.5] OK
Percent reads pairs mapping on different chromosomes: 0.68 [μ 1.2, σ 1, z -0.5] OK
Stdev of coverage: 8.1 [μ 8.1, σ 0.6, z 0] OK

Mapping distance distribution:



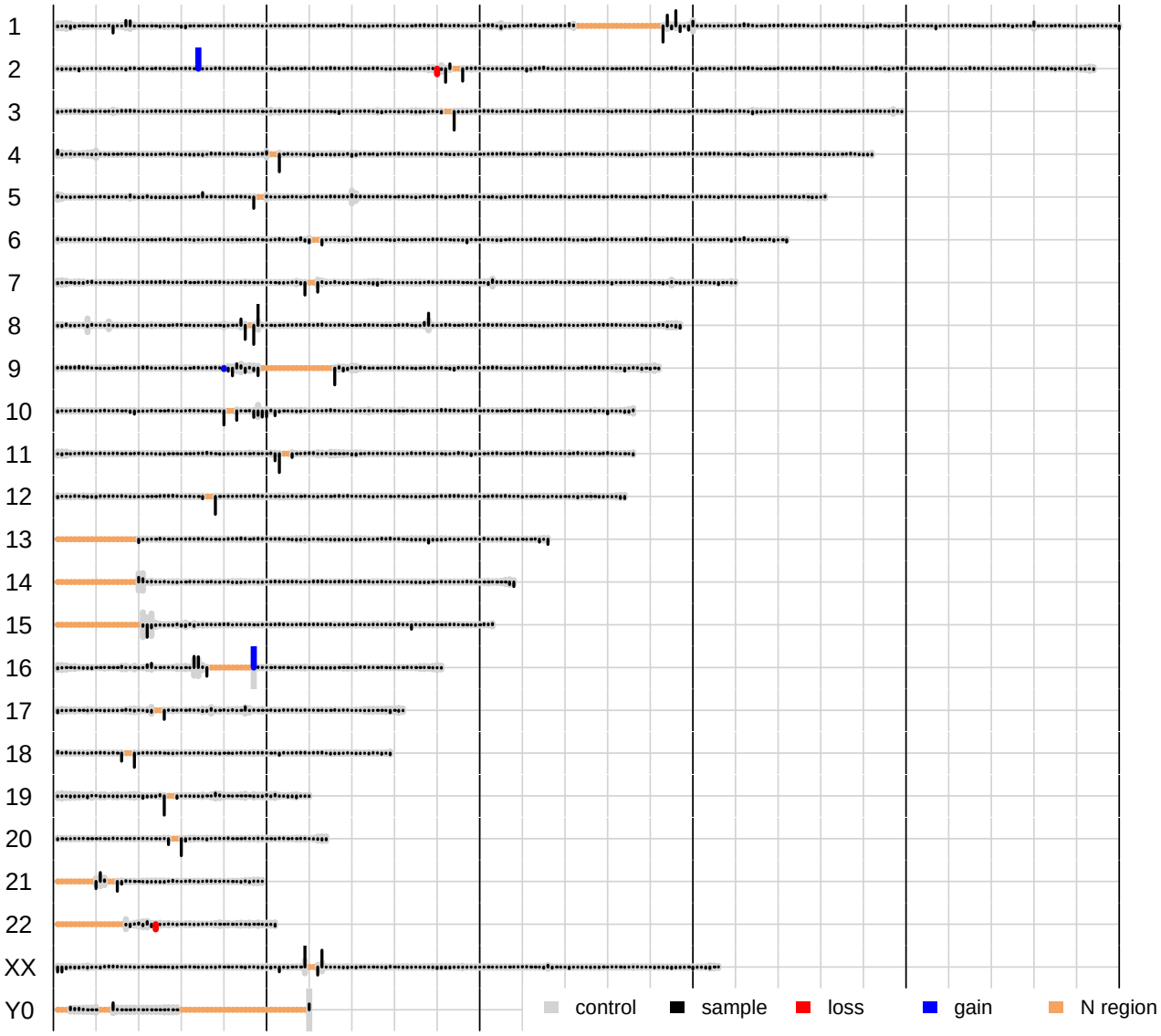
Concordant size min (green): 54 [μ 54, σ 6.7, z 0] OK
Concordant size max (green): 867 [μ 895, σ 60, z -0.5] OK
Mean mapping distance size (blue): 430 [μ 444, σ 26, z -0.5] OK
Stdev mapping distance size: 110 [μ 109, σ 7.4, z 0.1] OK

2 Input discordant pairs and split reads

Input discordant pairs: 10181760 [μ 12025111, σ 7866150, z -0.2] OK
Input split reads: 1032913 [μ 1122920, σ 174515, z -0.5] OK

3 Coverage by chromosome

Average sequence coverage: autosomes 38.4x, X 38.4x, Y 0x
Inferred gender: XX (XX female, XY male, XXY Klinefelter's Syndrome)



The coverage is normalised by the autosomal average. If the coverage of Y is close to 0, Y0 is plotted representing the background noise. This enables to detect a partial presence of Y.

4 Number called SVs

By call confidence:
High confidence variant: 2044 [μ 1844, σ 106, z 1.9] OK[b]
Pass and High confidence variants: 5724 [μ 5841, σ 323, z -0.4] OK
Low confidence variants: 4528 [μ 3983, σ 666, z 0.8] OK

By SV type [a]:
CNVs: 4634 [μ 4730, σ 187, z -0.5] OK
Loss: 3959 [μ 4025, σ 173, z -0.4] OK
Gains: 675 [μ 705, σ 54, z -0.6] OK
Balanced events including inversions and translocations: 1090 [μ 1110, σ 147, z -0.1] OK

By caller [a]:
Lumpy calls: 4154 [μ 3492, σ 255, z 2.6] OK[b]
CNVnator calls: 848 [μ 865, σ 37, z -0.5] OK[b]
In both: 722 [μ 1484, σ 87, z -8.8] !!![b]

Gene affecting / rare [a]:
Affecting genes: 2763 [μ 2832, σ 170, z -0.4] OK
Rare Pass and High confidence: 87 [μ 41, σ 23, z 2] OK[b]
Rare Pass, High confidence, affecting genes: 49 [μ 22, σ 12, z 2.2] OK[b]
Rare Pass, High confidence CNV, affecting genes: 41 [μ 17, σ 8, z 3] OK[b]

[a] Without low confidence variants
[b] OK if $|z| \leq 4$

5 Explanation QC and SV evaluation report

5.1 General QC

Low quality input data may lead to unexpected results. To guarantee the quality of the results several variables that have an impact on or are an indicator for the quality of the results are measured and compiled in this automated QC report. QC metrics measured for a particular sample are compared to the expected range obtained from analyzing 500 germline controls samples. The control samples represent previously analyzed healthy individuals (MGRB cohort) that passed QC. The expected range is generally defined as two times the standard deviation ($|z| \leq 2$) from the mean of the control cohort, unless specified otherwise. If a measured metric is within expectations, it is marked with a green OK, else with three orange exclamation marks. ClinSV is robust to a few metrics being outside the expected range, but within 4 times the standard deviation.

Re 1. QC from bam alignment file
Read pairs from Illumina paired end sequencing do not always align to the reference with their expected distance (roughly 450 bp, depending fragmentation size and size selection), regardless of the presence of structural variation. The sequencing process produces a small percentage of chimeric read pairs. These pairs originate from distant genomic locations. Despite these chimeric reads being randomly distributed; elevated numbers will impact the SV calling. Indicators for the relative abundance of chimeric pairs is the percentage of reads not mapping as proper pairs and the percentage of pairs mapping on different chromosomes. An uneven read coverage can affect CNVnator resulting in an elevated number of false CNV calls. The un-evenness of the read coverage is reflected by an increased standard deviation of the read coverage. The number of discordantly mapping pairs the prediction program Lumpy can handle is finite. The threshold for when the read mapping distance is considered discordant for pairs mapping with the expected read orientation is automatically determined (see online methods section), and results are shown here. The insert size distribution and resulting thresholds for concordant mapping distances have an impact on the smallest detectable deletions. To save computing time, metrics in this section are estimated for a 10 mega base pair region on chromosome 1 (chr1:20,000,001-30,000,000).

Re 2. Input discordant pairs and split reads
Number of discordant read pairs and split reads used as input for Lumpy. Deviating numbers could indicate library preparation or sequencing issues, deeper coverage, or samples with high numbers of structural variation, as expected for cancer samples.

Re 3. Coverage by chromosome
The average sequence coverage was determined for all chromosomes (see methods). The number of sex chromosomes is inferred from the sequence coverage. Sex chromosome aneuploidy is visible here. This section also displays the chromosome wide coverage in intervals of 1 Mega bases. Grey dots below the black dots represent the average coverage in 500 control samples plus minus two times the standard deviation. The black dots indicate the coverage of the current sample. Truncated alignment files will not cover all grey dots. One Mega base segments greater than five times the standard deviation of the control are colored blue, highlighting regions that have a copy number gain, and segments less than five times the standard deviation are colored in red, highlighting regions of copy number loss. The standard deviation is used, because regions close to the centromere tend to show a greater variation that is still considered normal, thus will not get highlighted in blue or red. Large deletions or duplications, that are likely clinical significant will be visible in this representation. N-regions usually correspond to centromeric or telomeric regions of the chromosome. The sex chromosomes will be compared to the expected coverage of X, XX, Y and/or Y0 (Y-zero), depending the average coverage of X and Y. Y0 (Y-Zero) indicates unspecific background read coverage of the Y chromosome and is helpful to reveal a partial presence of Y.

Re 4. Number of called SVs
Number of called variants by call confidence, SV type, caller, and number of variants affecting genes and being rare. Some metrics in this section show greater variation and are allowed 4 times the standard deviation from the control average. For

instance the number of rare variants could be increased in an individual of a race underrepresented in the control cohort.

5.2 NA12878 SV evaluation

The following two sections are to evaluate the SV recall rate of a NA12878 sample and allow assessing the fitness of the entire SV detection pipeline. Metrics are compared to average values of nine NA12878 control samples. Here z values greater or equal to -2 are acceptable, in order to not penalize a greater concordance than expected from the nine NA12878 control samples. This section appears when option -eval is set.

Re 5. Sensitivity
This section shows the sensitivity of detecting gold standard deletion calls, as published by GIAB (Parikh et al. 2016), excluding 12 false positives $>$ 500 bases (Minoche et al. 2017).

Re 6. Comparison to NA12878 sample
Concordance of SV between NA12878 control sample (FR05812662) and current test sample, shown in percent of FR05812662 calls or in percent of test sample calls. High confidence calls generally have a higher reproducibility compared to all pass variants. CNVs and all SVs are tested separately.