Framework for quality assessment of whole genome, cancer sequences

Justin P. Whalley[1,2], Ivo Buchhalter[3,4], Esther Rheinbay[5,6], Keiran M. Raine[7], Kortine Kleinheinz[3], Miranda D. Stobbe[1,2], Johannes Werner[3], Sergi Beltran[1,2], Marta Gut[1,2], Daniel Huebschmann[3,4,8], Barbara Hutter[9], Dimitri Livitz[5,6], Marc Perry[10], Mara Rosenberg[5,6], Gordon Saksena[5,6], Jean-Rémi Trotta[1,2], Roland Eils[3,4], Jan Korbel[11], Daniela S. Gerhard[12], Peter Campbell[7], Gad Getz[5,6,13], Matthias Schlesner[3], Ivo G. Gut*[1,2], PCAWG-Tech, PCAWG-QC & PCAWG Network

[1]*CNAG-CRG, Centre for Genomic Regulation (CRG), Barcelona Institute of Science and Technology (BIST), Barcelona, Spain*

[2]*Universitat Pompeu Fabra (UPF), Barcelona, Spain*

[3]*Division of Theoretical Bioinformatics (B080), German Cancer Research Center (DKFZ), Heidelberg, Germany*

[4]*Department for Bioinformatics and Functional Genomics, Institute for Pharmacy and Molecular Biotechnology (IPMB) and BioQuant, Heidelberg University, Heidelberg, Germany*

[5]*Massachusetts General Hospital Cancer Center and Department of Pathology, Boston, USA*

[6]*Broad Institute of Harvard and MIT, Cambridge, MA, USA*

[7]*Wellcome Trust Sanger Institute, Hinxton, UK*

[8]*Department of Pediatric Immunology, Hematology and Oncology, University Hospital Heidelberg, Heidelberg, Germany*

[9]*Division of Applied Bioinformatics (G200), German Cancer Research Center (DKFZ), Heidelberg, Germany*

[10]*Ontario Institute for Cancer Research, Toronto, Ontario, Canada*

[11]*Genome Biology Unit, European Molecular Biology Laboratory, Heidelberg, Germany*

[12]*Office of Cancer Genomics, National Cancer Institute, US National Institutes of Health, Bethesda, MD, USA*

[13]*Harvard Medical School, Boston, MA, USA*

29 *Corresponding author: ivo.gut@cnag.crg.eu

Abstract

Working with cancer whole genomes sequenced over a period of many years in different sequencing centres requires a validated framework to compare the quality of these sequences. The Pan-Cancer Analysis of Whole Genomes (PCAWG) of the International Cancer Genome Consortium (ICGC), a project a cohort of over 2800 donors provided us with the challenge of assessing the quality of the genome sequences. A non-redundant set of five quality control (QC) measurements were assembled and used to establish a star rating system. These QC measures reflect known differences in sequencing protocol and provide a guide to downstream analyses of these whole genome sequences. The resulting QC measures also allowed for exclusion samples of poor quality, providing researchers within PCAWG, and when the data is released for other researchers, a good idea of the sequencing quality. For a researcher wishing to apply the QC measures for their data we provide a Docker Container of the software used to calculate them. We believe that this is an effective framework of quality measures for whole genome, cancer sequences, which will be a useful addition to analytical pipelines, as it has to the PCAWG project.

Introduction

Combining whole genome sequencing data from individual projects has many advantages: increased statistical power, the ability to extend hypotheses across several projects and the possibility of asking biological questions covering a wider range of phenomena. However when the genome sequencing data comes from different centres, was sequenced at different times and under different protocols, great care must be taken to ensure that the sequencing data is of comparable quality, to avoid drawing false conclusions. The Pan-Cancer Analysis of Whole Genomes (PCAWG) project provided us with a great opportunity to assemble, test and finalise which quality control measures are important for comparing the quality of whole genome, cancer sequences.

The PCAWG project assembled a cohort of 48 projects encompassed in the International Cancer Genome Consortium (ICGC)[1] and The Cancer Genome Atlas (TCGA)[2] of which we analysed 2959 cancer genomes (normal-tumour genome pairs) from 2830 donors. The size of the dataset and the diversity of the samples, representing many different cancers from varied populations, allow the exploration of many fundamental questions of cancer. Although there were inclusion criteria based on the sequencing platform (Illumina) and minimum sequencing depth, there was a need to ascertain the quality of the sequencing data and how they compared to each other. There were 17 different sequencing centres involved and the sequencing was performed in over a five year time-span (2009-2014) (a time period of during which the sequencing methodology was still evolving rapidly), so to be able to perform analysis across the whole data set, it was necessary that the quality of the sequencing was carefully assessed.

There are advantages in developing a comprehensive set of quality measures. We will be able to exclude samples of low quality. This will save running downstream analyses, saving computational and the researchers' time. For researchers in PCAWG studying driver mutations, we can provide a sanity check. If the driver mutation is only found in low quality samples, it may not be a good candidate, compared to if it is supported by high quality samples. As PCAWG will release the data for community to use, our quality measures will provide a guide to the quality of the whole genome sequences within. For researchers who wish to assess the quality of their whole genome cancer sequences, we will also be releasing our methods, in a Docker Container for easy implementation.

To develop a framework needed to determine the quality of samples, we use the methods employed by the sequencing centres involved in PCAWG as well as results in the literature. TCGA marker papers (see references[3-5] for examples from 2014-16) all include quality control (QC) measures such as depth of coverage, batch effects and contamination levels, as calculated as part of the Firehose analysis infrastructure. Likewise a recent ICGC paper[6] with samples sequenced from three different centres relied on similar QC measures computed by the Picard toolkit. Lu et al.[7], carried out meta-analysis of exome data available from the TCGA for 12 cancer types which is similar, but not identical in scope, to the data set examined here. Their inclusion criteria were based on coverage depth and percentage of exome coverage for both the normal and tumour samples. Other cancer studies have also pointed to the importance of the percentage of the genome covered[8,9] as well as error rates for each of the paired reads[10] as QC measures. The scale and diversity of the PCAWG project provides a useful testing ground for QC measures,

both for selection of the measures and the thresholds to use in grading the sequences.

Here we present the results of the work by sequencing centres and research groups involved in PCAWG to define important quality control measures, and how best to combine the results from these measures. Based on the PCAWG data we selected measures covering five important features to assess the quality of cancer genome sequences: mean coverage, evenness of coverage, somatic mutation calling coverage, paired reads mapping to different chromosomes and the ratio of difference in edits between paired reads, an edit being a base in the read which is different to the reference genome. These measurements we computed for both the normal and tumour samples. To summarise the five QC measures, we established a star rating system to cover the range of the highest quality cancer genomes, passing the thresholds set for each measurement, to those that had many sequencing quality issues.

Results

All our analysis is based on the aligned sequences from the PCAWG core pipeline[11]. Within the aligned sequences we did not use duplicate reads, reads with a mapping quality of zero and ignored supplementary alignments. The first three quality control measures; mean coverage, evenness of coverage and somatic mutation calling coverage; are linked to different aspects of the coverage of the genomic sequence. The other two measures indicate discrepancies between the paired reads: mapping to different chromosomes and the ratio of edits between the paired reads compared to the reference genome. Finally we summarise these five measures into a star rating, for easy comparison of each of the sample pair's quality.

Mean Coverage When deciding on what depth to sequence cancer genomes to, a trade off has to be made between the advantages of sequencing deep to the cost of sequencing. The deeper the cancer genome is sequenced the greater the confidence in calling somatic events (see Tyler et al.[12] for a comparison of somatic mutation calling at depths up to 300X). A precondition for the inclusion of a patient in the PCAWG study was the availability of a whole genome sequence of the normal and tumour with 25X coverage or greater. We found that a number of the submitters calculated coverage differently. For standardization the mean number of reads covering each position in the genome was calculated, after low quality and duplicate reads were excluded so to not inflate the number of reads (see *Supplementary Methods* for exact methods used). As shown in *Supplementary Figure S1*, most commonly the normal samples were sequenced to around 30X, while there was a bimodal distribution for the tumour samples with maxima at 38X and 60X. To provide a meaningful guide to the quality of the genomes in PCAWG, we therefore set the thresholds for the mean coverage, after aligning, to 25X for normal samples and 30X for tumour samples. This resulted in 0.4% normal and 2.2% tumour samples not reaching these minimum criteria (see *Supplementary Figure S1*).

Evenness of Coverage To confidently identify germline variants and somatic mutations requires an even coverage across the target area[13], in this case the entire genome. Two different methods, in use by the sequencing centres involved for whole genomes, were chosen. One measure is to calculate the ratio of the median coverage over the mean coverage (MoM). An evenly covered sequence should have a ratio of one, with the mean value the same as the median value, not skewed by very low or high coverage in certain regions. To decide within what range of values a sample should fall to be regarded as

evenly covered, we used the whiskers of the boxplots in *Figure 1*, 1.5 × I.Q.R (interquartile range) of the data, which results in the range of 0.99 - 1.06 for a normal sample and the wider range of 0.92 - 1.09 for the tumour samples (see *Supplementary Figure S2*). For MoM coverage ratio (and for FWHM described below), there is a greater range of values for the tumour samples than normal samples, potentially due to biologically reasons valid for tumours, e.g. large deletions could lead to a more unevenly covered sample. If the normal sample is unevenly covered, it is more likely due to a sequencing artefact. Hence, we are more stringent for the normal than the tumour samples.

The second measure of evenness looks at the variation of the normalised coverage in ten kilobase (kb) genomic windows, after correction for GC-dependent coverage bias using the somatic CNV calling algorithm ACEseq[14] (*Figure 2*). The main cloud, which corresponds to the main copy number state of the sample, is determined (as shown by the red dots in *Figure 2*). The remaining coverage variation is measured as full width at half maximum (FWHM) of the main cloud. This measure is insensitive to copy number aberrations and GC-dependent coverage bias. To determine the thresholds, 1000 WGS samples from different tumour types were used. We chose the pruning values based on clustering of these samples and subsequent visual inspection of the "best" samples that exceeded the threshold to see whether they are valid. Using these results the thresholds chosen are 0.205 for the normal and the more lenient 0.34 for the tumour, above which the sample would be regarded as having an uneven coverage (see *Supplementary Figure S3*).

The two evenness measures tend to identify different samples as having uneven coverage

(see *Figure 3*). Spearman's correlation coefficient for the two measures suggests that these measures are not correlated for the normal ($\rho = 0.24$) and tumour ($\rho = -0.06$) samples. FWHM is insensitive to GC bias, as the CNV caller corrects for this while MoM identifies other evenness outliers.

A sample needs to be in the respective ranges of the MoM and FWHM for the normal and the tumour to pass the evenness quality measure, of which 6.28% and 5.81% respectively of the samples were not.

Somatic Mutation Calling Coverage Having both the measure of the depth of coverage in terms of the mean and evenness of coverage, our next QC measure look at the effect of these at each base in the cancer genome (so both the normal and the tumour sample). This measures gives a good summary of how much of the cancer genome is sufficiently covered to call a somatic mutation event. MuTect[15] with default settings calculates for each base in the genome, if it has sufficient coverage in both the normal and tumour sample (least fourteen reads are present in the tumour and eight reads in the matched normal sample). Based on those requirements, we had to establish the number of bases to consider the sample sufficiently covered. Ideally the threshold should be high enough to penalise the less well-sequenced samples, while not unduly penalising tumour samples that have had large deletions in the genome resulting in fewer bases to sequence. Taking into account the largest unambiguous mapping for a female donor (so not including the Y chromosome) would be 2,835,690,481 bases[16], 2.6 gigabases would best suit these two needs. This results in 5.95% of normal-tumour pairs with fewer bases, than this threshold (see *Supplementary Figure S4*).

Paired reads mapping to different chromosomes The two reads from a read pair should represent the ends of a contiguous DNA sequence that depending on the insert size should be a given distance apart (for PCAWG between 200 and 1,000 bases). Paired reads mapping to different chromosomes can be due to a rearrangement. However an excess of reads mapping to different chromosomes points to a technical artefact. So deciding a threshold based on percentage of paired reads mapping to different chromosomes, we should not penalise sequences with biological causes of the paired reads mapping to different chromosomes (such as chromothripsis[17], or more generally, interchromosomal rearrangements). We set the threshold to 3%, which even samples with confirmed high levels of rearrangements and chromothripsis do not exceed (which in our experience, do not have more than 1% of paired reads mapping to different chromosomes). Of the normal sequences 14.5% exceed the threshold, as do 13.0% tumour sequences (see *Supplementary Figure S5*). Interestingly there are more normal samples failing this measure, which cannot be explained by biological processes. A possible explanation may be that for lower quality samples in preparing libraries with PCR amplification, this amplification step causes an increase in two fragments of DNA from different parts of the genome being fused together, as has previously been noted[18]. Consequently, this translates to an increase in percentage of paired reads mapping to different chromosomes.

Ratio of difference in edits between paired reads Damage in sequencing runs has been linked to a global imbalance in edits (where the base in read is different compared to the reference) between read 1 and read 2 in paired end sequencing[19]. Therefore the ratio of the sum of edits between paired reads for a well-sequenced sample should be close to

one. We adjudged samples with a two-fold ratio of edits between the paired reads, or greater, as having something gone wrong in the sequencing cycle resulting in lower data quality. Based on this threshold 4.66% and 4.49% normal and tumour samples failed respectively.

Summary The five quality measures were selected to provide minimal redundancy in flagging quality issues in normal/tumour paired genome sequences - that each measure reflects a facet of sequencing quality that other measure does not. The best way to summarise these comparisons between the different measures is with a Venn diagram (*Figure 4*). There is some overlap between certain measures, for example 75 sample pairs are penalised by both having a high percentage paired reads mapping to different chromosomes and uneven coverage. However a much higher number of samples penalised by one of these measures and not the other. Having defined these five, non-redundant QC measures our next step was to summarise them, to give an overall score for quality for the other researchers in PCAWG to use.

Star rating system

We used the five quality measures to construct a star rating for each cancer genome (normal/tumour whole genome sequence). For each QC measure a star is awarded if both the normal and tumour sample pass the threshold. Half a star is awarded if only the normal passes the threshold for the respective QC measures. For somatic mutation calling coverage, a whole star is awarded for passing, none otherwise. The reasons for the extra weighting of the normal sample for the other four measures are that there is no biological reason for low quality in the normal sequence and a well-sequenced normal sample is

226 important for calling somatic mutations.

227 Summing the stars earned for each of the five QC measure results in 66.4% of the
228 normal/tumour sample pairs of the PCAWG being rated as 5 stars. Looking specifically
229 at the different projects (see *Figure 5*), a more nuanced picture is available. The quality
230 does not seem to be biased by tissue type (see *Supplementary Figure S7*) based on
231 detailed molecular subtypes of the tumours in PCAWG[20], the difference seems to be
232 more at the project level. Unfortunately, there is only limited project metadata on when
233 and which protocol was used to sequence the samples. Detailed metadata was available
234 for 95 donors of the CLLE-ES project (concerning Chronic Lymphocytic Leukaemia), so
235 it could be used as an example. Changes in protocol had an effect on the quality of the
236 sequencing over the four years in which CLLE-ES samples were sequenced. For the
237 CLLE-ES project, most notable was the change to a no PCR proband in 2012, which
238 resulted in improvements to the measures of paired reads mapping to different
239 chromosomes and evenness of coverage. This in turn resulted in a measurable change in
240 somatic mutation calling coverage and improvement in star ratings (*Supplementary*
241 *Figure S8*). We found similar results for a subset of 348 samples sequenced at the Broad
242 Institute (see *supplementary Figure S9*), which had metadata recorded in CGHub[21] about
243 the time and instruments used to sequence. We hypothesise that this will be true for other
244 projects as well.

245 Having calculated the star rating for the sequences, it was interesting to see how our QC
246 measures relate to the calling of somatic single nucleotide variants (SNVs)[11], somatic
247 insertion and deletions (indels)[11] and somatic structural variants (SVs)[22] in PCAWG. An
248 advantage of using these PCAWG datasets is that four callers were used for each.

Looking at the proportion of calls, which all four callers supported, gives us a good idea how the quality of sequencing influences the identification of unambiguous somatic mutations. While the proportion of calls supporting the four callers varies greatly by sample, we find that the samples with four stars or greater tended to have higher proportions than samples with less than four stars for SNVs, indels and SVs (with p-values of $\sim 10^{-5}$, $\sim 10^{-5}$, $\sim 10^{-18}$ respectively, using the Mann-Whitney-U test, also see *Figure 6*).

Taking this analysis further we used linear regression models, to further analyse the relation between the proportion of calls supported by four callers and the actual QC measures (see *Supplementary Tables S1-S3*). The results show that, significantly, an increasing percentage of paired reads mapping to different chromosomes in tumour samples, has a negative effect on the proportion of calls supported by four callers for SNVs, indels and SVs. More specifically, for SNVs an increasing mean coverage in tumours has a significant positive effect on the proportion of calls supported by four callers. While in indels there is a significant effect negative effect on the proportion of calls supported by four calls by increasing unevenness (as measured by FWHM) in tumours. As in indels, the unevenness effect is also true in SVs as well as significant negative effects by increasing percentage of paired reads mapping to different chromosomes in normal samples and ratio of difference in edits between paired reads in tumour samples.

The results from this analysis suggest the quality of sequencing as measured by star rating does have a measurable effect the downstream analyses. However the QC measures which make up the star rating effect the different downstream analyses in

different ways. As our QC measures reflect different aspects of sequencing quality, they also have varying levels of importance in using these sequences in the downstream analyses of calling SNVs, indels and SVs.

Discussion

The established star rating system allows grading the normal and tumour sample sequences by quality in absence of information on how sequencing was carried out, what protocols were used and what problems may have occurred during the sequencing process. The system is not designed to be all encompassing, instead using a small amount of computational resources and time (compared to the actual aligning of the sequences), we get a good snapshot of the quality of the normal-tumour sample pair sequences on which to call somatic mutations. Likewise having graded the cancer genomes with our five-star system, we do not intend researchers to necessarily exclude the lower ranked cancer genomes, just to be wary of any conclusions based solely on the lower scoring genomes.

With our star rating system, we sent several samples to the exclusion list due to their poor performance in one of the QC measures. Due to the timing, this did not prevent the downstream analyses being performed. Though anecdotally it would have saved 55 days computational runtime for our one star sample. For all samples that remained, the QC star rating was embedded in the header of the variant call format files for use of the researchers within PCAWG, and when the data is released, to all researchers.

For those projects in PCAWG, which we had metadata, we found that sequencing quality has definitely improved over the time period 2009-2014 in which the samples sequenced.

Our results for the CLLE-ES project suggest that in part a protocol change to PCR-free methods improved sequencing, as in line with best practices from a recent benchmarking exercise.[12]

Another advantage of our quality control is the link to the downstream analyses. In aggregate, the higher the quality of the sequences, had a higher proportion of the consensus somatic SNVs, indels, SVs called. These results suggest overall that higher quality sequence will identify the true positive somatic mutations with higher probability. Our data would suggest that when pre-amplification of DNA will be needed for WGS, e.g. DNA isolated from formalin fixed, paraffin embedded tissue, the star rating system will be helpful when the variants and mutations are interpreted.

We believe that our method can be adapted for similar projects that look to use whole genome sequences from a variety of sources. The thresholds we used based on our experience and applied to this dataset of 2959 cancer genomes can also be used as guide to quality of sequences. It is worth noting that they represent a trade-off of being severe enough to penalise poor quality while not discriminating against samples with valid biological causes. We also would recommend using our methods to ascertain the quality before downstream analyses by other groups. To enable others to use our approach, there is a Docker Container, which can be accessed at https://github.com/eilslabs/PanCanQC. We provide a framework for quality assessment, which opens the door to do large-scale meta-analysis in a more robust framework.

Acknowledgements

The authors would like to thank Jennifer Jennings and her colleagues at the Ontario Institute for Cancer Research (OICR) for their help in the administration of this working group.

JPW, MDS, SB, MG, JT and IGG are supported by the Ministerio de Economía, Industria y Competitividad and European Regional Development Fund (MINECO/FEDER BIO2015-71792-P), the Instituto de Salud Carlos III (ISCIII) and the Generalitat de Catalunya. In addition we have received funding from ELIXIR-EXCELERATE (EC H2020 #676559) and RD-Connect (EC FP7/2007-2013 #305444).

The work done by IB, KK, JW, DH, BH, RE and MS was supported by the BMBF-funded Heidelberg Center for Human Bioinformatics (HD-HuB) within the German Network for Bioinformatics Infrastructure (de.NBI) (#031A537A, #031A537C) and the BMBF-funded German ICGC-projects (ICGC-PedBrain: 109252 (German Cancer Aid), 01KU1201A,B; ICGC-MMML: 01KU1002B and ICGC-DE-MINING: 01KU1505E).

ER, DL, MR, GS and GG would like to acknowledge G.G. MGH startup package and Broad funds.

KMR and PC are members of the Cancer Genome Project supported by a Wellcome Trust grant (098051).

Author contributions

JPW, IB, ER, KMR, KK, MDS and JW wrote the manuscript, helped develop and apply the methods and analysed the results.

SB, MG, DH, BH, DL, MP, MR, GS and JT contributed to the development of the methods.

RE, JK, DSG, PC, GG, MS and IGG provided project supervision; through feedback and the reviewing of the work done, as well as editing of the manuscript.

IB and JW constructed the Docker Container with code contributions from KK and KMR.

PCAWG-Tech and PCAWG-Network provided the data, metadata and the framework for this research.

Competing financial interests

The authors declare no competing financial interests

References

1. International Cancer Genome Consortium et al. International network of cancer genome projects. Nature 464, 993–8 (2010).

2. Cancer Genome Atlas Research Network et al. The cancer genome atlas pan-cancer analysis project. Nat Genet 45, 1113–20 (2013).

3. Ceccarelli, M. et al. Molecular profiling reveals biologically discrete subsets and pathways of progression in diffuse glioma. Cell 164, 550–63 (2016).

4. Cancer Genome Atlas Network. Comprehensive genomic characterization of head and neck squamous cell carcinomas. Nature 517, 576–82 (2015).

5. Cancer Genome Atlas Research Network. Comprehensive molecular characterization of urothelial bladder carcinoma. Nature 507, 315–22 (2014).

6. Biankin, A. V. et al. Pancreatic cancer genomes reveal aberrations in axon guidance pathway genes. Nature 491, 399–405 (2012).

7. Lu, C. et al. Patterns and functional implications of rare germline variants across 12 cancer types. Nat Commun 6, 10086 (2015).

8. Liu, J. et al. Genome and transcriptome sequencing of lung cancers reveal diverse mutational and splicing events. Genome Res 22, 2315–27 (2012).

9. Ramkissoon, L. A. et al. Genomic analysis of diffuse pediatric low-grade gliomas identifies recurrent oncogenic truncating rearrangements in the transcription factor mybl1. Proc Natl Acad Sci U S A 110, 8188–93 (2013).

10. Berger, M. F. et al. The genomic complexity of primary human prostate cancer. Nature 470, 214–20 (2011).

11. Simpson, J. et al. Detecting Somatic Mutations in 2,834 Cancer Whole Genomes. In preparation.

12. Alioto, T. S. et al. A comprehensive assessment of somatic mutation detection in cancer using whole-genome sequencing. Nat Commun 6, 10001 (2015).

13. Mokry, M et al. Accurate SNP and mutation detection by targeted custom microarray-based genomic enrichment of short-fragment sequencing libraries. Nucleic Acids Res (2010).

14. Kleinheinz et al. Copy-number variants from ACESeq. In preparation.

15. Cibulskis, K. Et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat Biotechnol 31, 213–9 (2013).

16. Zook, J. M. et al. Integrating human sequence data sets provides a resource of benchmark snp and indel genotype calls. Nat Biotechnol 32, 246–51 (2014).

17. Korbel, J. O. & Campbell, P. J. Criteria for inference of chromothripsis in cancer genomes. Cell 152, 1226–36 (2013).

18. Oyola, S. O. *et al.* Optimizing illumina next-generation sequencing library preparation for extremely at-biased genomes. BMC Genomics 13, 1 (2012).

19. Chen, L., Liu, P., Evans, T. C. & Ettwiller, L. M. DNA damage is a pervasive cause of sequencing errors, directly confounding variant identification. Science 355, 752– 756 (2017).

20. Hoadley, K. et al. Supervised and unsupervised molecular classification of diverse tumour types from whole genome sequencing data. In preparation.

21. Wilks, C. et al. The cancer genomics hub (CGHub): overcoming cancer through the power of torrential data. Database (2014).

22. PCAWG-6. Patterns of structural variations, signatures, genomic correlations, retrotransposons, mobile elements. In preparation.

Figures

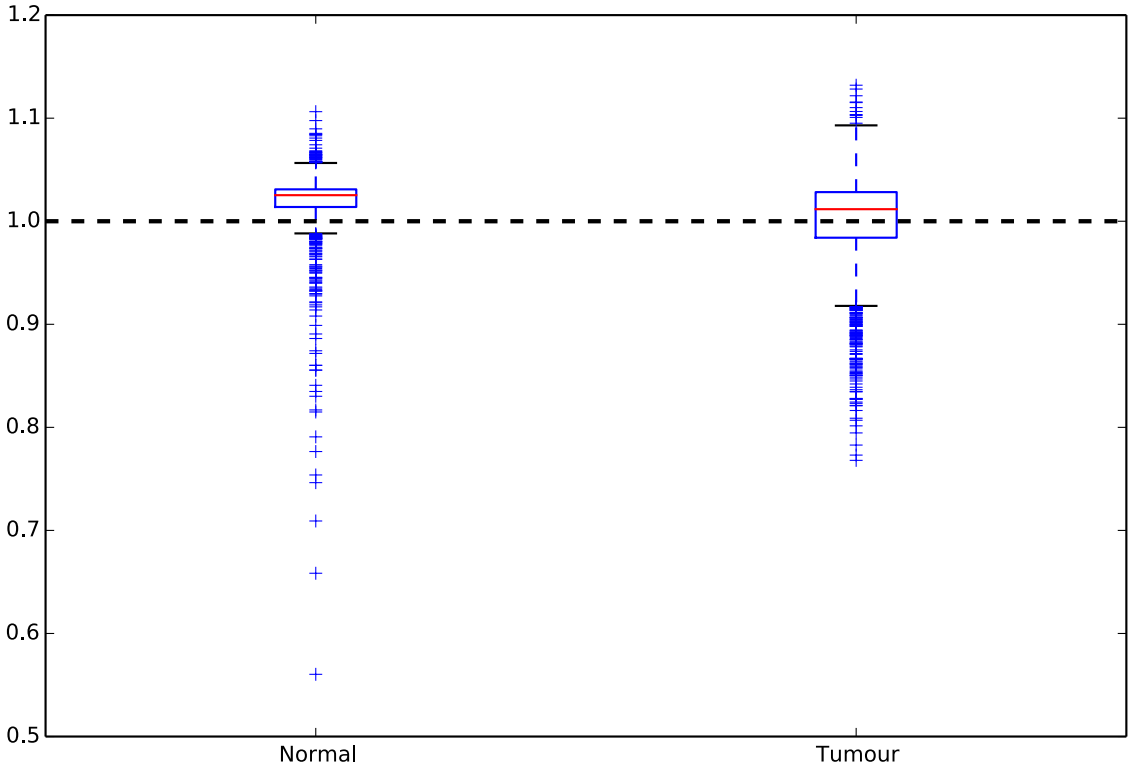
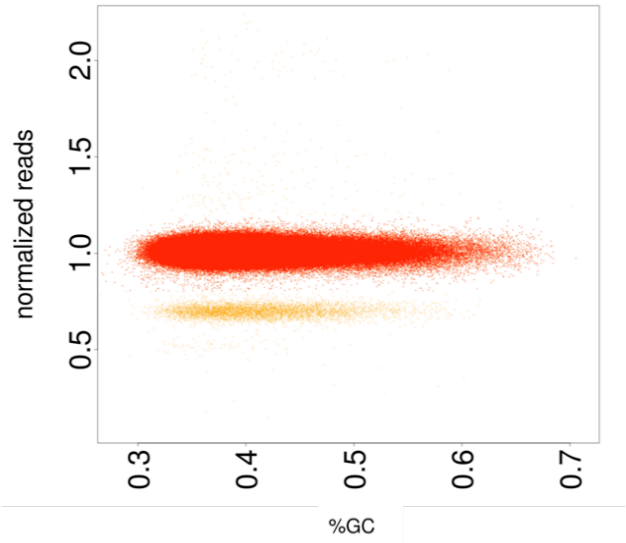


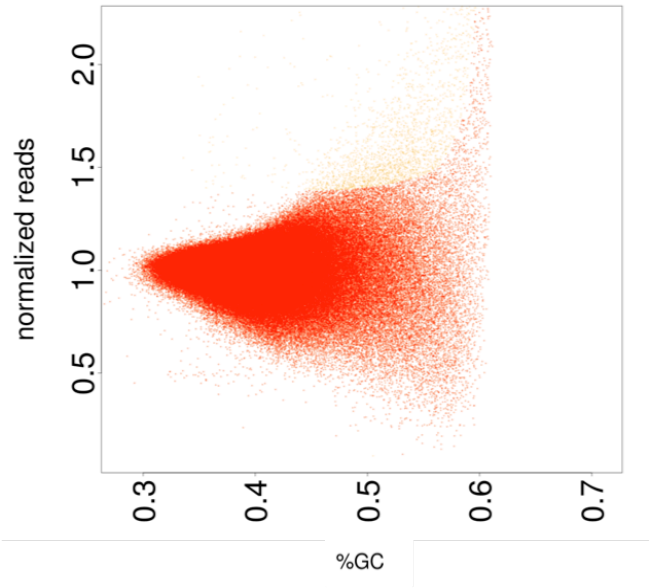
Figure 1: Distribution of the median coverage over mean coverage ratios for normal and tumour samples. The horizontal dashed bar at 1 represents the value of an evenly covered sample. As shown in the plot the tumour samples have a greater spread of values than the normal, we hypothesize this is to be expected as tumours are more likely to have deletions and structural rearrangements, which will lead to less evenly covered sequence. The whiskers on each of the boxplots (0.99-1.06 for the normal and 0.92-1.09 for the tumour) were taken as thresholds for this measure.

403 *a)*



404

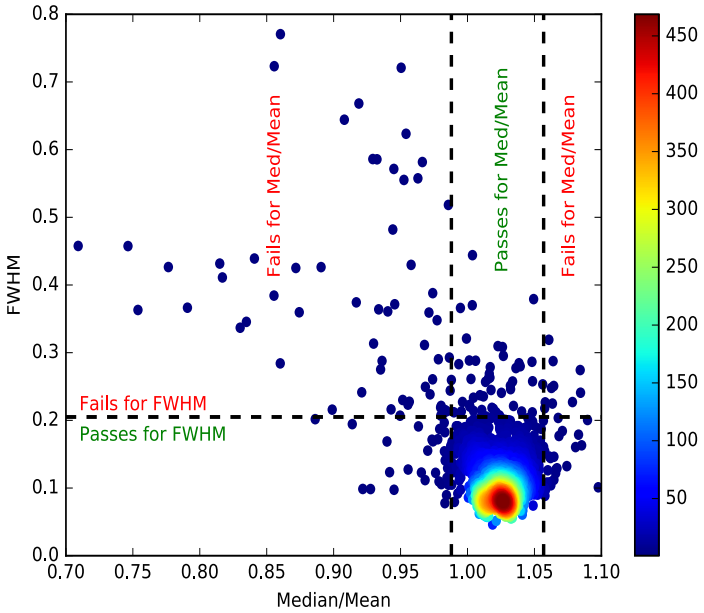
405 *b)*



406

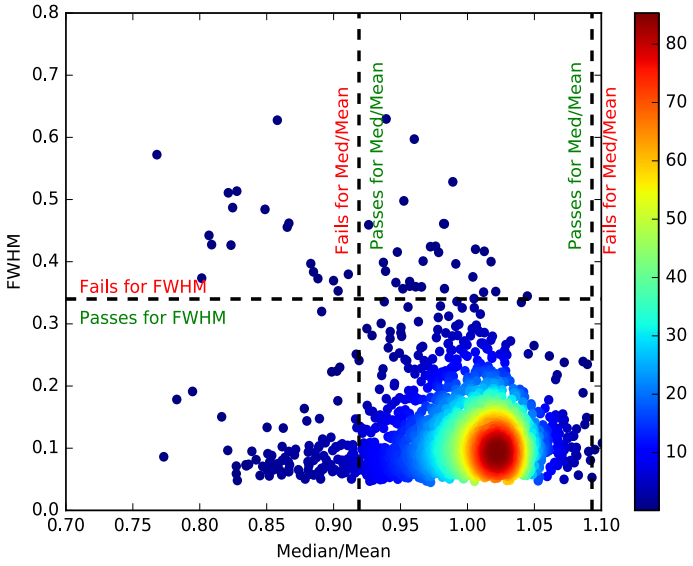
407 *Figure 2: GC content versus the normalised coverage for evenly covered sample (a) and*
408 *unevenly covered sample (b). The main cloud, corresponding to the main copy number*
409 *state of the samples, is indicated in red. Other clouds (yellow in the (a)) represent*
410 *different copy number states of copy- number aberrant regions. FWHM is calculated on*
411 *the main copy number state.*

412 *a)*



413

414 b)



415

416 *Figure 3: Density scatter plot comparing the two evenness of coverage measures for*
417 *normal (a) and tumour (b). The number of points overlapping is reflected by the colour at*
418 *that point as shown by the legend. The dashed lines reflect the thresholds for the evenness*
419 *measures. These graphs show while there both methods pick out certain samples as*
420 *unevenly covered, they also show individually samples which do not have even coverage.*

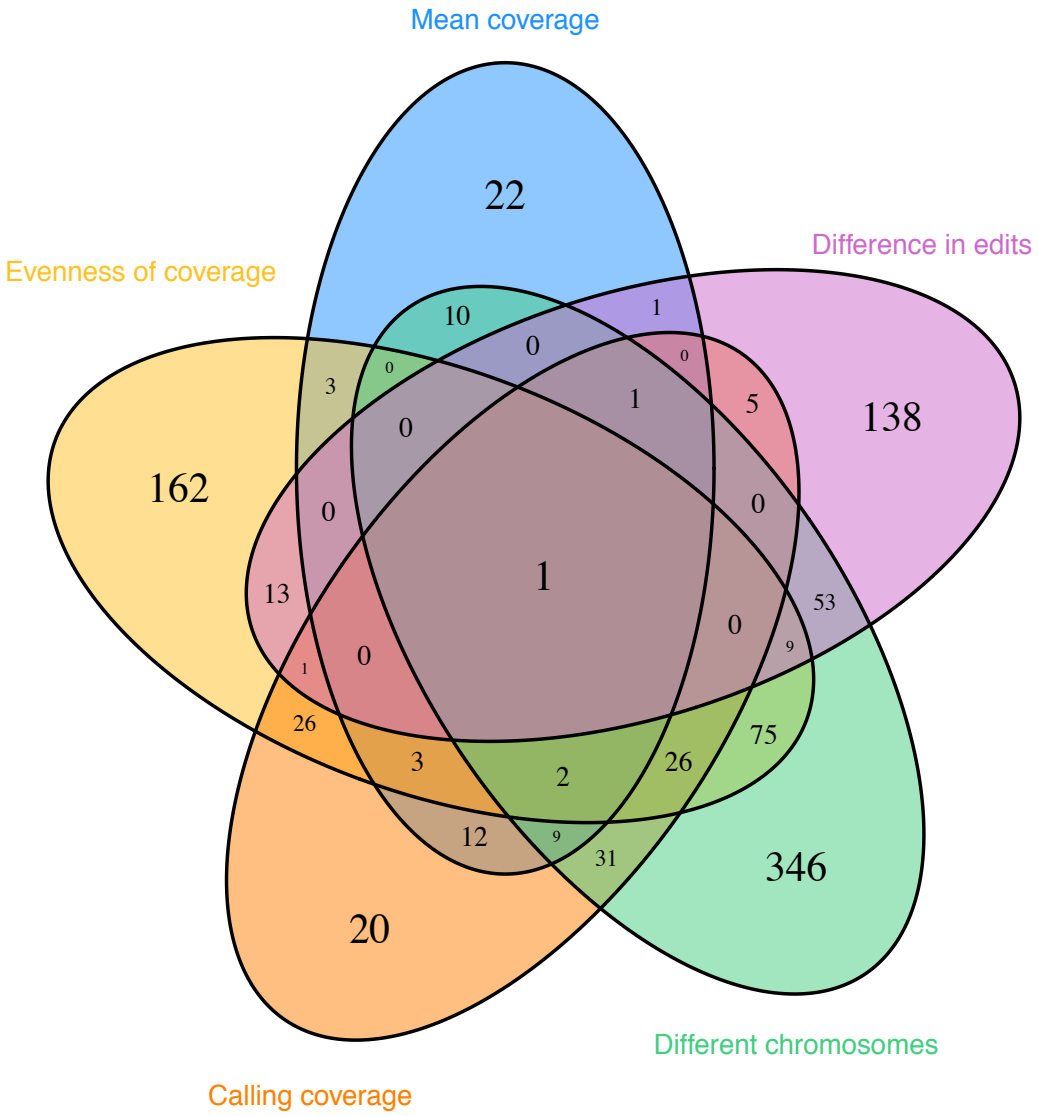


Figure 4: Venn diagram showing for which QC measure sample pairs were penalised for. The outside numbers show that each QC measures penalises a fair number of sample pairs uniquely. Looking at the overlaps between QC measures, while some measures are closer to each other than others, they all maintain a large degree of independence.

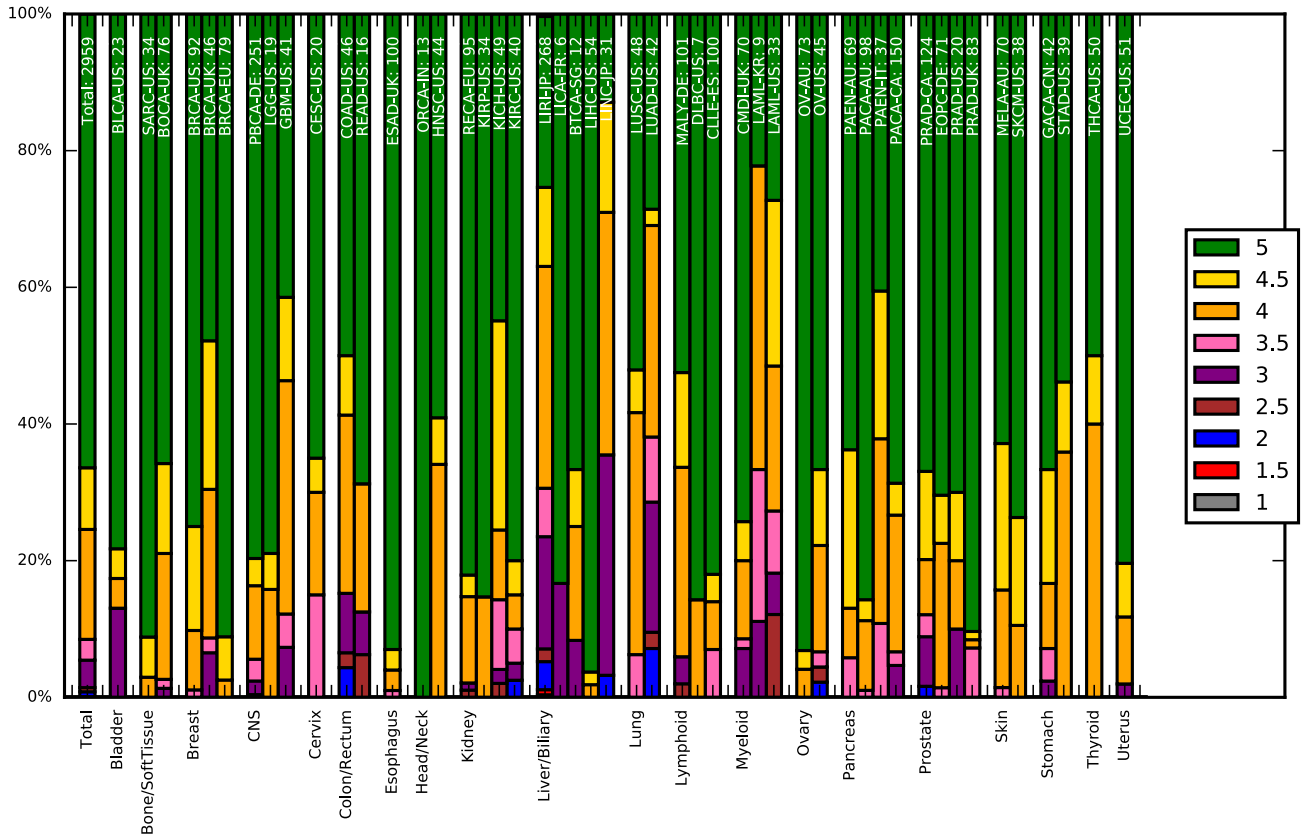


Figure 5: Distribution of the star ratings for the PCAWG genomes, grouped by tissue type (as labelled along the x-axis), and then project. The project name and number of samples in the project are labelled at the top of the bar. The colour of the bar reflects what percentage of samples in the project have that star rating (corresponding to the legend). The bar on the far left shows the results for all samples. The plot demonstrates the varying quality of different projects - differences we believe come from when the genome was sequenced and the sequencing protocol used.

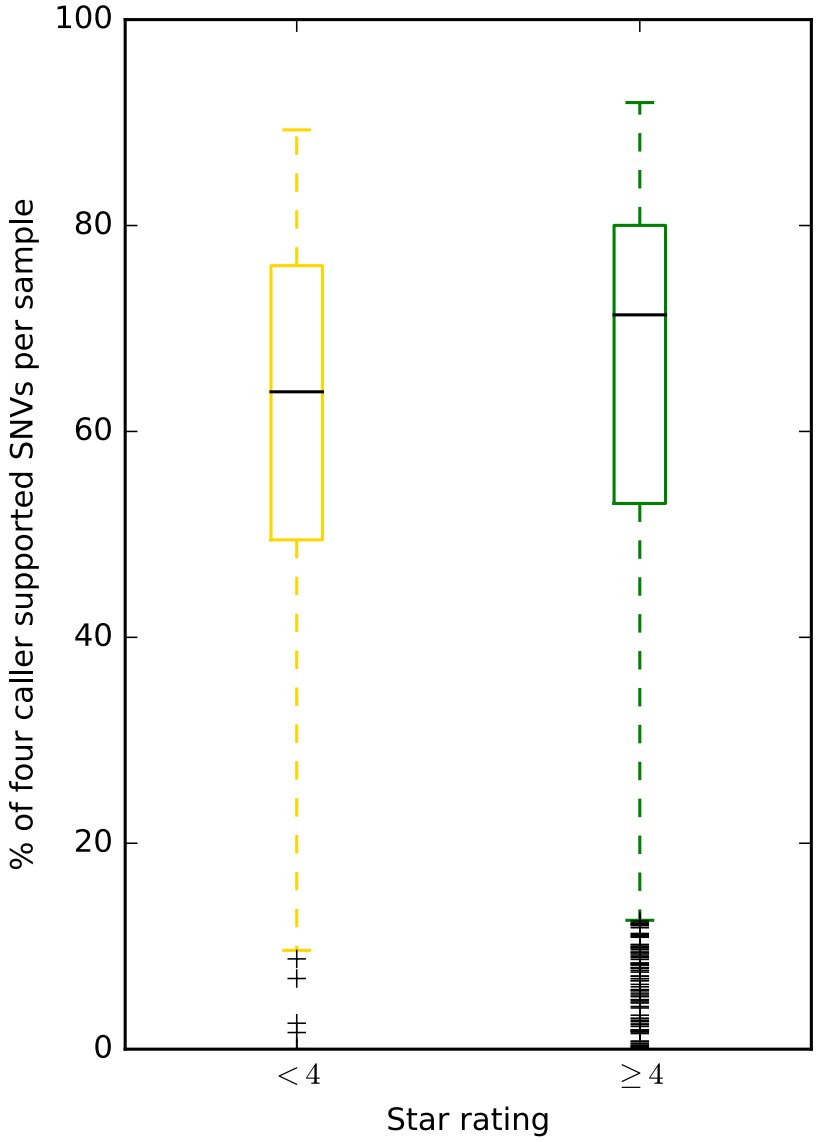
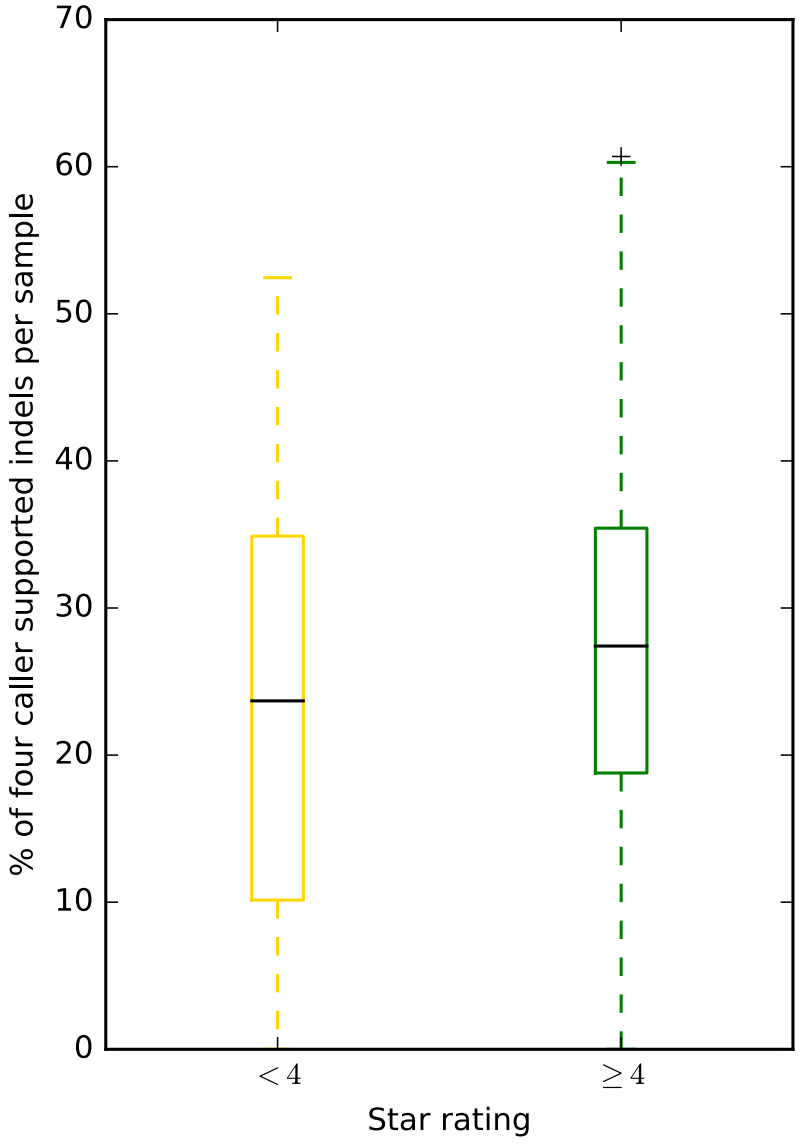


Figure 6a: Samples with four stars or greater tend to have a higher the proportion of somatic single nucleotide variants (SNV) calls supported by four callers than samples with fewer than four stars. This is significant using the Mann-Whitney U test, with p-value ~ 10^{-5}.



439

440 *Figure 6b: Samples with four stars or greater tend to have a higher the proportion of*
441 *somatic insertion and deletion (indel) calls supported by four callers than samples with*
442 *fewer than four stars. This is significant using the Mann-Whitney U test, with p-value ~*
443 10^{-5}.

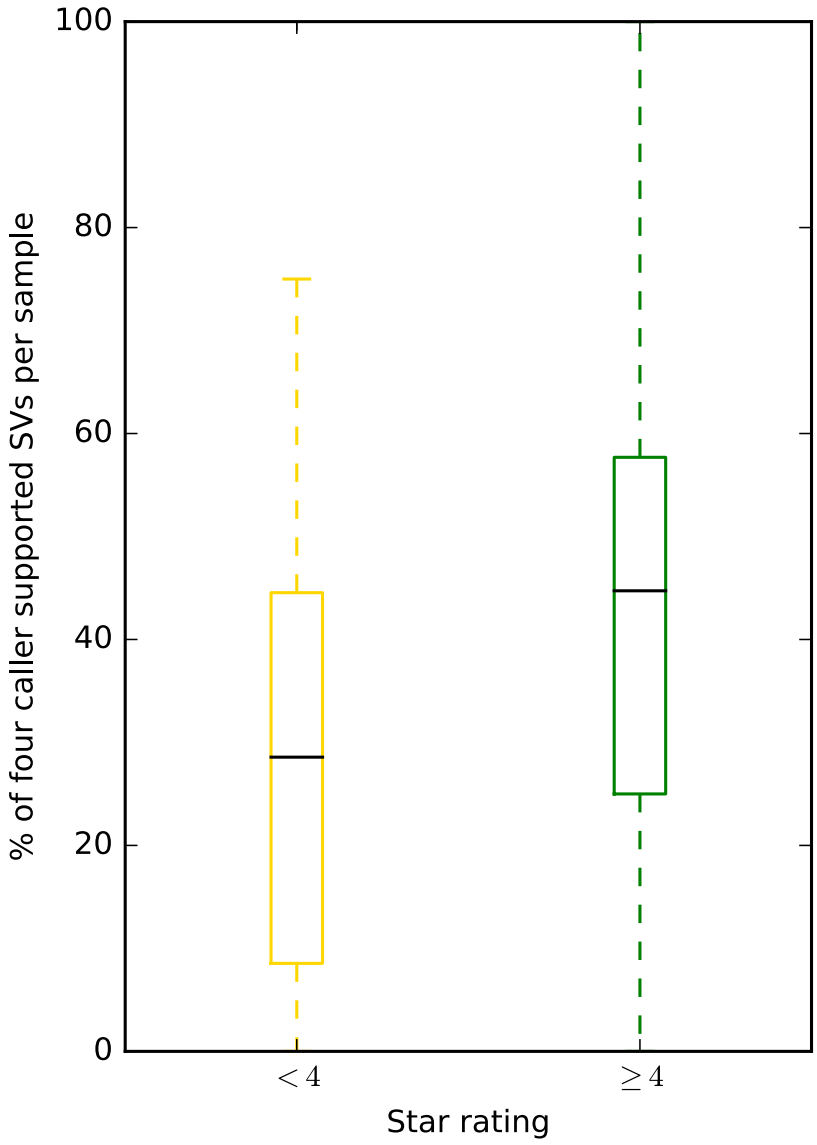


Figure 6c: Samples with four stars or greater tend to have a higher the proportion of somatic structural variant (SV) calls supported by four callers than samples with fewer than four stars. This is significant using the Mann-Whitney U test, with p-value ~ 10^{-8}.