Combining accurate tumour genome simulation with crowd-sourcing to benchmark somatic single nucleotide variant detection

Adam D. Ewing[1,2,11], Kathleen E. Houlahan[3,11], Yin Hu[4,11], Kyle Ellrott[1], Cristian Caloian[3], Takafumi N. Yamaguchi[3], J. Christopher Bare[4], Christine P'ng[3], Daryl Waggott[3], Veronica Y. Sabelnykova[3], ICGC-TCGA DREAM Somatic Mutation Calling Challenge Participants[5], Michael R. Kellen[4], Thea C. Norman[4], David Haussler[1], Stephen H. Friend[4], Gustavo Stolovitzky[6], Adam A. Margolin[4,7,8,12], Joshua M. Stuart[1,12], Paul C. Boutros[3,9,10,12]

1 Department of Biomolecular Engineering; University of California, Santa Cruz; Santa Cruz, CA, USA
2 Mater Research Institute; University of Queensland; Woolloongabba, QLD, Australia
3 Informatics and Biocomputing Program; Ontario Institute for Cancer Research; Toronto, Ontario, Canada
4 Sage Bionetworks; Seattle, WA, USA
5 A list of members and affiliations appears at the end of the paper
6 IBM Computational Biology Center; T.J. Watson Research Center; Yorktown Heights, NY, USA
7 Computational Biology Program; Oregon Health & Science University; Portland, OR, USA
8 Department of Biomedical Engineering; Oregon Health & Science University; Portland, OR, USA
9 Department of Medical Biophysics; University of Toronto; Toronto, Ontario, Canada
10 Department of Pharmacology & Toxicology; University of Toronto; Toronto, Ontario, Canada
11 These authors contributed equally
12 These authors jointly directed the work

Correspondence should be addressed to P.C.B. (Paul.Boutros@oicr.on.ca) or A.D.E. (Adam.Ewing@mater.uq.edu.au)

Nature Methods: doi:10.1038/nmeth.3407

Supplementary Note 1

Challenge Design

Due to the rapid pace of new technology introduction, algorithm development for interpreting NGS results has been forced to adapt quickly. This has led to a situation in which both the sequencing characterization and the analysis software have been poorly characterized in terms of their error profiles, which can confound their use in both discovery and clinical applications. Here we report the development of BAMSurgeon to generate robust *in silico* tumour-normal pairs, and its use with crowd-sourcing to provide the largest benchmark of somatic SNV-calling methods to date. Challenges incentivize collaboration, can lead to innovative solutions or to the identification of new problems that can become fodder for new Challenges, accelerate learning, help establish community-standards, allow objective prioritization of methods and help build a community of researchers around specific and timely problems. There has been a growing trend in the use of crowd-sourcing to stimulate research in specific areas[1], and DREAM (Dialogue for Reverse Engineering Assessment and Methods) has been a leader in promoting this approach across multiple problem domains. Recent and ongoing DREAM Challenges in systems biology are promoting rigorous performance assessment, development of standards and demonstrating how ensemble methods sampled across community predictions can improve upon the work of any individual group. Thus structuring benchmark development as a Challenge incentivized collaboration and rapid learning, and allowed the Challenge community to assess a broad cross-section of current methods efficiently.

The SMC Challenge includes two components (**Supplementary Figure 1**). To encourage participation from researchers in alternate fields, we simulated five synthetic tumours (sub-Challenges 2A-1 to 2A-5 and 2B-1 to 2B-5) of increasing difficulty and created corresponding leaderboards to provide real-time feedback. These five sub-Challenges allow for algorithm training prior to the main Challenge (Intel-10 SNV sub-Challenge and ITM1-10 SV sub-Challenge) in which we provided 10 tumour/normal pairs from real patients (five samples derived from prostate cancers and five derived from pancreatic cancers). It is ensured that participants have approval of data access by the ICGC Data Access Compliance Office. To validate performance on the real tumours, thousands of predicted variants will be sequenced using Ion Torrent, an independent sequencing technology. These two stages will allow for benchmarking of somatic single-nucleotide and structural variation prediction on synthetic and patient-derived datasets. Upon completion of the Challenge, the best performing methods will be made available to the community as validated open source pipelines.

The Challenge is run on the Synapse (www.synapse.org) open computational platform. Synapse serves not just as a data repository but also as a framework for conducting collaborative analysis and sharing and documenting data, models and analysis methods. Synapse enables researchers to seamlessly and transparently conduct, track and share their ongoing work – building up living research projects in real-time. GeneTorrent client, an open--source software developed by Annai Systems, is available for local data download. A comprehensive description of GeneTorrent features and operation is available on the CGHub website: https://cghub.ucsc.edu/docs/user/index.html. Google is offering Google Cloud Platform credits of $2,000 to approved Challenge participants, including free access to contest data in Google Cloud Storage. These credits can be used for Compute Engine VMs and other Cloud Platform services. Futhermore, free access to Challenge data is provided via a Google Cloud Storage bucket, so all computation and submissions can be performed on the Google Cloud Platform.

Overall Challenge Findings

While we have reported here only the results of the first three SMC Challenge tumours, participation remains high; to date 387 registrants have submitted 3,132 analyses of 14 genomes. Our analysis of the results from the first SMC Challenge tumour has yielded several important discoveries. First, it has confirmed the widely suspected inter-regional variability in error-rates, where variant-calling tool-chains have been optimized towards coding regions. As increasing numbers of functional non-coding SNVs are identified[2], algorithm-developers will be able to use tools like BAMSurgeon to develop algorithms with improved accuracy outside of coding regions. Second, the large number of submissions allowed for robust statistical modeling of the sequence-characteristics associated with errors. False-positives and false-negatives showed distinct characteristics, with only a few variables (*e.g.* mapping quality, normal coverage and base quality) being important for both -- this may guide the quality-evaluation of clinically-targeted sequencing. Third, our results provide clear evidence that ensemble-based approaches comprised of existing algorithms may be an effective way to improve prediction accuracies, as shown in several other areas of biology by other DREAM Challenges[3-5], and hinted at previously for somatic mutation calling[6]. Fourth, we have shown that sequencing-errors can closely resemble real biological discoveries. Ongoing stages of the challenge address structural variants and short indels in synthetic tumor-normal pairs. Analysis of the synthetic phases will be used to guide later stages of the Challenge when the algorithms are applied to real tumor/normal pairs. Finally, comparison of synthetic and real results will feed back into BAMSurgeon development efforts, improving the fidelity with which synthetic reads can be generated.

An unexpected outcome of this Challenge was an improvement in our ability to accurately simulate tumour-normal genomes. Challenge contestants continually offered suggestions (including source-code patches and detailed statistical analyses) to enhance BAMSurgeon's simulations. This highlights the value of open-science to foster incremental community-improvements that yield robust tools of broad benefit.

No algorithm perfectly predicts somatic SNVs on even the simplest tumour (IS1) -- the best team achieved an F-score of 0.975. This is high enough to lead to significant artifacts in downstream analyses as errors are non-randomly distributed. Surprisingly, reduced tumour-cellularity did not significantly alter error-rates (IS2), but sub-clonality did: the best methods achieved F-scores of ~0.95 in IS3. These data strongly suggest that there remains significant room to improve somatic SNV prediction algorithms.

Supplementary Note 2

Related Work

Existing methods for simulating cancer genomes generally fall into one of two categories: (1) simulating reads from a reference genome assembly or (2) spiking in sequence reads that support known *bona fide* mutations into an alignment that lacks the spiked-in mutations. The first approach is exemplified by a number of software tools, the most widely used is perhaps the wgsim utility (https://github.com/lh3/wgsim), built upon DWGSIM (https://github.com/nh13/DWGSIM). Further examples include pIRS[7], GemSIM[8], SInC[9], Mason[10], ART[11], among others. Each of these has varying parameters and some include simulation of error models for some subset of sequencing technologies. Examples of the second approach where reads from one sample are 'spiked-in' to another include the SomaticSpike tool used to evaluate the MuTect somatic mutation detection method[12], and the datasets generated for the SMaSH benchmarking toolkit[13]. These two general approaches (read simulation and spike-in of 'real' mutations) both have their merits and demerits; for example, simulating reads can simulate any underlying genome mutation or rearrangement as the reference from which the simulated reads are generated serves as the 'ground truth'. The primary drawback is that simulated reads cannot recapitulate biases and error profiles if they are not completely known for a given combination of sequence technology and sample preparation method - this is a reasonably serious drawback given that the sequencing method is a fundamental source of error in mutation calls and it is unlikely that the error profile of any given combination of sequencing method and sample preparation method is completely specified in a way amenable to simulation approaches. Using reads from actual sequencing results that support known mutations provides a clear route around this drawback, but with the disadvantage that the sites of spiked-in mutations must come from known mutations: any arbitrary site in the genome is unlikely to be the site of a mutation present in dbSNP[14], COSMIC[15] (cancer.sanger.ac.uk), or other sources of *bona fide* validated mutations. Put another way, any 'spike-in' mutation must have been detectable by some means, therefore simulations using this method could conceivably be biased towards mutations already detectable by existing mutation callers, thereby limiting the development of callers with improved sensitivity. BAMSurgeon bridges these two general approaches to mutation simulation by providing a third alternative: modifying pre-existing alignments and realigning the modified reads. Through this approach, any arbitrary site with adequate read coverage (as defined by the user) can be mutated, and the underlying error profile stemming from the sequencing technology and sample preparation method will be realistic.

Supplementary References

1 Costello, J. C. & Stolovitzky, G. Seeking the wisdom of crowds through challenge-based competitions in biomedical research. *Clin Pharmacol Ther* **93**, 396-398, doi:10.1038/clpt.2013.36 (2013).

2 Vinagre, J. *et al.* Frequency of TERT promoter mutations in human cancers. *Nature communications* **4**, 2185, doi:10.1038/ncomms3185 (2013).

3 Cozzetto, D., Kryshtafovych, A. & Tramontano, A. Evaluation of CASP8 model quality predictions. *Proteins* **77 Suppl 9**, 157-166 (2009).

4 Margolin, A. A. *et al.* Systematic analysis of challenge-driven improvements in molecular prognostic models for breast cancer. *Science translational medicine* **5**, 181re181, doi:10.1126/scitranslmed.3006112 (2013).

5 Marbach, D. *et al.* Wisdom of crowds for robust gene network inference. *Nat Methods* **9**, 796-804, doi:10.1038/nmeth.2016 (2012).

6 Kim, S. Y., Jacob, L. & Speed, T. P. Combining calls from multiple somatic mutation-callers. *BMC Bioinfo.* **15**, 154, doi:10.1186/1471-2105-15-154 (2014).

7 Hu, X. *et al.* pIRS: Profile-based Illumina pair-end reads simulator. *Bioinformatics* **28**, 1533-1535, doi:10.1093/bioinformatics/bts187 (2012).

8 McElroy, K. E., Luciani, F. & Thomas, T. GemSIM: general, error-model based simulator of next-generation sequencing data. *BMC Genomics* **13**, 74, doi:10.1186/1471-2164-13-74 (2012).

9 Pattnaik, S., Gupta, S., Rao, A. A. & Panda, B. SInC: an accurate and fast error-model based simulator for SNPs, Indels and CNVs coupled with a read generator for short-read sequence data. *BMC Bioinfo.* **15**, 40, doi:10.1186/1471-2105-15-40 (2014).

10 Holtgrewe, M., Emde, A. K., Weese, D. & Reinert, K. A novel and well-defined benchmarking method for second generation read mapping. *BMC Bioinfo.* **12**, 210, doi:10.1186/1471-2105-12-210 (2011).

11 Huang, W., Li, L., Myers, J. R. & Marth, G. T. ART: a next-generation sequencing read simulator. *Bioinformatics* **28**, 593-594, doi:10.1093/bioinformatics/btr708 (2012).

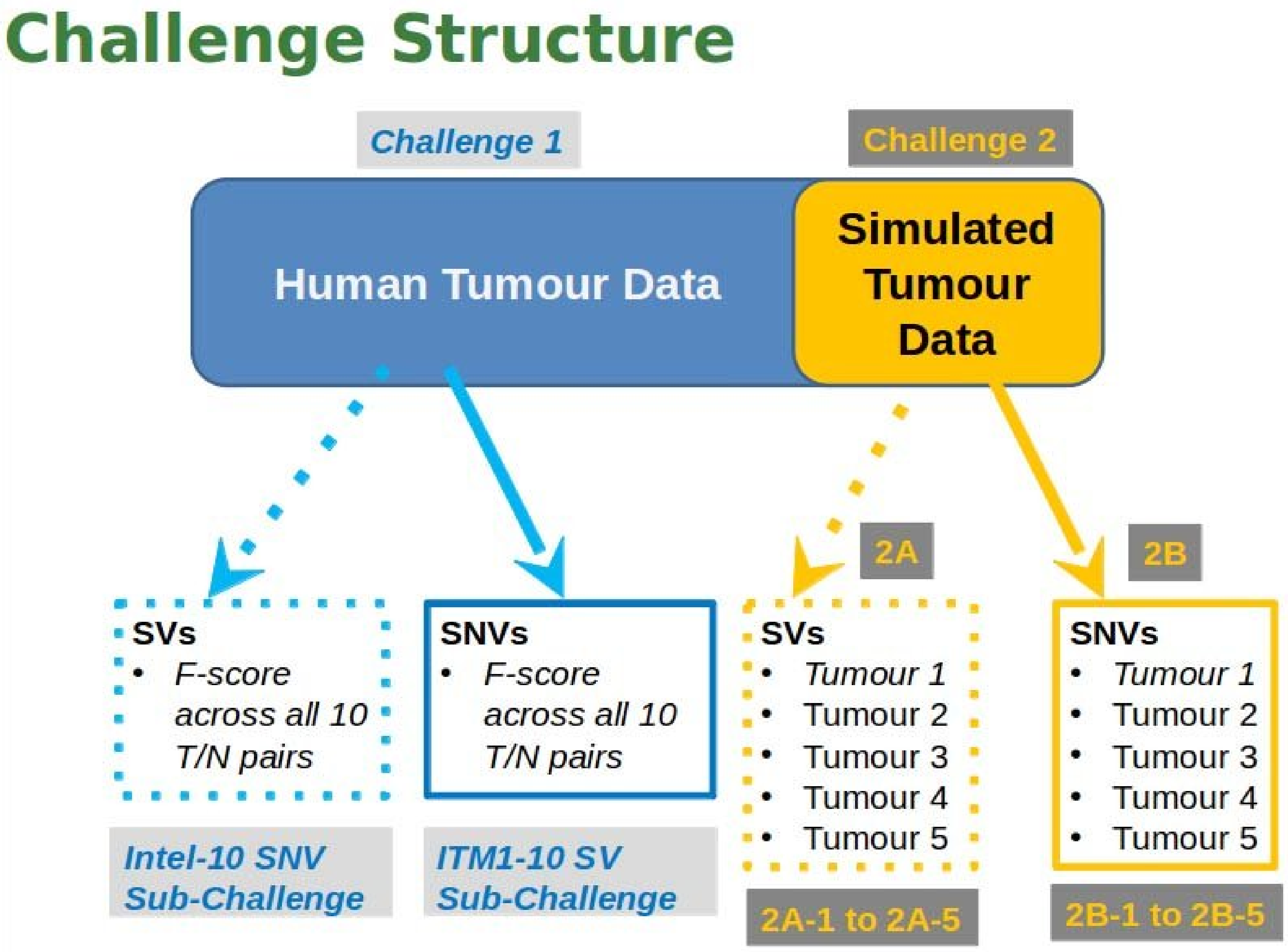
12 Cibulskis, K. *et al.* Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol.* **31**, 213-219, doi:10.1038/nbt.2514 (2013).

13 Talwalkar, A. *et al.* SMaSH: a benchmarking toolkit for human genome variant calling. *Bioinformatics* **30**, 2787-2795, doi:10.1093/bioinformatics/btu345 (2014).

14 Sherry, S. T. *et al.* dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* **29**, 308-311 (2001).

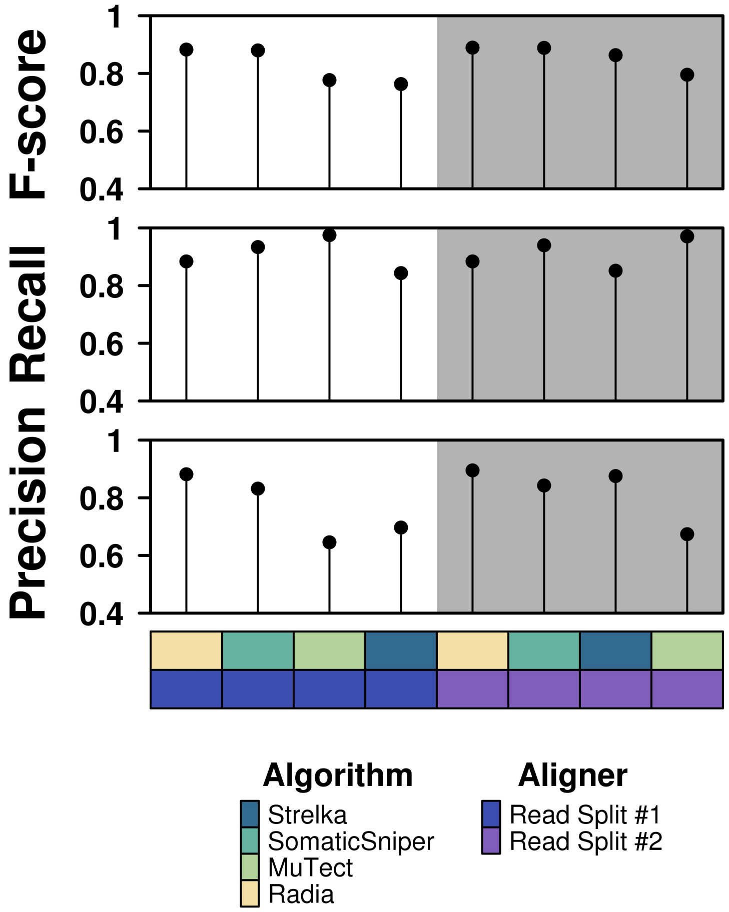
15 Forbes, S. A. *et al.* COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Res.*, doi:10.1093/nar/gku1075 (2014).

Supplementary Figure 1: Challenge Structure



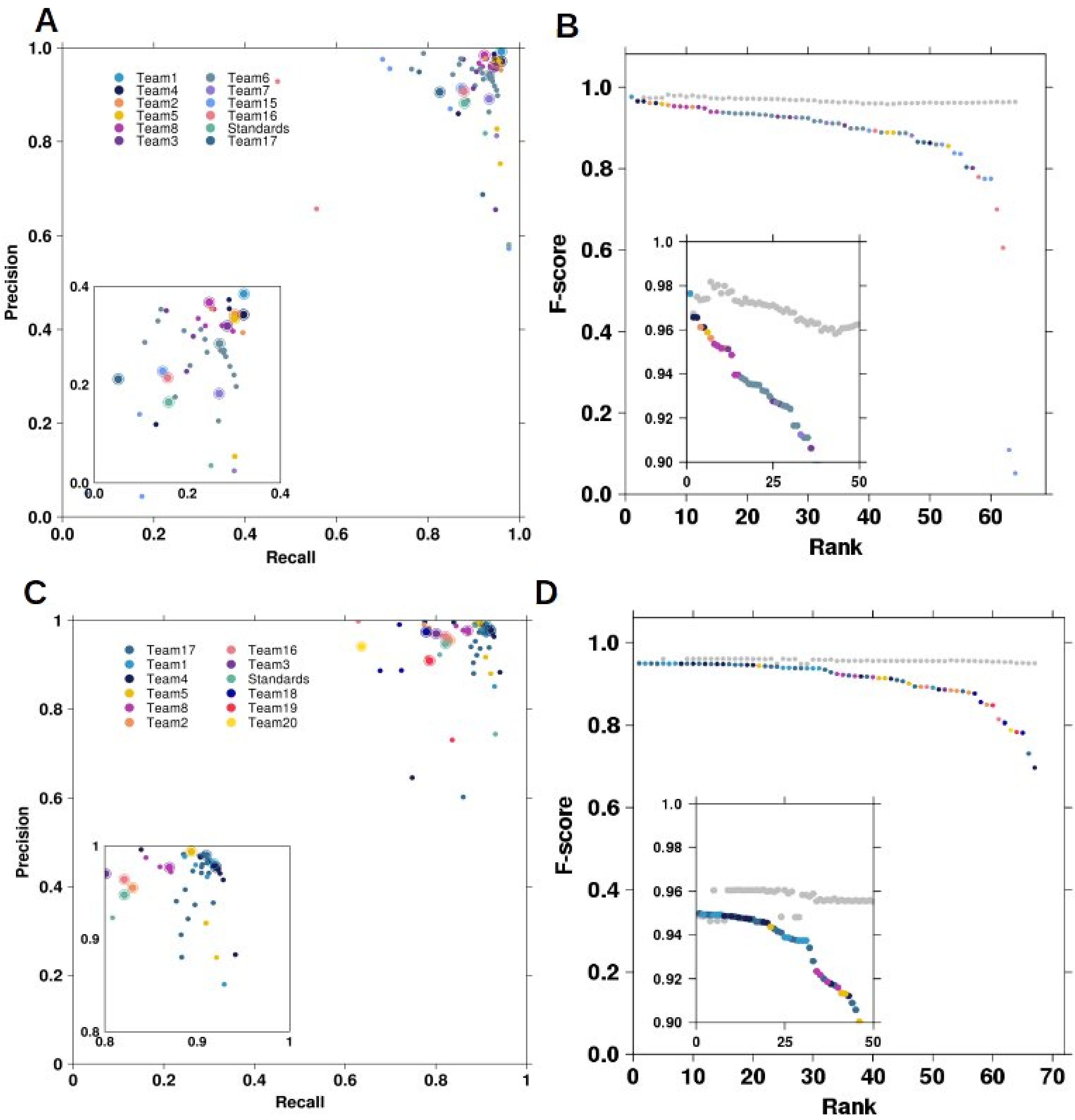
The SMC Challenge consists of two parts. The main challenge (Intel-10 SNV Sub-Challenge and ITM1-10 SV Sub-Challenge) consists of human tumour data derived from real patients - ten tumour/normal paired samples. The second challenge (2A-1 to 2A-5 and 2B-1 to 2B-5), proceeding the real data in timeline, consists of five synthetic datasets increasing in difficulty to allow participants to train their tools prior to the real tumour challenge.

Supplementary Figure 2: Effect of Read Split on Tool Performance



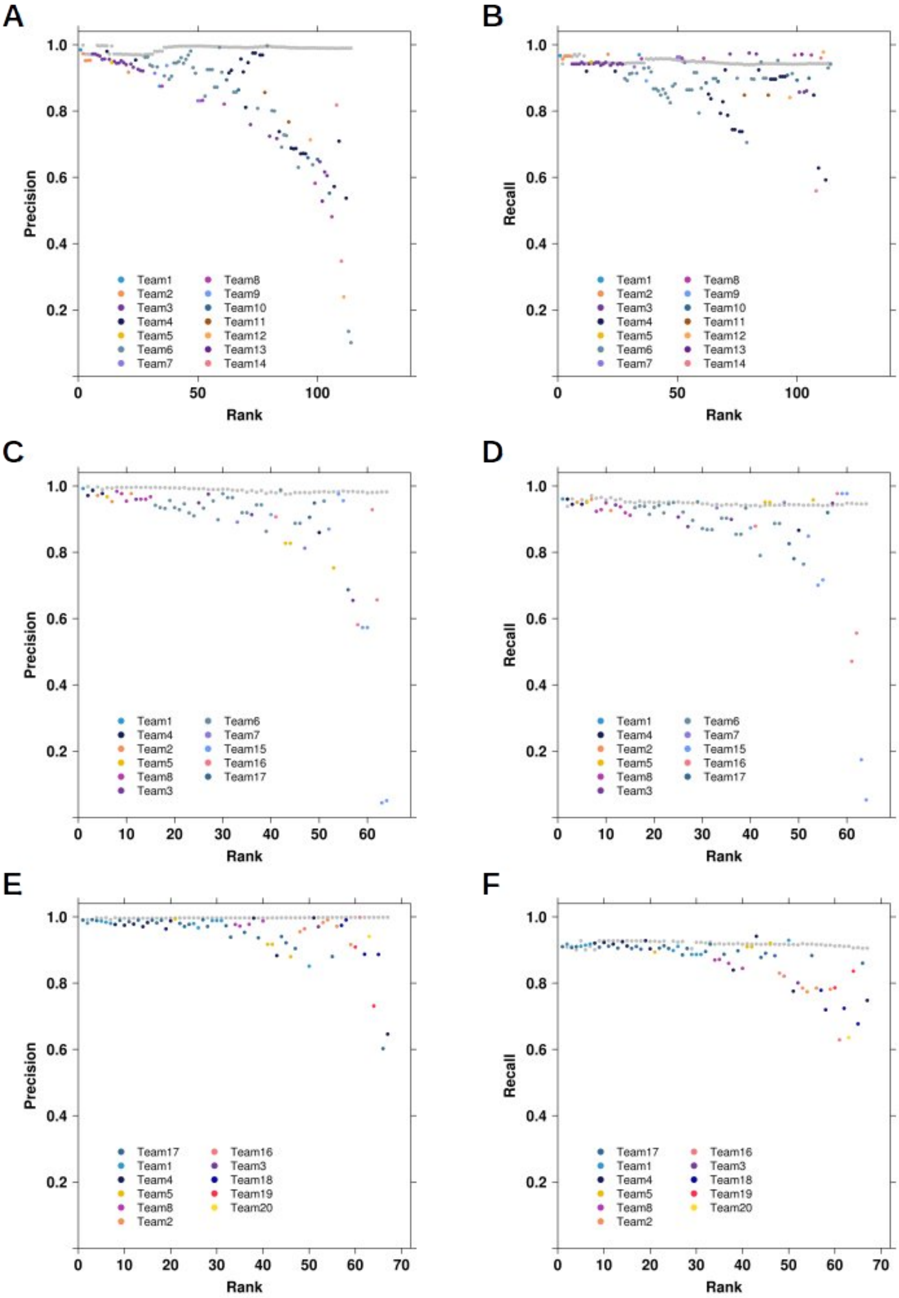
To test the robustness of BAMSurgeon to read split, we compared the rank of Radia (yellow), MuTect (dark blue), SomaticSniper (light blue) and Strelka (light green) on a tumour/normal paired dataset with alternate read splitting. Radia and SomaticSniper retained the top two positions while MuTect and Strelka remained third and fourth, regardless of read split.

Supplementary Figure 3: Overview of SMC-DNA *in silico* Challenges 2 and 3 Datasets



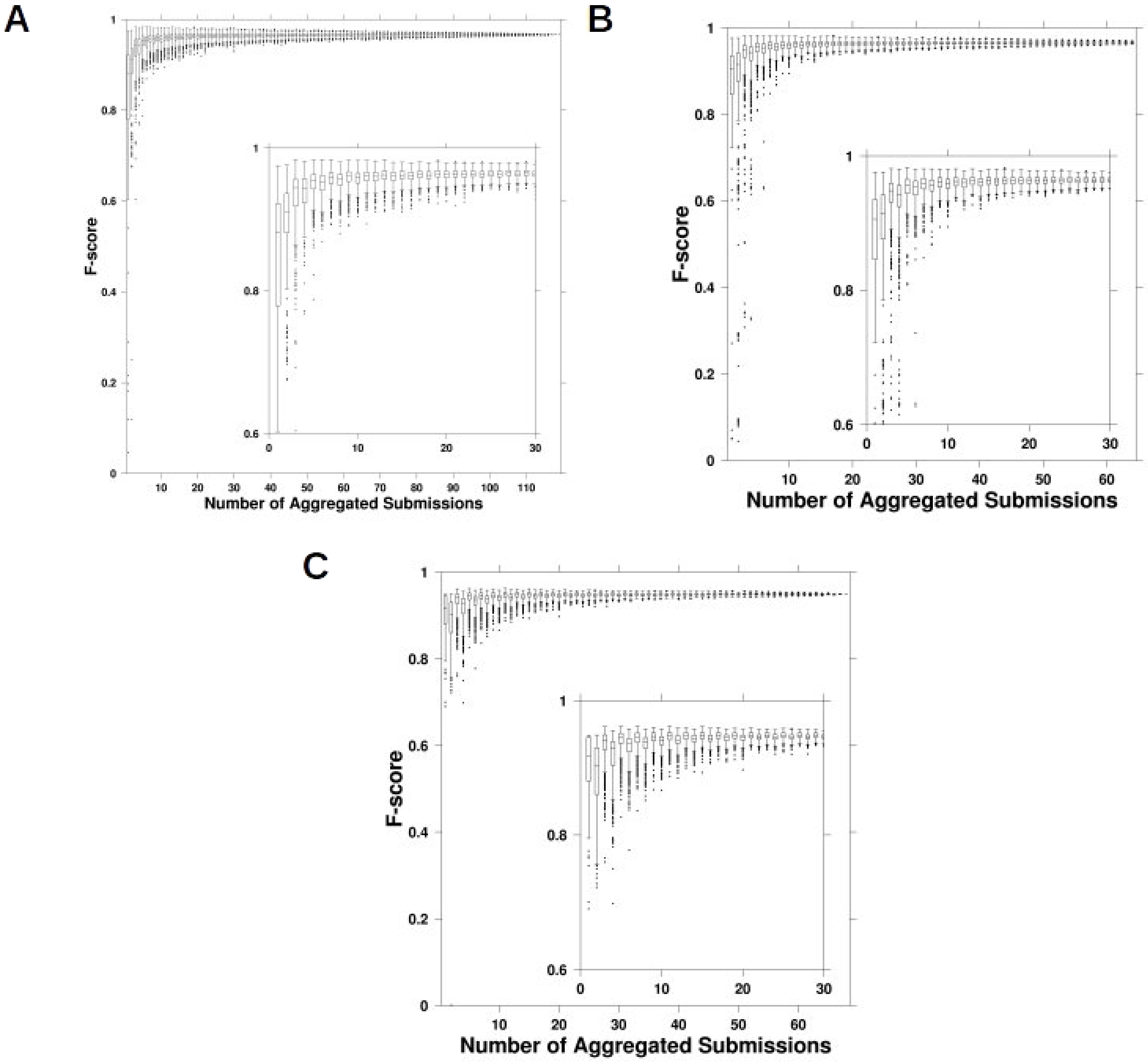
Precision-recall plot for all entries to IS2 - colours represent individual teams, and the best submission from each team by F-score is outlined **(A)**. We evaluated the performance of an ensemble somatic SNV predictor by taking the majority vote of calls made by a subset of the top performing submissions on IS2 and IS3. Ensemble models created and tested on IS2 – colours represent individual submissions while the gray dots represent the ensemble model **(B)**. Precision-recall plot for all entries to IS3 **(C)**. Ensemble models created and tested on IS3 **(D)**.

Supplementary Figure 4: Precision and Recall of Ensemble Classifier



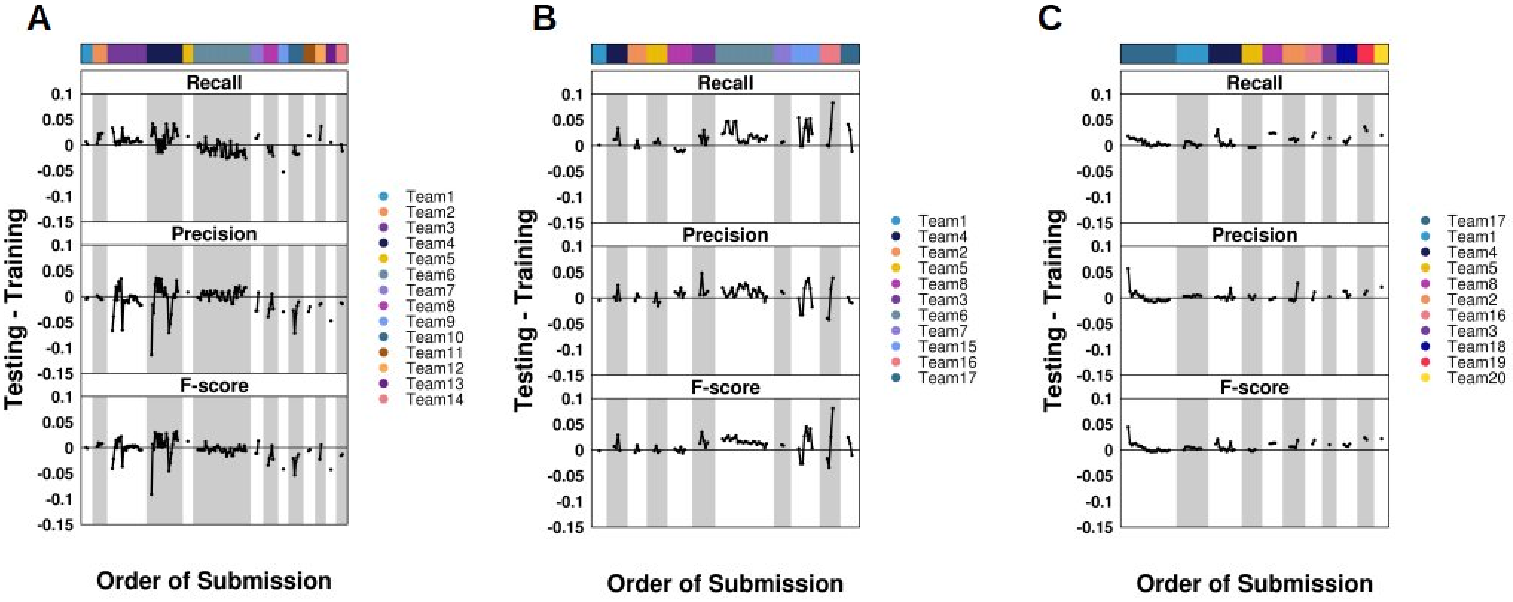
An ensemble classifier of subsets ranging from 1 to 119 algorithms selected from the IS1 submissions was developed taking calls with the majority vote across incorporated algorithms. The precision **(A)** and recall **(B)** of the ensemble classifier (grey) was compared to the values of the individual submissions (coloured). Dot colour reflects the submitting team. The ensemble classifier was found to have higher recall and precision than majority of the individual submissions. Similar plots are shown for IS2 precision **(C)** and recall **(D)**, as well as IS3 precision **(E)** and recall **(F)**.

Supplementary Figure 5: Permutation Analysis of Ensemble Robustness



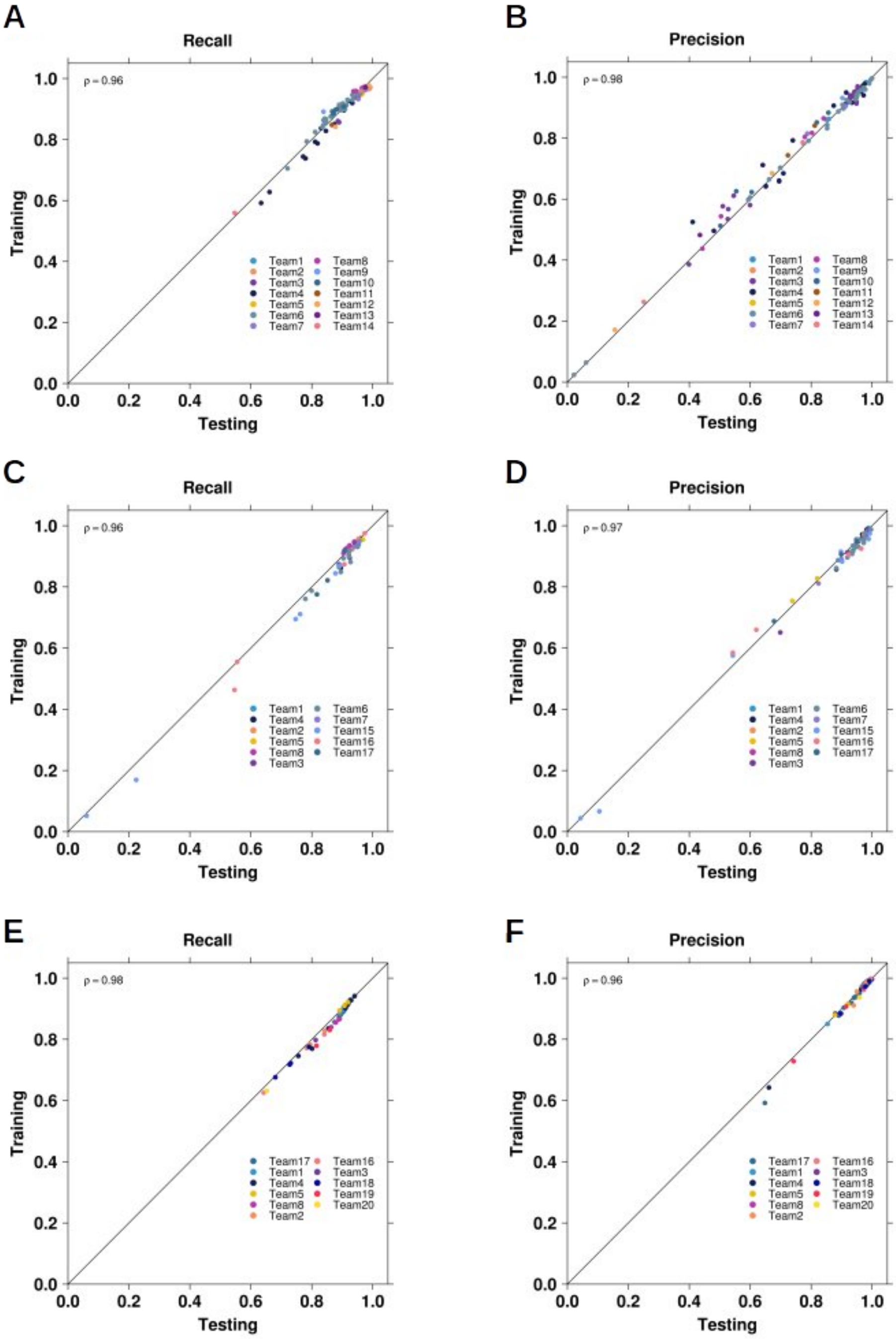
To evaluate the robustness of the IS1 ensemble classifier we randomly sampled algorithms at each subset size 1,000 times and evaluated performance **(A)**. The distribution of performance at each size threshold reflected the performance seen by subsetting the top scoring algorithms giving evidence for the robustness of the method. **(B)** IS2 and **(C)** IS3.

Supplementary Figure 6: Evaluation of Overfitting



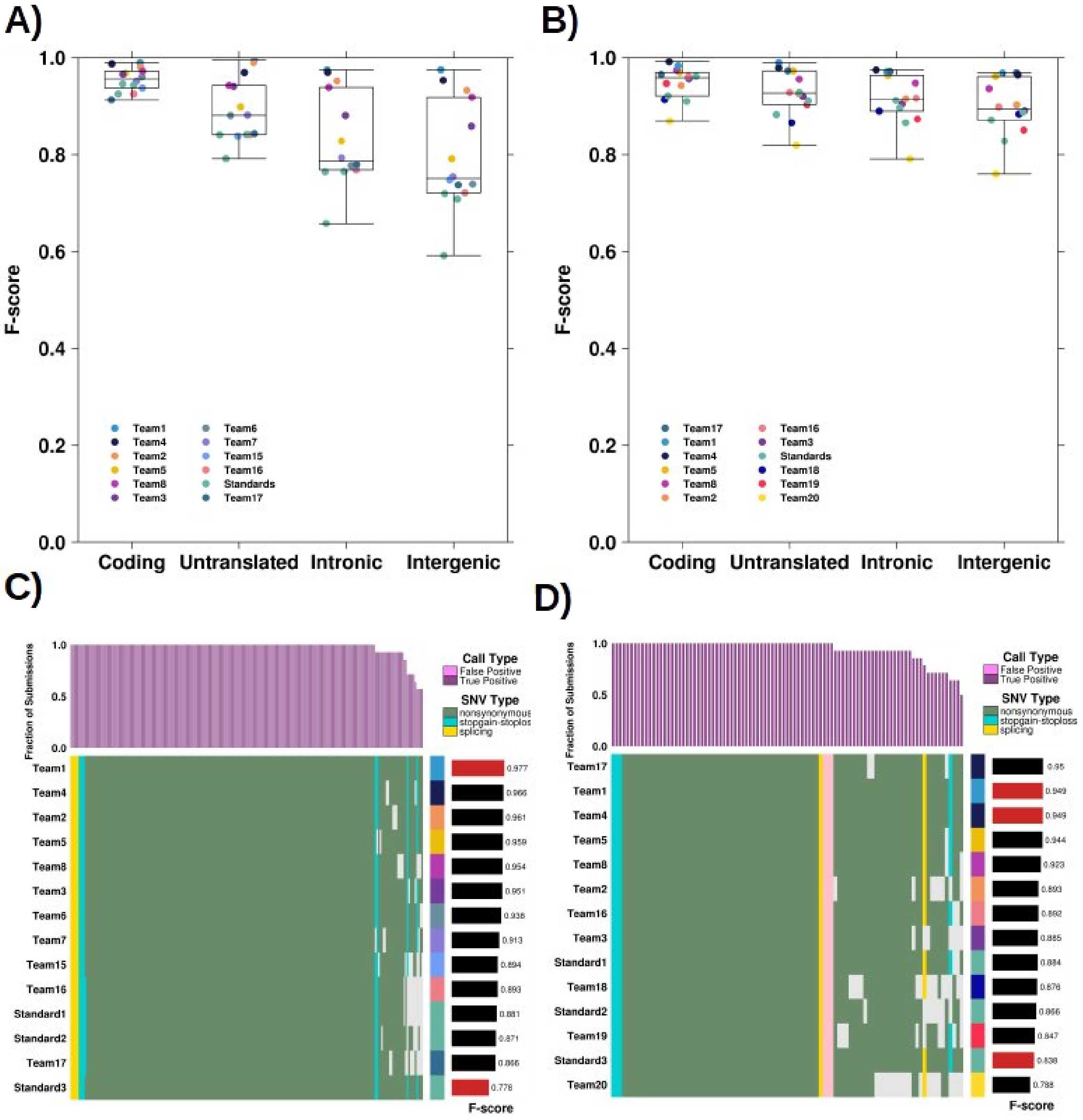
The delta between recall, precision and F-scores on training (all chromosomes but chromosome one) and testing (chromosome one only) datasets were plotted for each submission. All delta values varied around 0 and never exceed a difference greater than 0.15. This shows evidence of little overfitting. **(A)** IS1, **(B)** IS2 and **(C)** IS3.

Supplementary Figure 7: Correlation of Training and Testing Scores



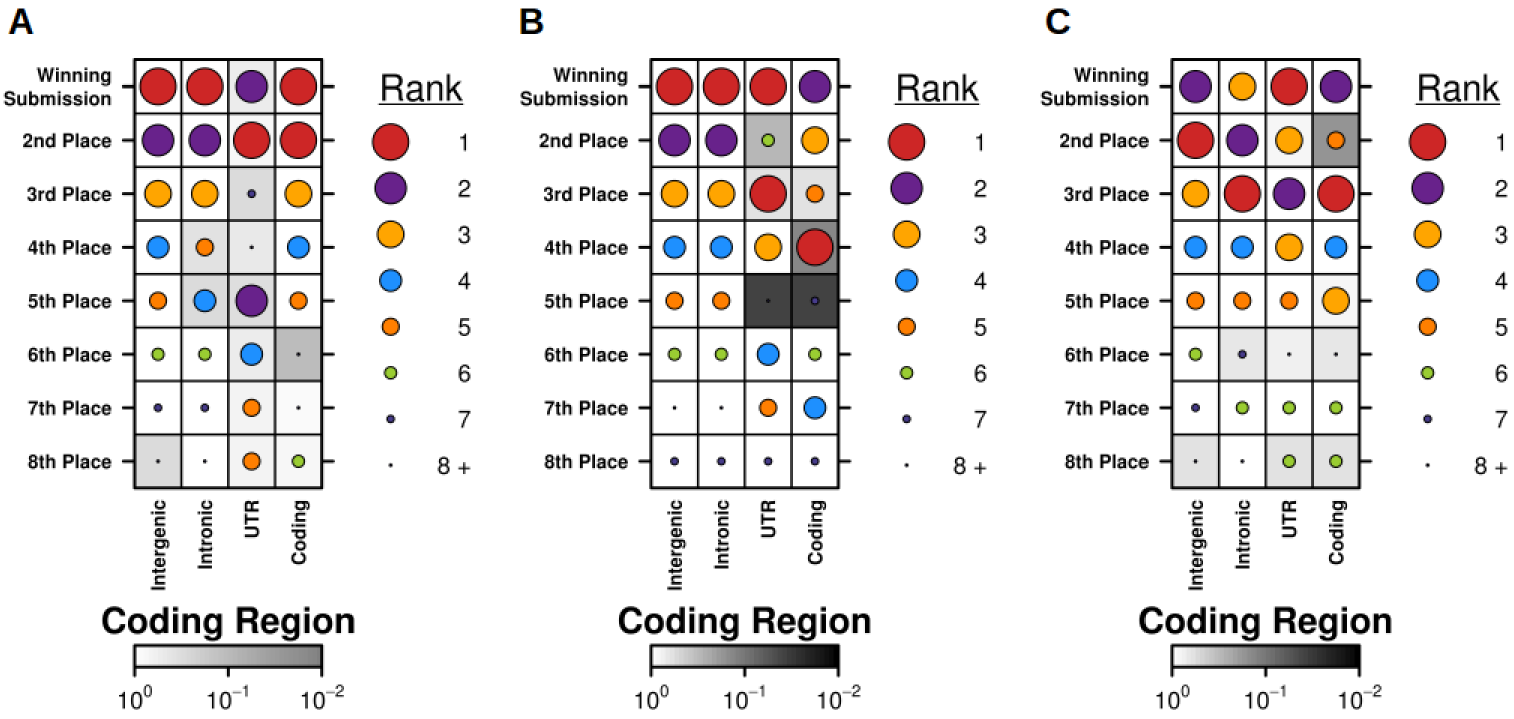
Scatterplots were created showing the relationship between training and testing recall **(A)** and precision **(B)** for each IS1 submission compared to the y=x line. Both showed high degree of correlation with Spearman correlation values of 0.96 and 0.98. This is further evidence of little overfitting. Colours indicate the submission team. Similar results are observed for IS2 recall **(C)** and precision **(D)** and IS3 recall **(E)** and precision **(F)**.

Supplementary Figure 8: Effects of Genomic Localization on *in silico* 2 and 3 Datasets



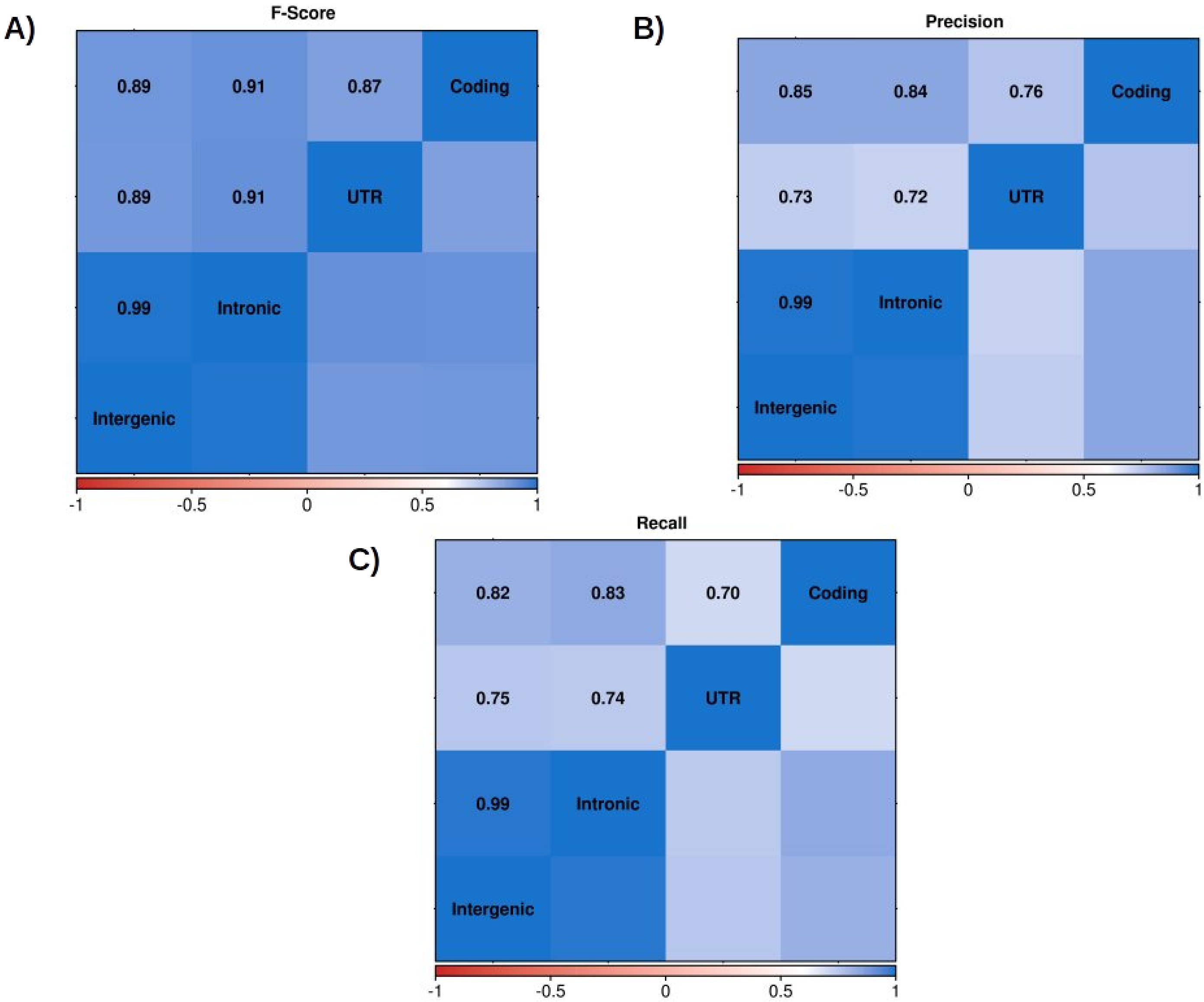
Boxplot gives median (line), inter-quartile range (box) and ± 1.5 IQR. F-scores were highest in coding and untranslated regions (UTR) and lowest in introns and intergenic in both (**A**) IS2 (*P* = 3.68 x 10^{-8}; Friedman Rank Sum Test) and (**B**) IS3 (*P* = 1.96 x 10^{-5}; Friedman Rank Sum Test). Dot colours represent individual teams. Rows show individual submissions to IS2 (**C**) and IS3 (**D**), columns show genes with non-synonymous SNV calls. The upper barplot indicates the fraction of submissions agreeing on these calls, and the colour indicates if these are FPs (light purple) or true-positives (dark purple). The barplot located to the right gives the F-score of the submission over the whole genome. The right-hand side covariate shows the submitting team.

Supplementary Figure 9: Rank in Genomic Elements



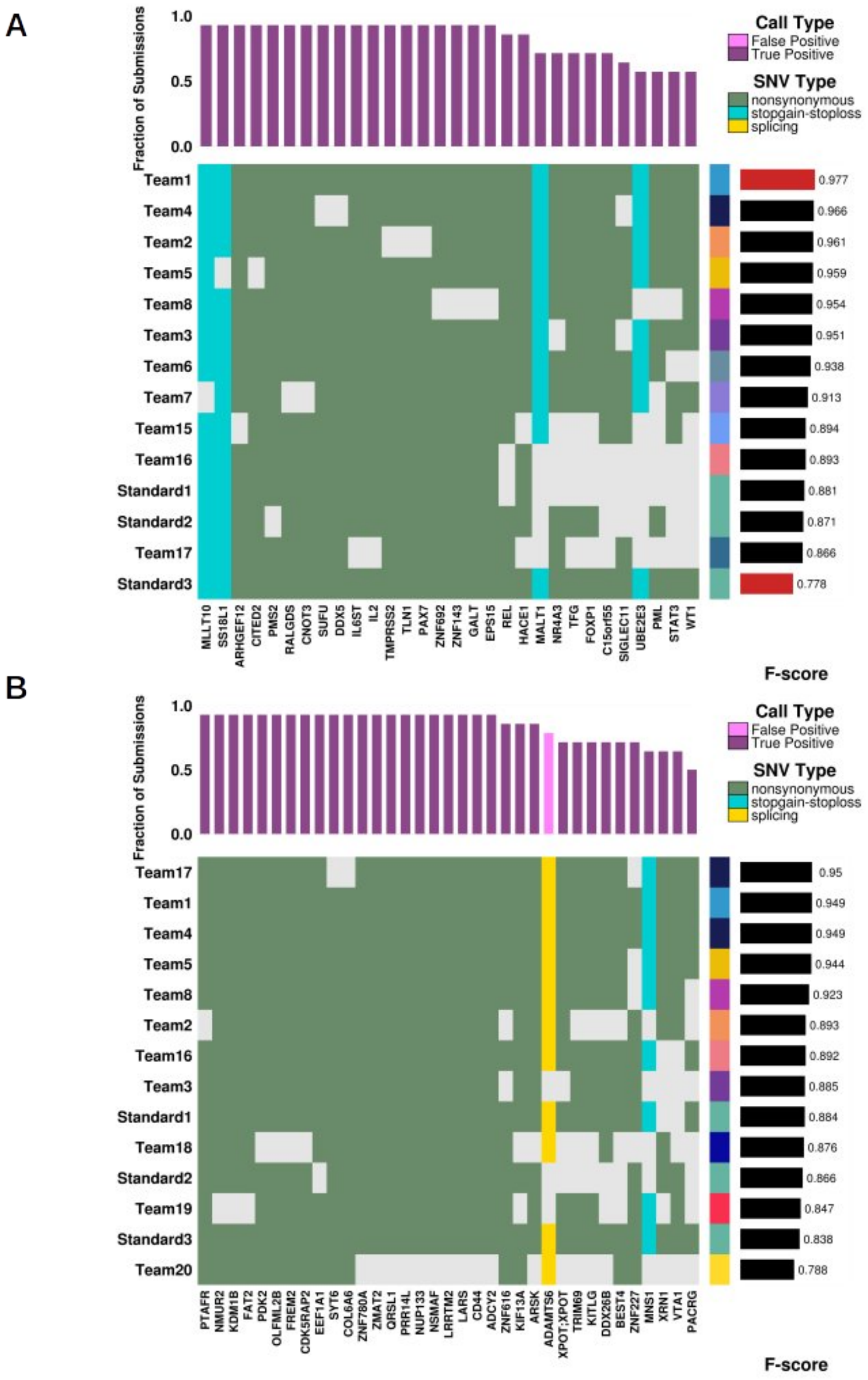
The rank of each submission on each genomic element was compared to the submission's overall rank. Dot size and colour reflect the rank, as determined by the F-score, of that submission in that genomic element. Ranks in intergenic, intronic and coding show a high degree of consistency, however, more variation is seen in untranslated regions (UTR). Background shading reflects significance of variation as compared to chance alone. **(A)** IS1, **(B)** IS2 and **(C)** IS3.

Supplementary Figure 10: F-score, Recall and Precision Correlation in Genomic Elements



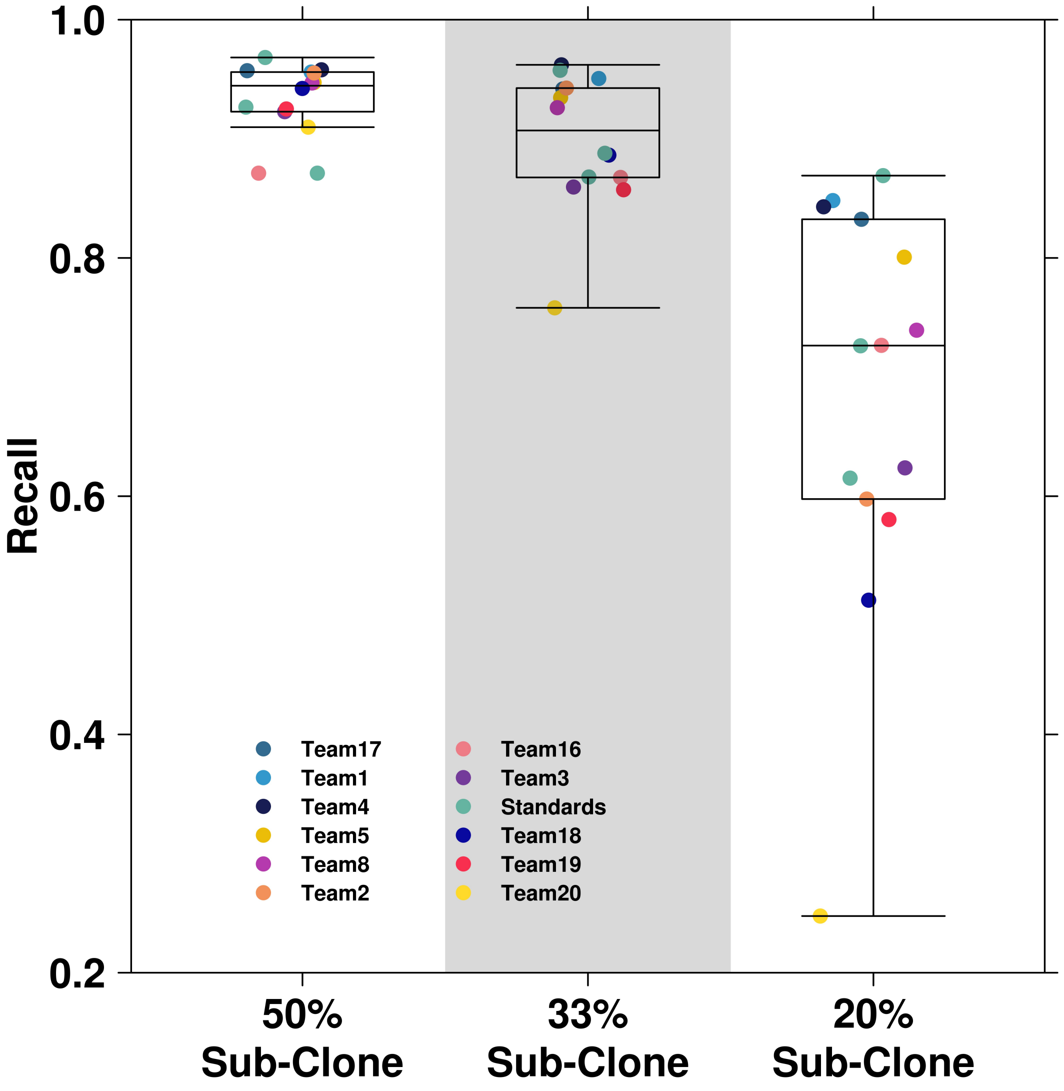
Heatmaps show the Spearman correlation of F-scores **(A)**, precision **(B)** and recall **(C)** between genomic elements in IS1. Intergenic and intronic regions show high correlation in all three scores.

Supplementary Figure 11: Prediction in Exonic Regions of All Submissions



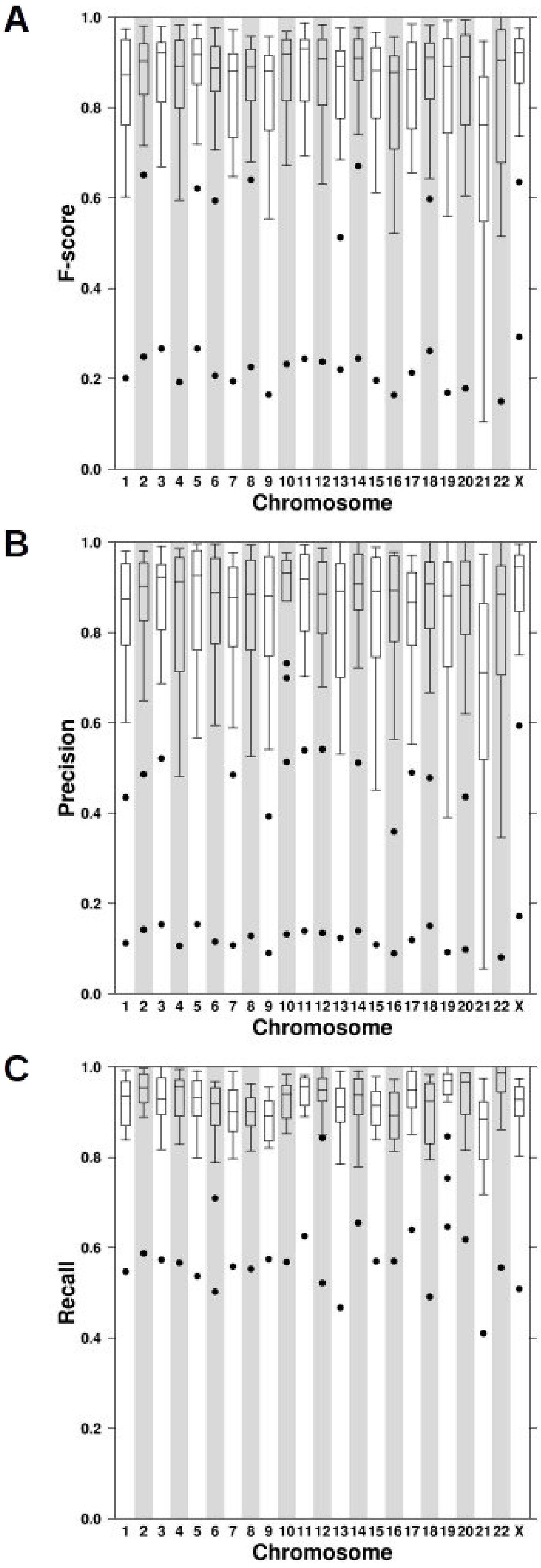
SNV calls in exonic regions corresponding to known genes (x-axis) were plotted to show the number of submissions that called each SNV (highlighted in green) - calls made by the lowest scoring submission only were omitted from this subset. The barplot along the top indicates the fraction of submissions that called each position, while the colour indicates whether the position is a true positive (dark purple) or a false positive (light pink). The covariate along the right of the plot indicates the team while the barplot on the right shows the overall F-score of that submission. **(A)** IS1, **(B)** IS2 and **(C)** IS3.

Supplementary Figure 12: Recall in Sub-Clones



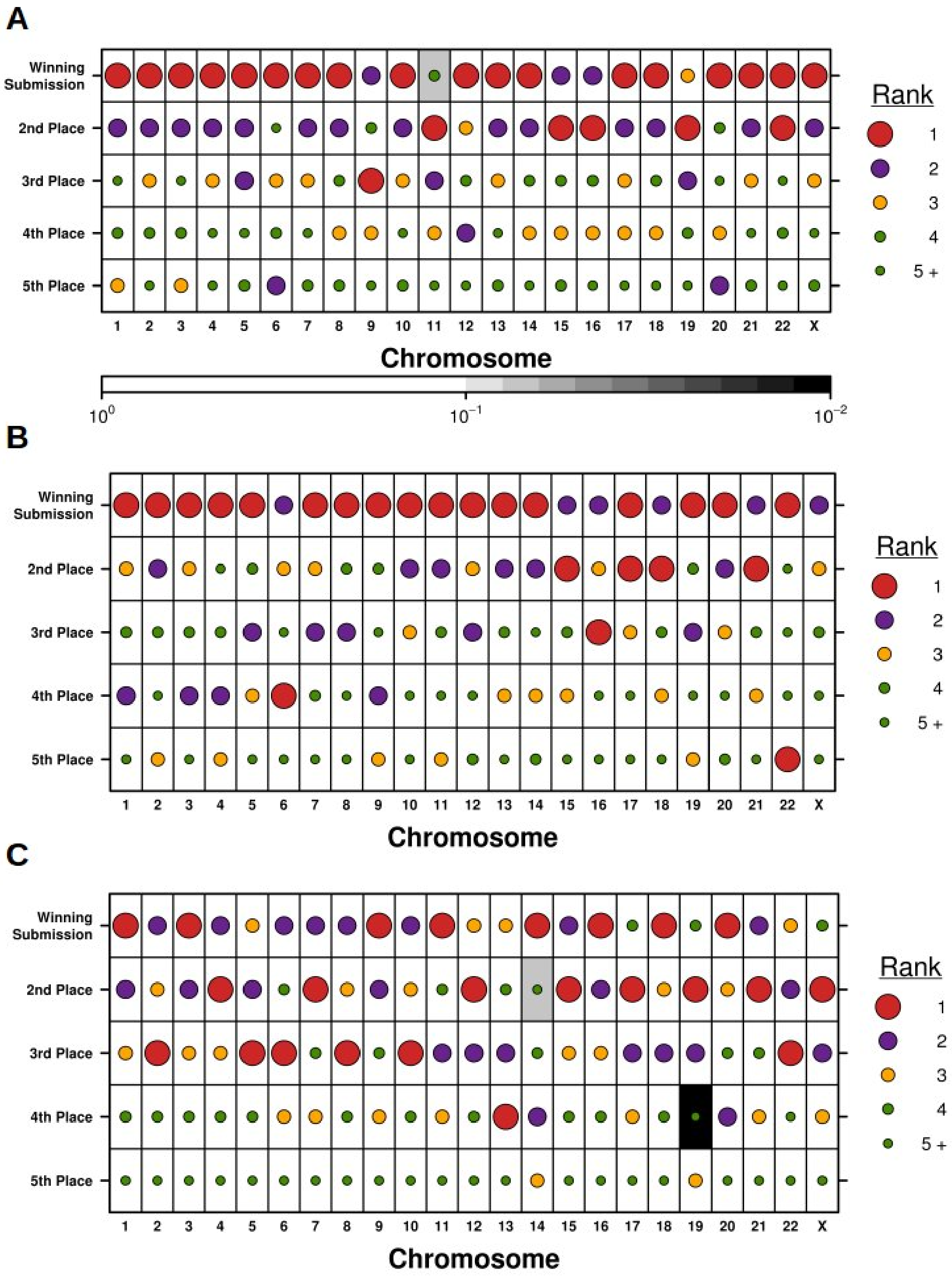
Boxplot gives median (line), inter-quartile range (box) and ± 1.5 IQR for recall of submission in 50%, 33% and 20% sub-clones within the IS3 dataset. Dots represent the individual recall scores for each submission. Submissions showed higher recall in 50% sub-clones.

Supplementary Figure 13: Chromosome Bias in Precision and Recall for IS1



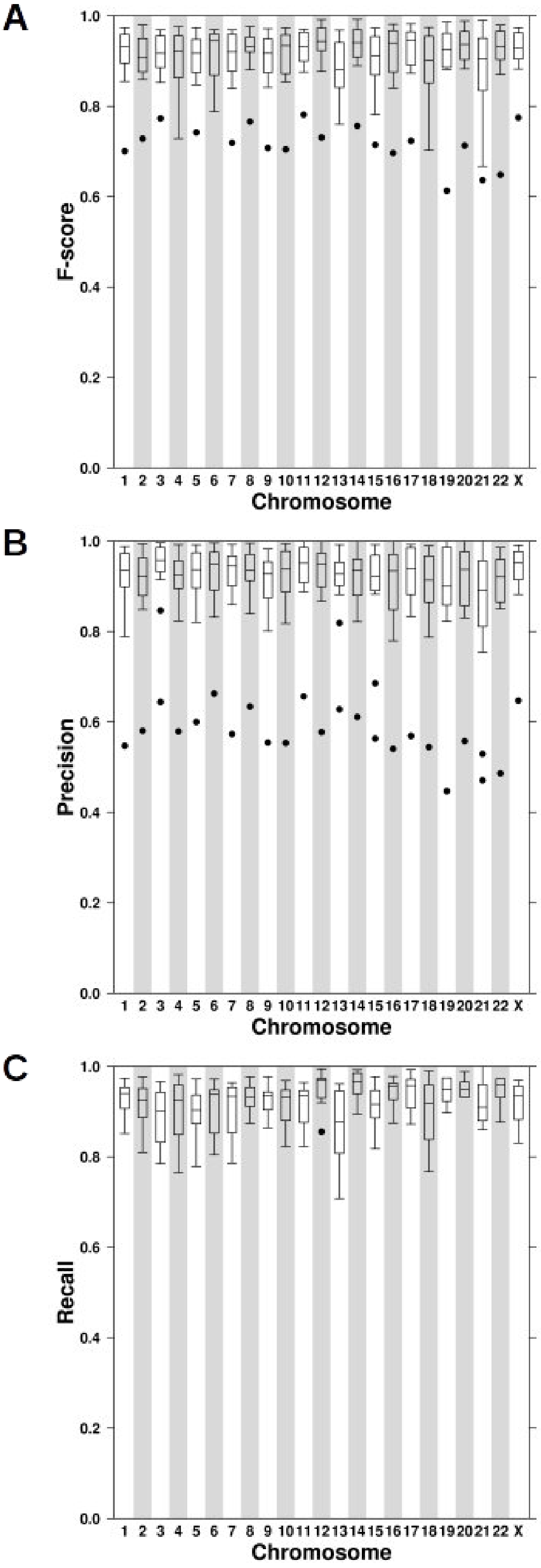
Boxplots show the distribution of F-Score **(A)**, precision **(B)** and recall **(C)** on each chromosome of IS1. A noticeable decrease in both precision and recall was seen on chromosome 21.

Supplementary Figure 14: Variation in Chromosome Rank



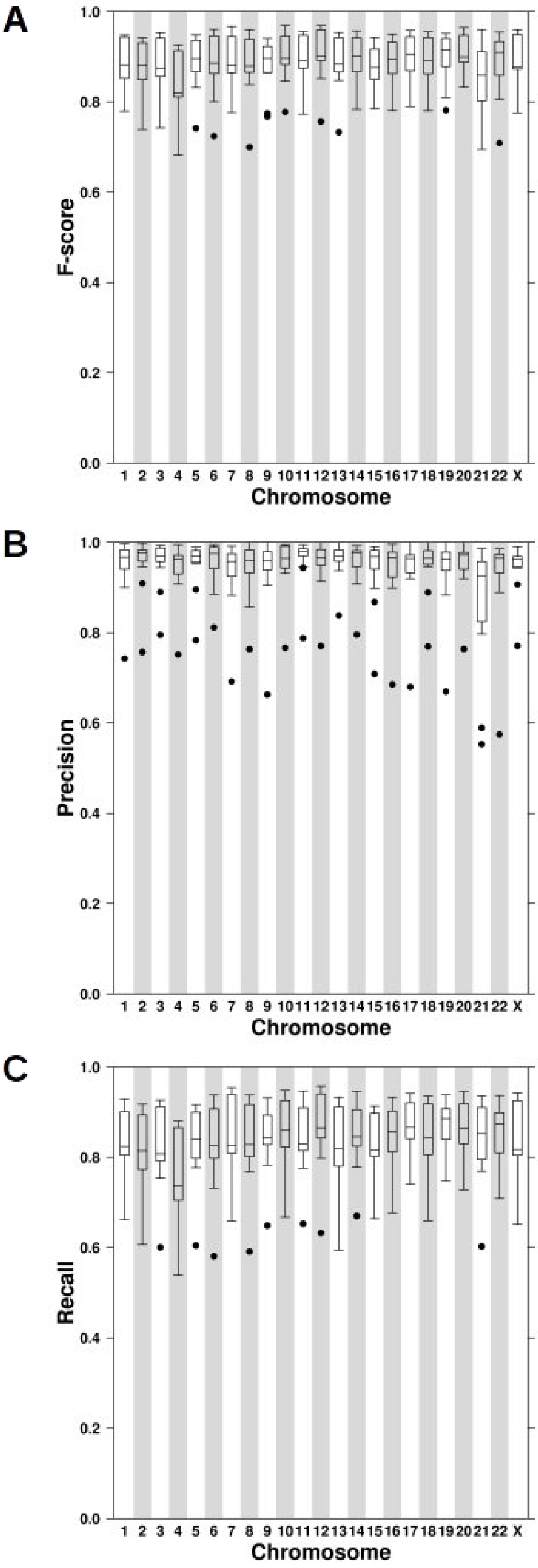
Variation in chromosome rank (indicated by dot size and colour) and overall rank was observed, most noticeably on chromosome 11, where the best scoring algorithm ranked fourth. P-values reflecting the probability of this variation being seen by chance alone are shown submission and chromosome-wise in the background shading, chromosome-wise in the top barplot, and submission-wise in the right hand side barplot. **(A)** IS1, **(B)** IS2 and **(C)** IS3.

Supplementary Figure 15: Chromosome Bias in Precision and Recall for IS2



Boxplots show the distribution of F-Score **(A)**, precision **(B)** and recall **(C)** on each chromosome of IS2.

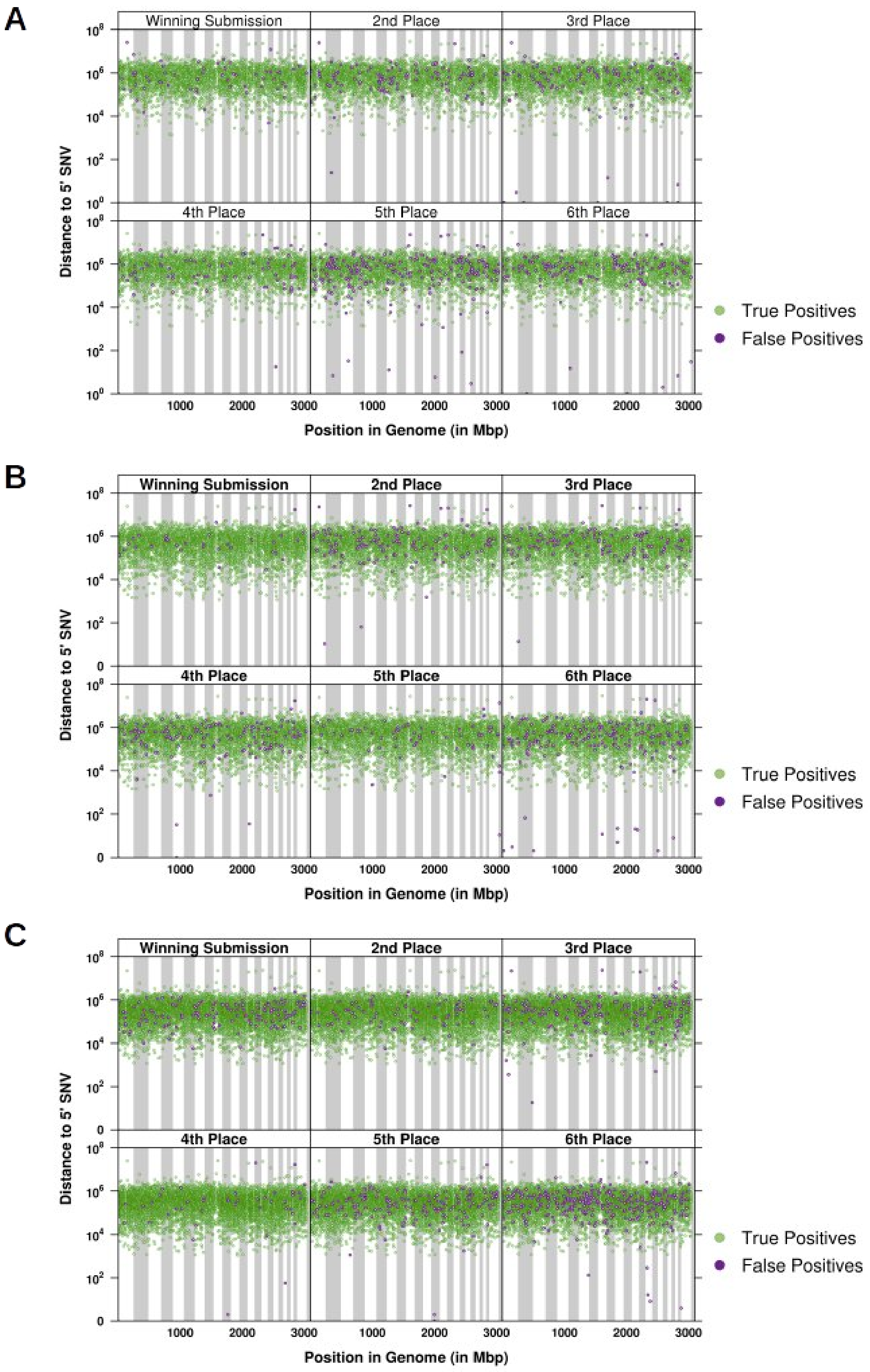
Supplementary Figure 16: Chromosome Bias in Precision and Recall for IS3



Boxplots show the distribution of F-Score **(A)**, precision **(B)** and recall **(C)** on each chromosome of IS3.

Nature Methods: doi:10.1038/nmeth.3407

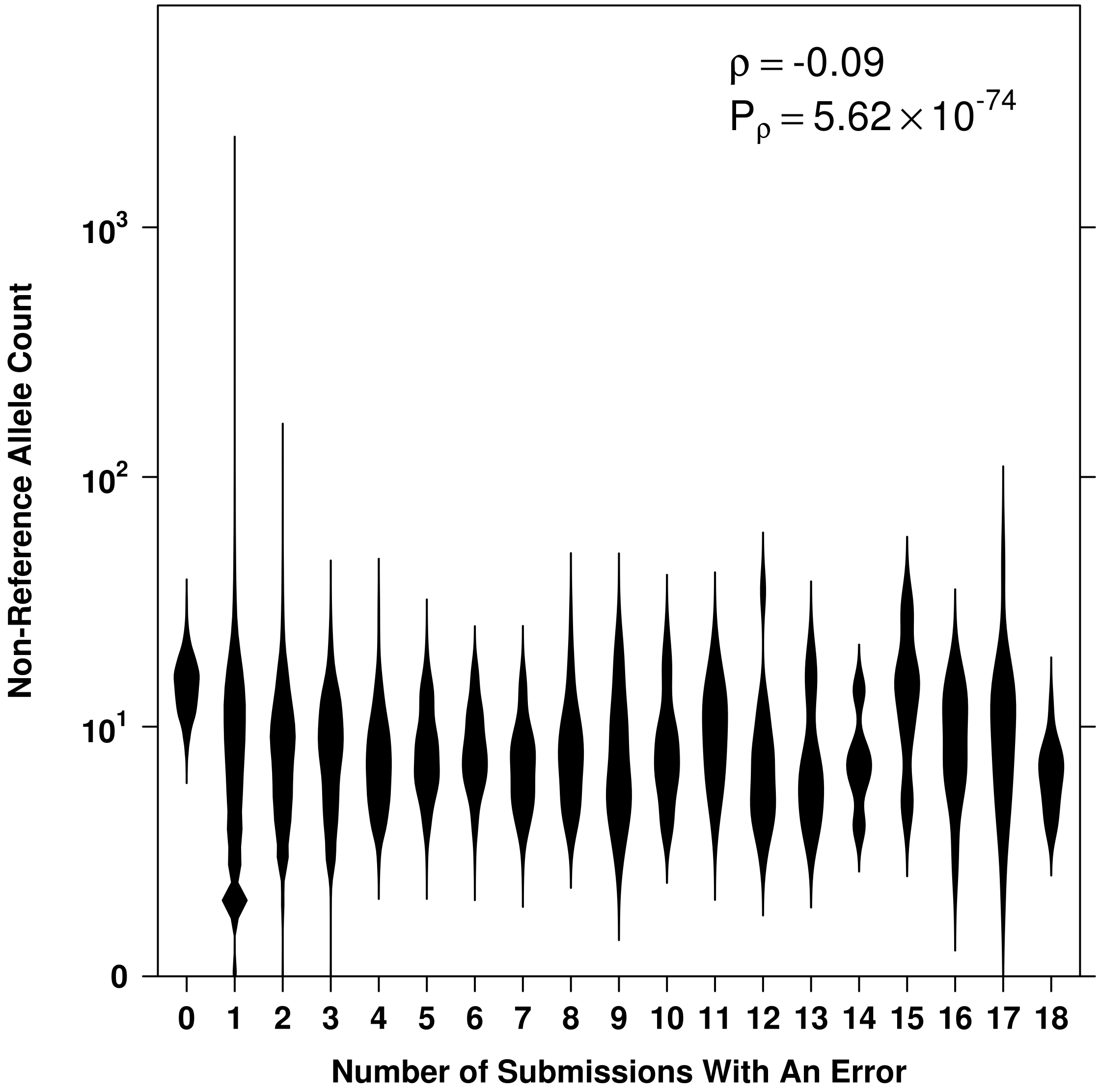
Supplementary Figure 17: Distribution of Calls in Genome



The genomic location of each call made by the top 6 algorithms was plotted against the distance of the the call to the closest 5’ SNV. True positives were plotted in green while false positives were plotted in purple. The lack of points clustering towards the bottom of each plot is evidence that kataegis is not occurring in SNV prediction as the calls appear to spread out through the genome. **(A)** IS1, **(B)** IS2, **(C)** IS3.

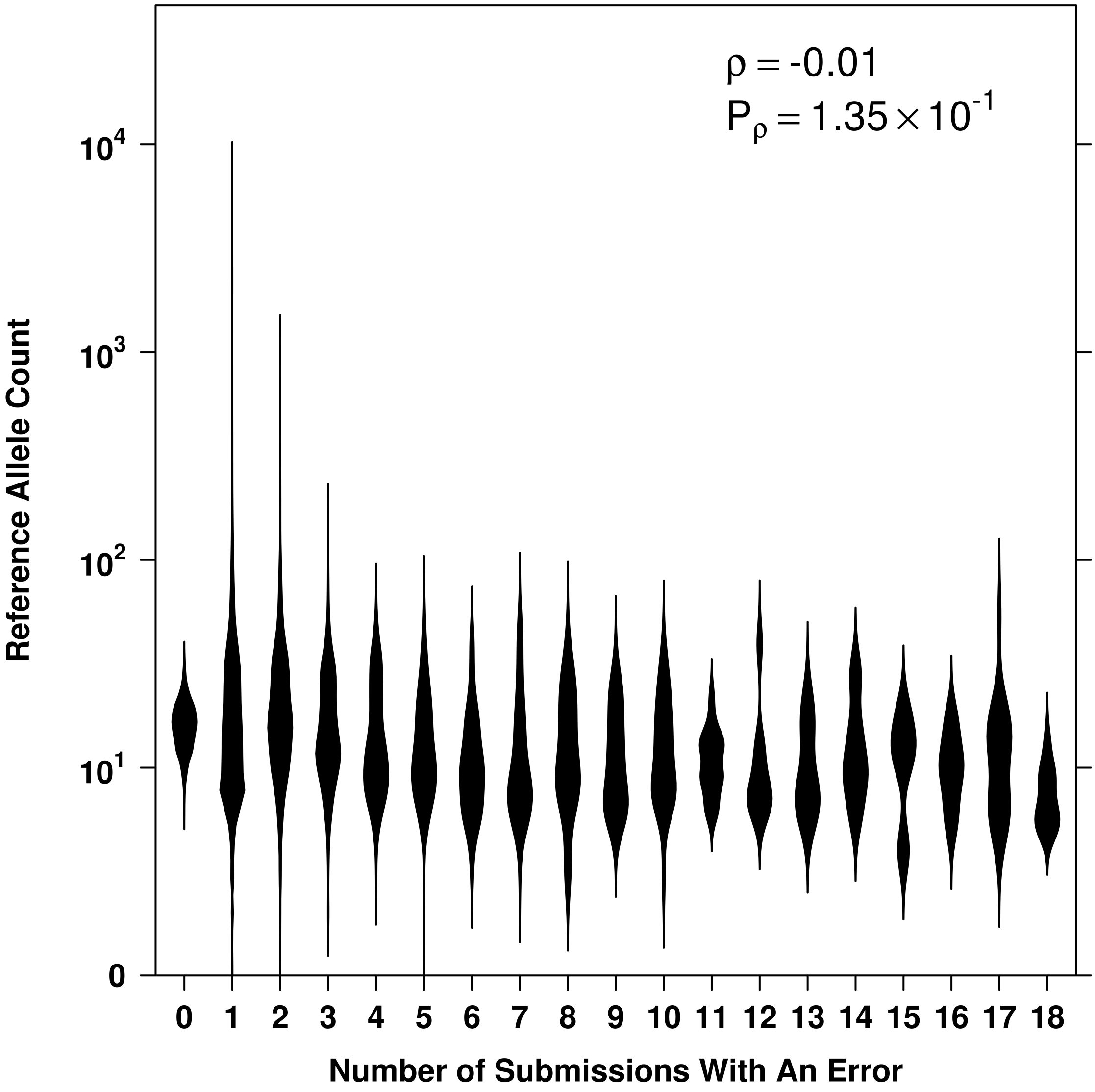
Nature Methods: doi:10.1038/nmeth.3407

Supplementary Figure 18: Univariate Analysis of Genomic Factors – Non-Reference Allele Count



Thirteen genomic variables - non-reference allele count **(18)**, reference allele count **(19),** base quality **(20)**, mapping quality **(21)**, tumour coverage **(22)**, normal coverage **(23)**, distance to nearest germline SNP **(24)**, homopolymer rate **(25)**, GC content **(26)**, read position **(27)**, trinucleotide sequence **(28)** and genomic element **(29)** - were selected to analyze their effect on the number of submissions that made an error at each position for IS1. Violin plots show the relationship between continuous variables and number of submissions, while heatmaps show the relationship between categorical variables and number of submissions. Spearman correlation and corresponding p-values were calculated for continuous variables, while, one-way ANOVAs were run on categorical variables.

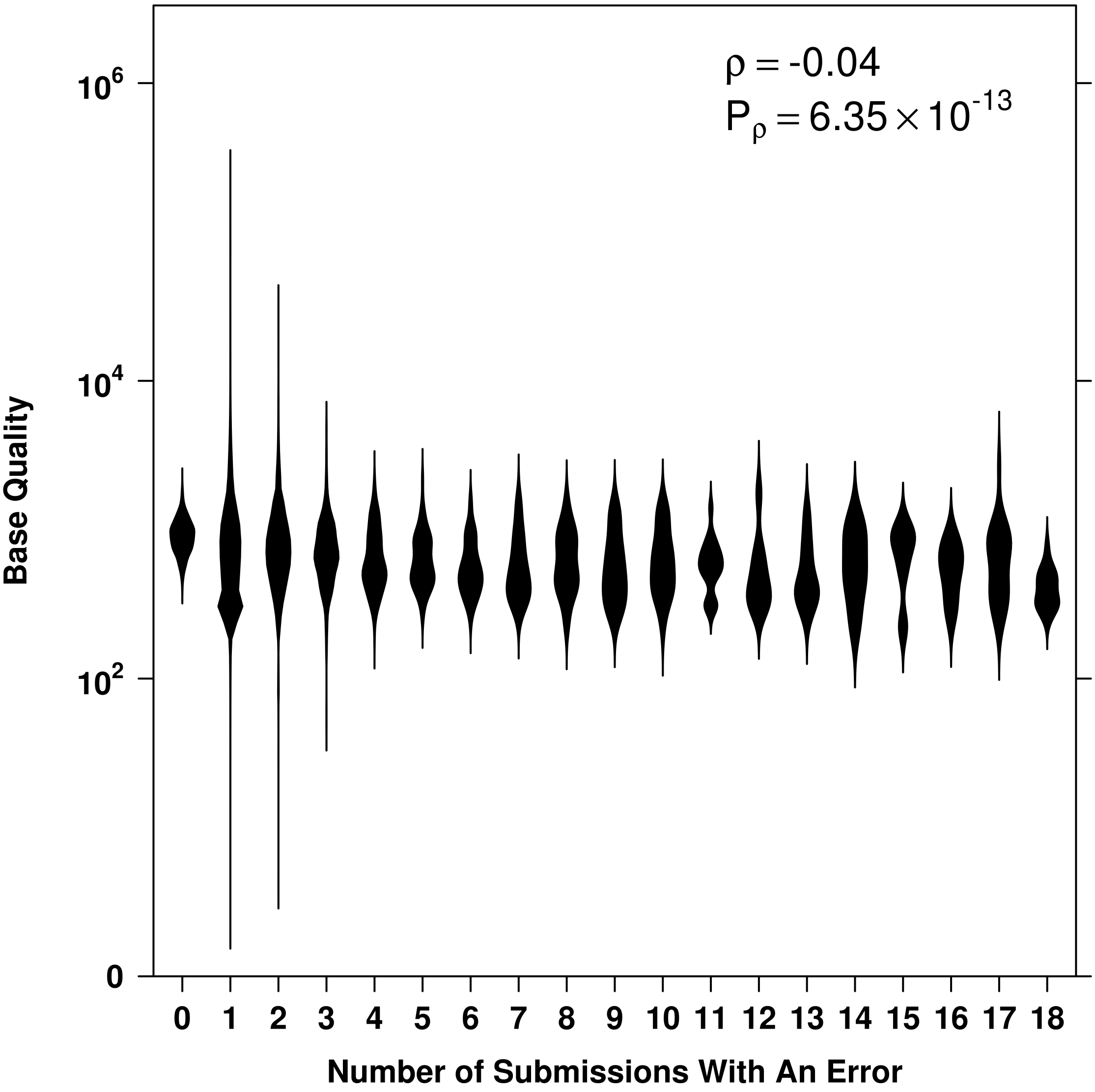
Supplementary Figure 19: Univariate Analysis of Genomic Factors – Reference Allele Count



Violin plot shows the relationship between reference allele count and number of submissions. Spearman correlation and corresponding p-values were calculated.

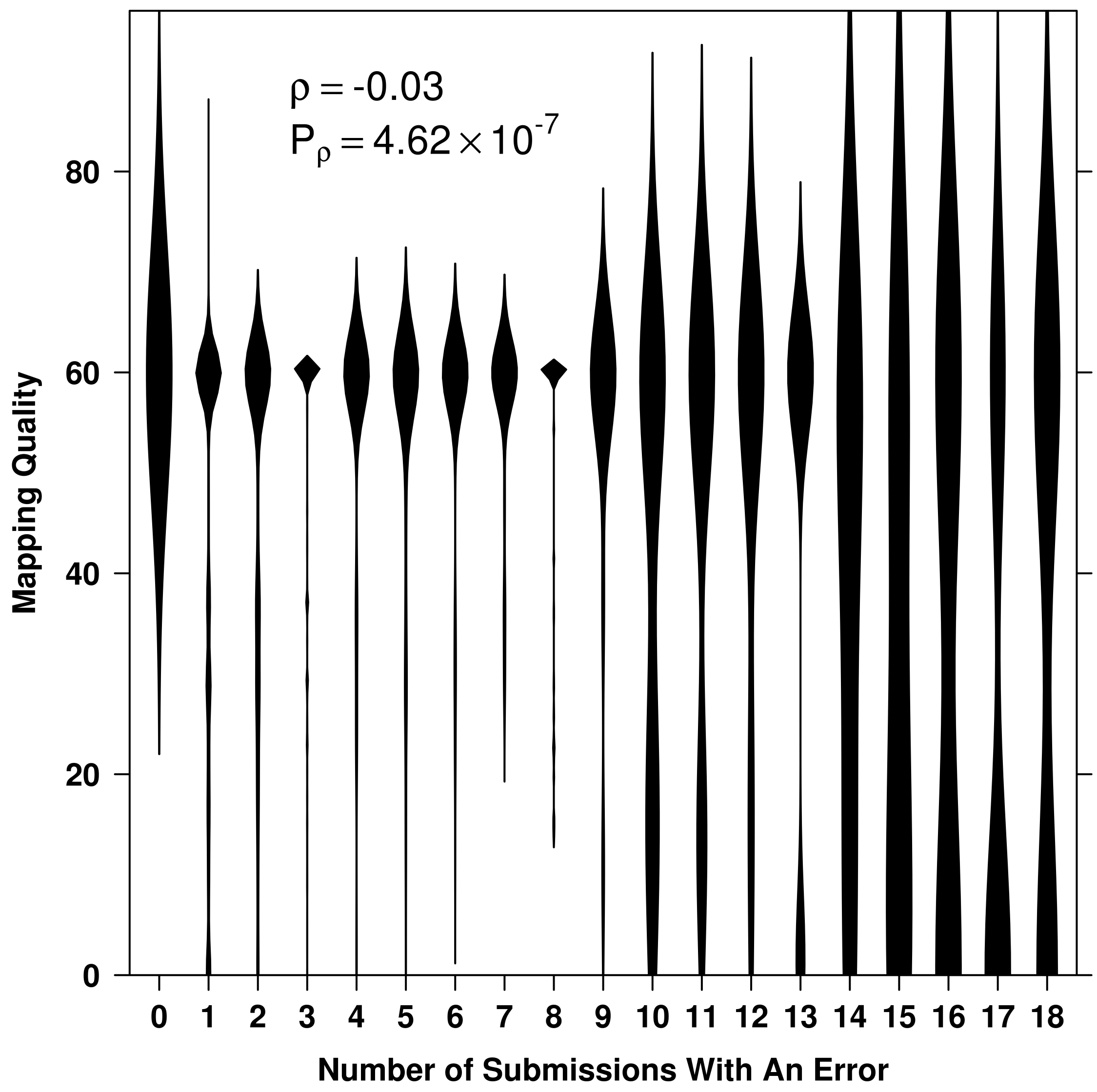
Nature Methods: doi:10.1038/nmeth.3407

Supplementary Figure 20: Univariate Analysis of Genomic Factors – Base Quality



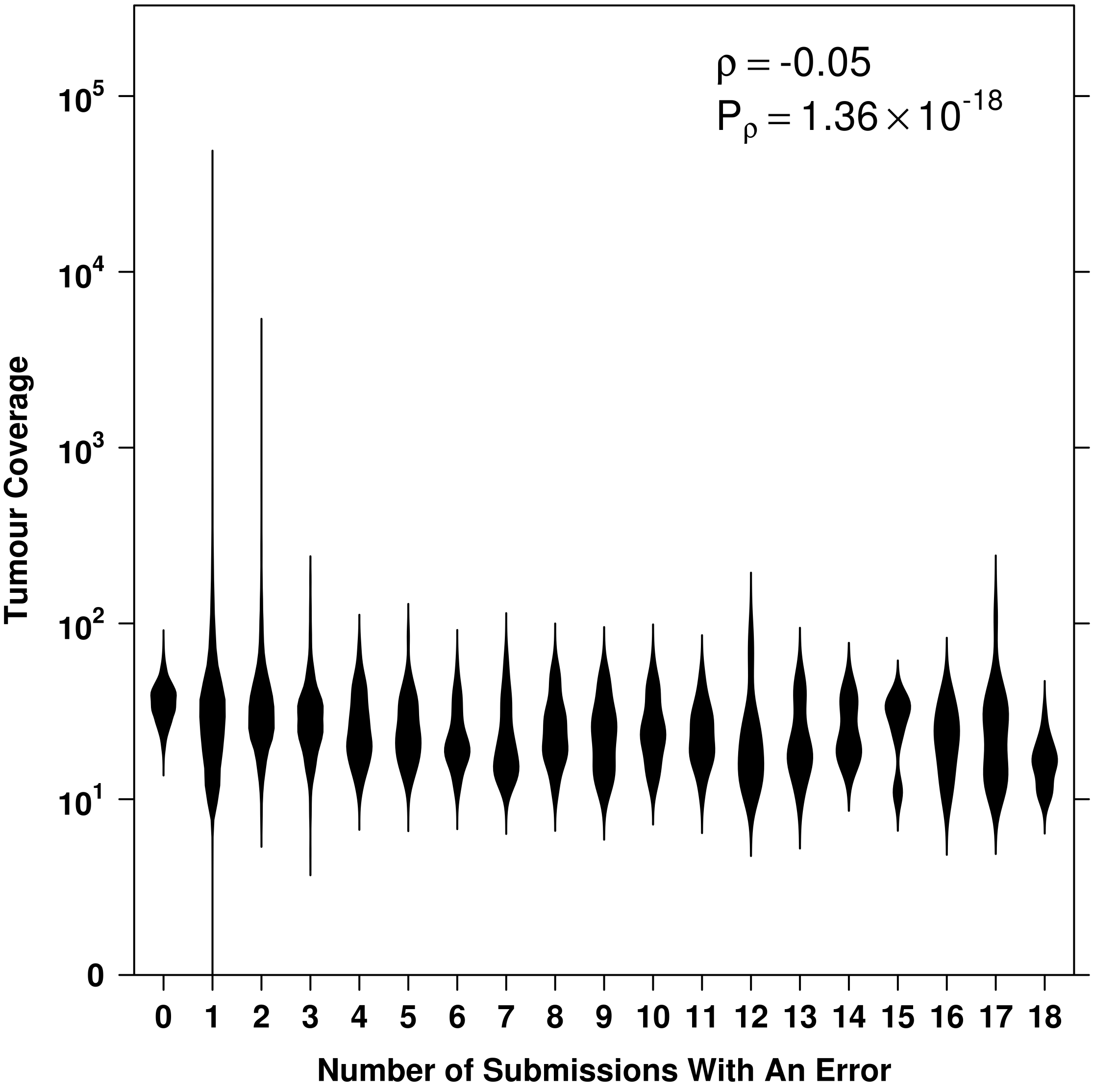
Violin plot shows the relationship between base quality and number of submissions. Spearman correlation and corresponding p-values were calculated.

Supplementary Figure 21: Univariate Analysis of Genomic Factors – Mapping Quality



Violin plot shows the relationship between mapping quality and number of submissions. Spearman correlation and corresponding p-values were calculated.

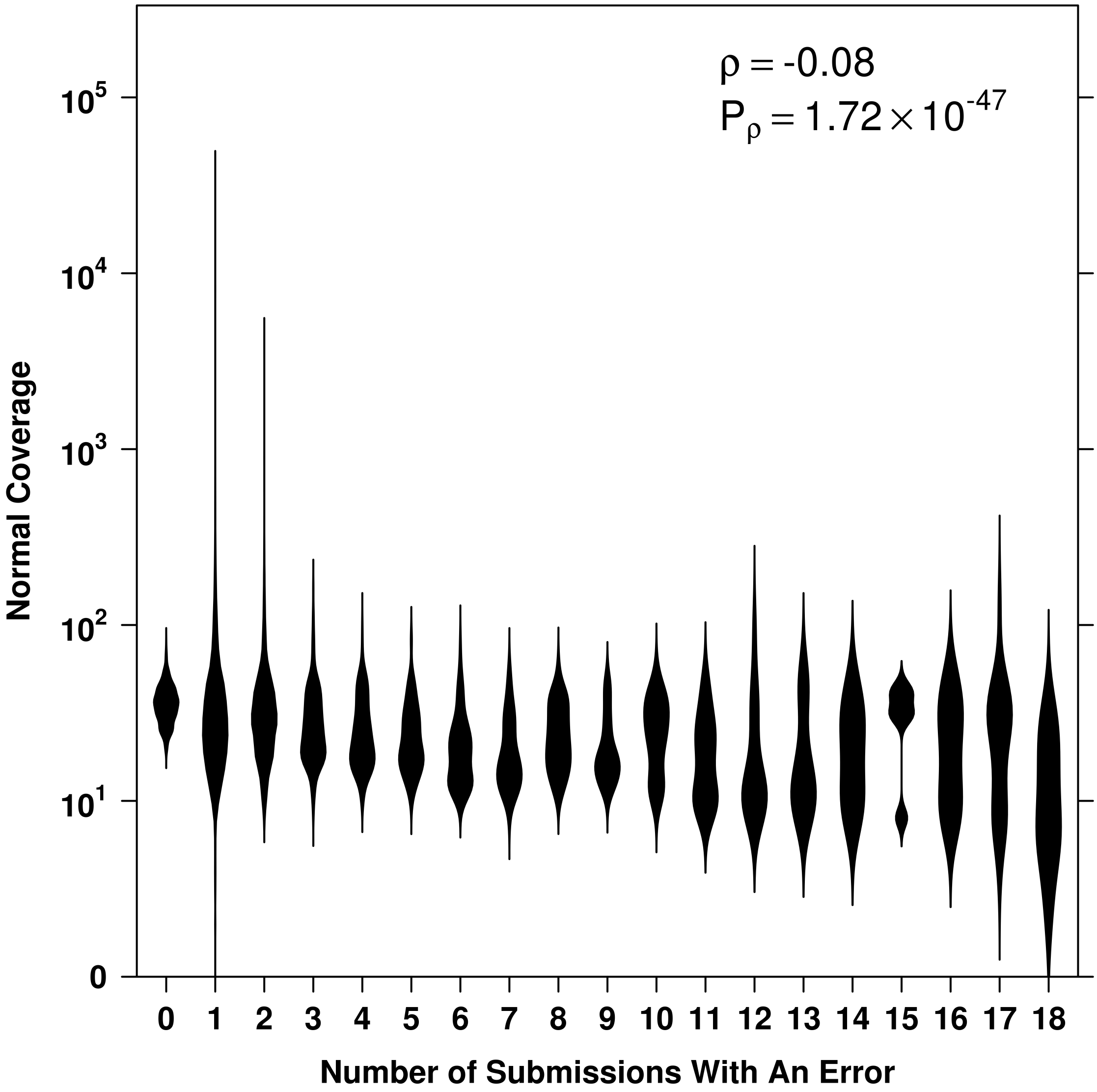
Supplementary Figure 22: Univariate Analysis of Genomic Factors – Tumour Coverage



Violin plot shows the relationship between tumour coverage and number of submissions. Spearman correlation and corresponding p-values were calculated.

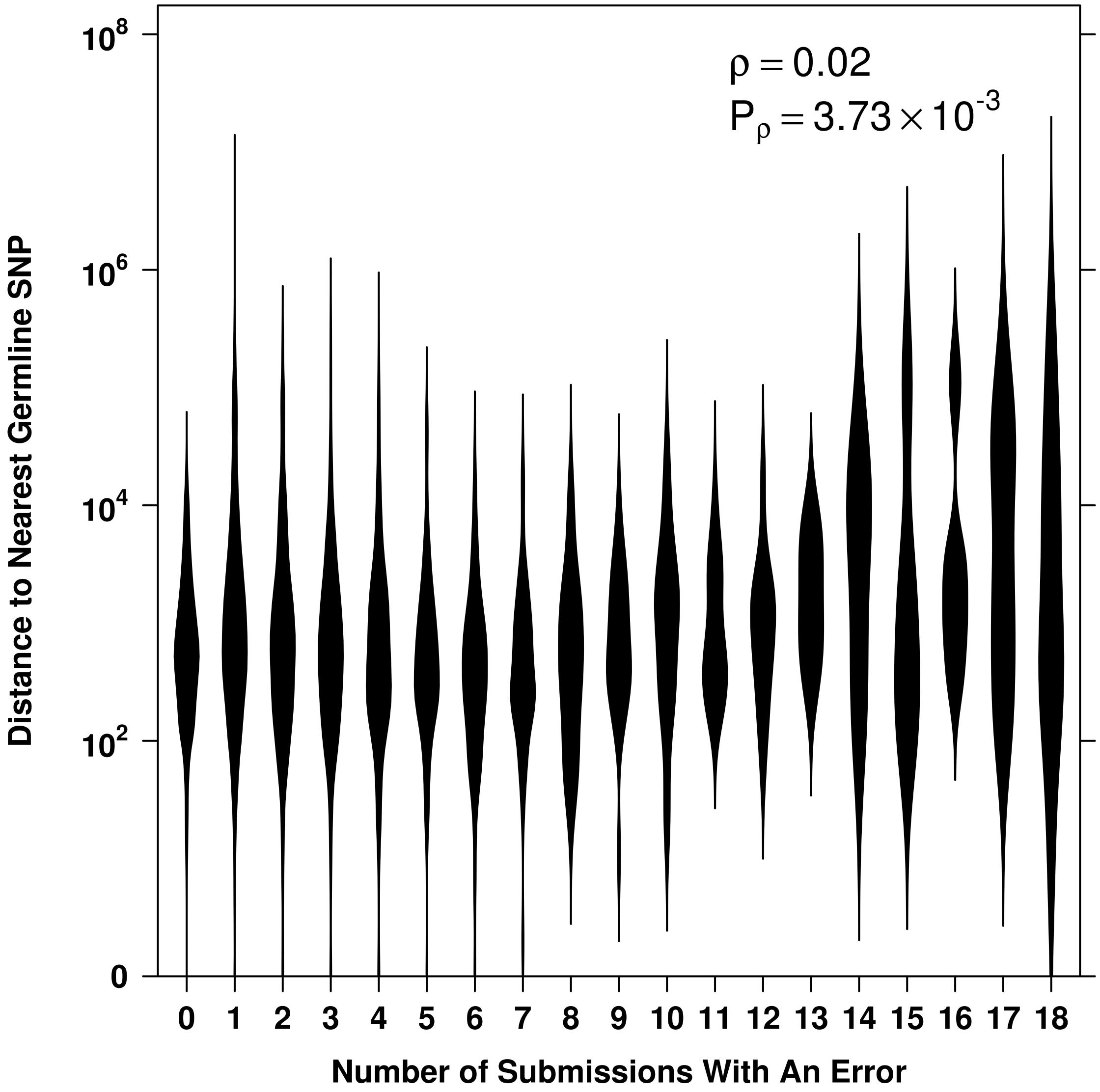
Nature Methods: doi:10.1038/nmeth.3407

Supplementary Figure 23: Univariate Analysis of Genomic Factors – Normal Coverage



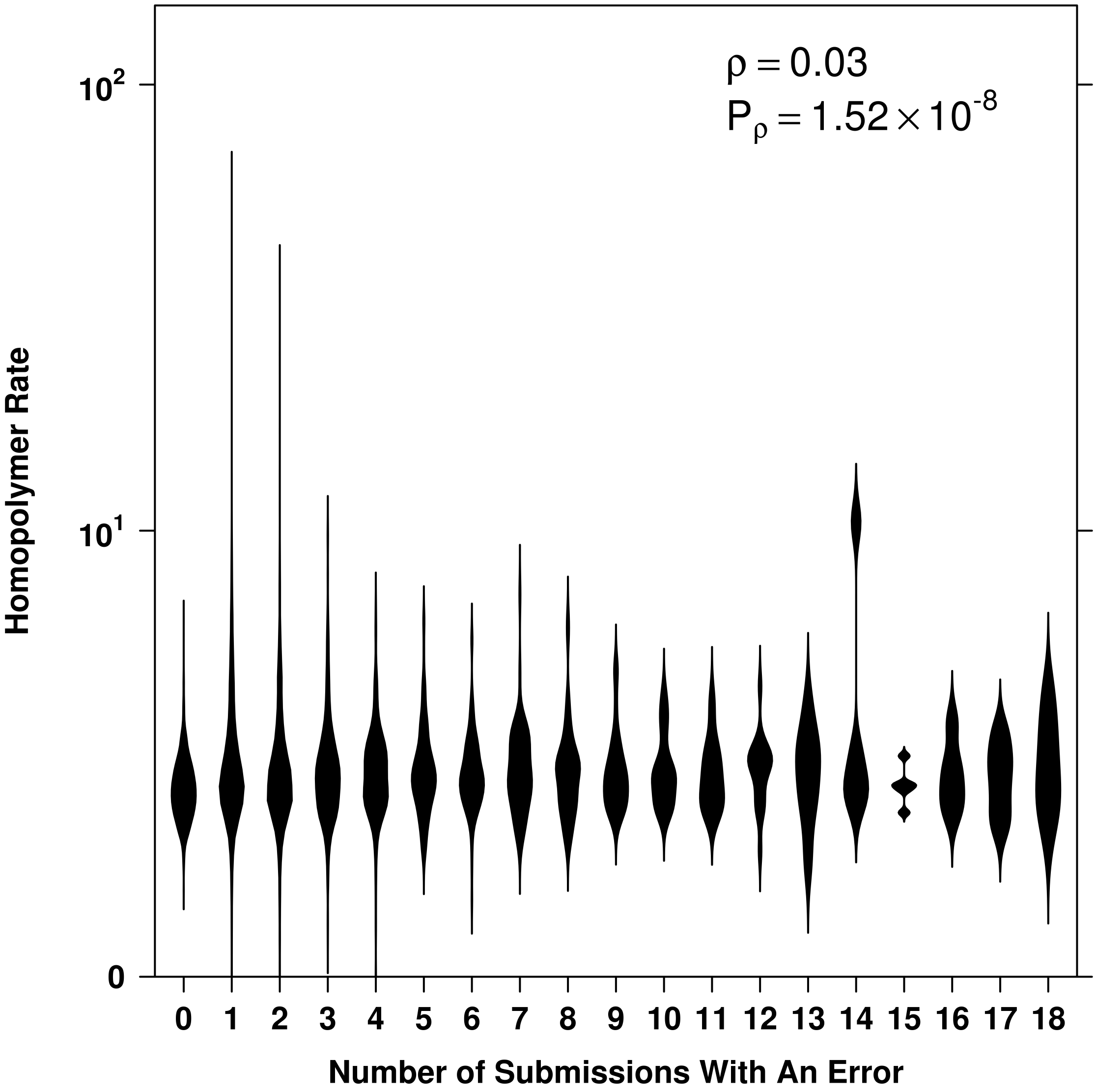
Violin plot shows the relationship between normal coverage and number of submissions. Spearman correlation and corresponding p-values were calculated.

Supplementary Figure 24: Univariate Analysis of Genomic Factors – Distance to Nearest Germline SNP



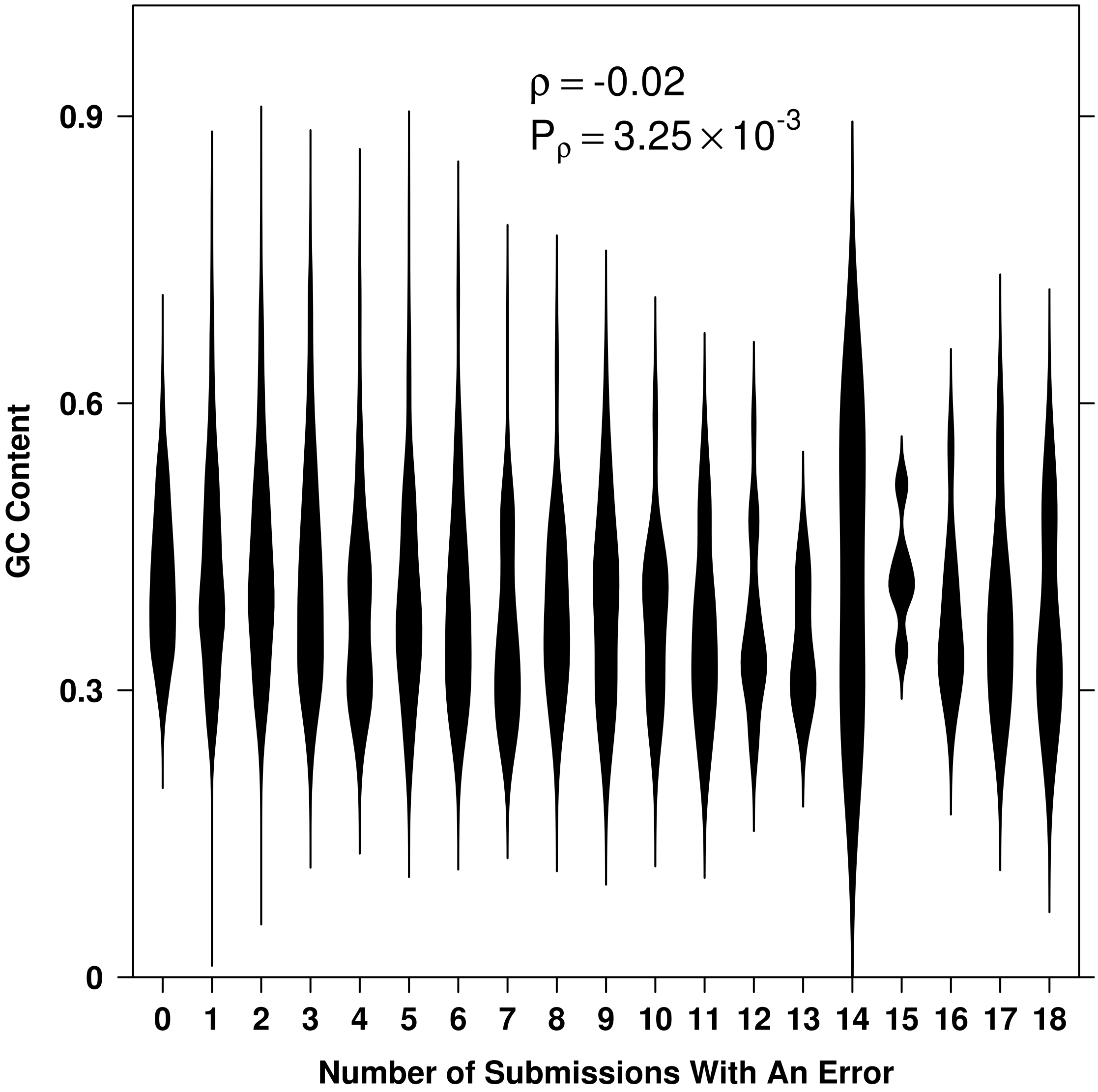
Violin plot shows the relationship between distance to nearest germline SNP and number of submissions. Spearman correlation and corresponding p-values were calculated.

Supplementary Figure 25: Univariate Analysis of Genomic Factors – Homopolymer Rate



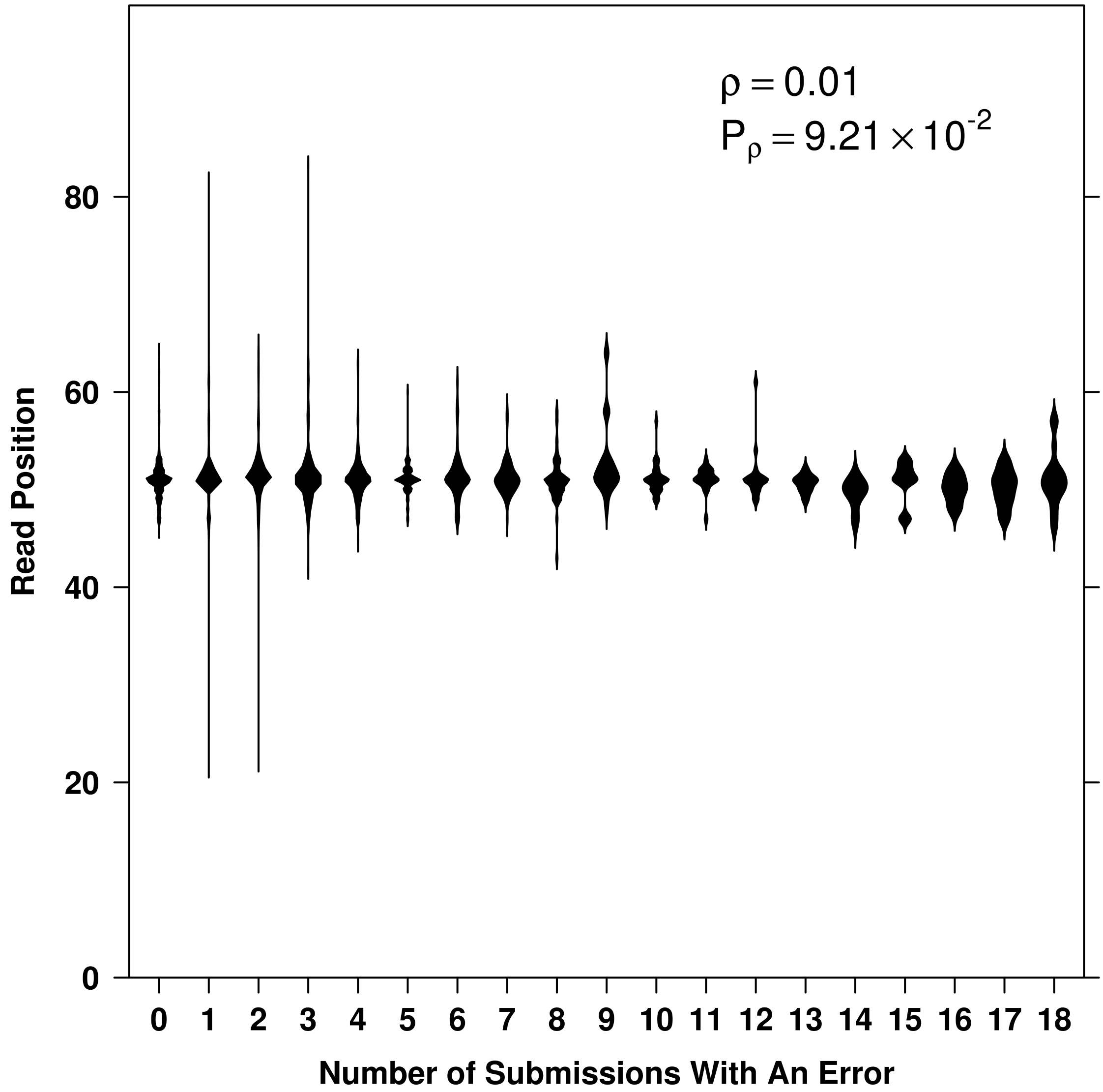
Violin plot shows the relationship between homopolymer rate and number of submissions. Spearman correlation and corresponding p-values were calculated.

Supplementary Figure 26: Univariate Analysis of Genomic Factors – GC Content



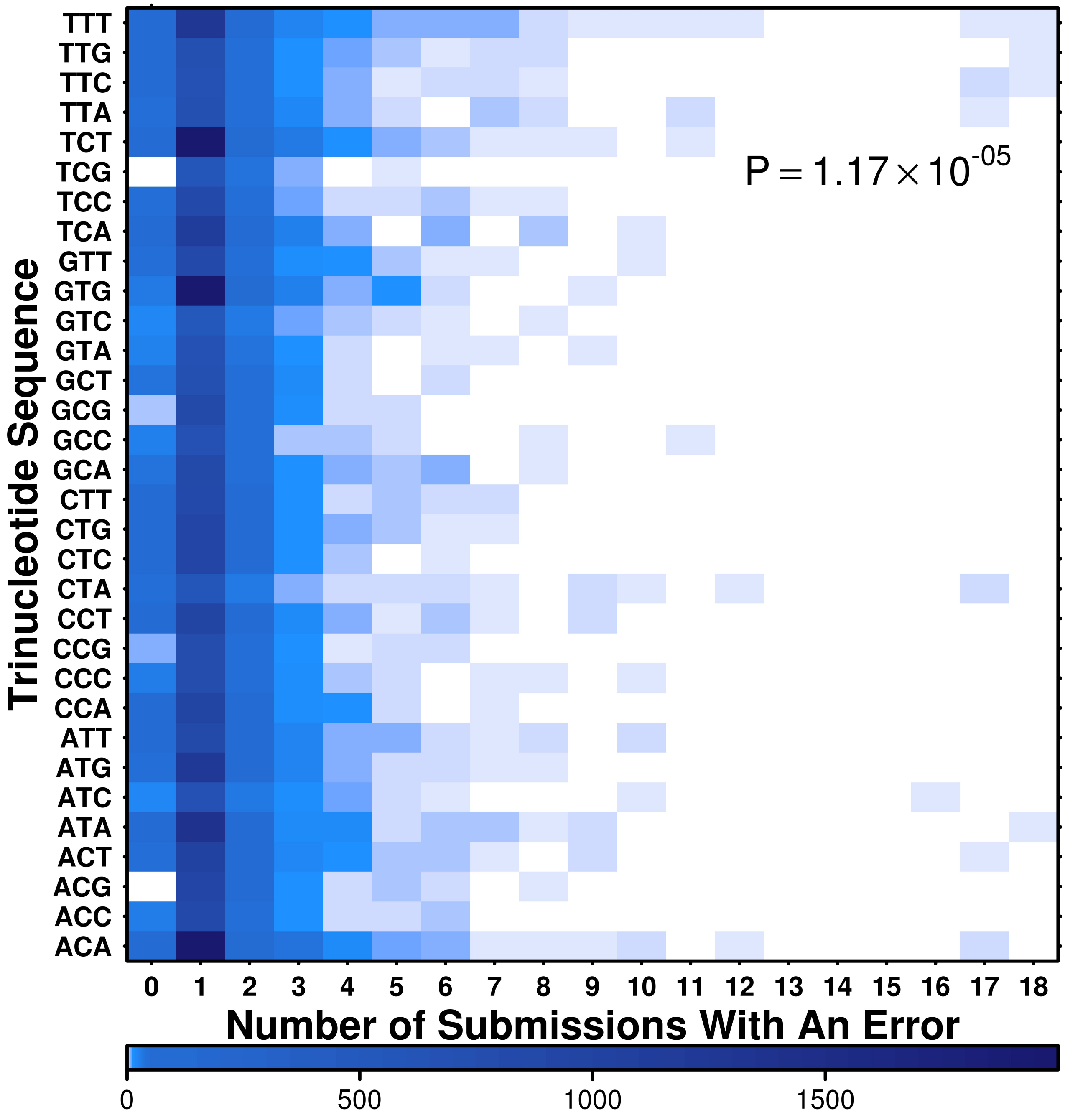
Violin plot shows the relationship between GC content and number of submissions. Spearman correlation and corresponding p-values were calculated.

Supplementary Figure 27: Univariate Analysis of Genomic Factors – Read Position



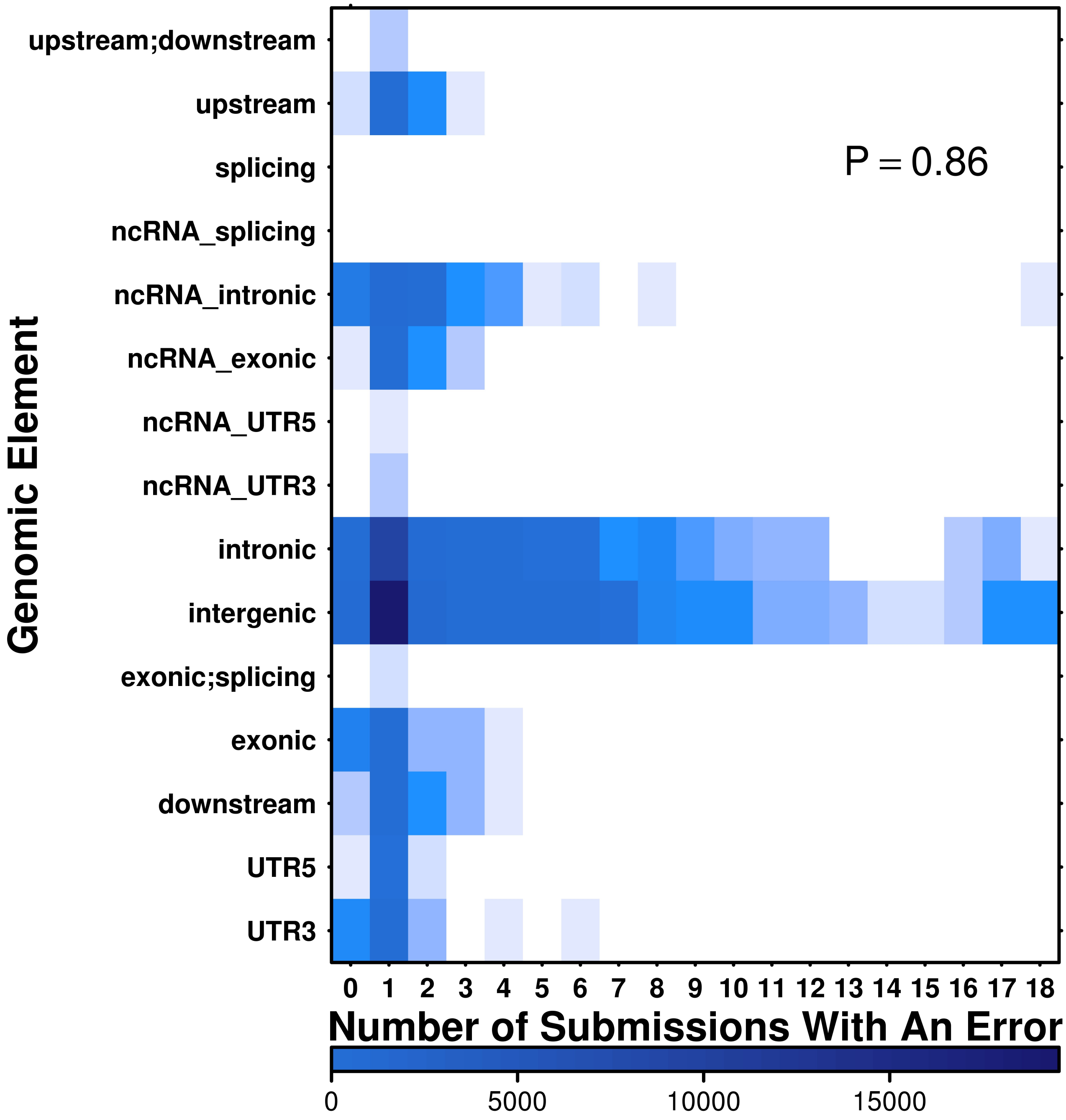
Violin plot shows the relationship between read position and number of submissions. Spearman correlation and corresponding p-values were calculated.

Supplementary Figure 28: Univariate Analysis of Genomic Factors –Trinucleotide Sequence



Heatmap shows the relationship between trinucleotide sequence and number of submissions. P-values were generated from one-way ANOVA.

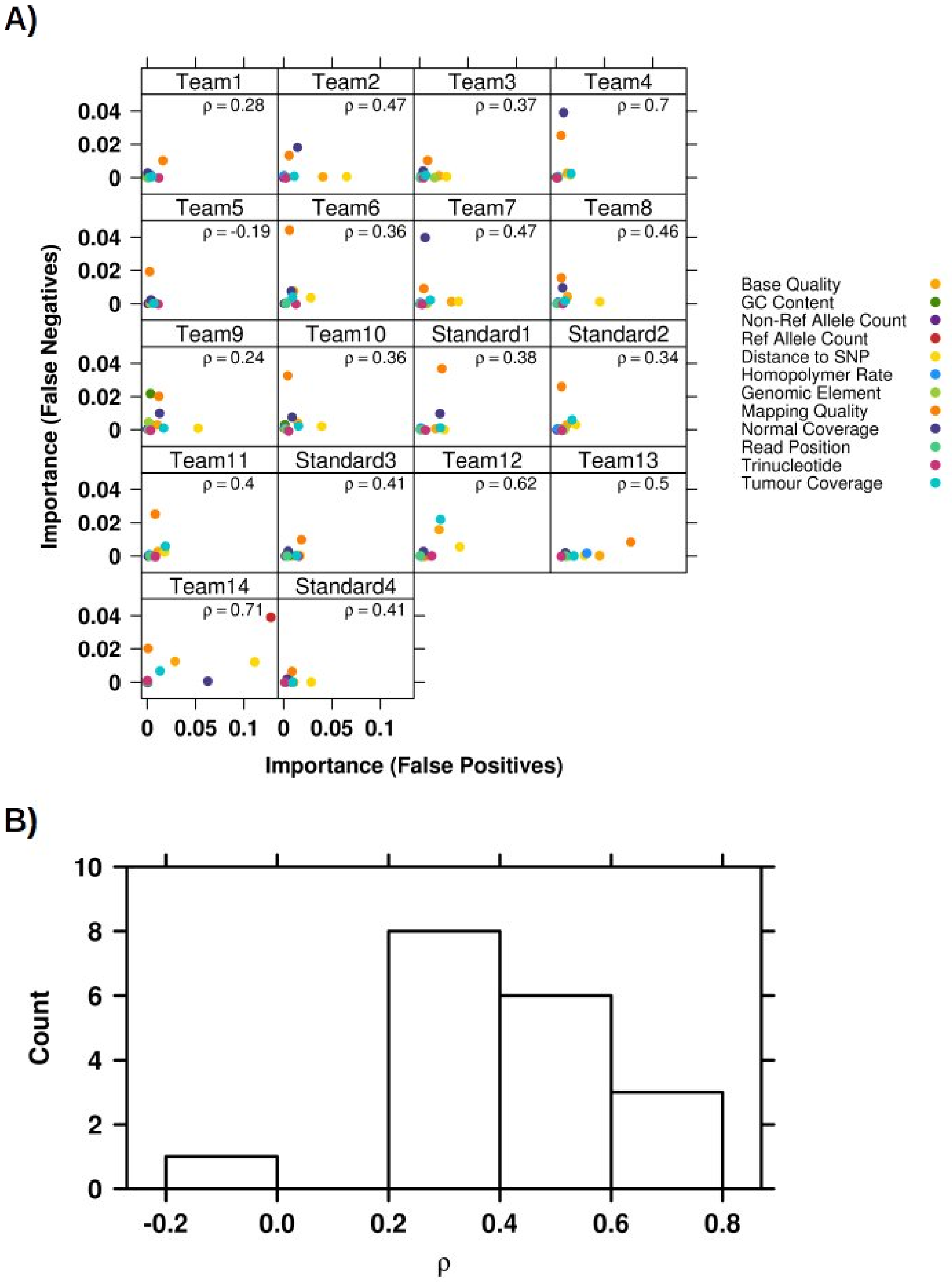
Supplementary Figure 29: Univariate Analysis of Genomic Factors – Genomic Element



Heatmap shows the relationship between genomic element and number of submissions. P-values were generated from one-way ANOVA.

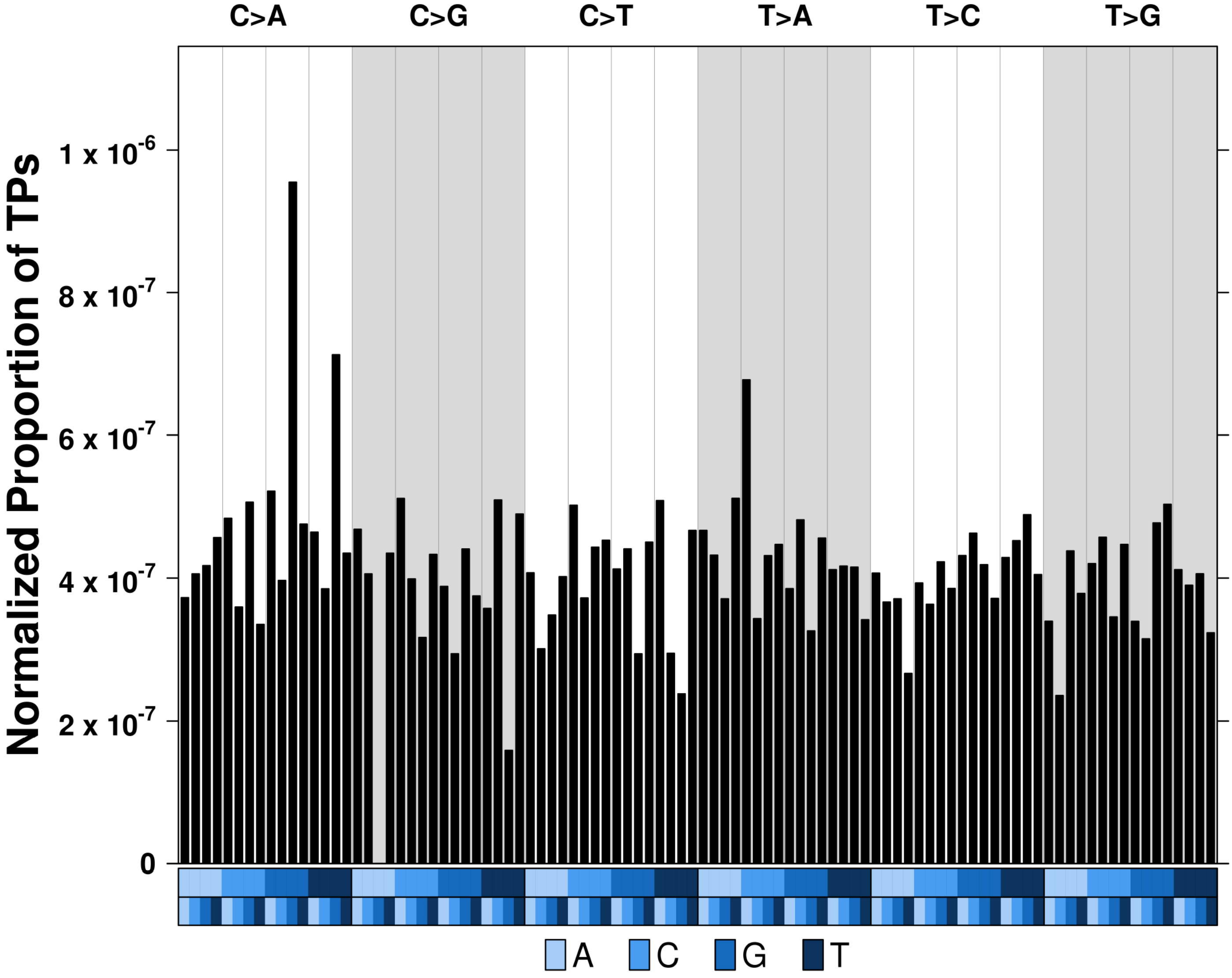
Nature Methods: doi:10.1038/nmeth.3407

Supplementary Figure 30: False Positive and False Negative Error Profiles



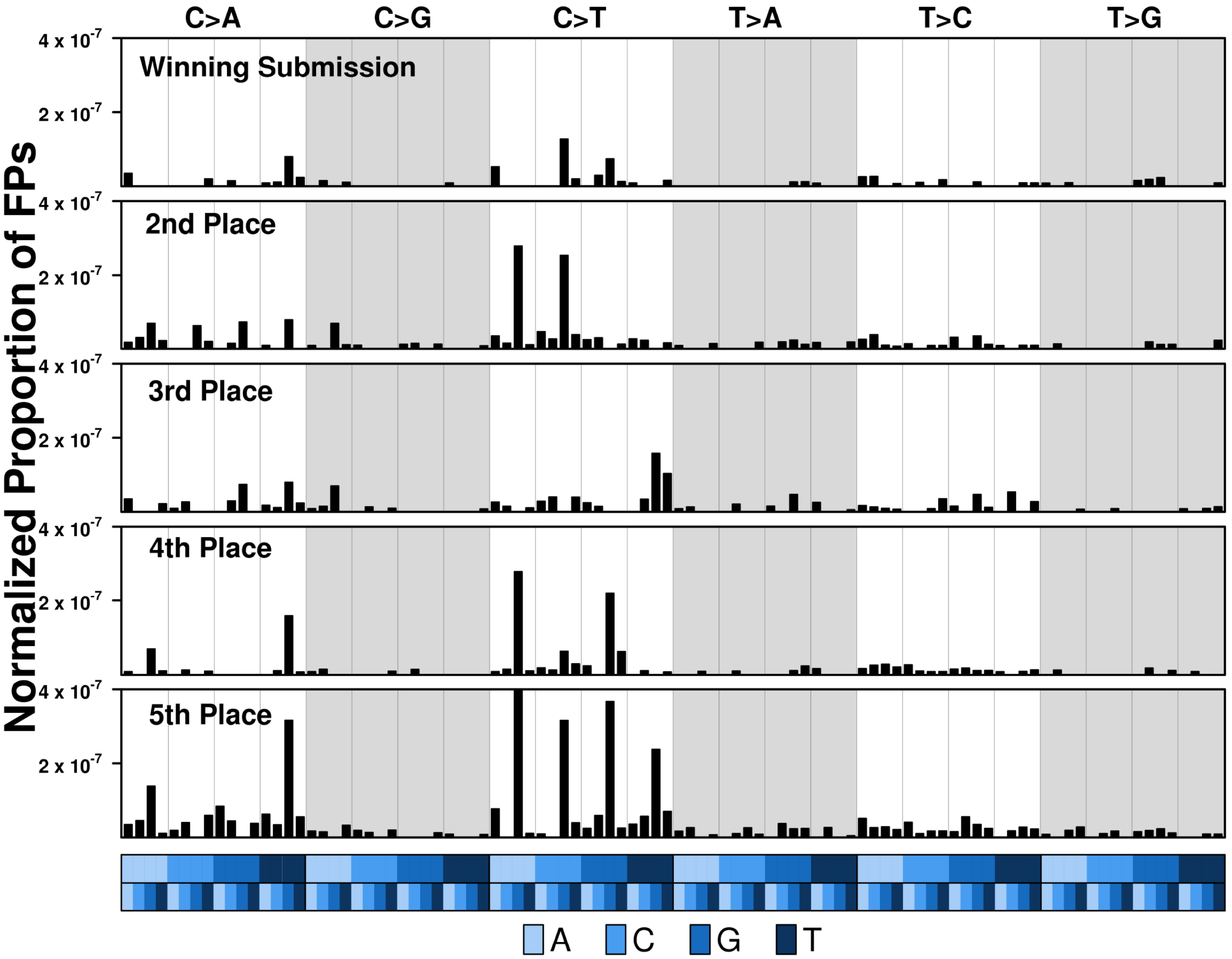
Variable importance measures showed low correlation between false positive and false negative positions for all submissions in IS1 - ranging from 0.26 to 0.71. This low concordance indicates largely different error profiles for false positive and false negative positions. To better visualize false positive and false negative relationship, some points that exceeded the limits of the scatterplots were omitted.

Supplementary Figure 31: Trinucleotide Profile of True SNV Positions



The trinucleotide context of each true SNV was determined for IS1. Trinucleotide patterns were stratified by SNV base change and counts were normalized by the genomic trinucleotide distribution. Base change is indicated at the top of the plot while the 5’ and 3’ bases are indicated by the bottom covariates, respectively. As expected, the true positive profile is more consistent than the false positive profile as mutations were spiked in at random, independent of trinucleotide pattern. The true positive trinucleotide profile was statistically indistinguishable from that of the entire genome (*P* = 0.41; Pearson’s χ^2 test).

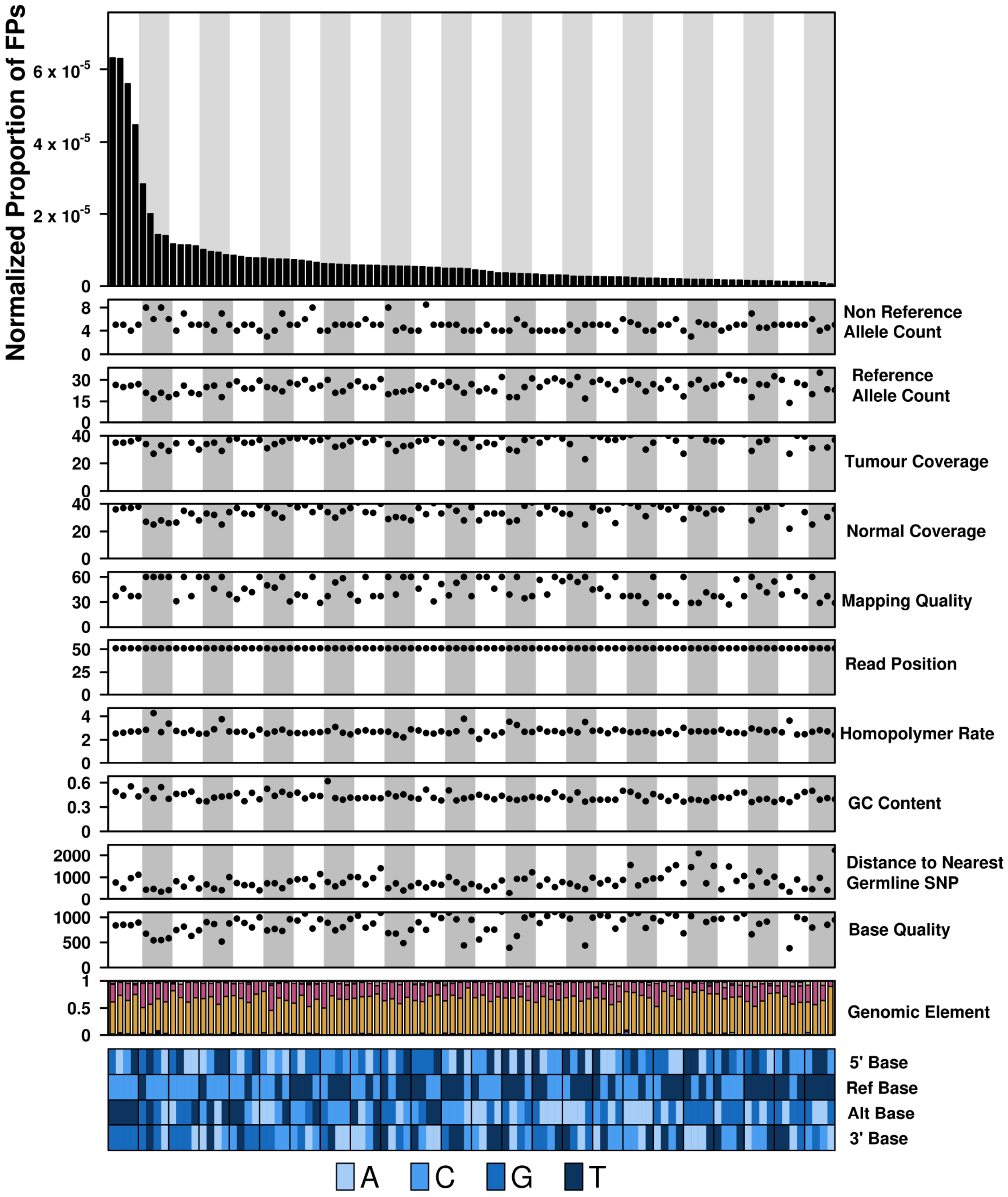
Supplementary Figure 32: Trinucleotide Profile of Top Five Algorithms



The trinucleotide profiles of false positives in the top five algorithms for IS1 were plotted. Some trinucleotide elevations were unique to a subset of algorithms (*i.e.* A<u>C</u>G-to-A<u>T</u>G) while some elevations were seen in all submissions (*i.e.* T<u>C</u>G-to-T<u>A</u>G). The XCG elevation seen in the overall false positive profile (**Figure 4B**) is a combination of the error profiles of multiple submissions and not heavily dominated by only one algorithm.

Nature Methods: doi:10.1038/nmeth.3407

Supplementary Figure 33: Effect of Genomic Factors on Trinucleotide Profile



We binned the values of ten genomic variables - non-reference allele count, reference allele count, tumour coverage, normal coverage, mapping quality, read position, homopolymer rate, GC content, distance to nearest germline SNP and base quality - by their corresponding trinucleotide context. We then plotted the median value for each variable in IS1 for each trinucleotide bin. The proportion of calls located in each genomic element was also plotted. There was no clear pattern in any of the eleven variables corresponding with trinucleotide elevation.