Detailed Comparison of Two Popular Variant Calling Packages for Exome and Targeted Exon Studies

Charles D Warden[1*], Aaron W Adamson[2], Susan L Neuhausen[2], and Xiwei Wu[1*]

[1]Integrative Genomics Core, Department of Molecular and Cellular Biology, [2]Department of Population Sciences, City of Hope National Medical Center, Duarte, CA 91010

* To whom correspondence should be addressed. Tel: 626-256-4673; Fax: 626-471-3708; Email: cwarden@coh.org or xwu@coh.org

ABSTRACT

The Genome Analysis Toolkit (GATK) is often considered to be the “gold standard” for variant calling of single nucleotide polymorphisms (SNPs) and small insertions and deletions (indels) from short-read sequencing data aligned against a reference genome. There have been a number of variant calling comparisons against GATK, but an adequate comparison against VarScan may have not yet been performed. More specifically, we compared four lists of variants called by GATK (using the UnifiedGenotyper and the HaplotypeCaller algorithms, with and without filtering low quality variants) and three lists of variants called using VarScan (with varying sets of parameters). Variant calling was performed on three datasets (1 targeted exon dataset and 2 exome datasets), each with approximately a dozen subjects. We found that running VarScan with a conservative set of parameters (referred to as “VarScan-Cons”) resulted in a high quality gene list, with high concordance (>97%) when compared to high quality variants called by the GATK UnifiedGenotyper and HaplotypeCaller. These conservative parameters result in decreased sensitivity, but the VarScan-Cons variant list could still recover 84-88% of the high-quality GATK SNPs in the exome datasets. We also accessed the impact of pre-processing (e.g., indel realignment and quality score base recalibration using GATK). In most cases, these pre-processing steps had only a modest impact on the variant calls, but the importance of the pre-processing steps varied between datasets and variant callers. More broadly, we believe the metrics used for comparison in this study can be useful in accessing the quality of variant calls in the context of a specific experimental design. As an example, a limited number of variant calling comparisons are also performed on two additional variant callers.

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.403v2 | CC-BY 4.0 Open Access | received: 6 Jun 2014, published: 6 Jun 2014

INTRODUCTION

Multiple studies have previously compared variant callers for short-read sequencing data (Bauer 2011; Cheng et al. 2014; Liu et al. 2013; O'Rawe et al. 2013; Pabinger et al. 2014; Yi et al. 2014; Yu & Sun 2013), which have often indicated that the variant callers available in the Genome Analysis ToolKit (GATK, DePristo et al. 2011; McKenna et al. 2010) show the best performance (Bauer 2011; Liu et al. 2013; Yi et al. 2014). This is in accordance with the popular use of GATK for variant calling, especially for Illumina sequencing data (Boland et al. 2013; Li et al. 2014; Linderman et al. 2014; Worthey 2013). The use of multiple-variant callers has also been proposed (Lam et al. 2012; O'Rawe et al. 2013; Pabinger et al. 2014; Yu & Sun 2013), but this will increase the run-time (or at least computational resources) necessary for analysis (which can be especially important for large patient cohorts).

Most comparisons did not directly compare GATK variant callers against VarScan (Koboldt et al. 2009; Koboldt et al. 2012). One study indicated that VarScan was less accurate than the other variant callers (Cheng et al. 2014). Another study showed VarScan as being more similar to the other variant callers but still ranked GATK as the best option (Yi et al. 2014). However, VarScan has been used for variant calling in a large number of studies (Worthey 2013), and we hypothesize that the simple, intuitive parameters can be helpful in establishing an optimal set of variants for a given dataset. Also, as emphasized in this study, the run-time for VarScan should typically be shorter than GATK. Therefore, we wished to determine if 1) the previous VarScan benchmarks can be reproduced in our own analysis and 2) if use of non-standard parameters can improve the quality of the VarScan predictions.

Additionally, we wished to determine the relative impact of pre-processing steps in GATK (specifically, the indel realignment and quality score base recalibration steps). Therefore, we compared variant calls with GATK and VarScan for each step separately, with both pre-processing steps, as well as without either pre-processing step. There has been at least one previous study to showing that filtering can improve the quality of variant calls (Carson et al. 2014), beyond the GATK quality score base recalibration. We assessed whether running VarScan with different sets of parameters (using three different parameters settings: see Methods) can also increase the accuracy of the resulting variant calls. Additionally, we have used a simple filtering strategy for GATK variants (looking at all variants called versus filtering out variants with a low quality flag), so there are two sets of variant lists for both GATK UnifiedGenotyper and GATK HaplotypeCaller.

In order to avoid any bias that could come from studying only a limited number of samples, variant calls were performed on 14 targeted exon (for 1000 genes) samples from the 1000 Genomes project (The 1000 Genomes Project Consortium 2012), 12 exome samples from the 1000 Genomes project (The 1000 Genomes Project Consortium 2012), and 15 Illumina exome samples from SRP019719 (O'Rawe et al. 2013). We believe this helps yield robust results both in terms of the number of samples studied per cohort as well as variations in study design (i.e. the method of targeted sequencing). The 1000 Genomes study was specifically chosen in order to test recovery of validated variants as well as to compare concordance between samples subject to both targeted exon and exome sequencing.

In short, this study presents a detailed characterization between GATK and VarScan on 41 samples (with varying target designs), where each sample has 28 variant lists for comparison. Variant lists are compared based upon the number of variants called, the proportion of unknown frequency variants in the variant list, and the reproducibility of variant calls using different methodologies. A limited number of additional comparisons are also performed in order to help illustrate how these metrics can be used to select the optimal variant caller for a given dataset. This analysis demonstrates that a conservative set of parameters can be used to produce a robust, high-quality list of variants from VarScan, and this study also presents evidence that the GATK HaplotypeCaller may have a higher false positive rate in calling indels compared to the GATK UnifiedGenotyper.

MATERIAL AND METHODS

Sample Selection

All datasets were downloaded as .fastq files from the European Nucleotide Archive (Leinonen et al. 2011). Illumina exome samples were downloaded from SRP019719 (O'Rawe et al. 2013). 1000 Genomes Project (The 1000 Genomes Project Consortium 2012) data, abbreviated as 1KG in this manuscript, was selected on the basis of having 1) exome data, 2) targeted exome data, 3) Omni

SNP chip data, and 4) validated SNPs. Among samples meeting those criteria, 12 samples were selected based upon 1) their presence in disparate populations (CEU: Northern and Western European Ancestry, CHB: Han Chinese, JPT: Japanese, and YRI: Yoruba/African) and 2) maximum number of validated SNPs within each of the four selected populations.

1000 Genomes validated SNPs and Omni SNP chip .idat files were downloaded from the 1000 Genomes FTP site (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/). Likewise, target design files (for targeted exon and exome samples) were downloaded from this site. More specifically, the phase 3 design files were used to calculate coverage statistics for the exome samples and the validated variants come from ALL.chr20.exome_consensus_validation_454.20120118.snp.exome.sites.vcf.gz. At the time this dataset was downloaded, the hg19 reference location for validated SNPs was off-set by one (similar to indels in .vcf files), and this was taken into consideration during analysis. Targeted gene co-ordinates (for hg18) from P3_consensus_exonic_targets.bed were converted to hg19 coordinates using the LiftOver function in Galaxy (Blankenberg et al. 2001; Giardine et al. 2005; Goecks et al. 2010).

Data Processing

Reads were aligned to a karotype-sorted hg19 reference (necessary for running GATK (McKenna et al. 2010)) using BWA (Li & Durbin 2009). Prior to variant calling, singletons were filtered out via samtools (Li et al. 2009), .bam files were coordinate sorted via Picard (http://picard.sourceforge.net/), and duplicates were removed via Picard. Prior to running GATK, read groups were added via Picard, and .bam file was re-ordered according to chromosome karyotype. Prior to running VarScan (Koboldt et al. 2012), a pileup file was created via mpileup in samtools (Li et al. 2009) and positions without any aligned reads were filtered out. These represent the minimum pre-processing sets and are labelled as “No Preprocessing”. Alignment statistics for these samples are shown in supplementary tables (Table S1 for 1KG Targeted Exon, Table S2 for 1KG Exome, and Table S3 for SRP019719).

There are three additional possible pre-processing pipelines that were considered for VarScan and GATK comparisons. “Realign Only” runs an indel realignment using GATK (using the RealignerTargetCreater and IndelRealigner functions). “Recalibrate Only” uses GATK to recalibrate quality scores (using the BaseRecalibrator and PrintReads functions). "Full Pipeline" runs the indel realignment functions and then performs base recalibration.

The Human610 SNP chip data for individual K8101-49685s (paired with SRP019719 Illumina exome sample SRX265476) was reported in the Illumina "TOP" format, so the reverse complement of the allele was used when the IlmnStrand and RefStrand did not match (as defined in the human610-quadv1_h.csv manifest file). Allele sequences were provided without respect to a reference sequence, so SeattleSeq Variant Annotation (http://snp.gs.washington.edu/SeattleSeqAnnotation138/) was used to determine the reference sequence to focus on variants that differ from the reference sequence (to make results comparable to the Illumina sequencing variant calls). This is the latest version of the manifest file, but it was designed using dbSNP 131; so, some discordant SNPs are due to outdated annotations where the forward strand may vary from the hg19 reference sequence. However, this only affected a relatively small minority of SNPs (<5% of variants; see Results).

Raw .idat files for 1000 Genomes Omni SNP chip samples were processed in Illumina® Genome Studio^{TM} (V2011.1). A pre-defined clustering file (HumanOmni2.5-4v1-Multi_H.egt) was used to call genotypes. Variants were exported in the "Plus" format (so, no genotype conversion as necessary). Samples were annotated using the HumanOmni2.5-4v1-Multi_B.bpm manifest file, including the genomic position in hg18 coordinates. Coordinates were converted to hg19 via liftOver in Galaxy (Blankenberg et al. 2001; Giardine et al. 2005; Goecks et al. 2010) and reference sequences were determined for all on-target probes (n=2257) via SeattleSeq Variant Annotation (http://snp.gs.washington.edu/SeattleSeqAnnotation138/). Each sample had 2179-2188 genotyped SNPs recognized by SeattleSeq, with 477-616 non-reference alleles per sample.

Concordance Definitions

Concordance was defined as recovery of a set of lower-throughput variants (validated, SNP chip, targeted exon), except for the targeted exon technical replicates where concordance was defined as (2 * Number of Overlapping Variants) / (Sum of Variants from both samples). When defining variant concordance between the SNP chip data and exome data, recovery of the known variant is counted as a concordant variant (even if multiple variants are called at a given position).

Unlike most comparisons in this study, the SNP chip comparisons do not specifically focus on the coding variants. As such, extraction of SNP chip variants within the target regions currently includes some non-coding variants (such as intronic variants) that would falsely be called discordant if focusing only on coding variants. SNP chip variants have been filtered to only include variants that vary from the reference sequence.

Calling Variants

Variants were called using GATK (v.2.8.1) based upon established best practices (DePristo et al. 2011; Van der Auwera et al. 2002). Variants were called using both the UnifiedGenotyper (UG) and the HaplotypeCaller (HC). For variant characterization, the set of all variants was considered (labelled as UG-all in figures for the UnifiedGenotyper and HC-all for the HaplotypeCaller) as well as only the high-quality variants that didn't contain the "LowQual" flag in the .vcf file (labelled as UG-HQ in figures for the UnifiedGenotyper and HC-HQ for figures for the HaplotypeCaller). In addition to the parameters described in the GATK best practices, variant calling for SRP019719 also required some additional parameters due to the quality scores for the ENA reads (these extra parameters were not necessary for calling variants from 1000 Genomes Project data). More specifically, SRP019719 variants were called with the parameters "--fix_misencoded_quality_scores -fixMisencodedQuals" for HaplotypeCaller (but only for "No Preprocessing" variants), UnifiedGenotyper (but only for "No Preprocessing" variants), RealignerTargetCreator, IndelRealigner, and BaseRecalibrator (except for the "Full Pipeline" variants where IndelRealigner has been run already). SRP019719 variants were called with the parameter "-allowPotentiallyMisencodedQuals" for PrintReads (following BaseRecalibrator, except for the "Full Pipeline" variants where IndelRealigner has been run already)

VarScan (v.2.2.8) variants were called using pileup2snp and pileup2indel. Three different sets of parameters were used for calling variants. "VarScan: Default" specifies no additional parameters beyond the minimal requirements. "VarScan: P-value" sets a minimum p-value threshold of 0.05, but specifies no additional parameters. "VarScan: Conservative" uses the following parameters to stringently call variants: minimum 10 total reads at the position of interest, minimum of 4 supporting reads to call variant, minimum average quality of 20, and minimum variant allele frequency of 0.3. Please see Table 1 for a summary of parameters used for the VarScan comparisons.

Additional variant callers were also applied to the "No Preprocessing" alignments and the "Full Pipeline" alignments. Freebayes (v0.9.14, Garrison & Marth 2012) was applied to the 1000 Genomes Exon Targeted samples, using default parameters. The Bayesian variant caller in the bcftools function (in the samtools package (v0.1.19, Li et al. 2009) was applied to all three datasets, using default parameters (followed by applying vcfutils.pl with a maximum read depth of 200). Unlike GATK and VarScan, samtools has a unique indel format to represent ambiguous indels. Although the ANNOVAR file conversion program can remove all nucleotides that are not part of the indel, the genomic position used to represent this indel is not necessarily the same as GATK and VarScan. For this reason, we only present the samtools SNP results in this manuscript. All analysis was performed on a shared Linux server with concurrent usage (x64, CentOS Red Hat 4.1, 264 GB RAM, 6 x 2.27 GHz processors).

Because a publication on the 1000 Genomes exome and targeted exon datasets has not yet been published, we are only reporting variant frequencies for a single chromosome (chr20) in order to comply with 1KG publication requirements. Thus, only 1,140,996 base pairs of targeted sequence is considered for variant call benchmarks in the 1000 Genomes exome datsets, and only 35,309 base pairs are considered for variant call benchmarks in the 1000 Genomes exon targeted samples. In contrast, the SRP019719 are made genome-wide (with a targeted design covering 46,401,093 base pairs).

Annotating Variants

After variants were called, ANNOVAR (Wang et al. 2010) was used to determine the population frequency for each variant. Variants were defined as "low frequency" if the 1000 Genomes Project (The 1000 Genomes Project Consortium 2012) frequency and NHLBI Exome Sequencing Project (Fu et al. 2013) frequencies were both less than 0.01. Variants were defined as "unknown frequency" if they were not present in any 1KG or ESP samples as well as undefined in dbSNP (Sherry et al. 2001). Variants were predicted as damaging if the SIFT (Ng & Henikoff 2003) score greater than or equal to 0.95 or the PolyPhen (Adzhubei et al. 2010) score was greater than or equal to 0.85 (which are the thresholds used by ANNOVAR to flag a variant as damaging, for those two programs). Although population frequency and damaging frequencies can be provided for both SNPs and indels, there were very few indels within on-target regions for the targeted exon samples (so, ANNOVAR characterization is only present for SNPs). ANNOVAR also provides a common format that can compare all variant callers (and it allows easy determination of variants included within the targeted exon panel), so all between-sample comparisons were performed using the ANNOVAR exome summary table (except for the SNP chip comparison, which uses the genome summary table).

RESULTS

Run-Times Scale Differently for GATK versus VarScan

Figure 1 shows run-times for the entire variant calling pipeline. The run-time for each step is reported in Table S4-S6, with the average run-times for the variant calling step itself shown in Table 2 (for the “Full Pipeline” samples). Among the 1000 genome (1KG) targeted exon samples (n=14), the GATK UnifiedGenotyper had the longest run-time (Figure 1A). However, the run-time for all samples was less than 5 hours, so run-time is not a severe limiting factor for either VarScan or GATK (for targeted exon samples). Adding a p-value filter did not significantly affect the run-time for VarScan; in fact, the run-time for VarScan is essentially the same regardless of what parameters are used for analysis. In contrast, the 1KG exome samples showed a wide range of run-times (Figure 2A), with the HaplotypeCaller having a very noticeably longer run-time compared to the other pipelines. Quality score recalibration also caused a noticeable increase in run-time.

In addition to the 1000 Genomes Exome and Targeted Exon samples, the same analysis was performed on an independent dataset of 14 exomes (SRP019719). The most obvious trend is that the run-time for the GATK HaplotypeCaller was considerably longer when running base recalibration without prior indel realignment (Figure S1). Two factors are mostly likely causing this increased-run time are 1) quality scores were on a different scale than the 1000 Genomes samples (see Methods) and 2) the servers running the variant callers had considerable concurrent usage (which may have exacerbated a sensitivity to running the function in a way that deviates from the GATK “best practices” and uses non-standard parameters). Therefore, we do not expect this trend to apply to most samples. However, it is worth noting that the pipeline described by the GATK “best practices” (in this case, the

"Full Pipeline" pre-processing strategy) is probably subjected to the most troubleshooting: therefore, the a practical advantage to using the "Full Pipeline" pre-processing over indel realignment and/or base recalibration in isolation is that there is a decreased chance of getting less accurate results due to technical problems with the software. Likewise, total variant counts and population frequencies among variants (Figure S2 and S3) indicate that the variants called using base recalibration alone were probably not reliable (for these particular samples), but the variants called using the full pipeline (indel realignment + base recalibration) were more likely to have a lower false positive rate. Please see the following section for a more detailed description of the rationale for using these metrics.

Estimation of Accuracy for GATK and VarScan Variant Calls

Although 1KG samples were selected on the basis of having some validated positive controls, it is useful to have quality control metrics that will be correlated with the true sensitivity and specificity for a given variant caller. Each strategy has notable caveats for interpretation, but we think it is useful to have quality control metrics that can be used to select an optimal variant caller for a given dataset.

First, we assume that the total number of variants is correlated with sensitivity. Of course, the accuracy of this assumption would depend upon the false positive rate for the variant caller. Nevertheless, we believe that the total number of called variants is a useful benchmark to compare variant callers. SNP and indel frequencies are shown for 1KG exome and targeted exon samples in Figure 2. In most cases, the pre-processing pipeline has a minimal effect on the size of the resulting list of SNPs, with the notable exception of the GATK UnifiedGenotyper (although the impact is significantly decreased if only focusing on high-quality SNPs). However, it is difficult to tell how common this trend is for all datasets: for example, the exome samples from SRP019719 do not show this same difference (Figure S2) and instead show a relatively greater difference in the number of SNPs called by VarScan with default parameters. The size of the SNP lists are notably variable with different VarScan parameters: the p-value threshold considerably decreases the number of SNPs called, the conservative parameter set is even more restrictive, and both results run with non-default setting produce significantly fewer SNPs than GATK. Among the high-quality calls, the number of SNPs is similar for GATK HaplotypeCaller versus GATK UnifiedGenotyper. The relative number of indels vary from the relative number of SNP calls: however, there are only a limited number of indel calls on chr20 for the targeted exon panel, so we think the number of 1KG exome indel calls are more reliable for assessing general trends. There is a modest but noticeable increase in the number of indels called if the indel realignment was run, and the GATK HaplotypeCaller includes a much larger number of indels than the GATK UnifiedGenotyper. Similar to the SNP calls, VarScan produces fewer indel calls when using more stringent parameters.

Second, we assume that most variants should have known frequencies, and that an over-representation of variants of unknown frequencies should correspond to an increased false positive rate. Of course, some unknown frequency variants will in fact be accurate and the proportion of true unknown frequency variants will vary with demographics (subjects from ethnicities that have been characterized in greater detail will likely show fewer unknown frequency variants, and vice versa). However, some unknown frequency rates are clearly unacceptable: for example, more than 50% of the variants called by VarScan (using default parameters) are unknown frequency variants, and the majority of these unknown frequency variants are predicted to be damaging (Figure 3 and Figure S3). In contrast, there is only a minority of variant calls (0.4-9.7% for 1KG exome; 0.8-4.2% for 1KG targeted exon; 0.1-2.1% for SRP019719 exome samples) are known to be present at low frequency (<1%; see Methods) in the overall population, and this is true for all samples using all variant calling strategies. A high proportion of unknown frequency variants is especially dubious in the 1000 Genomes samples (Figure 3), where these sequencing of these particular individuals was used in the determination of variant frequencies. Likewise, the unknown variant frequencies universally high for the “Recalibrate Only” SRP019719 samples for all variant callers; in this case, the abnormally high run-time (Figure S1) and number of variants called (Figure S2) corroborate the conclusion that there was a technical problem in producing these variant calls (as predicted by the large number of unknown frequency variants). Additionally, results in the subsequent section will show that reproducibility is worse for variant lists with a very high proportion of unknown frequency variants.

Most GATK unknown variant distributions look similar, except when the GATK UnifiedGenotyper is run without quality score recalibration in the 1000 Genomes samples (Figure 3). However, trend doesn’t apply to the SRP019719 samples (Figure S3), emphasizing that there are other factors that can influence these results. The pre-processing steps have a modest impact on the VarScan frequencies, but the VarScan parameters have a very strong impact on the results. Namely, the unknown variant frequency is extremely high when running VarScan with default parameters (much higher than observed for low frequency variants), and we expect these SNPs may have a high false positive rate. However, the distribution for VarScan variants called with conservative parameters looks very similar to the GATK distributions, so these are likely reliable variant calls.

ANNOVAR can also annotate variant frequencies for indels. However, small indels have not been characterised as well as SNPs (for example, there are considerably fewer indels in dbSNP (Sherry et al. 2001), compared to SNPs) and damaging predictions focus primarily on SNPs. Additionally, there are almost no indels in the coding regions of chromosome 20 for the 1000 Genomes targeted exon dataset. Nevertheless, Figure S4 shows the variant frequencies for indel calls in the 1000 Genome and SRP019719 datasets. Similar to the SNP distributions, VarScan-Cons contains the least number of unknown frequency indels (among the VarScan comparisons), which is likely associated with a lower false positive rate. Also, the unknown frequency indels are more common in GATK HaplotypeCaller variants than GATK UnifiedGenotyper variants. Arguably, this could indicate that the higher number of indels called by the HaplotypeCaller is also associated with a higher false positive rate. However, the variant frequencies alone are not sufficient to prove this to be the case because there may have been technical limitations in discovering the indels uniquely called by the GATK HaplotypeCaller. However, there is at least one study showing the validation rate for HaplotypeCaller indels was lower than UnifiedGenotyper indels (Lescai et al. 2014), which is consistent with the hypothesis that the HaplotypeCaller indels may be lower quality (and the VarScan indels may be similar quality to the GATK UnifiedGenotyper indels).

Reproducibility of GATK and VarScan Variant Calls

The recovery rate for validated SNPs is similar for each variant calling pipeline (Figure 3, Table S7). Statistics are only provided for the 1KG exome samples, because no validated SNPs occur in the targeted regions for the targeted exon samples (for the samples selected for this study). Out of the total 35 validated SNPs present among the 12 exome samples, recovery rates varied between 80-94%. VarScan-Default had the highest overall sensitivity and VarScan-Cons had the lowest overall sensitivity. The pre-processing steps had little impact on the results: in fact, the only effect was that base recalibration caused one fewer validated SNP to be recovered when using the

GATK UnifiedGenotyper. There is certainly more similarity in the recovery rate among variant callers than in the total number of SNPs produced by the different strategies, which would make sense if the larger variant lists has a proportionately higher false positive rate: unfortunately, this is a limited number of validated SNPs, so it is difficult to say how closely these are tied to the true sensitivity rates.

The recovery of targeted exon variants (within the set of targeted genes) among the exome samples was also typically quite high (Figure 5). The main exception is for variants called using VarScan with default parameters, and this is also true for the VarScan calls with the p-value filter, to a lesser extent. In this case, it is important to note that the targeted exon calls are not truly a gold standard. In other words, we expect these calls to have a high false positive rate based upon the abnormally high proportion of unknown frequency variants (Figure 3), and false positives in the exon targeted dataset are not likely to be recovered in the exome dataset. For example, position 17933286 on chromosome 20 for individual NA18566 is covered by 189 reads in the targeted exon sample and 109 reads in the exome sample: 90% of reads match the reference sequence in the targeted exon sample and 99% percent of the reads match the reference sequence in the exome sample. Using the default parameters, VarScan calls a variant “G” allele in the targeted exon sample (which is present in 12/189 reads). However, this is likely a normal diploid individual with the true genotype of T/T at this position, with deviations from the reference sequence that are probably due to technical error (where the proportion of errors found at a particular site can randomly fluctuate between samples).

Although there were only a limited number of validated SNPs to test, there are a much larger list of variants present on the Omni SNP array. However, it should be noted that these will mostly be common SNPs, which importantly means that recovery of SNP array variants may not represent the accuracy of rare variant calls. Nevertheless, it is useful to to see how the SNP chip recovery compared to the validated SNPs and the total SNP calls. Every 1000 Genomes sample in this study has a paired Omni SNP chip sample: in most cases, all variant calling algorithms could recover >85% of Omni SNP chip variants (Figure 6A). Similarly, we compared variant calls for a subject from the SRP019719 dataset that had both Illumina exome and SNP chip data: again, the majority (88-96%) of SNP chip variants were recovered in the paired exome dataset (Figure 6B). In both cases, recovery was restricted to SNP chip variants within targeted regions for the exome sample. It is also worth noting that SNP chip alleles matching the reference sequence are not considered (in order to treat the

SNP chip data more like variant calls from sequencing data), and concordance would of course be higher if these sites were considered. Unlike the validated variants (Figure 4), VarScan (using default parameters) no longer had the highest sensitivity; instead, the average recovery of SNP chip variants was slightly higher for GATK. In short, we believe that all variant calling strategies show a similar recovery rate for validated SNPs and SNP chip variants, with a false negative rate for these common variants likely being less than 15%.

Additionally, there are two individuals that had two targeted exon samples (NA18637: SRR013654 + SRR013709, and NA18510: SRR017908 + SRR018122). If we treat these samples as technical replicates, then this provides another method to access the reproducibility of the variant calls. The SNP concordance between samples was clearly higher for NA18637 than NA18510 (Figure S5A). These trends hold true when only coding variants are considered (Figure S5B) or only coding variants within targeted genes are considered (Figure S5C), although the level of concordance increases when using a more focused set of SNPs. Similar statistics were provided for indels (Figure S6), but the sample size was too small to compare indels with more focused variant sets. It is unlikely that one NA18510 sample was simply mislabelled: the concordance between exome and targeted exon samples was more similar (Table S8 and S9), and we would expect worse concordance between two random individuals. One possible factor is that the percent duplicates was similar for both NA18637 samples but SRR018122 has twice as many duplicates as SRR017908 (Table S1), and in turn the total number of unique, on-target reads was much more similar for the NA18637 samples than the NA18510 samples. Also, it is worth noting that the concordance of indel calls was better for the GATK UnifiedGenotyper (and usually VarScan) than the GATK HaplotypeCaller (Figure S7), and this is true for both samples. This may complement the observation that GATK HaplotypeCaller produces a larger number of unknown frequency variants (Figure S4) and it has been reported that the GATK HaplotyperCaller indels have a lower validation rate than the GATK UnifiedGenotyper indels (Lescai et al. 2014).

Overlap of High-Quality GATK and VarScan Variant Calls

Given the previous comparisons, we believe the highest quality variant calls can be made by excluding GATK variants with LowQual flags (e.g. UG-HQ and HC-HQ) and using the conservative parameters defined in this manuscript when running VarScan (e.g. VarScan-Cons). Because the number of variants called increases when running the indel realignment and base recalibrator steps, we think the most useful variant lists to compare between variant callers is the “full pipeline” coding variants. In the 1000 Genomes samples, the variant caller concordance is quite high (Figure 7 and Figure S7). For example, all SNPs called using VarScan-Cons were also called using the GATK HaplotypeCaller. Additionally, all indels called using VarScan-Cons were also called by either the GATK HaplotypeCaller or the GATK UnifiedGenotyper. Because only chr20 variants can be reported for the 1000 Genomes dataset in this manuscript, it may also be useful to see the overlap for the genome-wide variant lists from SRP019719 (Figure 7). This time, the VarScan-Cons variants were not completely recovered by using one or both GATK variant callers, but there was still only a minority of variants that were not represented in either list of GATK variants (2.9% of VarScan-Cons SNPs and 1.5% of VarScan-Cons indels). This emphasizes that VarScan-Cons likely calls robust, reliable variants.

The high recovery of VarScan-Cons variants may result from high specificity with decreased sensitivity, which is potentially the biggest drawback to this strategy. The degree to which VarScan-Cons recovers the set of high-quality GATK variants varies between datasets. Among the 1000 Genome variants on chromosome 20, 87% of UnifiedGenotyper SNPs and 88% of HaplotypeCaller SNPs were also called by VarScan-Cons in the exome samples. For the targeted exon samples, 62% of UnifiedGenotyper SNPs and 68% of HaplotypeCaller SNPs were called by VarScan-Cons in the targeted exon samples, but this lower percentage may be due to the smaller number of SNPs and/or lower on-target coverage (Table S1-S3). For indels in the 1000 Genomes exome dataset, 69% of UnifiedGenotyper SNPs but only 38% of HaplotypeCaller SNPs were also called by VarScan-Cons. We would expect indel calls to have a higher false positive rate, but this could also be due to the small number of coding indels on chromosome 20 as well as the potentially higher false-positive rate for the HaplotypeCaller (Figure S4). Accordingly, the SRP019719 exome samples show similar statistics for SNP concordance (84% recovery for UnifiedGenotyper SNPs, 86% recovery for HaplotyperCaller SNPs) but much better results for UnifiedGenotyper indels (81% recovery). There is still only 51% recovery for HaplotypeCaller indels among VarScan-Cons indels, but it is possible that the indels uniquely called by GATK have a higher false positive rate. This has been previously reported for loss-of-function indels (Lescai et al. 2014). Either way, it appears that VarScan-Cons produces a reliable set of SNP calls.

Application of QC Metrics to Other Variant Callers

Although we were primarily interesting in comparing VarScan (with various sets of parameters) to GATK, we also examined similar metrics for other variant callers. Given that the distribution of rare and unknown variant frequencies was similar for the 1000 Genomes targeted exon and exome datasets, we used the targeted exon samples to test additional variant callers. In particular, we tested freebayes (DePristo et al. 2011) and samtools (Li et al. 2009), directly using the BWA alignment (with duplicates removed) as well as the "full pipeline" (GATK indel realignment + quality score recalibration). It is clear that freebayes yields too many unknown frequency variants (Figure S8), so it was not tested on any further datasets. Of course, changes in parameters and/or downstream filtering of variants may result in a decrease in the unknown variant frequency; however, we were using these two lists of variants (called with default parameters) as examples of how the quality control metrics in this dataset could be applied more broadly. Interestingly, the pre-processing steps appeared to have minimal effect on the 1000 Genomes targeted exon and exome datasets, but running the "full pipeline" for pre-processing considerably decreased the proportion of unknown frequency variants for samtools variant calls in the SRP019719 exome dataset (Figure S8).

SNP overlap is also shown for these four sets of variant calls (Figure S9). For the exome datasets, adding VarScan-Cons only reduced the number of variants otherwise called by GATK UnifiedGenotyper, GATK HaplotypeCaller, and samtools by ~10%. In other words, we would expect that variant calling with VarScan-Cons to have sensitivity similar to requiring variants to be called by GATK UnifiedGenotyper, GATK HaplotypeCaller, and samtools.

The ANNOVAR frequency plots (like those shown in Figure 3, Figure S3, and Figure S8) are a useful metric, but they only show average trends across the dataset and do not show how the frequency of unknown variants compares to the total number of variants called. If the number of conditions is reduced to only those using the "full pipeline" for pre-processing, then one can visualize the comparison between variant callers (Figure 8). There is more variability in the proportion of unknown frequency variants for the targeted exon dataset (compared to either exome dataset), but we suspect this may relate to the lower number of total SNPs and/or the lower on-target coverage (Table S1-S3). While it is not necessarily safe to assume that the false positive rate substantially varies when the proportion of unknown frequency variants is between 1 and 4% (which is the observed range for these variant callers in the exome datasets), the proportion of unknown frequency variants is universally low for all VarScan-Cons calls in all three datasets while there is more variability in the proportion of unknown frequency variants in the GATK and samtools variant calls.

DISCUSSION

Although running VarScan with default parameters (with the functions defined in the Methods section) was shown to result in an unacceptably high false positive rate (in accordance with a previous publication, Cheng et al. 2014; also see Figure 3 and Figure S3), running VarScan with a conservative set of non-standard parameters (referred to as VarScan-Cons in this study) can produce a reliable set of variants that should show a high validation rate (especially for SNPs). Almost all variants called with VarScan-Cons were also called using the GATK HaplotypeCaller or GATK UnifiedGenotyper, with a potential modest decrease in sensitivity for SNPs (Figure 6). GATK HaplotypeCaller did call a substantial number of indels not called using VarScan-Cons (as well as GATK UnifiedGenotyper), even after removing variants that were flagged as low quality. However, it is possible that this set of indels also has an increased false positive rate (as reported in Lescai et al. 2014; also see Figure S4 and S7), so it will be important to see if this is independently validated in other studies.

Use of GATK (both in terms of variant calling as well as pre-processing) can considerably increase the run-time for analysis (Figure 1 and Figure S1), especially when applying HaplotypeCaller on samples with high-coverage reads covering a significant proportion of the genome. Of course, run-times will vary based upon computational resources. If users only have a limited number of samples to analyze, cloud-based solutions may be more efficient than analyzing the data locally (Asmann et al. 2012; Fischer et al. 2012; Karczewski et al. 2014; Langmead et al. 2009). Additionally, parallelizing jobs by chromosome could also assist with decreasing the run-time for certain tasks. Also, the newest version of GATK is should decrease the runtime for GATK HaplotypeCaller (GATK 3.0 Features Overview).

This study focused on variant calling in (most likely) normal human samples due to the availability of a large amount of public validation data. However, the strategies described in this study may not apply equally well in all circumstances. For example, sometimes variants may be present in a minority of cells in a sample (such as a heterogeneous tumor), and it may not be safe to make assumptions about the ploidy of the sample (which might affect the usefulness of the GATK HaplotypeCaller, for example). Likewise, somatic variant calling is also an area of interest (Wang et al. 2013) that would typically utilize different variant calling tools. Additionally, the indel metrics in this study only apply to small indels: large indels and structural variants require a different set of algorithms. Nevertheless, we think the strategies described in this study (comparing the proportion of unknown frequency variants, using filters for high-quality variants, etc.) can be useful to help other scientists prioritize variant calling strategies for their own data.

ACKNOWLEDGEMENTS

We would like to thank Laura Clarke for providing technical assistance with the 1000 Genomes data, Jason O'Rawe and Gholson Lyon for answering questions about the O'Rawe et al. 2013 exome dataset and providing access to the Human610 SNP chip data, and Illumina technical assistance for answering questions about the Illumina SNP chips.

REFERENCES

Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, and Sunyaev SR. 2010. A method and server for predicting damaging missense mutations. *Nat Meth* 7:248-249.

Asmann YW, Middha S, Hossain A, Baheti S, Li Y, Chai H-S, Sun Z, Duffy PH, Hadad AA, Nair A, Liu X, Zhang Y, Klee EW, Kalari KR, and Kocher J-PA. 2012. TREAT: a bioinformatics tool for variant annotations and visualizations in targeted and exome sequencing data. *Bioinformatics* 28:277-278.

Bauer D. 2011. Variant calling comparison CASAVA1.8 and GATK. *Nature Precedings* http://dx.doi.org/10.1038/npre.2011.6107.1.

Blankenberg D, Kuster GV, Coraor N, Ananda G, Lazarus R, Mangan M, Nekrutenko A, and Taylor J. 2001. Galaxy: A Web-Based Genome Analysis Tool for Experimentalists. *Current Protocols in Molecular Biology*: John Wiley & Sons, Inc.

Boland J, Chung C, Roberson D, Mitchell J, Zhang X, Im K, He J, Chanock S, Yeager M, and Dean M. 2013. The new sequencer on the block: comparison of Life Technology's Proton sequencer to an Illumina HiSeq for whole-exome sequencing. *Human Genetics* 132:1153-1163.

Carson A, Smith E, Matsui H, Braekkan S, Jepsen K, Hansen J-B, and Frazer K. 2014. Effective filtering strategies to improve data quality from population-based whole exome sequencing studies. *BMC Bioinformatics* 15:125.

Chen K, Wallis JW, McLellan MD, Larson DE, Kalicki JM, Pohl CS, McGrath SD, Wendl MC, Zhang Q, Locke DP, Shi X, Fulton RS, Ley TJ, Wilson RK, Ding L, and Mardis ER. 2009. BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nat Meth* 6:677-681.

Cheng AY, Teo Y-Y, and Ong RT-H. 2014. Assessing single nucleotide variant detection and genotype calling on whole-genome sequenced individuals. *Bioinformatics*.

DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, McKenna A, Fennell TJ, Kernytsky AM, Sivachenko AY, Cibulskis K, Gabriel SB, Altshuler D, and Daly MJ. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* 43:491-498.

Fischer M, Snajder R, Pabinger S, Dander A, Schossig A, Zschocke J, Trajanoski Z, and Stocker G. 2012. SIMPLEX: Cloud-Enabled Pipeline for the Comprehensive Analysis of Exome Sequencing Data. *PLoS ONE* 7:e41948.

Fu W, O/'Connor TD, Jun G, Kang HM, Abecasis G, Leal SM, Gabriel S, Altshuler D, Shendure J, Nickerson DA, Bamshad MJ, Project NES, and Akey JM. 2013. Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature* 493:216-220.

Garrison E, and Marth G. 2012. Haplotype-based variant detection from short-read sequencing. *arXiv preprint* arXiv:1207.3907 [q-bio.GN].

GATK 3.0 Features Overview. GATK 3.0 Features Overview. *Available at* *http://www.broadinstitute.org/gatk/blog?id=38172014)*.

Giardine B, Riemer C, Hardison RC, Burhans R, Elnitski L, Shah P, Zhang Y, Blankenberg D, Albert I, Taylor J, Miller W, Kent WJ, and Nekrutenko A. 2005. Galaxy: A platform for interactive large-scale genome analysis. *Genome Research* 15:1451-1455.

Goecks J, Nekrutenko A, Taylor J, and Team TG. 2010. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biology* 11:R86.

Karczewski KJ, Fernald GH, Martin AR, Snyder M, Tatonetti NP, and Dudley JT. 2014. STORMSeq: An Open-Source, User-Friendly Pipeline for Processing Personal Genomics Data in the Cloud. *PLoS ONE* 9:e84860.

Koboldt DC, Chen K, Wylie T, Larson DE, McLellan MD, Mardis ER, Weinstock GM, Wilson RK, and Ding L. 2009. VarScan: variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics* 25:2283-2285.

Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, Miller CA, Mardis ER, Ding L, and Wilson RK. 2012. VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Research* 22:568-576.

Krumm N, Sudmant PH, Ko A, O'Roak BJ, Malig M, Coe BP, NHLBI Exome Sequencing Project N, Quinlan AR, Nickerson DA, and Eichler EE. 2012. Copy number variation detection and genotyping from exome sequence data. *Genome Research*.

Lam HYK, Pan C, Clark MJ, Lacroute P, Chen R, Haraksingh R, O'Huallachain M, Gerstein MB, Kidd JM, Bustamante CD, and Snyder M. 2012. Detecting and annotating genetic variations using the HugeSeq pipeline. *Nat Biotech* 30:226-229.

Langmead B, Schatz M, Lin J, Pop M, and Salzberg S. 2009. Searching for SNPs with cloud computing. *Genome Biology* 10:R134.

Leinonen R, Akhtar R, Birney E, Bower L, Cerdeno-Tárraga A, Cheng Y, Cleland I, Faruque N, Goodgame N, Gibson R, Hoad G, Jang M, Pakseresht N, Plaister S, Radhakrishnan R, Reddy K, Sobhany S, Ten Hoopen P, Vaughan R, Zalunin V, and Cochrane G. 2011. The European Nucleotide Archive. *Nucleic Acids Research* 39:D28-D31.

Lescai F, Marasco E, Bacchelli C, Stanier P, Mantovani V, and Beales P. 2014. Identification and validation of loss of function variants in clinical contexts. *Molecular Genetics & Genomic Medicine* 2:58-63.

Li H, and Durbin R. 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* 25:1754-1760.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, and Subgroup GPDP. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25:2078-2079.

Li J, Doyle MA, Saeed I, Wong SQ, Mar V, Goode DL, Caramia F, Doig K, Ryland GL, Thompson ER, Hunter SM, Halgamuge SK, Ellul J, Dobrovic A, Campbell IG, Papenfuss AT, McArthur GA, and Tothill RW. 2014. Bioinformatics Pipelines for Targeted Resequencing and Whole-Exome Sequencing of Human and Mouse Genomes: A Virtual Appliance Approach for Instant Deployment. *PLoS ONE* 9:e95217.

Linderman M, Brandt T, Edelmann L, Jabado O, Kasai Y, Kornreich R, Mahajan M, Shah H, Kasarskis A, and Schadt E. 2014. Analytical validation of whole exome and whole genome sequencing for clinical applications. *BMC Medical Genomics* 7:20.

Liu X, Han S, Wang Z, Gelernter J, and Yang B-Z. 2013. Variant Callers for Next-Generation Sequencing Data: A Comparison Study. *PLoS ONE* 8:e75619.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, and DePristo MA. 2010. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research* 20:1297-1303.

Ng PC, and Henikoff S. 2003. SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Research* 31:3812-3814.

O'Rawe J, Jiang T, Sun G, Wu Y, Wang W, Hu J, Bodily P, Tian L, Hakonarson H, Johnson WE, Wei Z, Wang K, and Lyon G. 2013. Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing. *Genome Medicine* 5:28.

Pabinger S, Dander A, Fischer M, Snajder R, Sperk M, Efremova M, Krabichler B, Speicher MR, Zschocke J, and Trajanoski Z. 2014. A survey of tools for variant analysis of next-generation genome sequencing data. *Briefings in Bioinformatics* 15:256-278.

Sherry ST, Ward M-H, Kholodov M, Baker J, Phan L, Smigielski EM, and Sirotkin K. 2001. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Research* 29:308-311.

The 1000 Genomes Project Consortium. 2012. An integrated map of genetic variation from 1,092 human genomes. *Nature* 491:56-65.

Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, del Angel G, Levy-Moonshine A, Jordan T, Shakir K, Roazen D, Thibault J, Banks E, Garimella KV, Altshuler D, Gabriel S, and DePristo MA. 2002. From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. *Current Protocols in Bioinformatics*: John Wiley & Sons, Inc.

Wang K, Li M, and Hakonarson H. 2010. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research* 38:e164.

Wang Q, Jia P, Li F, Chen H, Ji H, Hucks D, Dahlman K, Pao W, and Zhao Z. 2013. Detecting somatic point mutations in cancer genome sequencing data: a comparison of mutation callers. *Genome Medicine* 5:91.

Worthey EA. 2013. Analysis and Annotation of Whole-Genome or Whole-Exome Sequencing–Derived Variants for Clinical Diagnosis. *Current Protocols in Human Genetics*: John Wiley & Sons, Inc.

Yi M, Zhao Y, Jia L, He M, Kebebew E, and Stephens RM. 2014. Performance comparison of SNP detection tools with illumina exome sequencing data—an assessment using both family pedigree information and sample-matched SNP array data. *Nucleic Acids Research*.

Yu X, and Sun S. 2013. Comparing a few SNP calling algorithms using low-coverage sequencing data. *BMC Bioinformatics* 14:274.

FIGURE LEGENDS

Figure 1: Run Times for Variant Calling Pipelines for GATK versus VarScan (1KG) A. Run times for selected 1000 Genomes targeted exon samples (n=14, Table S1). Run-times are displayed for the entire variant calling pipeline when using both GATK indel realignment and quality score recalibration (“Full Pipeline” - purple), indel realignment only (“Realign Only” - red), quality score recalibration only (“Recalibrate Only” - green), or neither (“No Preprocess” – blue). Run times are provided for the GATK UnifiedGenotyper (“UG”), GATK HaplotypeCaller (“HC”), and VarScan using 3 sets of parameters (see Methods). “VarScan-Cons” is the most conservative set of parameters. **B.** Same as A, but for selected 1000 Genomes exome samples (n=12, Table S2).

Figure 2: Number of Variants Called for GATK versus VarScan (1KG – chr20) A. Number of SNPs called for selected 1000 Genomes (1KG) targeted exon samples (n=14, left) and exome samples (n=12, right). The number of SNPs called is displayed for variants called using both GATK indel realignment and quality score recalibration (“Full Pipeline” - purple), indel realignment only (“Realign Only” - red), quality score recalibration only (“Recalibrate Only” - green), or neither (“No Preprocess” – blue). Variant counts are provided for the GATK UnifiedGenotyper (“UG”), GATK HaplotypeCaller (“HC”), and VarScan using 3 sets of parameters (see Methods). UnifiedGenotyper

and HaplotypeCaller variants are then divided into the set of all variants ("UG-all" and "HC-all") and higher-quality variant calls where variants flagged as low quality have been removed ("UG-HQ" and "HC-HQ"). "VarScan-Cons" is the most conservative set of parameters for VarScan. These are the total number of SNPs called (not just coding SNPs or SNPs within targeted regions). **B.** Same as A, for indels instead of SNPs.

Figure 3: Distribution of ANNOVAR Annotations for Coding SNP Variants (1KG – chr20) A. Distribution of variant types defined in the ANNOVAR exome report for selected 1000 Genomes (1KG) targeted exon samples (n=14). Variants are classified based upon population frequency and damaging prediction (see Methods). Low frequency variants (MAF < 0.01) that are displayed in orange if they are predicted to be damaging and are displayed in green if they are not predicted to be damaging. Unknown frequency variants are displayed in red if they are predicted to be damaging and are displayed in blue if they are not predicted to be damaging. Although all samples should contain some unknown frequency variants, a high proportion of unknown frequency variants are expected to correlate with a high false positive rate (for example, ideally, you would expect fewer unknown frequency variants than low frequency variants). Seven variant calling strategies were tested (GATK UnifiedGenotyper and HaplotypeCaller, with and without filtering low quality variants; VarScan with 3 sets of parameters, see Methods). "VarScan-Cons" is the most conservative set of parameters for VarScan. Each variant caller was also tested with 4 preprocessing conditions, corresponding to the colored boxes under the bar plot: variants called using both GATK indel realignment and quality score recalibration (purple), indel realignment only (red), quality score recalibration only (green), or neither (blue). **B.** Same as A, but for selected 1000 Genomes exome samples (n=12, Table S2).

Figure 4: Recovery of 1KG Validated SNPs (chr20) A pooled set of 35 validated variant from the 1000 Genomes exome sample characterized in this study (n=12) was used to assess the sensitivity of various variant calling algorithms. The 1000 Genomes targeted exon samples were not compared because no validated variants were covered in targeted regions for that design. Seven variant calling strategies were tested (GATK UnifiedGenotyper and HaplotypeCaller, with and without filtering low quality variants; VarScan with 3 sets of parameters, see Methods). "VarScan-Cons" is the most conservative set of parameters for VarScan. Each variant caller was also tested with 4 preprocessing conditions: variants called using both GATK indel realignment and quality score recalibration ("Full

Pipeline" - purple), indel realignment only ("Realign Only" - red), quality score recalibration only ("Recalibrate Only" - green), or neither ("No Preprocess" – blue). Publically available validated variants are only available for chr20, so this is the maximum number of validated SNPs that can be characterized for these samples (in fact, these samples were selected based upon their ability to cover a maximum number of validated SNPs). The validation status for each individual SNP under each variant calling condition is provided in Table S7. Validated variants were never called for chr20:3193991 for individual NA18505 (exome sample SRX237141, covered 81x with the reference allele in all reads) or for chr20:57769739 for individual NA18532 (exome sample ERR031956, not covered by any reads but located in the coding sequence for ZNF831)

Figure 5: Recovery of Coding SNPs from Targeted Exon Samples (1KG Exome – chr20) SNP calls from paired targeted exon and exome datasets were compared to test the robustness of the calls made in the targeted exon data. Indel calls are not presented because there are practically no coding indels on chr20 for the targeted exon datasets. Two subjects has two targeted exon datasets (Table S1-S2), and concordance with exome steps was reported for both targeted exon datasets separately (resulting in 14 concordance values per variant calling strategy). Seven variant calling strategies were tested (GATK UnifiedGenotyper and HaplotypeCaller, with and without filtering low quality variants; VarScan with 3 sets of parameters, see Methods). "VarScan-Cons" is the most conservative set of parameters for VarScan. Each variant caller was also tested with 4 preprocessing conditions: variants called using both GATK indel realignment and quality score recalibration ("Full Pipeline" - purple), indel realignment only ("Realign Only" - red), quality score recalibration only ("Recalibrate Only" - green), or neither ("No Preprocess" – blue). Concordance is reported as recovery of SNPs called in the targeted exon data, but these cannot be treated as "gold standard" variant calls. Most clearly, there was a high false positive rate when running VarScan with default parameters, so a high proportion of those variants called in the targeted exon samples could not be recovered in the exome dataset. In fact, on-target coverage is typically lower for the targeted exon samples than the exome staples (Table S1-S2)

Figure 6: Recovery of SNP Chip Variants in Exome Samples A. Recovery of variants called on the Omni 1KG for 1000 Genomes exome samples (n=12). Only alleles that varied from the reference sequence and were located within the regions targeted in the exome sequencing design were

considered for this analysis (this varied to some extent between samples, between a minimum of 477 variants and a maximum of 616 variants). Seven variant calling strategies were tested (GATK UnifiedGenotyper and HaplotypeCaller, with and without filtering low quality variants; VarScan with 3 sets of parameters, see Methods). “VarScan-Cons” is the most conservative set of parameters for VarScan. Each variant caller was also tested with 4 preprocessing conditions: variants called using both GATK indel realignment and quality score recalibration (“Full Pipeline” - purple), indel realignment only (“Realign Only” - red), quality score recalibration only (“Recalibrate Only” - green), or neither (“No Preprocess” – blue). **B.** Same as A, but for 6437 variants on a different SNP chip design (Human 610) compared to exome variant calls for sample SRX265476.

Figure 7: Exome Variant Caller Overlap (Coding Variants) All coding variants were tabulated for all exome samples (1KG n=12, left; SRP019719 n=15, right), keeping track of the samples in which the variants were called. Only coding variants on chromosome 20 were considered for the 1000 Genomes (1KG) samples, but all coding variants were considered for the SRP019719 samples. In order to simplify presentation of these results, we focused on the highest quality variant calls for each variant calling strategy: GATK UnifiedGenotyper with low-quality variants removed (UG-HQ, blue), GATK HaplotypeCaller with low-quality variants removed (HC-HQ, green), and VarScan using a custom set of conservative parameters (VarScan-Cons, red). Similarly, only variants subject to GATK indel realignment and quality score recalibration (“Full Pipeline”) are considered for this comparison. To show the different concordance rates, SNPs are presented at the figure and indels are presented at the bottom of the figure. Almost all VarScan-Cons variants were also called by GATK (either HaplotypeCaller or UnifiedGenotyper). All three variant callers called a similar number of SNPs, but GATK HaplotyperCaller called more indels than either GATK UnifiedGenotyper or VarScan-Cons.

Figure 8: QC Metrics to Estimate Specificity Versus Sensitivity for Variant Callers (Coding SNPs) For each of the three datasets characterized in this study (1KG targeted exon, n=14; 1KG exome, n=12; SRP019719 exome, n=15), the number of SNPs called per sample is plotted along the x-axis and the proportion of unknown frequency variants is plotted on the y-axis. In order to simplify presentation of these results, we focused on the highest quality variant calls for each variant calling strategy: GATK UnifiedGenotyper with low-quality variants removed (UG-HQ, blue), GATK HaplotypeCaller with low-quality variants removed (HC-HQ, green), and VarScan using a custom set

of conservative parameters (VarScan-Cons, red). Additionally, an unfiltered set of variants called via samtools are plotted in black. Only variants subject to GATK indel realignment and quality score recalibration ("Full Pipeline") are considered for this comparison. The shape of the data point corresponds to the depth of on-target coverage: <50x coverage is represented as an X in an open-circle, 50-100x is represented as an open circle, and >100x is represented as a filled circle. If the unknown frequency percentage was tightly correlated with the actual false positive rate and the number of variants was tightly correlated with the actual sensitivity of the variant caller, than the ideal variant caller would show a cluster of data points in the bottom-right hand corner of the plot.

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.403v2 | CC-BY 4.0 Open Access | received: 6 Jun 2014, published: 6 Jun 2014

Table 1: Parameter Settings for VarScan Comparisons

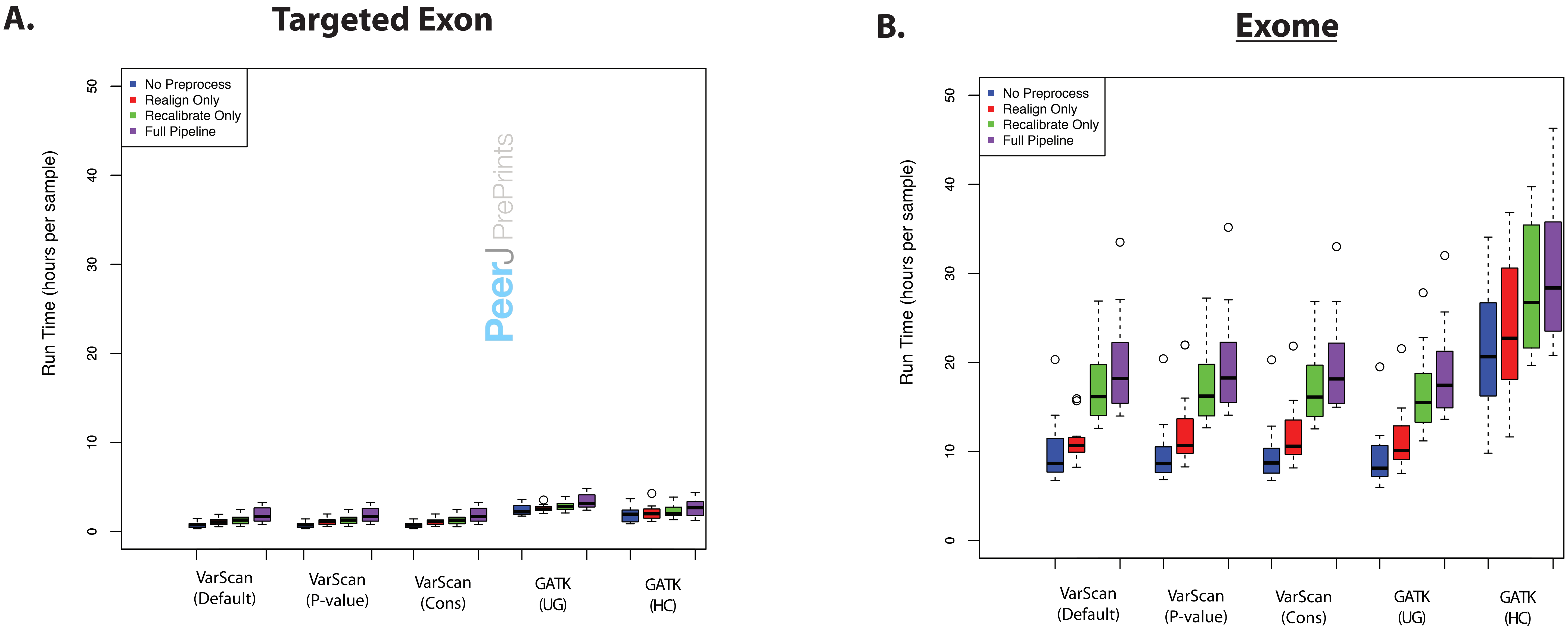
	VarScan-Default	VarScan-Pvalue	VarScan-Cons
Minimum Coverage	8	8	10
Minimum Supporting Reads	2	2	4
Minimum Average Quality	15	15	20
Minimum Variant Frequency	0.01	0.01	0.3
Minimum P-Value	0.99	0.05	0.99

Table 2: Average Run Times for Variant Calling Step for Various Variant Callers

	1KG Targeted Exon (n=14)	1KG Exome (n=12)	SRP019719 Exome (n=15)
VarScan	0:10	2:14	1:52
GATK UnifiedGenotyper	1:55	4:32	4:10
GATK HaplotypeCaller	1:16	15:44	11:27
samtools	0:26	7:04	4:51
freebayes	0:43	▨	▨

Average run-times for are “Full Pipeline” variants (with indel realignment and/or quality score recalibration). VarScan run-time is for VarScan using the default setting. 1KG = selected 1000 genomes project samples.

Figure 1: Run Times for Variant Calling Pipelines for GATK versus VarScan (1KG)



PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.403v2 | CC-BY 4.0 Open Access | received: 6 Jun 2014, published: 6 Jun 2014

Figure 2: Number of Variants Called for GATK versus VarScan (1KG - chr20)

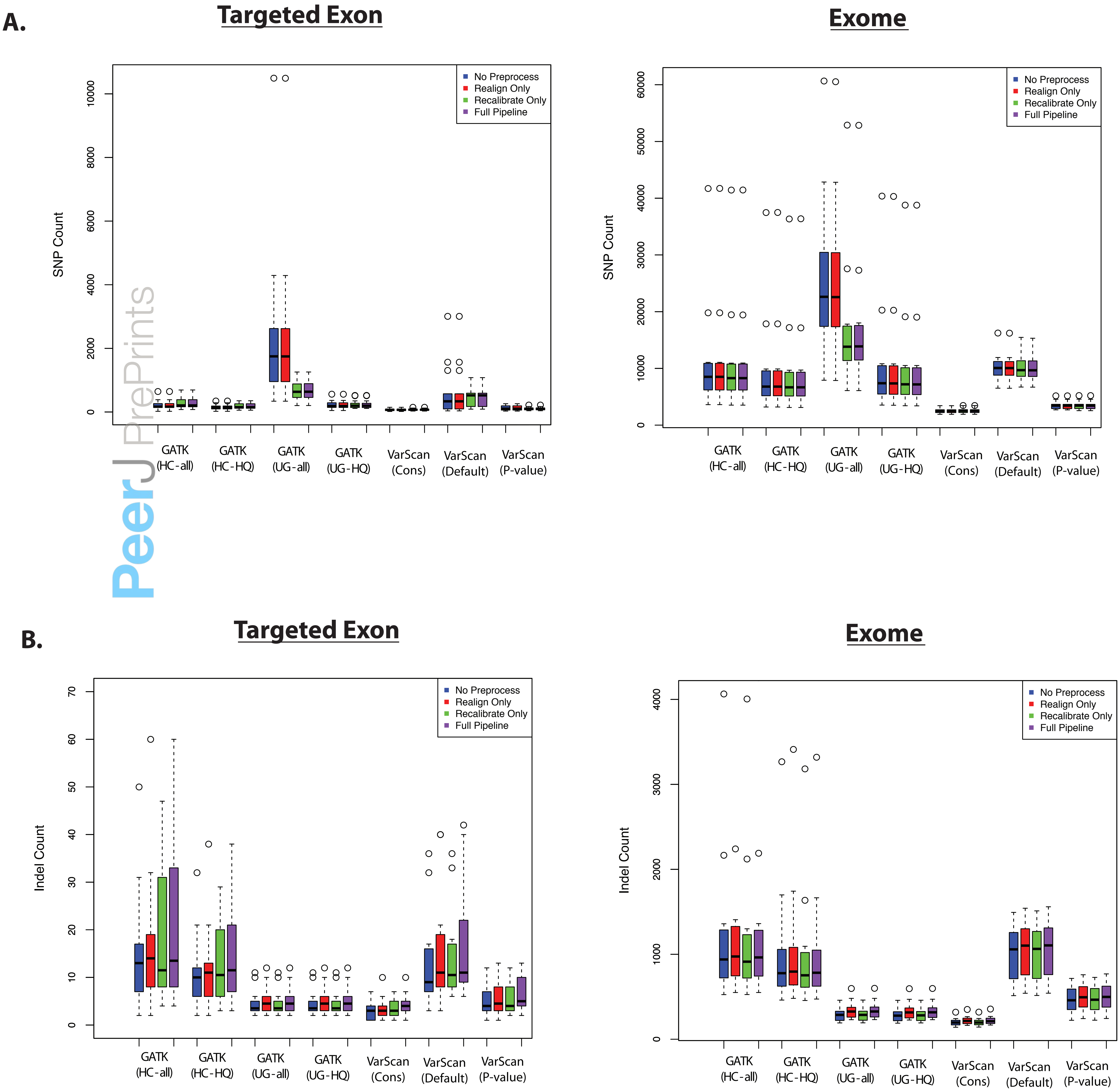
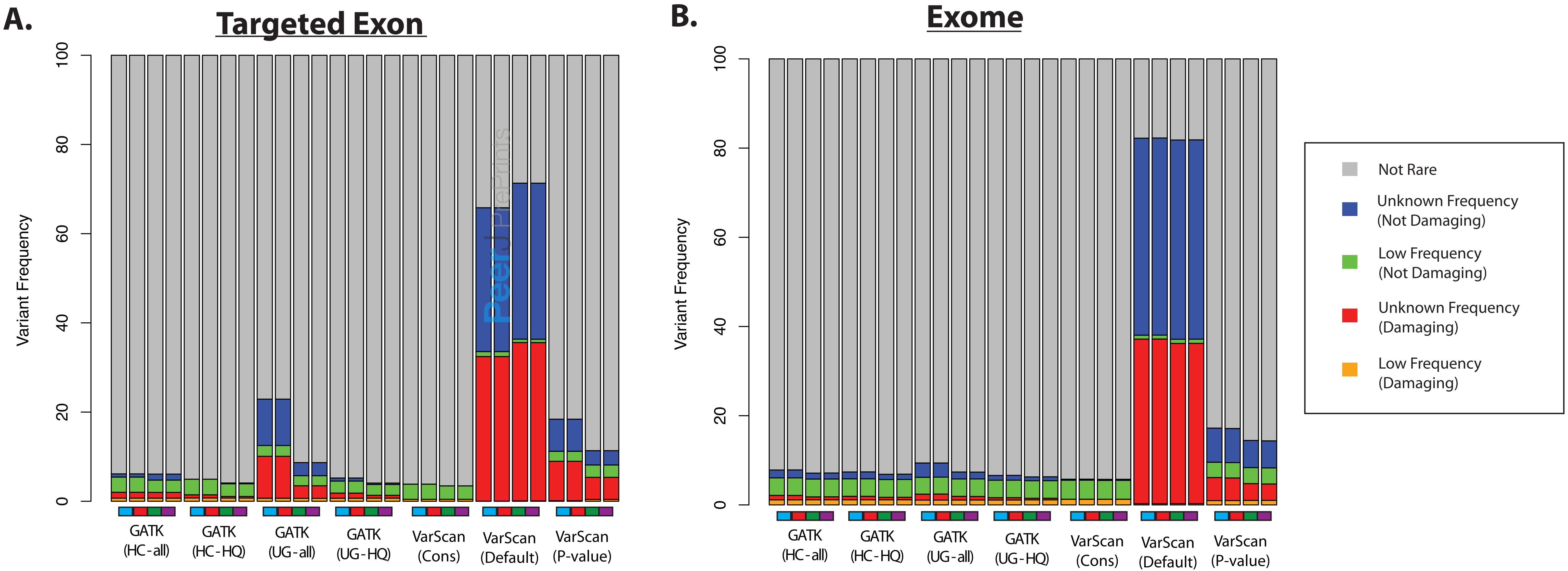
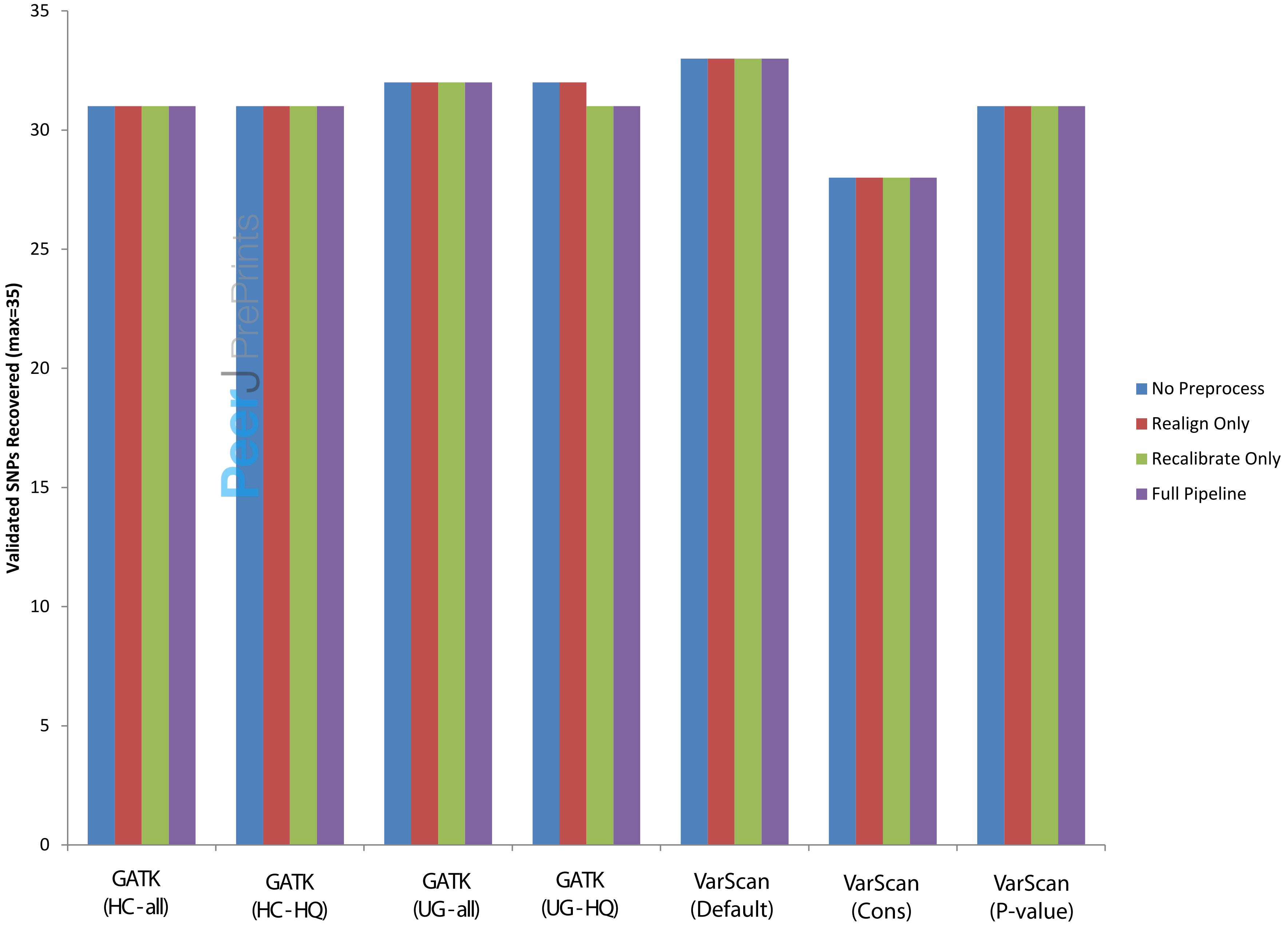


Figure 3: Distribution of ANNOVAR Annotations for Coding SNP Variants (1KG - chr20)



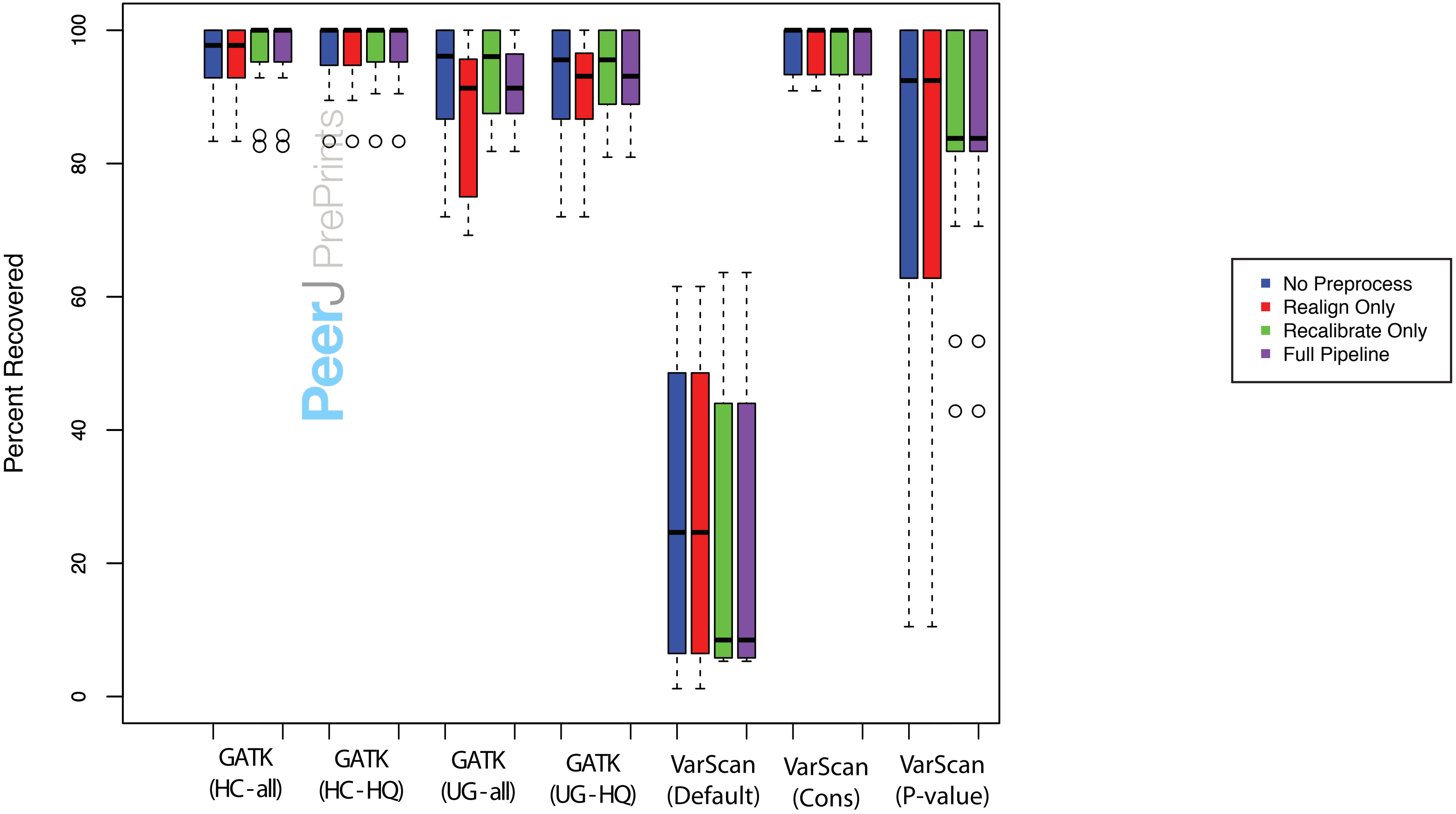
PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.403v2 | CC-BY 4.0 Open Access | received: 6 Jun 2014, published: 6 Jun 2014

Figure 4: Recovery of 1KG Validated SNPs (chr20)



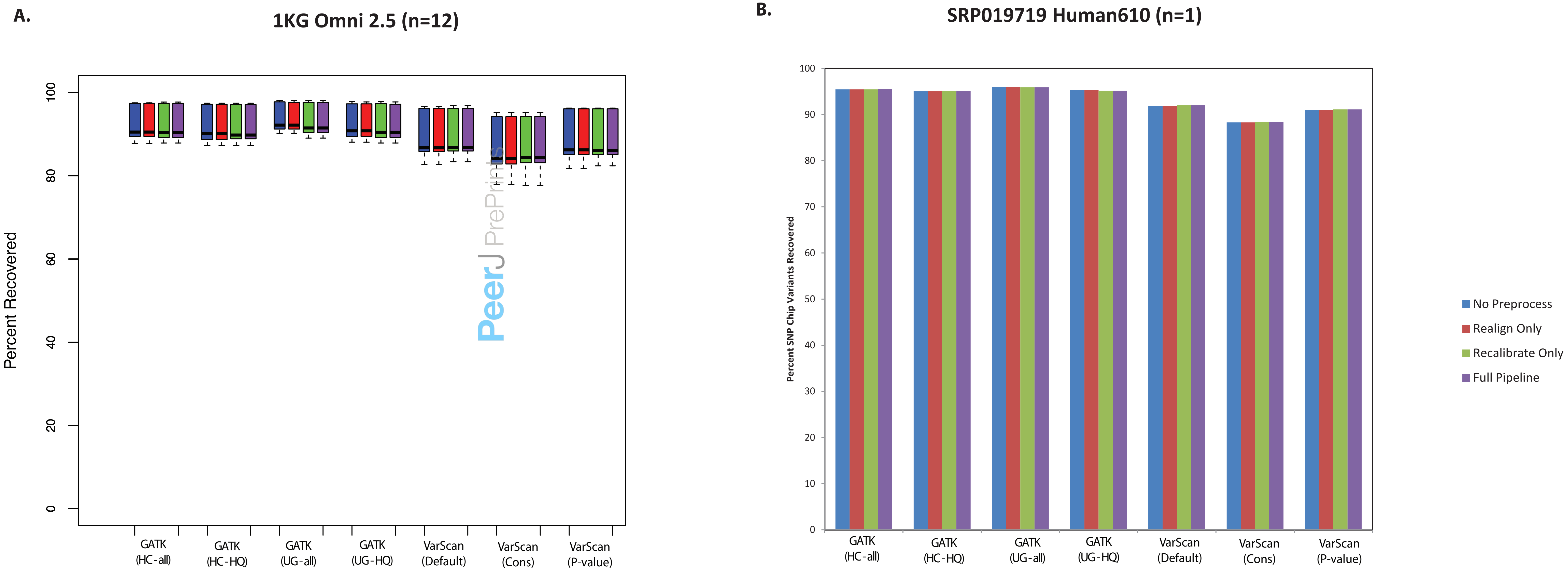
PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.403v2 | CC-BY 4.0 Open Access | received: 6 Jun 2014, published: 6 Jun 2014

Figure 5: Recovery of Coding SNPs from Targeted Exon Samples (1KG Exome - chr20)



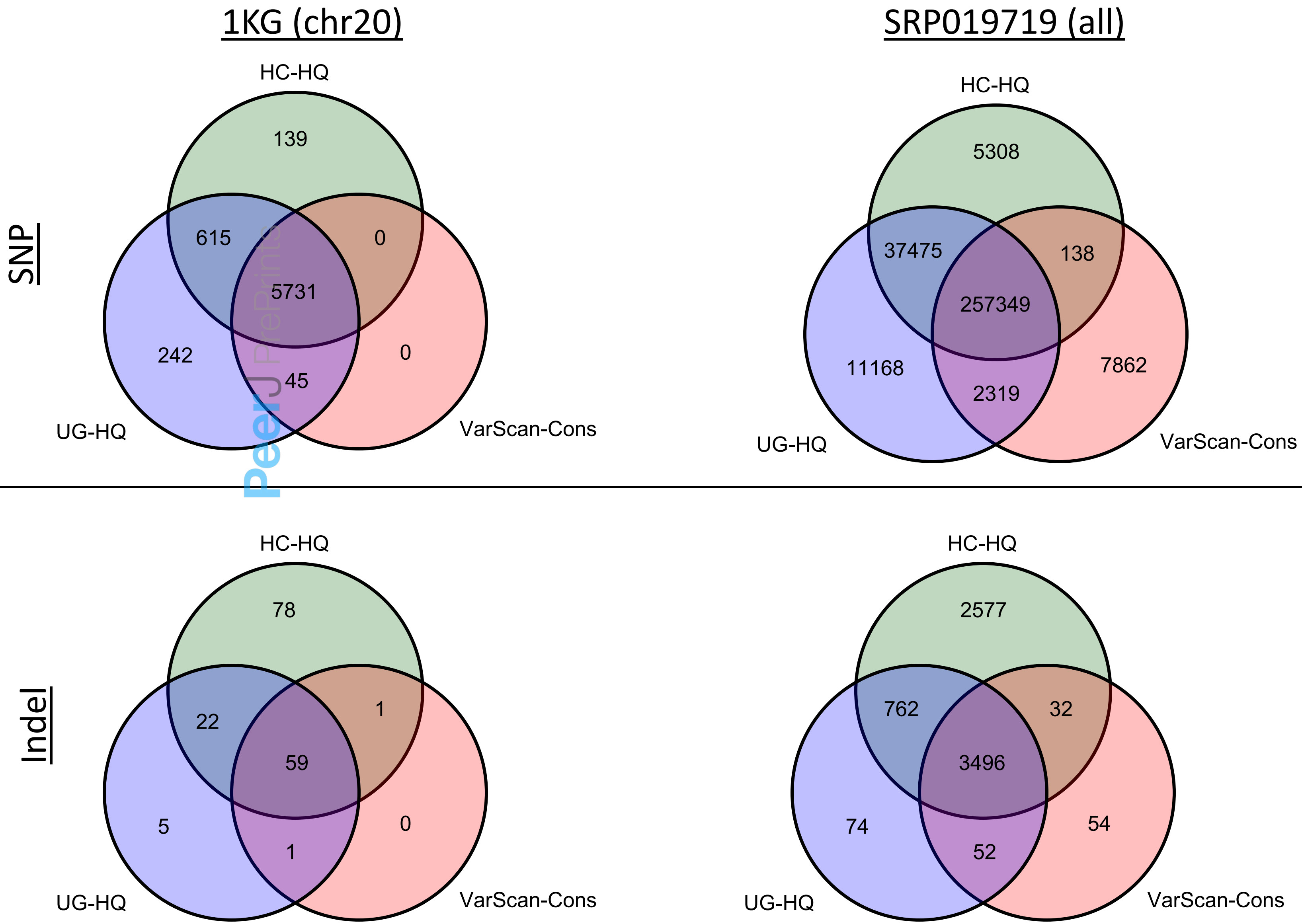
PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.403v2 | CC-BY 4.0 Open Access | received: 6 Jun 2014, published: 6 Jun 2014

Figure 6: Recovery of SNP Chip Variants in Exome Samples



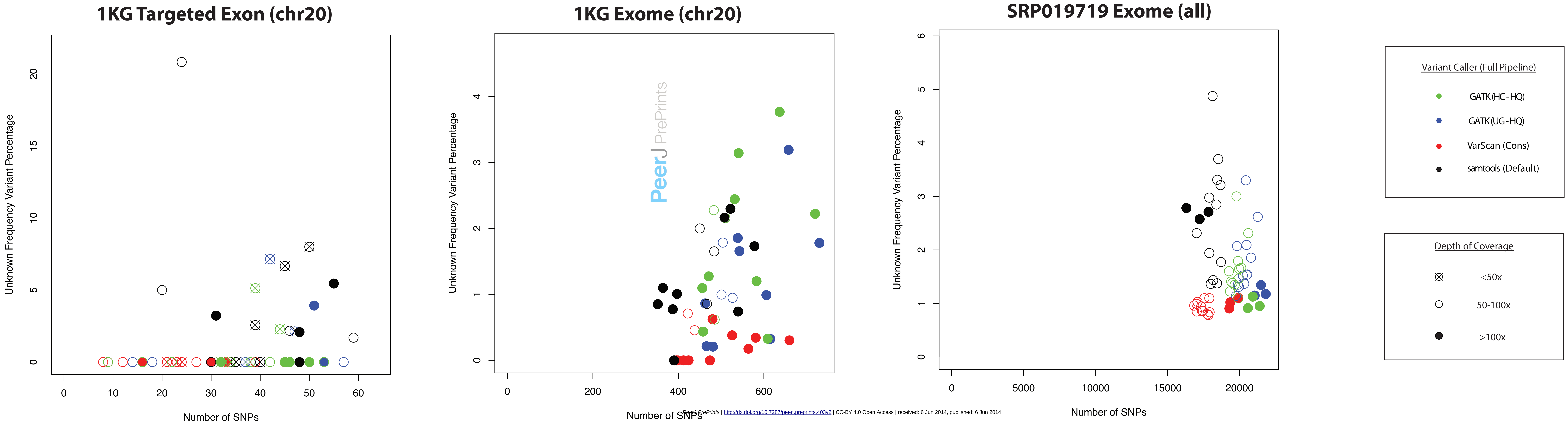
PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.403v2 | CC-BY 4.0 Open Access | received: 6 Jun 2014, published: 6 Jun 2014

Figure 7: Exome Variant Caller Overlap (Coding Variants)



PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.403v2 | CC-BY 4.0 Open Access | received: 6 Jun 2014, published: 6 Jun 2014

Figure 8: QC Metrics to Estimate Specificity Versus Sensitivity for Variant Callers (Coding SNPs)



PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.403v2 | CC-BY 4.0 Open Access | received: 6 Jun 2014, published: 6 Jun 2014