Title

A Rigorous Interlaboratory Examination of the Need to Confirm NGS-Detected Variants by an Orthogonal Method in Clinical Genetic Testing

Authors

Stephen Lincoln [1]
Rebecca Truty [1]
Chiao-Feng Lin [2,3]
Justin M Zook [4]
Joshua Paul [1]
Vince Ramey [1]
Marc Salit [4,5]
Heidi Rehm [2,3,6,7,8]
Robert Nussbaum [1,9]
Matthew Lebo [2,3,7,8]

1. Invitae, San Francisco, CA
2. Partners HealthCare Laboratory for Molecular Medicine, Cambridge, MA
3. Brigham and Women's Hospital, Boston, MA
4. National Institute of Standards and Technology, Gaithersburg, MD
5. Joint Initiative for Metrology in Biology, Stanford, CA
6. Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA
7. Harvard Medical School, Boston, MA
8. The Broad Institute of MIT and Harvard, Cambridge, MA
9. Department of Medicine, University of California San Francisco, San Francisco, CA

Corresponding Author

Stephen E. Lincoln
1400 16th Street, San Francisco, CA, 94103, USA
Mobile: 301-312-1725
email: steve.lincoln@me.com

Contents - This Document

Abstract
Main Text
- Introduction
- Methods
- Results
- Discussion
Acknowledgements
Disclosures
Oversight
Main Figure Legends
References
Main Figures 1-4
Main Table 1

Contents - Other Documents

Supplemental Figures 1-16
Supplemental Figure Legends
Supplemental Tables 1,2

Abstract

Background: The confirmation of genetic variants identified by next-generation sequencing (NGS) using orthogonal assays (e.g., Sanger sequencing) is standard practice in many laboratories. Published studies have examined this issue, concluding that confirmation of the highest-quality NGS calls may not always be necessary. However, these studies are generally small, omit statistical justification, and explore limited aspects of the underlying data. Defining criteria that separate high accuracy NGS calls that do not benefit from confirmation from those that do remains a critical and pressing issue.

Methods: We conducted a rigorous examination of NGS data from two clinical laboratories. Five Genome in a Bottle reference samples and over 80,000 clinical patient specimens were analyzed, with the combined data providing insights that neither data type alone could provide. In total, almost 200,000 variant calls with orthogonal data were examined including 1684 false–positives detected by confirmation.

Results: We used these data to identify criteria that flag 100% of false positives as requiring confirmation (CI lower bound: 98.5 to 99.8% depending on variant type) while minimizing the number of flagged true positives. Rather than relying on one or two quality metrics, as current publications do, a battery of criteria proved superior, consistent with the metrics recommended by recent practice guidelines. Indeed, our expanded criteria identify some false positives as requiring confirmation that the currently published criteria miss. We also find that historical performance (observing a variant as a true positive some number of times) without other strict quality criteria can also lead to false positives escaping confirmation.

Discussion: Although we found limitations with the currently published criteria, our large, multi-laboratory study reaffirms prior findings that high accuracy variant calls can be separated from those of lower confidence. Our methodology for determining test and laboratory-specific criteria can be generalized into a practical approach that can help reduce the cost and time burden of confirmation without impacting clinical accuracy.

Introduction

The use of orthogonal assays (e.g., Sanger sequencing) to confirm variants identified by next-generation sequencing (NGS) is standard practice in many laboratories to reduce the risk of delivering false–positive (FP) results. Clinical NGS tests can inform significant medical decisions [1,2], and thus confirmation is recommended by medical practice guidelines [3,4] although the details are generally left up to the laboratory [3,5]. Because the methods used in clinical NGS often emphasize sensitivity (to avoid missing clinically important variants), FP rates can be elevated compared with those in research NGS [6]. Moreover, pathogenic variants are often technically challenging (e.g., many are located within repetitive sequences or segmental duplications), which can further increase FP rates [7–9]. Confirmation assays have a monetary cost, however, and can increase the time needed to deliver results, a critical factor in many clinical situations.

Published studies examining this issue have concluded that confirmation of the highest-quality NGS calls may not always be necessary [10–13]. Some of these studies [10,12] propose specific criteria to separate high-confidence true-positive (TP) variant calls from those that are possibly but not confidently true (Figure 1). These criteria differ from those used in filtering, the separate process of removing calls confidently believed to be false or unsupportable. The remaining intermediate-confidence calls are those that indeed benefit from confirmatory assays, manual data review, or both to determine which are TPs and which are FPs. Unfortunately, these prior studies are generally small, in some cases proposing quality thresholds using only a single example FP (Table 1, Figure 1). At first glance, the presence of few FPs may seem reassuring. However, this leads to significant limitations in these studies. First, because their statistical power to characterize the FP population is limited, these studies do not address the question of whether future FPs are likely to resemble the few observed in

the study. Quite possibly, additional FPs could be different and thus missed by the proposed criteria. Secondly, these studies use identical datasets for training and evaluating the proposed criteria, likely making the results subject to overfitting [14]. Few of these prior studies provide statistical justifications for their proposed criteria. Finally, these prior studies use data from individual laboratories and do not examine whether their methodologies could be generalized across settings.

We set out to examine the role of confirmation using a much larger set of variant calls than had been published previously. We combined five reference samples characterized by the Genome in a Bottle Consortium (GIAB) [15–17] with confirmatory data from over 80,000 clinical tests. Our methodology was applied in two clinical laboratories that use similar but not identical NGS methods. Like prior studies, we found it possible to identify high-confidence NGS calls that do not benefit from confirmation. However, we found that a battery of criteria was necessary to capture all FPs, in contrast with the one or two metrics used in prior studies. Indeed we found that the specific criteria proposed by prior studies would miss FPs in our, much larger, datasets. We also found that observations of a variant as a TP can say little about its chance of being an FP in a different run, which indicates that historical performance can be an ineffective quality metric. Approaches such as ours can be used by many laboratories to provide efficient, effective, and statistically justified criteria for prioritizing variant calls for confirmation.

Methods

We compiled eight component datasets from two laboratories (Table 1) following the process illustrated in Figure 1. Five GIAB specimens were used: NA12878, NA24385, NA24143, NA24149, and NA24631 (Coriell Institute, Camden, NJ) with reference data version 3.3.2 downloaded from www.genomeinabottle.org. Both laboratories' sequencing and bioinformatics methods have been described previously [8,18,19]. In brief, NGS was performed using Illumina (San Diego, CA) 2x150 base pair (bp) reads. Seven (Lab 1) and three (Lab 2) custom hybridization-based targeting assays were used, each targeting the exons of 100-1000 genes. Average coverage in targeted regions was minimally 300x and typically 500–1000x or more depending on the assay and laboratory. Replicates of the GIAB samples were included.

Both laboratories used the Genome Analysis Toolkit (GATK) Haplotype Caller software [20,21], although there were many differences between the two laboratories' bioinformatics pipelines. Copy number and structural variants were excluded from this analysis. Variant filtering and manual review were performed according to each laboratory's clinically validated criteria, and variant calls that were filtered or manually removed were excluded from this analysis. Orthogonal confirmation of clinical NGS calls that passed manual review was performed using Sanger (Thermo Fisher, Waltham, MA) or Pacific Biosciences (Menlo Park, CA) sequencing. Multiple rounds of confirmation using different primers were performed as needed.

VCF (variant call format) files were annotated using the stratification BED (browser extensible data) files produced by the Global Alliance for Genomics and Health (GA4GH) benchmarking workgroup [22] (Supplemental Table 1). These files delineate genomic regions by technical complexity (e.g., repetitive sequences). These regions were padded by 10 bp on each side to ensure that all affected positions were appropriately annotated. The sample-specific GIAB "high-confidence" BED files were

used to define regions (called GIAB-HC) where the reference calls were known to be highly accurate [17]. In patient specimens, the union of the five GIAB-HC files was used to determine regions that were considered GIAB-HC —i.e., if a region was considered high confidence in any of the five GIAB reference datasets, it was considered to be GIAB-HC in all patients. This approach prevents specific low-confidence variant calls in a particular GIAB specimen from inappropriately annotating that site as low confidence in general. Both laboratories defined clinical target regions to include, with rare exceptions, protein-coding exons plus immediate flanking sequences (10–20 bp on each side).

GIAB reference calls were used as confirmation data for variants uncovered in the GIAB specimens. Manual review was not performed for these calls, although analysis was restricted to the GIAB-HC regions. VCFeval [23] was used to compare our variant calls to the reference data and determine which calls were TPs and which were FPs. VCFeval can match variant calls even when the same diploid sequence is represented in different ways, an important factor when comparing indels and complex variants [22,23]. VCFeval also detects partial matches, e.g. zygosity errors (falsely calling a homozygous variant as heterozygous or vice versa) or sites at which one of two called alleles matches and one does not. These cases, collectively reported by VCFeval as "FP_CA", were considered FPs in our study. This study did not examine false negatives in detail.

Confidence intervals (CIs) were computed at 95% using the Jeffreys method. The Wilson score method was also used for comparison in Figure 2.

Results

Our data for this study included (a) almost 14,000 NGS variant calls subjected to confirmation during clinical testing, and (b) more than 184,000 calls with high-quality reference data from the GIAB samples (Table 1). We observed a total of 1684 FPs within these datasets. We calculated the false discovery rate (FDR) of calls in each set, defined as the number of FPs divided by the number of FPs plus TPs. Note that FDR is 1 minus the analytic positive predictive value (aPPV), a metric recommended by the AMP/CAP guidelines, which consider aPPV preferable to computing specificity for multigene sequencing tests [24]. FDR varied considerably among our datasets owing to differences in the variant types, assay targets and methods. As expected, the FDR of indels was consistently higher than that of SNVs (single nucleotide variants). In the clinical data, Lab 2 applied less stringent filtering and manual review compared to Lab 1, which was reflected in the FDRs. For both laboratories, the indel FDR was particularly high in the GIAB datasets because (a) manual review had removed certain FPs from the clinical datasets, and (b) off-target regions containing low-coverage, repetitive, and complex sequences were included in the GIAB data. Such regions are present but less common in clinical targets.

Analysis of individual quality metrics from Lab 1

We examined a variety of quality metrics to determine which were informative for identifying (or "flagging") FPs. We considered a variety of thresholds for each metric and computed the fraction of FPs and TPs flagged—these indicate the "value" and "cost," respectively, of using each metric and threshold. We also computed FDRs, which indicated how likely flagged variants are to be FPs. Plots for each metric are in Supplemental Figures 1–4 and the metrics are defined in Supplemental Table 1. In the text below, we refer to the subplot within Supplemental Figures 1-4 for each metric by letter (A–U). In summary, we

We identified a number of informative metrics, although no single metric proved adequate alone. For example, consider the quality score (QUAL) assigned to each SNV call in the GIAB results (Figure 2, Supplemental Figures G). This same metric was used by Strom et al. [10] to identify SNVs that require confirmation, although Strom had only a single FP to use in defining thresholds. In our dataset, no threshold value for QUAL could be chosen that flagged all, or even most, of the FPs without also flagging many of the TPs. A QUAL cutoff of 800, for example, flagged only 56% of FPs but also flagged 17% of TPs. These flagged variants had an FDR of 0.28%, indicating that this subset is not highly enriched for errors compared with the overall FDR of 0.15%. Indels fared similarly: 55% of FPs and 18% of TPs had QUAL<800, with an FDR of 16% compared with the baseline rate of 6.3%.

This observation was also true of our clinical dataset (Supplemental Figures 3-4G) where 39% of FP SNVs and 2.9% of TPs had QUAL<800. These variants had a 3.6% FDR compared with 0.16% overall – 20-fold enrichment – although most low-quality calls were still TPs. There is little precision in these measurements, however, because the clinical dataset contained only 10 FP SNVs, and only four of those had QUAL<800. Flagging all 10 FP SNVs required a QUAL threshold of 6400 (using logarithmic binning), which would flag almost 30% of TPs. Overfitting would certainly be an issue when analyzing the clinical dataset alone: If the single clinical FP SNV with the highest QUAL score (5516) were absent from these data, then thresholds could be driven by the next, much lower score (2567). There is little statistical confidence in any such threshold: If 10 of 10 FPs are flagged, the point estimate of sensitivity for FPs would be 100%, but the lower bound of the CI is only 78%, demonstrating that thresholds determined using such a small dataset could miss many FPs.

Other quality metrics were also examined, including those recommended by the AMP/CAP guidelines [24] (Supplemental Figures 1–8). Approximately 89% of

heterozygous FP SNVs had an allele balance (AB) outside of the 40–60% range, whereas only 9.5% of heterozygous TPs had ABs outside of this range, showing that AB is informative although insufficient alone (Supplemental Figures D–F). Another metric, the ratio of the two distinct depths reported by GATK for each variant (INFO_DP and GT_DP) empirically also proved to be highly informative (Supplemental Figures M). Variant calls that were near other variant calls in the same specimen were often FPs, regardless of quality metrics or filtering status for the nearby variant call (Supplemental Figures N). Nearby calls in a VCF file can indicate a complex (delins) or het-alt call (i.e., a heterozygous variant in which neither allele is present in the reference human genome) – both variant types have relatively high FDRs. Nearby variants can also indicate misalignment caused by indels or paralogous loci (e.g. pseudogenes).

The genomic context of each variant was found to be informative (Supplemental Figures P, R) with important caveats. Repetitive sequences were enriched for FPs, although many TPs were also located in repeats, reflecting the fact that these sequences are highly polymorphic. Regions of poor mappability are generally not considered high-confidence (GIAB-HC), and we were thus unable to systematically examine error rates within these regions. Segmental duplications (segdups), were difficult to study for both similar and different reasons, including oversensitivity for marking regions as segdups using some methods.

Most other metrics proved to be informative for only a modest population of FPs. For example, neither read depth nor the quality-depth score (QD) performed well alone—indeed, some FPs had very high depth and QD (Supplemental Figures A–C, H). Strand bias metrics showed high efficiency, capturing very few TPs at useful thresholds (Supplemental Figures I–L); however, most FPs did not demonstrate significant strand bias, limiting the utility of these metrics.

Of the variants observed more than once, approximately half of those with at least one FP call were also correctly called as TPs in different specimens (Figure 3, Supplemental Figures S, T). This was true for both SNVs and indels. It was even sometimes true across replicates of the same sample, a situation made possible because partial matches were considered FPs in our analysis. Most examples were in the GIAB data, although we observed a few examples in our smaller clinical datasets despite manual review having removed many FPs from those datasets. We calculated the observed FDR for each variant with multiple calls, where 100% indicates a fully systematic FP and 0% a consistent TP. Variants with low FDRs between 1–40% were prevalent. Indeed, we suspect that many more such variants exist, as our power to detect low FDRs is limited by the small number of GIAB specimens available.

Many (85%) of the GIAB variant calls used in this analysis were located in off-target regions. GIAB indel FDRs were similar on and off-target (5.3% vs. 6.4%, respectively) although SNV FDRs differed somewhat (0.04% vs. 0.17%). The GIAB analyses described above were repeated using only target regions for comparison (Supplemental Figures 13-14) producing generally consistent results albeit with far less confidence and power owing to the smaller dataset size.

Combining metrics

As no single metric proved adequate, we investigated whether combining metrics might work better. One precedent for this approach is the Mu et al. study [12], in which the authors suggested that requiring depth>100 and AB between 40-60% will identify high confidence TPs that do not require confirmation. In our, much larger dataset this is not the case: More than 3% of FPs, both SNVs and indels, failed to be flagged as requiring confirmation using Mu's criteria. Nevertheless, we suspected that a larger battery of metrics might prove effective.

A heuristic algorithm was developed to explore this hypothesis. Briefly, this algorithm incrementally adds flags to a proposed set with the primary aim of capturing 100% of FPs using the combination of flags. This algorithm secondarily prioritizes the addition of flags that minimize the number of TPs also captured. Multiple candidate thresholds for quantitative metrics were provided from which the algorithm could choose. The algorithm does not attempt to minimize the number of flags used as a larger set of flags may prove most efficient, capturing the fewest TPs. This algorithm uses the combined set of FPs from the GIAB and clinical samples to leverage the power of this much larger and more diverse set compared to clinical FPs alone. However, it uses only TPs from clinical samples in order to provide a representative count of unnecessary confirmations that would be performed in clinical practice. We ran this algorithm separately for SNVs and indels. To avoid overfitting, we ran the algorithm on hundreds of randomly sampled data subsets, in each case evaluating the chosen flags on the remaining data. This step suggested important refinements to our approach which we implemented (Supplemental Text).

The final criteria chosen by this process are shown in Figure 4, Supplemental Figures 9–10, and Supplemental Table 2. The selected criteria included depth, AB, QUAL, SB, nearby variants, and others. Combined, these criteria flagged 100% of all 201 FP SNVs (CI 98.9–100%) and 100% of all 987 FP indels (CI 99.8–100%) in the GIAB-HC regions. Only 5.8% of clinical TP SNVs and 6.8% of TP indels in GIAB-HC regions were flagged by the same criteria. Most FPs were captured by two or more flags, which we found reassuring. If we maximized such double-flagging, the fraction of TPs flagged increased to 14.9% of SNVs and 18.1% of indels.

Application of these methods to data from another laboratory

We examined whether a similar approach would work in the data from Lab 2, even though it used somewhat different NGS methods and bioinformatics. As of this analysis, Lab 2 had only sequenced one GIAB sample (NA12878) in replicate in addition to clinical data. FDRs in the GIAB data were roughly comparable between the two laboratories, which reflects specificity of the calling pipelines. The Lab 2 clinical FDRs were however much higher than those of Lab 1 (see Table 1) owing to the different criteria used in filtering, manual review and selection of variants for confirmation.

We examined individual quality metrics (Supplemental Figures 5-8) observing similar patterns to those from Lab 1. Allele balance remained one of the most informative metrics, separating many FPs from TPs, but not all. Quality scores (QUAL and QD), read depth and and strand bias similarly identified some but not all FPs. Many indel FPs were located in or near repeats, although many TPs were also in these regions. Het-alt variant calls and variants with other calls nearby were often erroneous. These themes were generally consistent between the GIAB data and the clinical dataset, and were also true in the GIAB data when we limited analysis to clinical target regions (Supplemental Figures 15-16). Similar to Lab 1, no single metric, such as QUAL, nor the AB/depth combination could flag all FPs in the full dataset without also flagging many TPs. As for Lab 1, variants with both FP and TP calls in different runs were observed.

We combined metrics using the same heuristic approach detailed above. We saw general consistency with the Lab 1 results, as depth, allele balance, QUAL score, strand bias, repeats, nearby variants, and other similar criteria were chosen by our algorithm (Supplemental Figures 11-12). There were differences in specific criteria and thresholds however (Supplemental Table 2). By combining criteria, 100% of FPs in

GIAB-HC regions could be flagged at a modest cost in clinical TPs also flagged (13.2% of SNVs, 15.4% of indels). CIs for the fraction of FPs flagged were 98.5–100% for SNVs and 99.1–100% for indels.

In comparison with Lab 1, the criteria chosen for Lab 2 appear equally effective however somewhat less efficient (i.e. more TPs are flagged as requiring confirmation in Lab 2 versus Lab 1). There are multiple reasons for this, one of which results from a difference in laboratory workflow. Lab 2 uses less aggressive filtering and manual removal of “bad” clinical variant calls prior to confirmation compared to Lab 1. This does not present an accuracy problem for the test, but may cause the algorithm in this study to broaden criteria as it strives to flag the full spectrum of FPs it’s presented with. Tightening up Lab 2’s lowerer filtering thresholds, where possible, and adding more manual review should allow more optimal criteria to be chosen, narrowing the intermediate confidence (yellow) variant category illustrated in Figure 1. In this case the CIs would become wider as the number of FPs drops, although these CIs would become more representative of the truly intermediate confidence variants.

Another difference is that Clinical FPs, rather than GIAB FPs, comprised a large fraction of the Lab 2 dataset and played a significant role in determining both criteria and CIs (Table 1). We do not consider this a problem although we note that clinical data can have significant biases, such as over-representation of specific relatively frequent pathogenic variants. Diversity of variants in a study such as ours is certainly advantageous. The unique variant count for Lab 2 will increase substantially, and the CIs will tighten when sequencing and analysis of the four additional GIAB samples is completed.

Discussion

Our study investigated whether a large and diverse dataset could be used to develop statistically robust criteria to guide the application of confirmation assays in clinical genetic testing. Our datasets combined clinical confirmation results with data from the sequencing of five GIAB specimens. Similar to prior studies [10–12], we found it possible to separate intermediate confidence calls that require confirmation (i.e., to determine which are TPs and which are FPs) from calls that are already high confidence from NGS alone and would not benefit from confirmation. We cannot guarantee that the criteria chosen by our method (or any such method) will capture all FPs. However, we can state that the probability of missing an FP is below a measurable level in datasets containing, collectively, almost 200,000 diverse variant calls with confirmatory data.

Our results differ from those of prior studies in important ways. Neither quality score (as suggested by Strom et al. [10]) nor the combination of allele balance and read depth (as suggested by Mu et al. [12]) efficiently captured all of the FPs in our datasets even if we re-optimized thresholds. Although we acknowledge that criteria must be established independently for each NGS workflow—and indeed, workflows varied among these studies—we believe that this discrepancy results from the small datasets used in these prior studies, which may have left them underpowered and subject to overfitting (Table 1, Figure 2). A study by Baudhuin et al. [11] reported no FPs, a seemingly excellent result. However, our data (Supplemental Figures 1–8) and other studies [6,12] show that FPs are abundant at the threshold boundaries of most NGS quality indicators and that this marginal population contains a significant number of TPs as well. Thus, with current NGS methods, achieving high specificity with a single set of quality thresholds must sacrifice sensitivity. Indeed, a separate study showed that the specific methods used by Baudhuin had certain sensitivity limitations [7].

In our datasets, a battery of criteria, including read depth, strand bias, quality score, allele balance, genomic context, nearby calls, and other metrics were required to identify all FPs. This list closely follows the metrics recommended by the most recent AMP/CAP guidelines [24]. Many of these metrics are commonly used in variant filtering, albeit with different thresholds that emphasize sensitivity rather than specificity. In theory, a single quality score that captures all of these factors would be simpler to use than a battery of criteria. Unfortunately, we know of no current approach for computing such a score that will identify all FPs without capturing many TPs.

There were numerous benefits to using the large datasets we generated. These data not only provided a diverse set of FP calls that we could use in establishing criteria, but also offered ample data with which to establish separate training and test sets and identify outliers to minimize overfitting. Furthermore, we had adequate sample sizes for the computation of statistical measures of confidence, a crucial part of any laboratory validation study, albeit one that is not always used [24]. The statistical metric we used (CI lower bound given number of FPs) is rigorous but simple. More sophisticated approaches are feasible and ideal topics for future work. Prior studies report cumulative PPV, a metric which can be misleading in this context because FPs near the edge of quality thresholds can be diluted by a large number of very high confidence TPs. Marginal error rates are far more informative for selecting thresholds and were extensively used in our study.

A key question is how large a dataset is adequate? Basic guidance can be found using the calculations described in recent guidelines [24,25]. For example, if certain criteria flag 59 of 59 FP variants, then the point estimate of FP sensitivity is 100%, but rigorously these criteria have only been demonstrated to capture at least 95% of FPs at a level of $p = 0.05$ (i.e., 19 out of 20 times). Similar values are shown in Figure 2. Whether demonstrating 95% at $p = 0.05$ is adequate is not a statistical question but a

medical policy question, and the answer could vary among clinical applications. Our laboratories aimed higher, and achieved CI lower bounds between 98.5 and 99.8% (Table 1).

Obtaining a large number of FPs for study can be challenging, and we were careful not to do so in artificial and problematic ways. For example, FPs could be collected by simply lowering filtering thresholds. However, the resulting criteria might be good only at flagging clearly erroneous calls as opposed to defining the intermediate confidence set (Figure 1) for which confirmation matters most. Another approach would be to run many samples containing the same FP variants, either replicates or different patients with systematic errors. In this case, the resulting criteria might be good only for limited classes of FPs. To avoid these problems, we used the same filtering thresholds that we use in clinical practice and we considered both the number of unique variants and the number of variant calls when designing our study (Table 1). Running the set of five GIAB specimens greatly increased the diversity of variants with confirmation data.

One might argue that our study artificially increased the number of FPs by including the off-target regions in the GIAB specimens. To the contrary, we consider it valuable to deliberately challenge our approach by including these regions. Doing so increased our ability to observe many different types of FPs, many of which have features that are infrequent, yet certainly present, in coding exons (e.g., low depth, repeats, and complex variants). Because our algorithm required 100% of the combined FPs (clinical and GIAB, on- and off-target) to be flagged, we know that the criteria it selects will be effective on any subset of these FPs. We also ensured validity of our results by examining GIAB quality metrics in clinical target regions alone (Supplemental Figures 13–16). Although the FP rates in off-target regions may be higher than those observed in clinical practice, the criteria will remain applicable.

Our study design has important limitations. We relied on the GIAB reference call sets, which are highly accurate but imperfect [17]. For example, false–negative variant clusters are known to exist in the GIAB reference call sets, which will incorrectly appear as FPs in our laboratory data. In addition, when sequencing the Coriell GIAB specimens, there can be genetic differences compared with the original DNA samples used to develop the GIAB call sets. In general, these two issues will cause our methodology to establish broader criteria than might otherwise be necessary. Such errors do not increase the risk of an FP being missed, but they do increase the number of TPs that are flagged. Given our study's objective, this is the "better" problem to have. In our study, outliers (FPs in the GIAB data with unusually high confidence in our data) were examined in detail to ensure that the results appeared correct. We did not confirm any sites in the GIAB specimens using an orthogonal assay, although studies following ours may find it valuable to do so.

A further limitation results from our focus on high confidence regions in the GIAB reference call sets. We have little data regarding our ability to flag FPs outside of the GIAB-HC regions, which are heterogeneous and present diverse issues that can increase FP rates [15]. Thus, the criteria described above may inadequately capture FPs in these regions. We conclude that clinical variants outside of the GIAB-HC regions should always require confirmation unless additional, statistically valid studies are performed, likely on a variant-specific basis. In the Lab 1 dataset, 7.0% of clinical SNVs and 15.4% of indels were outside of the GIAB-HC regions. In combination with the flagged variants in GIAB-HC regions (see results), a total of 12.8% of SNVs and 22.2% of indels would thus require confirmation.

Our results have specific implications for the current New York state guidelines, which state that confirmation may be waived after "at least 10 positive samples per target gene" have been confirmed [4]. In our data, approximately half of all variants that

appeared as FPs also appeared as TPs. These variants run a high risk of being confirmed TPs in a series of tests and then, after confirmation is no longer considered necessary, being called falsely. Statistically, this is highly likely to happen with multi-gene panel, exome, or genome sequencing which can detect and report many different variants. Moreover, in our datasets, different variants within a gene often exhibited remarkably different properties that correlated with remarkably different FDRs. Observing some of these variants as TPs provides little information about whether the other variants in that gene are confidently called. In summary, our data argue against using the NY criteria for confirmation with multi-gene sequencing tests.

The ACMG guidelines for NGS recommend that laboratories have “extensive experience with NGS technology and be sufficiently aware of the pitfalls … before deciding that result confirmation with orthogonal technology can be eliminated” [3]. Our methodology provides a practical and rigorous way to follow this recommendation using a statistically significant dataset that can be readily generated by many laboratories. Of course, changes in NGS processes, whether biochemical or bioinformatic, could change the criteria for selecting variants for confirmation. Per guidelines, the process described in this manuscript would need to be repeated as part of validating any significant process change.

Our methodology does not address other roles that confirmation assays serve: e.g., verifying the identity of a specimen or determining the exact structure of indels and complex variants. Our framework also does not address the critical issue of how to handle conflicting results between NGS and confirmation assays. As Beck et al. [13] elegantly showed, naively assuming that a confirmation assay is always correct can introduce more errors into NGS data than confirmation fixes. Laboratories need to address these issues separately.

Our data show that confirmation assays can be limited to a small fraction of variants without any measurable effect on analytic specificity. An adequately large and diverse dataset is required to do this with confidence. Combining GIAB and clinical specimen data is a practical way to create such a data set, and the analysis of such data proved highly informative. Limiting confirmation assays in this careful manner may help reduce costs and improve turnaround time without compromising the quality of clinical genetic tests.

Acknowledgements

We thank the Invitae and Partners HealthCare Laboratory for Molecular Medicine laboratory staff for producing the data used in this study. We thank all of the collaborators in the Genome in a Bottle consortium for their excellent work on the reference specimens we used. We thank Nancy Jacoby for help on this manuscript. All of the authors' work on this project was funded by their respective employers.

Disclosures

Authors CL, ML, SL, RN, JP, HR, VR, and RT work for laboratories that provide clinical genetic testing services. SL, RN, JP, VR, and RT own stock and/or stock options in Invitae. SL owns stock in Illumina and Thermo Fisher.

Certain commercial equipment, instruments, or materials are identified in this paper only to specify the experimental procedure adequately. Such identification is not intended to imply recommendation or endorsement by the National Institute of Standards and Technology, nor is it intended to imply that the materials or equipment identified are necessarily the best available for the purpose.

Oversight

The use of de-identified clinical data in this study was subject to the requirements of each laboratory's institutional review board. The GIAB specimens were used under the terms of their material transfer agreements.

Main Figure Legends

Figure 1: This study's methodology.

(A) Variant calls can be separated into high confidence true positive (green) and intermediate confidence (yellow) categories using strict thresholds intended to maximize specificity of the high confidence set. The intermediate set will contain a mixture of true and false positive calls. Variant calls that are confidently false positives (red) or that are unsupportable are typically filtered out using different criteria that emphasize sensitivity. The analogy to a traffic light is illustrative.

(B) Process for collecting true positive and false positive calls for both clinical and Genome in a Bottle specimens used in our study.

Figure 2: Quality score data and statistics.

(A) Variant call quality scores (QUAL, Y-axis) for true positive (TP, blue) and false positive (FP, red) single nucleotide variant (SNV) calls in the Lab 1 Genome in a Bottle (GIAB) specimen dataset. The X-axis position of each point is randomly assigned. To make the relative point density visible, we used a random selection of 1/5th of the TP set along with all FP calls. Corresponding histograms with the full dataset and no downsampling are shown in Supplemental Figure 1G.

(B) Random sample of 1000 data points from the Lab 1 GIAB SNV data. All points are displayed. The arrows indicate the two FPs present. A thousand such samples were generated and many would lead to very different conclusions about QUAL thresholds (dashed line) compared to the larger data set in plot (A).

(C) Full dataset from Strom *et al*. Note that quality scores are not calibrated between these data and plots (A) and (B), although no Strom-like threshold would prove effective at separating TPs and FPs in the plot (A) dataset.

(D) Lower bound of the 95% confidence interval (CI) on the fraction of false positives captured (Y-axis) as a function of number of false positives used to determine criteria (X-axis) assuming 100% success is observed. This calculation used the Jeffreys method (blue), and the Wilson score method (red) which both produce similar but not identical results to the tolerance calculation used in the AMP/CAP guidelines [24] (this method produces a value between that of Wilson and Jeffreys). Specific points highlighted show 95%, 98%, 99%, of FPs captured. These levels require at least 49, 124, and 249 FPs, respectively, using the Jeffreys method. Many prior studies did not achieve such significance (Table 1).

Figure 3: Historical performance as a quality metric.

These two plots examine all variants that (i) were observed more than once in our Lab 1 dataset, and for which (ii) one or more of those calls was a false positive (FP). SNVs and indels are combined. Each unique variant (for example, a particular base change) is counted once. A detailed breakdown of these data and the corresponding Lab 2 results are in the supplement.

(A) Approximately half of all variants that were falsely called were also correctly called as positives in different specimens or runs.

(B) Many of these variants have low false discovery rates (FDRs). Calls of low FDR variants will usually but not always be correct, and repeated TP observations of such a variant provide little information about the accuracy of any following observation of that

same variant. Our study is underpowered to measure FDRs near 0% or near 100%, owing to the small number of GIAB specimens, and many more such variants may exist than are shown here. We conclude that historical performance is a ineffective quality metric in our datasets.

Figure 4: Combining flags.

These plots show the sequential effect (top to bottom) of combining flags chosen by the heuristic algorithm. Plot (A) shows this result for SNVs and plot (B) indels in the Lab 1 dataset. Red bars show the cumulative fraction of false positives (FPs) captured, blue bars the cumulative fraction of true positives (TPs), and grey bars show the cumulative fraction of false positives captured by two or more flags. The dotted lines illustrate the flags needed to capture 100% of false positives with at least one flag. Details can be found in Supplemental Tables 1-2 and Supplemental Figures 9-12.

References

1. Korf BR, Rehm HL. New approaches to molecular diagnosis. JAMA. 2013;309: 1511–1521.

2. Rehm HL. Disease-targeted sequencing: a cornerstone in the clinic. Nat Rev Genet. 2013;14: 295–300.

3. Rehm HL, Bale SJ, Bayrak-Toydemir P, Berg JS, Brown KK, Deignan JL, et al. ACMG clinical laboratory standards for next-generation sequencing. Genet Med. 2013;15: 733–747.

4. New York State Department of Health. Guidelines for Validation Submissions of Next Generation Sequencing (NGS) assays under the NYS Testing Category of Genetic Testing [Internet]. 2015. Available: health.ny.gov

5. Aziz N, Zhao Q, Bry L, Driscoll DK, Funke B, Gibson JS, et al. College of American Pathologists' laboratory standards for next-generation sequencing clinical tests. Arch Pathol Lab Med. 2015;139: 481–493.

6. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 2011;43: 491–498.

7. Lincoln SE, Zook JM, Chowdhury S, Mahamdallie S, Fellowes A, Klee EW, et al. An interlaboratory study of complex variant detection [Internet]. bioRxiv. 2017. p. 218529. doi:10.1101/218529

8. Lincoln SE, Kobayashi Y, Anderson MJ, Yang S, Desmond AJ, Mills MA, et al. A Systematic Comparison of Traditional and Multigene Panel Testing for Hereditary Breast and Ovarian Cancer Genes in More Than 1000 Patients. J Mol Diagn. 2015;17: 533–544.

9. Mandelker D, Schmidt RJ, Ankala A, McDonald Gibson K, Bowser M, Sharma H, et al. Navigating highly homologous genes in a molecular diagnostic setting: a resource for clinical next-generation sequencing. Genet Med. 2016;18: 1282–1289.

10. Strom SP, Lee H, Das K, Vilain E, Nelson SF, Grody WW, et al. Assessing the necessity of confirmatory testing for exome-sequencing results in a clinical molecular diagnostic laboratory. Genet Med. 2014;16: 510–515.

11. Baudhuin LM, Lagerstedt SA, Klee EW, Fadra N, Oglesbee D, Ferber MJ. Confirming Variants in Next-Generation Sequencing Panel Testing by Sanger Sequencing. J Mol Diagn. 2015;17: 456–461.

12. Mu W, Lu H-M, Chen J, Li S, Elliott AM. Sanger Confirmation Is Required to Achieve Optimal Sensitivity and Specificity in Next-Generation Sequencing Panel Testing. J Mol Diagn. 2016;18: 923–932.

13. Beck TF, Mullikin JC, NISC Comparative Sequencing Program, Biesecker LG. Systematic Evaluation of Sanger Validation of Next-Generation Sequencing Variants. Clin Chem. 2016;62: 647–654.

14. Alpaydin E. Introduction to Machine Learning. MIT Press; 2014.

15. Zook JM, Chapman B, Wang J, Mittelman D, Hofmann O, Hide W, et al. Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. Nat Biotechnol. 2014;32: 246–251.

16. Zook JM, Catoe D, McDaniel J, Vang L, Spies N, Sidow A, et al. Extensive sequencing of seven human genomes to characterize benchmark reference materials. Sci Data. 2016;3: 160025.

17. Zook J, McDaniel J, Parikh H, Heaton H, Irvine SA, Trigg L, et al. Reproducible integration of multiple sequencing datasets to form high-confidence SNP, indel, and reference calls for five human genome reference materials [Internet]. bioRxiv. 2018. p. 281006. doi:10.1101/281006

18. Pugh TJ, Kelly MA, Gowrisankar S, Hynes E, Seidman MA, Baxter SM, et al. The landscape of genetic variation in dilated cardiomyopathy as surveyed by clinical DNA sequencing. Genet Med. 2014;16: 601–608.

19. Alfares AA, Kelly MA, McDermott G, Funke BH, Lebo MS, Baxter SB, et al. Results of clinical genetic testing of 2,912 probands with hypertrophic cardiomyopathy: expanded panels offer limited additional sensitivity. Genet Med. 2015;17: 880–888.

20. Poplin R, Ruano-Rubio V, DePristo MA, Fennell TJ, Carneiro MO, Van der Auwera GA, et al. Scaling accurate genetic variant discovery to tens of thousands of samples [Internet]. bioRxiv. 2017. p. 201178. doi:10.1101/201178

21. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20: 1297–1303.

22. Krusche P, Trigg L, Boutros PC, Mason CE, De La Vega FM, Moore BL, et al. Best Practices for Benchmarking Germline Small Variant Calls in Human Genomes [Internet]. bioRxiv. 2018. p. 270157. doi:10.1101/270157

23. Cleary JG, Braithwaite R, Gaastra K, Hilbush BS, Inglis S, Irvine SA, et al. Comparing Variant Call Files for Performance Benchmarking of Next-Generation Sequencing Variant Calling Pipelines [Internet]. bioRxiv. 2015. p. 023754. doi:10.1101/023754

24. Roy S, Coldren C, Karunamurthy A, Kip NS, Klee EW, Lincoln SE, et al. Standards and Guidelines for Validating Next-Generation Sequencing Bioinformatics Pipelines: A Joint Recommendation of the Association for Molecular Pathology and the College of American Pathologists. J Mol Diagn. 2018;20: 4–27.

25. Jennings LJ, Arcila ME, Corless C, Kamel-Reid S, Lubin IM, Pfeifer J, et al. Guidelines for Validation of Next-Generation Sequencing-Based Oncology Panels: A Joint Consensus Recommendation of the Association for Molecular Pathology and College of American Pathologists. J Mol Diagn. 2017;19: 341–365.

Fig 1A

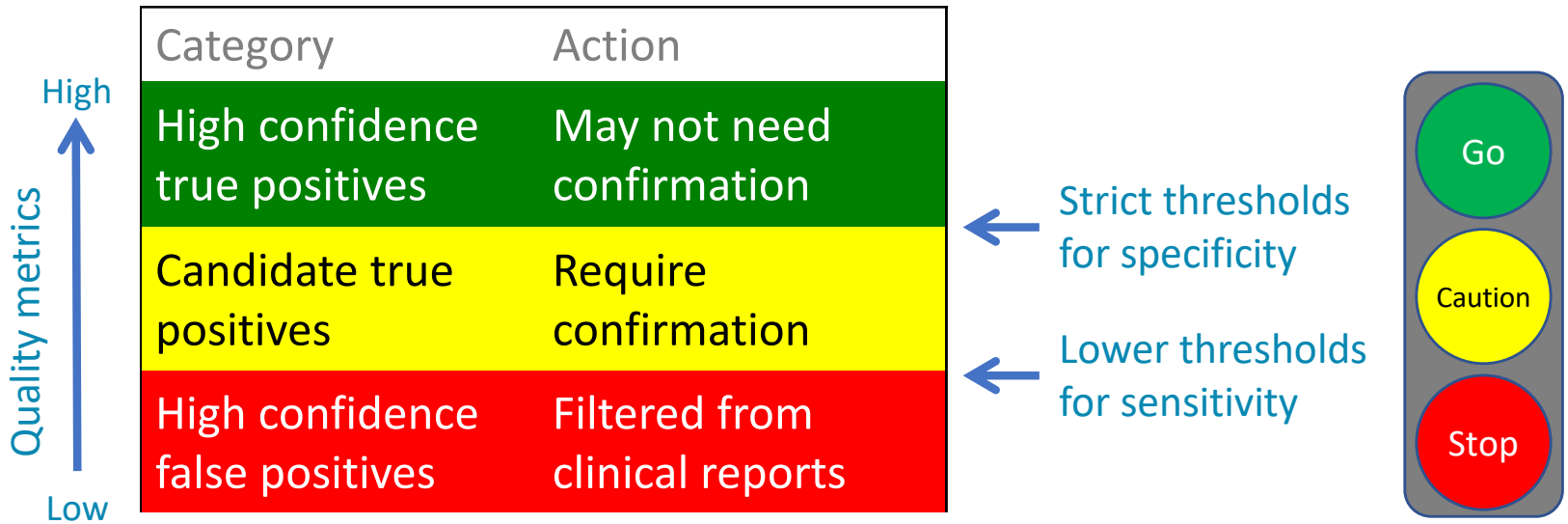

Quality metrics
High
Low
Category
Action
High confidence true positives
May not need confirmation
Candidate true positives
Require confirmation
High confidence false positives
Filtered from clinical reports
Strict thresholds for specificity
Lower thresholds for sensitivity
Go
Caution
Stop

Fig 1B

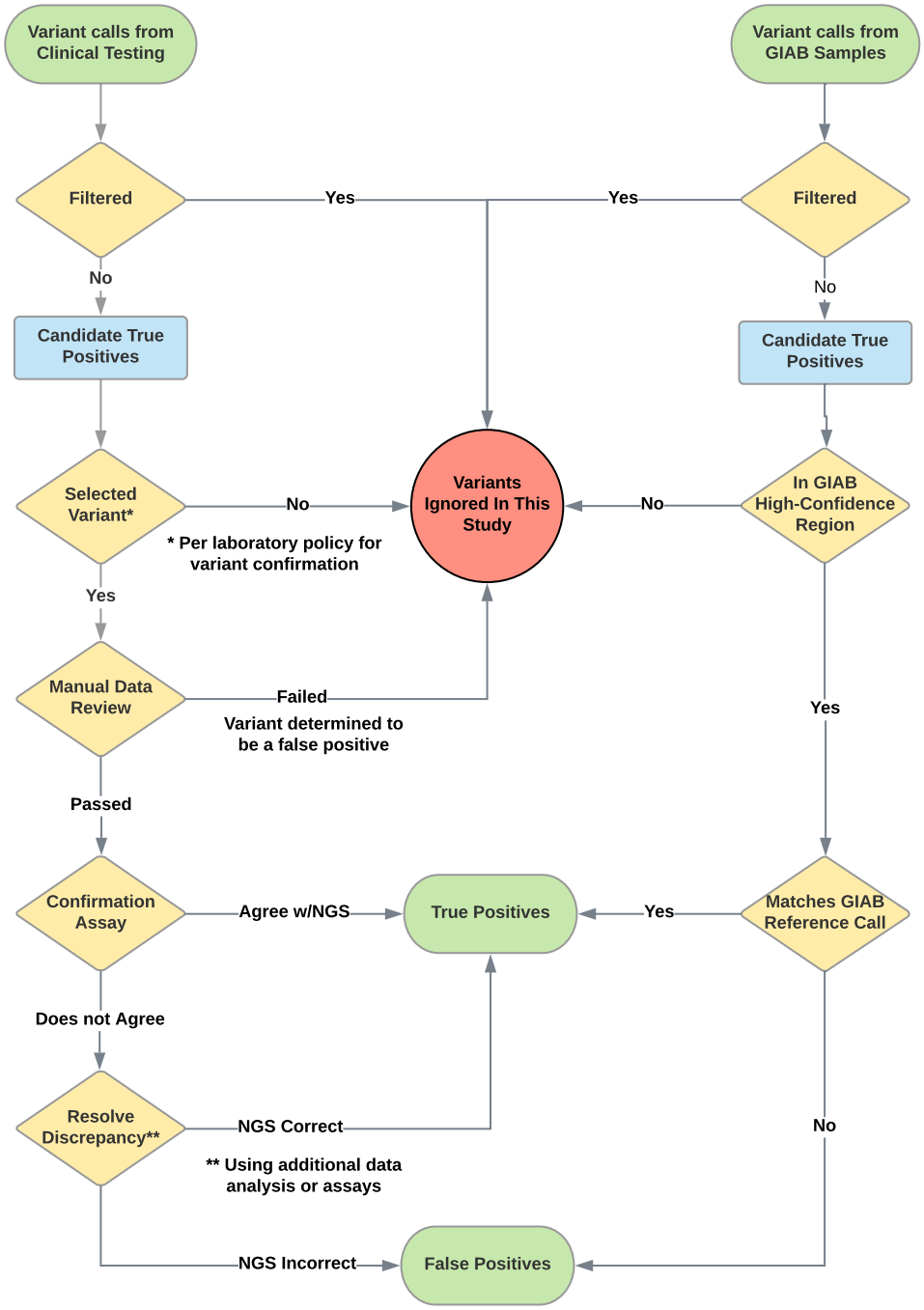

Variant calls from Clinical Testing
Variant calls from GIAB Samples
Filtered
Yes
Yes
Filtered
No
No
Candidate True Positives
Candidate True Positives
Selected Variant*
No
Variants Ignored In This Study
No
In GIAB High-Confidence Region
* Per laboratory policy for variant confirmation
Yes
Manual Data Review
Failed
Variant determined to be a false positive
Yes
Passed
Confirmation Assay
Agree w/NGS
True Positives
Yes
Matches GIAB Reference Call
Does not Agree
Resolve Discrepancy**
NGS Correct
No
** Using additional data analysis or assays
NGS Incorrect
False Positives

Fig 2 A-B

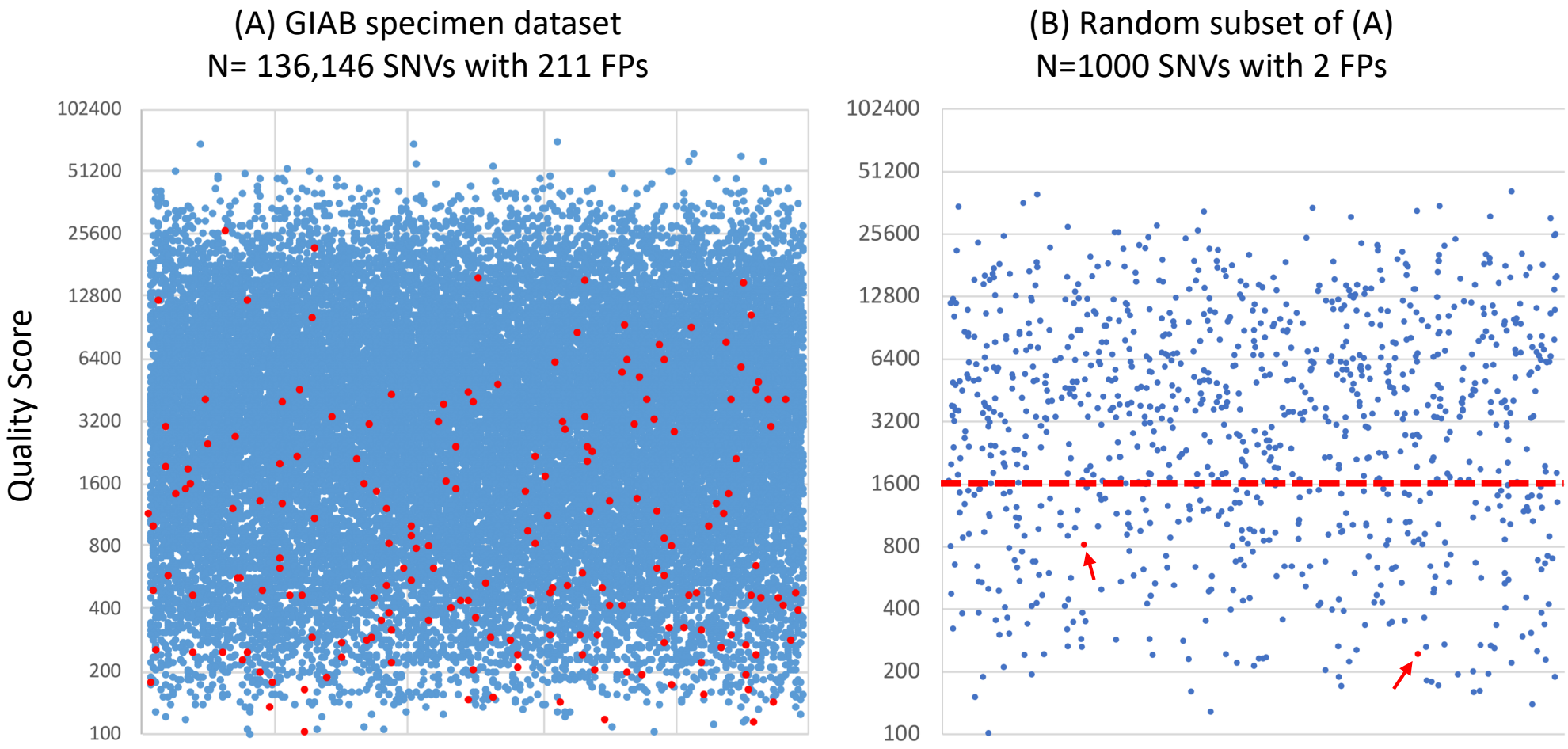
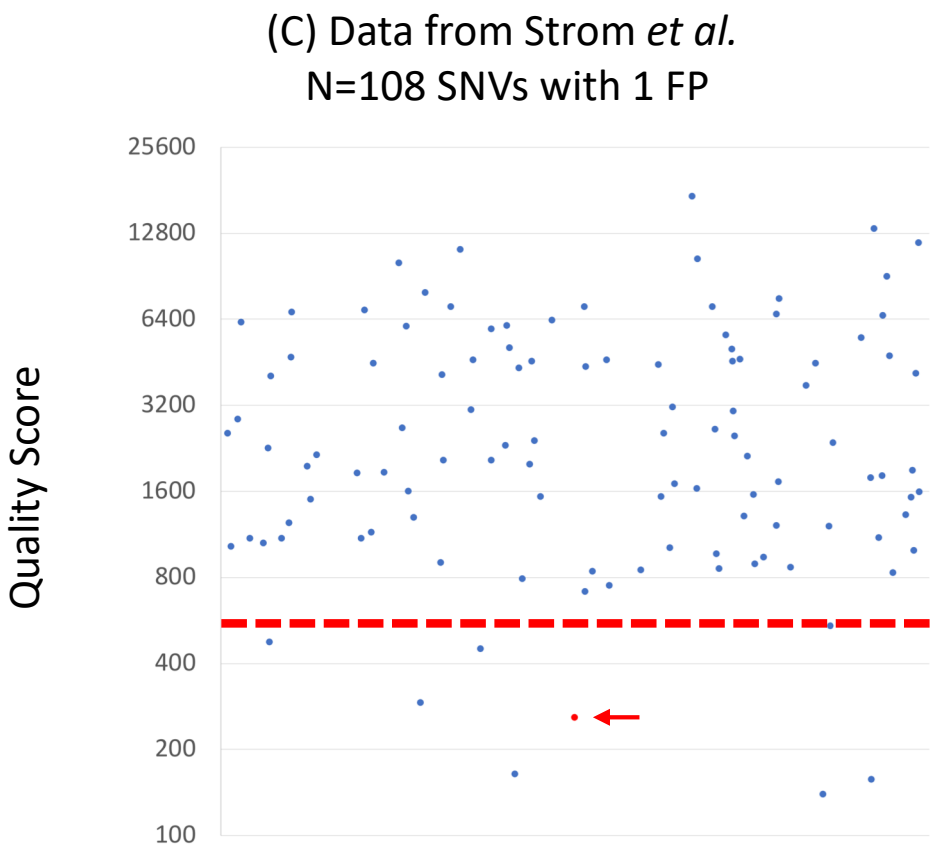

(A) GIAB specimen dataset
N= 136,146 SNVs with 211 FPs
(B) Random subset of (A)
N=1000 SNVs with 2 FPs
Quality Score
102400
51200
25600
12800
6400
3200
1600
800
400
200
100

Fig 2 C-D


(C) Data from Strom *et al.*
N=108 SNVs with 1 FP
Quality Score
25600
12800
6400
3200
1600
800
400
200
100

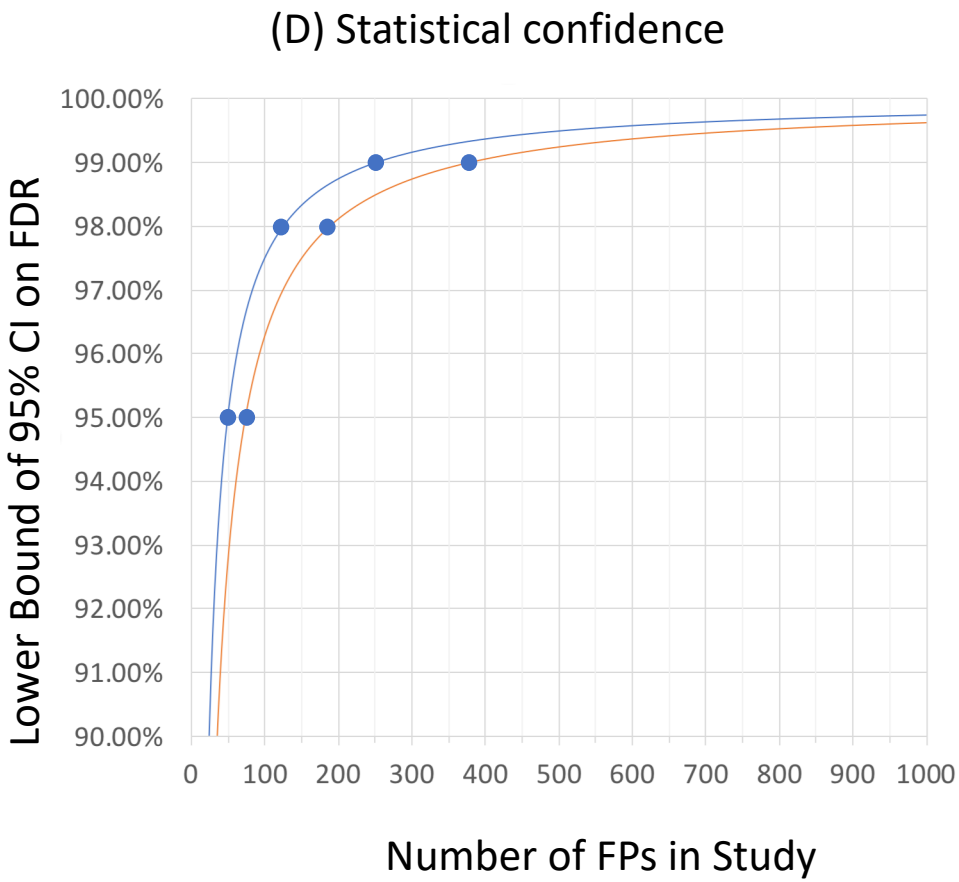

(D) Statistical confidence
Lower Bound of 95% CI on FDR
100.00%
99.00%
98.00%
97.00%
96.00%
95.00%
94.00%
93.00%
92.00%
91.00%
90.00%
0
100
200
300
400
500
600
700
800
900
1000
Number of FPs in Study

Fig 3 A-B

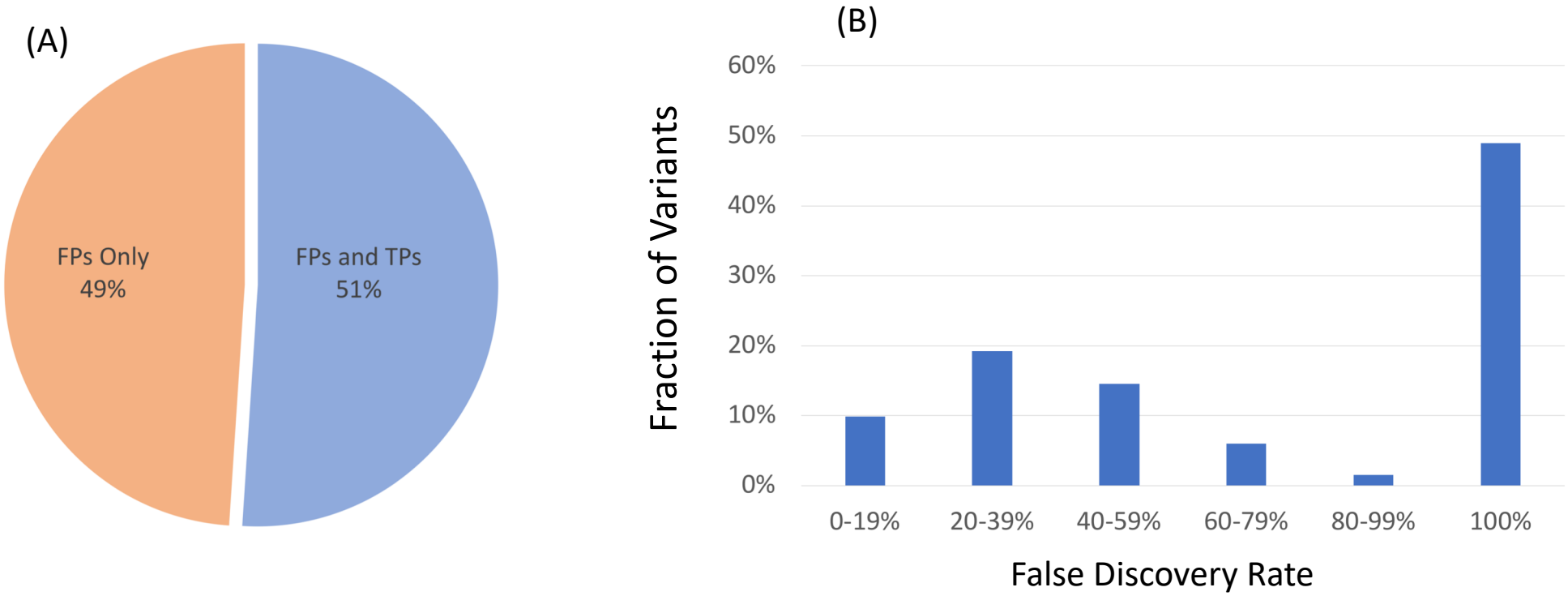
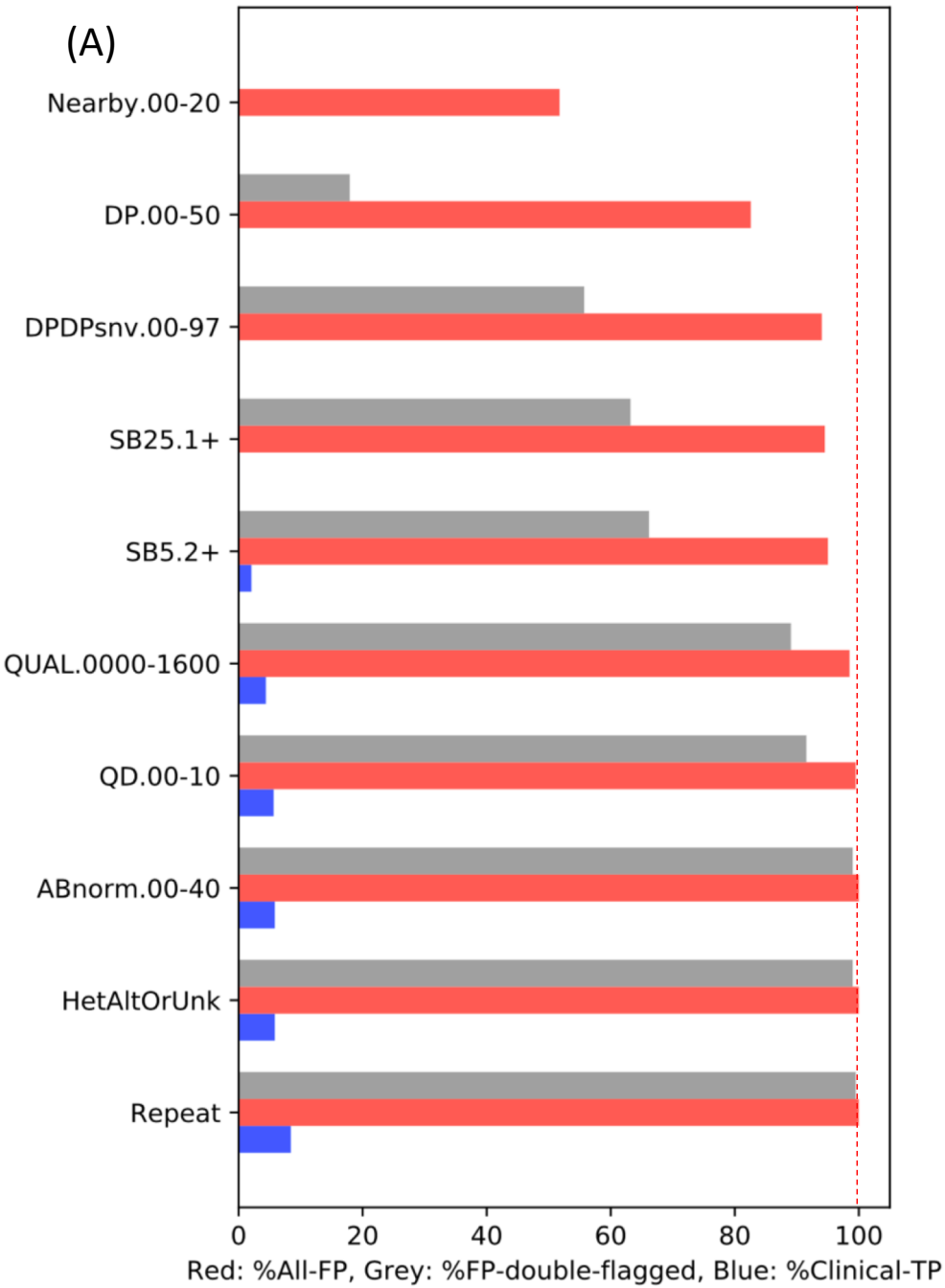
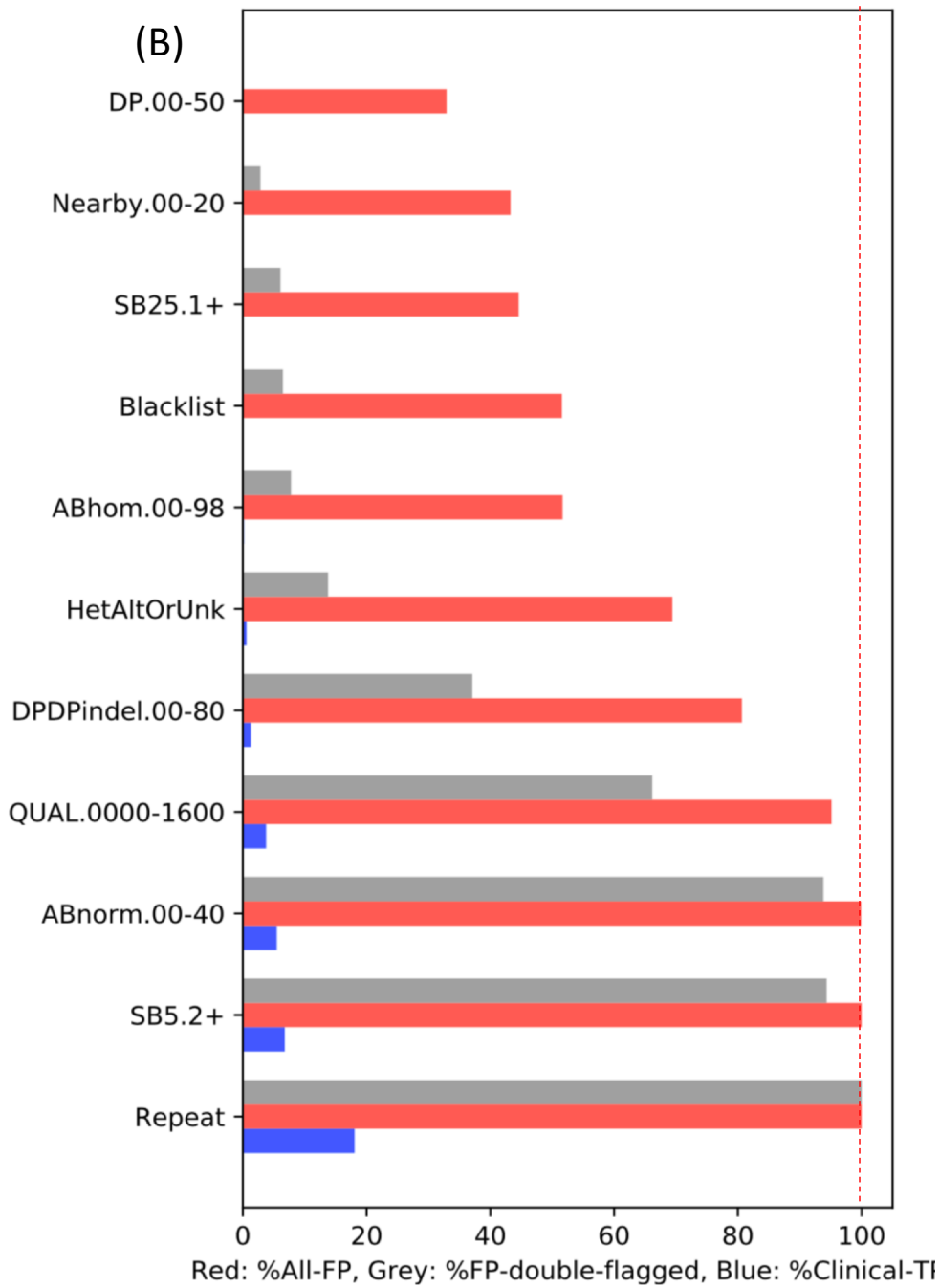

(A)
FPs Only
49%
FPs and TPs
51%
(B)
Fraction of Variants
60%
50%
40%
30%
20%
10%
0%
0-19%
20-39%
40-59%
60-79%
80-99%
100%
False Discovery Rate

Fig 4 A-B


(A)
Nearby.00-20
DP.00-50
DPDPsnv.00-97
SB25.1+
SB5.2+
QUAL.0000-1600
QD.00-10
ABnorm.00-40
HetAltOrUnk
Repeat
0
20
40
60
80
100
Red: %All-FP, Grey: %FP-double-flagged, Blue: %Clinical-TP


(B)
DP.00-50
Nearby.00-20
SB25.1+
Blacklist
ABhom.00-98
HetAltOrUnk
DPDPindel.00-80
QUAL.0000-1600
ABnorm.00-40
SB5.2+
Repeat
0
20
40
60
80
100
Red: %All-FP, Grey: %FP-double-flagged, Blue: %Clinical-TP

Main Table 1 **Study Data**

Source	Lab	Variant Type	Samples	Unique Variants	Variant Calls with Orthogonal Data	FP Calls	FDR	Total Calls	Total FPs	FPsens	CI lower bound	Notes
This study	Lab1	SNVs	GIAB	27,202	136,146	211	0.15%					
			Patients	2,840	6,217	10	0.16%	142,363	221	100%	98.9%	A,B
		Indels	GIAB	3,715	15,574	992	6.37%					
			Patients	1,749	2,262	12	0.53%	17,836	1004	100%	99.8%	A,B
	Lab2	SNVs	GIAB	5,816	29,148	38	0.13%					
			Patients	4,359	4,934	130	2.63%	34,082	168	100%	98.5%	A,B
		Indels	GIAB	1,185	3,617	274	7.58%					
			Patients	267	389	17	4.37%	4,006	291	100%	99.1%	A,B
Strom et al.	only one	SNVs	Patients		108	1	0.93%	108	1	100%	5.1%	B,D
Mu et al.	only one	SNVs	Patients		6,912	94	1.36%	6,912	94	100%	97.4%	B,D
		Indels	Patients		933	5	0.54%	933	5	100%	62.1%	B,D
Baudhuin et al.	only one	SNVs	Patients	380	797	0	0%					
			1KG	736	736	0	0%	1,533	0	N/A	N/A	E,F
		Indels	Patients	63	122	0	0%					G
			1KG	26	26	0	0%	148	0	N/A	N/A	E,F

Notes:

A	Clinical and GIAB data were combined in the analysis
B	Clinical data sets typically used manual review which can remove false positives prior to confirmation, reducing the FDR observed here. We did not do so in the GIAB dat
C	A large fraction of the FPs were in challenging regions of OTOA and CFTR. Withour these regions, FDRs wopuld be much lower.
D	We computed CIs based on data in the publication. No such statistics were provided by the study authors.
E	Baudhuin et al. analyzed only unique variants in their 1KG analysis.
F	Although discordances between NGS and reference calls were initially observed, improved 1KG reference data eliminated these.
G	Most of the indels in patients were intronic and homopolymer-associated

Abbreviations and definitions:

SNV	Single nucleotide variant
Indel	Insertion or deletion
Unique Variant	A particular alteration which may be present in one or more speimens
FP	False Positives from NGS. All of these false positives were detected by confirmation or reference data. In clinical practice these variants are removed from reports.
FDR	False discovery rate: Number of FPs divided by FPs plus TPs (true positives)
FPsens	The percentage of FPs captured using the study's proposed criteria
CI	95% confidence interval on FPsens
GIAB	Genome in a Bottle
1KG	1000 Genomes Project