To Confirm or Not To Confirm, That is the Question: A Rigorous Approach to Evaluating the Importance of Sanger Confirmation of Clinical NGS Findings

Stephen Lincoln[1], Birgit Funke[2,3], Shan Yang[1], Mark Bowser[3], Justin Zook[4], Shahikant Kulkarni[5], Swaroop Aradhya[1], and Heidi Rehm[2,3]

[1]Invitae, [2]Harvard Medical School, Department of Pathology, [3]Harvard Partners Laboratory for Molecular Medicine,
[4]National Institute for Standards and Technology, [5]Washington University

Introduction

When next-generation sequencing (NGS) first entered clinical use, analytic false positives were a significant and real concern. Laboratories thus implemented procedures to rule out false positives by confirming NGS results with an orthogonal method such as Sanger sequencing. While effective, these procedures significantly increase both the cost of testing and turn around time.

NGS technology has evolved, and new biochemical and bioinformatics methods together (not implemented in all labs) promise not only to improve the accuracy of NGS, but also to improve NGS quality metrics that can help identify the highest confidence calls. With these new methods a lab may never see a confirmation failure on thousands of high confidence calls, but nevertheless many high-quality labs, including ours, still confirm positive findings before reporting.

The ACMG NGS guidelines[1] anticipated this situation, instructing labs to confirm positives until they had established adequate experience and validated methods to otherwise ensure the same level of quality. However these guidelines do not provide specific requirements. We invite additional collaboration, data, feedback and discussion on this important topic.

[1]Rehm *et al., Genet Med* 2013

Methods

Our multi-laboratory collaboration has developed a general framework for addressing this challenge, specifically in germline DNA testing. This framework can be used with each lab's different assay targets, NGS protocols, bioinformatics algorithms, and QC thresholds. Our framework involves:

1. Each laboratory choosing, for its own NGS tests, **highly conservative QC criteria** which are believed to define a homogeneous population of high-confidence variant calls. With **multiple thresholds** these strict criteria do not have to pass all true positives - a separate, lower QC threshold can help ensure sensitivity.
2. Each laboratory assembling a large and clinically representative set of **confirmation data** across variants that meet these strict QC criteria.
3. Each laboratory uses **proper statistical measures** to quantify, by variant class, the accuracy of NGS variant calls that meet these criteria. If the demonstrated accuracy is adequate, variants meeting the strict QC criteria may not require confirmation in the future, depending on other clinical and operational factors.
4. Laboratories may expand the criteria to include additional variants, but only when adequate confirmation data for the newly added variants justify that expansion using the same statistical approaches.

We use **FDR** (**False Discovery Rate**, the fraction of reported positives that are false) as our key metric providing a simple, clinically relevant view of the importance of confirmation[1]. We use the upper bound of the 95% confidence interval[2] (CI) on FDR as a conservative estimate of performance at a standard scientific level of rigorous proof.

1: **FDR** is defined as FP/(TP + FP) the and is 1 – the analytic PPV (positive predictive value). Recall that FDR and FPR (False Positive Rate) are different: FPR takes into account the number of true negative results, which is often not thoroughly known in such studies. FPR is also often reported per base-pair tested (as 1 – specificity) which is not intuitive for this purpose.

2: We used the **Jeffreys method**, a Bayesian approach more appropriate than the traditional (Wald) CI method for error rates at or near zero. It also works well over a wide range of N.

Results

Laboratory 1 (Invitae): A combined data set from (a) validation, (b) clinical studies, and (c) clinical testing was assembled for approx. 10,000 specimens run on a 216 gene NGS assay. Calls were strictly filtered by multiple criteria:

	QC Criiteria
Per Run	Cluster density, %Q30 bases, Read yield
Per Library	Insert size, PCR duplicates, %on-target, Coverage, Reference agreement, AT/GC dropout, Contamination estimates
Per Variant Call	Min coverage, Allele balance, Strand bias, Call quality scores, Variant is not a known false positive in other samples (blacklist)
Genomic Region	High complexity, Highly mappable, not Segmental duplication, 25-65% GC content

Note: For QC details see Lincoln et al., J Mol Diag 2015 (Table S6). Genomic regions defined by Global Alliance for Genomics and Health, Benchmarking Work Group (email Justin.Zook@nist.gov).

Of 4190 calls with confirmatory data available, 2797 calls (for 1302 unique DNA alterations) met these strict QC criteria. **No false positives** were observed:

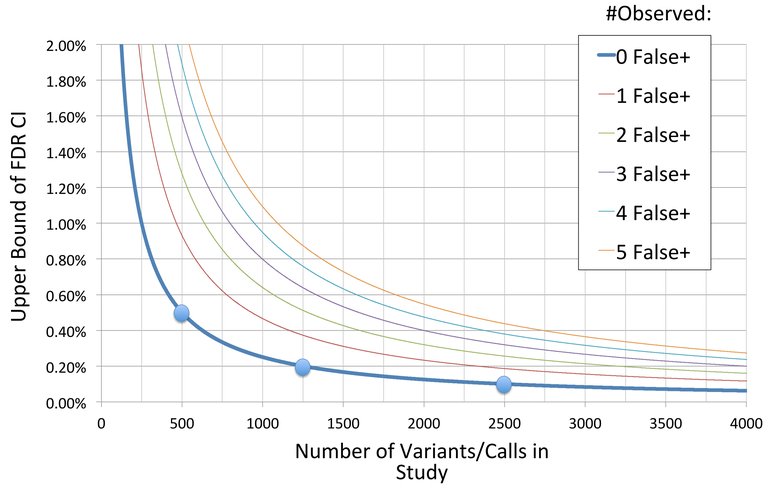
	False+	Calls	FDR CI	Unique	FDR CI
SNVs	**0**	2259	0.00% – **0.11%**	942	0.00% – **0.27%**
Indels ≤ 5bp	**0**	464	0.00% – **0.54%**	307	0.00% – **0.81%**
Indels > 5bp	**0**	74	0.00% – 3.33%	53	0.00% – 4.61%

For example, for SNVs meeting strict criteria, 0 false positives (0.0% FDR) were observed, rigorously demonstrating an FDR no higher than 0.27%. We believe this is the best metric to inform future decisions about confirmation of such calls. Indeed 0 false positives were observed in the larger and more heterogeneous set of 4190, but we are not drawing conclusions for all of these alterations here.

Laboratory 2 (Harvard LMM): 4286 calls for different panels were assembled which passed similarly strict (but not identical) QC criteria. As for the Invitae data, no false positives were observed:

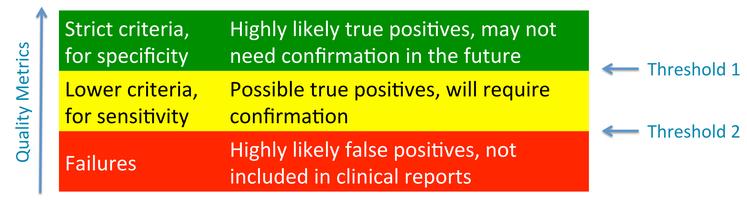
	False+	Calls	FDR CI	Unique	FDR CI
SNVs	**0**	4251	0.00% – 0.06%	408	0.00% – 0.61%
Indels ≤ 5bp	**0**	33	0.00% – 0.54%	7	0.00% – 29.2%
Indels > 5bp	**0**	2	0.00% – 3.33%	2	0.00% – 66.7%

Note: These data include multiple replicates of a control sample and fewer clinical specimens, hence the different ratio of unique variants to total calls and relatively small number of indels.



Recommended Practices

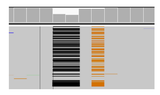
a. Conservative QC Criteria must be set based on lab director NGS experience and the lab's own large validation data set, per ACMG NGS guidelines.



b. Positive Sample Tracking Required: Sanger confirmation plays a role not just in detecting false positives, but also in ensuring that pathogenic variants are reported in the correct patient. An alternate mechanism must be in place whenever Sanger confirmation is not employed.

c. "Spelling differences": Sanger confirmation also helps ensure that the specific description of any alteration is correct when NGS read alignments have ambiguity. Alternate means must be in place if confirmation is not used.

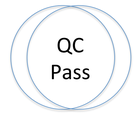
Initially:
c.6_7delAT and
c.4G>C *low-confidence*



Confirmed as:
c.4_6delinsC

d. Genomic Features and **Operational Metrics** should be used together. While NGS calling pipelines produce useful quality scores, certain known genomic regions can produce false positives which can "sneak through" and appear high confidence. These regions can be separately identified.

Run, sample, and call level QC metrics all pass.



No Challenging Genomic Features:
- Low Complexity
- High Homology
- Very High/Low GC

e. The **Genome in a Bottle** data are useful for this purpose, but are highly biased toward "easy" genomic regions and variant types. They must be carefully used in any performance study.

f. FDR for both **Unique** and **Non-unique variant counts** should be considered as False+ rates can be highly site and region specific.

g. No Overfitting: The confirmation data used in this analysis should **not** be identical to the data used in validation or to choose QC thresholds.

Conclusions

Our framework is the first effort to combine data across clinical NGS labs to help evaluate the value of orthogonal confirmation and determine the appropriate burden of proof to potentially change practice in distinct cases. We believe these results and this framework can contribute to the ongoing community dialog on this subject.