To Confirm or Not 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], Justin Zook[4], Shan Yang[1], Mark Bowser[3], Swaroop Aradhya[1], Drew Watson[1], Adam Rosendorff[1], Shashikant Kulkarni[5], and Heidi Rehm[2,3]

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

Introduction

Clinical next-generation sequencing (NGS) is rapidly becoming an established diagnostic tool. Because genetic test results can influence significant medical decisions, many labs rule out false positive NGS calls by confirming variants with an orthogonal method such as Sanger sequencing. The collective experience of many laboratories doing so is that the great majority of NGS calls indeed confirm, at least when using the best available NGS methods. Thus, in these labs, the need to broadly apply orthogonal confirmation has been questioned, consistent with the ACMG NGS guidelines[1]. However, identifying the highest confidence NGS calls and quantifying the degree of confidence in these calls with different NGS protocols can be challenging.

[1]Rehm et al., Genet Med 2013

Methods

Our cross-laboratory collaboration has developed a framework for addressing this challenge which can be used with each lab's assay targets, NGS protocols, bioinformatics algorithms, and QC thresholds. Our framework involves:

1. Each laboratory chooses, 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 admit all true positives in order to ensure sensitivity.
2. Each laboratory assembles 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 QC criteria. If the demonstrated accuracy is adequate, variants meeting these 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.

All data in this poster use the Illumina HiSeq 2500 or MiSeq, v3+ SBS chemistry and 2x100 or 2x150 reads. Target enrichment is is by multiple hybridization based methods.

Laboratory Results

Laboratory 1 (Invitae): A combined data set from (a) validation, (b) clinical studies, and (c) clinical testing was assembled for a 216 gene NGS assay. Calls in this data set were strictly filtered by:

	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 mapable, not Segmental duplication (self-chain), 25-65% GC

Note: For 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 orthogonal genetic data, 2797 met all of these criteria. No false+ were observed:

	Variant Calls	Unique Variants	False+
SNVs	2259	942	0
Indels ≤ 5bp	464	307	0
Indels > 5bp	74	53	0

The Jeffreys confidence intervals on these FDRs are:

	Variant Calls	Unique Variants
SNVs	0.00% – **0.11%**	0.00% – **0.27%**
Indels ≤ 5bp	0.00% – 0.54%	0.00% – 0.81%
Indels > 5bp	0.00% – 3.33%	0.00% – 4.61%

For SNVs that meet these strict criteria, a 0% FDR was observed and a rate no higher than 0.3% is estimated, informing decisions about confirmation of such calls.

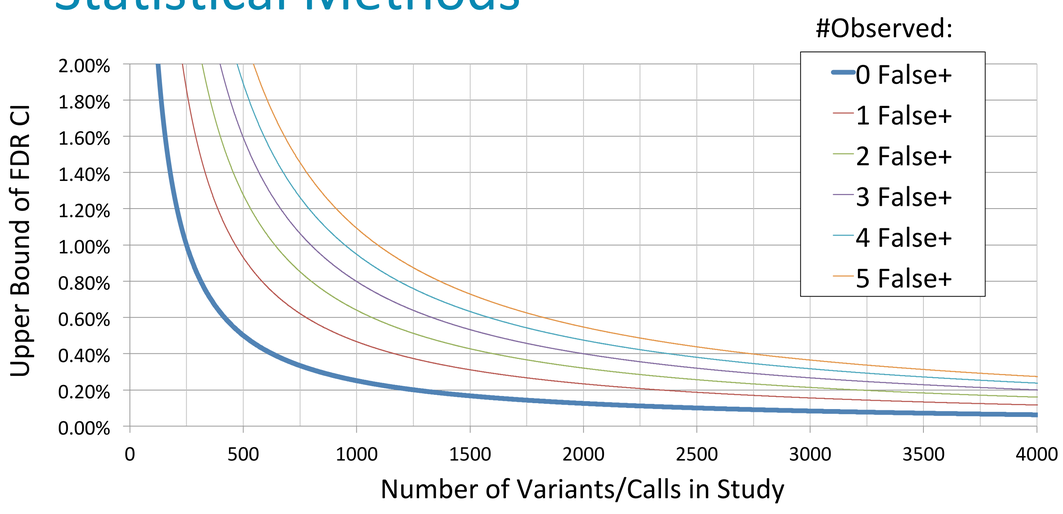
Laboratory 2 (Harvard LMM): A data set of 4521 calls for different panels was assembled passing similar (but not identical) QC criteria. No passing false+ were observed. The count and FDR CI upper bound was:

	Variant Calls		Unique Variants	
SNVs	4251	(**0.06%**)	408	(**0.61%**)
Indels ≤ 5bp	33	(7.28%)	7	(29.2%)
Indels > 5bp	2	(66.7%)	2	(66.7%)

Note: These data include multiple replicates of NA12878 and fewer clinical specimens, hence the different ratio of unique variants to total calls and relatively small number of indels.

Statistical Methods

We use FDR (False Discovery Rate) as our key metric providing a simple, clinically relevant view of the importance of confirmation[1]. We use the upper bound of the Jeffreys confidence interval[2] (CI) on FDR as a conservative estimate of performance at a standard scientific level of proof.



1: **FDR** is the fraction of reported positives that are false**.** Recall that FDR and FPR (False Positive Rate) are different: FPR takes into account the number of true (vs. false) negative results, which varies and is often not reliably known in such studies. As 1 – specificity, FPR is also often reported per base-pair tested, which may not be intuitive to clinicians for this purpose.

2: We used the **Jeffreys method** to calculate 95% binomial proportion confidence intervals. This Bayesian approach is 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.

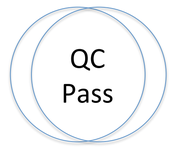
Best Practices

a. Conservative Starting Criteria must be set based on lab director NGS experience and the lab's own validation data.

Strict criteria, for specificity	Highly likely true positives, may not need confirmation
Lower criteria, for sensitivity	Possible true positives, will require confirmation
Failures	Highly likely false positives, not included in reports

b. Use **Genomic Regions** and **Operational QC Metrics** together**.** While Picard and GATK produce useful quality metrics, certain genomic regions produce false positives which can "sneak through" and appear high confidence.

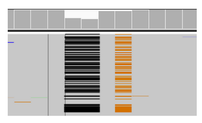
Run, sample, and call level QC metrics all pass.



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

c. Positive Sample Tracking: Sanger confirmation plays a role not just in detecting false positives, but also in ensuring that reported variants are indeed in the reported patient. An alternate mechanism should be in place if Sanger confirmation is not employed.

d. "Spelling" differences. Sanger confirmation also helps ensure that the specific description of the variant is correct when local read alignments have ambiguity. Alternate means should be in place if confirmation is not used.



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

Confirmed as: c.4_6delinsC

e. The **Genome in a Bottle (GIAB)** data are useful for this purpose, but are currently biased toward "easy" genomic regions. If GIAB is the dominant source of data in this analysis then FDR claims should be restricted to the GIAB high-confidence regions.

NA12878 Genome in a Bottle Consortium

f. 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. We believe these results and this framework can contribute to the ongoing community dialog on this subject.