# "SMRTer Confirmation": Scalable Clinical Read-through Variant Confirmation Using the Pacific Biosciences RSII SMRT® Sequencing Platform

Sarah A. McCalmon, Karel Konvicka, Nithin R. Reddy, John Whittaker, Curtis Kautzer, Jan Risinger, Jon Sorenson, Rachel Lewis, Michael Kennemer, Eric C. Olivares, and Adam Rosendorff
Invitae Corporation, San Francisco, CA

*Disclosure statement: All authors are employees and stockholders of Invitae Corporation.*

## Introduction

Next-generation sequencing (NGS) has significantly improved the cost and turnaround time for diagnostic genetic tests. ACMG recommends variant confirmation by an orthogonal method, unless sufficiently high sensitivity and specificity can be demonstrated using NGS alone (1). Most NGS laboratories make extensive use of Sanger sequencing for secondary confirmation of SNVs and indels, representing a large fraction of the cost and time required to deliver high quality genetic data. Despite its established data quality, Sanger is not a high-throughput method by today's standards from either an assay or analysis standpoint, as it involves manual review of Sanger traces and is not amenable to multiplexing. Toward a scalable solution for confirmation, Invitae has developed a fully automated and LIMS-tracked assay and informatics pipeline that utilizes the Pacific Biosciences RSII SMRT® sequencing platform. Using a barcoded-amplicon multiplexing approach, individual variants for confirmation are pooled from library prep through sequencing to reduce laboratory hands-on time and sequencing burden, in turn significantly lowering the cost of variant confirmation. Utilizing long (4hour) RSII movie lengths to sequence 96 barcoded-variants per SMRTcell, individual amplicons can be sequenced at a depth of 100-500X with an average of 50 CCS passes per molecule to achieve sufficiently high accuracy for variant confirmation. In a feasibility data set of 243 confirmations, we demonstrate the superior percentage of variants confirmed by PacBio (96.4% vs. 81.7% by Sanger). Further, we performed a clinical validation of this approach using a representative set of 30 previously Sanger-confirmed variants to demonstrate the equivalence of PacBio to Sanger for variant confirmation with 100% accuracy. Finally, we demonstrate the reduced cost per amplicon and enhanced scalability of the PacBio-driven confirmation pipeline.

## "SMRTer Confirmation" Workflow

### 1. Assay Workflow

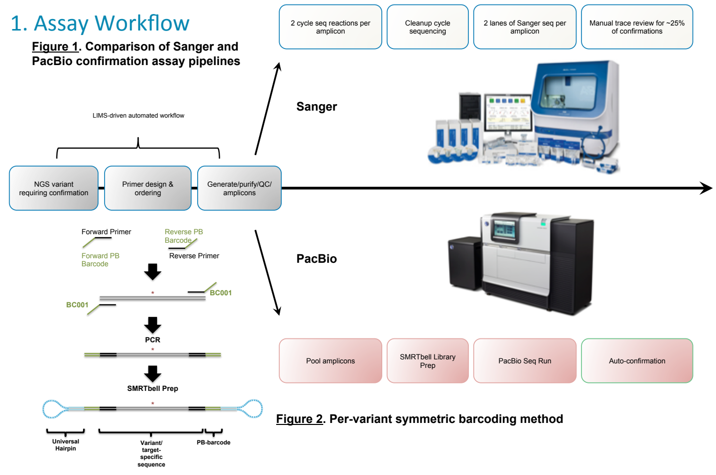

**Figure 1. Comparison of Sanger and PacBio confirmation assay pipelines**

**Figure 2. Per-variant symmetric barcoding method**

### 2. Invitae's PacBio-driven Variant Calling & Confirmation Pipeline

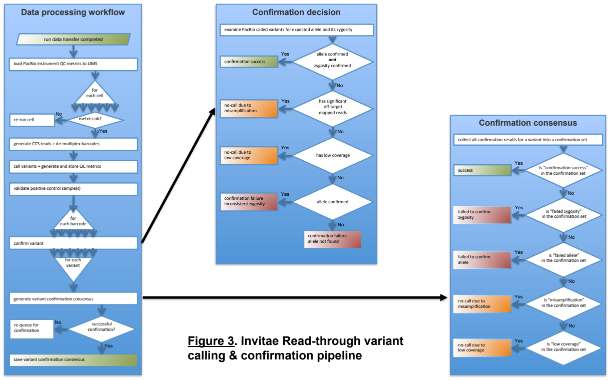
Data processing workflow
run data transfer completed
load PacBio instrument QC metrics to LIMS
for each cell
metrics ok?
No
re-run cell
Yes
generate CCS reads + de-multiplex barcodes
call variants + generate and store QC metrics
validate positive control sample(s)
for each barcode
confirm variant
for each variant
generate variant confirmation consensus
successful confirmation?
No
re-queue for confirmation
Yes
save variant confirmation consensus
Confirmation decision
examine PacBio called variants for expected allele and its zygosity
allele confirmed and zygosity confirmed
Yes
confirmation success
No
has significant off-target mapped reads
Yes
no-call due to misamplification
No
has low coverage
Yes
no-call due to low coverage
No
allele confirmed
Yes
confirmation failure inconsistent zygosity
No
confirmation failure allele not found
Confirmation consensus
collect all confirmation results for a variant into a confirmation set
is "confirmation success" in the confirmation set
Yes
success
No
is "failed zygosity" in the confirmation set
Yes
failed to confirm zygosity
No
is "failed allele" in the confirmation set
Yes
failed to confirm allele
No
is "misamplification" in the confirmation set
Yes
no-call due to misamplification
No
is "low coverage" in the confirmation set
Yes
no-call due to low coverage

**Figure 3. Invitae Read-through variant calling & confirmation pipeline**

## Methods

**Variants:** Insertions or deletions eligible for read-through variant confirmation were ≤ 50nt. Feasibility variants were those run in real-time through Invitae's Sanger confirmation pipeline, then run in parallel with the PacBio method. No prior knowledge of true-positive or true-negative status was determined. Samples chosen for the clinical validation were previously tested and determined to have SNV's or small insertions or deletions (≤50nt). These variants were all discovered via Invitae's NGS assay and were all previously confirmed via Sanger sequencing. Each of the 30 variants was repeated in triplicate within each PacBio pool and sequencing run. These were run alongside a positive and a negative control (also run in triplicate), for a total batch size of 96. The entire batch of 96 was repeated in triplicate for inter-run reproducibility. Each variant selected was only confirmed with one primer pair (selected from previous successful Sanger confirmation results). **PCR primers:** Primers were designed using Primer3 software (2). Primer pairs for each target were designed in symmetric fashion (see Fig.2) with 1 of the 384 pre-validated 16nt sequences from PacBio. Amplicon designs ranged from 200bp-1kb. **Template prep & sequencing:** Equimolar-pooled, barcoded amplicons were converted to SMRTbell templates using the SMRTbell Template Prep Kit 1.0. Primer annealing, binding, and dilution was performed according to the PacBio Binding Calculator recommendations. DNA Polymerase Binding Kit P6 v2 was used for binding chemistry. PacBio RSII DNA Internal Control Complex, DNA Sequencing Reagent Kit 4.0 v2, and PacBio RSII SMRTcells 8Pac v3 were employed for sequencing. RSII sequencing was performed using 240min movie lengths with standard diffusion loading.

## Methods

**Confirmation pipeline:** An automated pipeline was developed for the PacBio confirmation workflow (Fig. 3). The pipeline processes PacBio output and generates de-multiplexed fastq files with circular consensus sequencing (CCS) reads, then uses BWA-MEM and GATK HaplotypeCaller to align the reads and call variants. For each barcode, the variants are compared to the variant being confirmed using CGA Tools 'testvariants'. Since each variant can be confirmed using more than one amplicon, the individual amplicon confirmation results need to be aggregated into a "confirmation consensus" call. If the "confirmation consensus" call is not a no-call, the variant confirmation is considered complete and saved to Invitae's internal variant database.

## Results

### 1. Assay Scalability & Cost Comparison

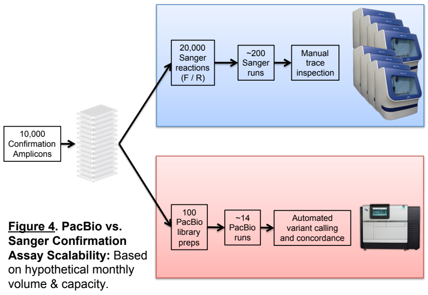
10,000 Confirmation Amplicons
20,000 Sanger reactions (F / R)
~200 Sanger runs
Manual trace inspection
100 PacBio library preps
~14 PacBio runs
Automated variant calling and concordance

**Figure 4. PacBio vs. Sanger Confirmation Assay Scalability:** Based on hypothetical monthly volume & capacity.

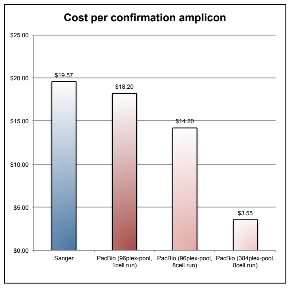

**Figure 5. Per amplicon confirmation cost comparison.** Includes all reagent & labor costs.

### 2. Feasibility

**PER-AMPLICON CONFIRMATION**

| | | SANGER | | |
|---|---|---|---|---|
| | | Confirmed | False_Positive | No_Call |
| PACBIO | Confirmed | 428 | 1 | 233 |
| | Failed_Allele | 2 | 16 | 43 |
| | Failed_Zygosity | 0 | 0 | 2 |
| | Nocall_Misamp | 0 | 0 | 5 |

**PER-AMPLICON**

| Status | # |
|---|---|
| Confirmed in both | 428 |
| PB confirmed, Sanger failed | 234 |
| Sanger confirmed, PB failed | 2 |
| Failed in both | 66 |
| Total | 730 |
| PacBio Confirmed (%) | 90.7% |
| Sanger confirmed (%) | 58.9% |

**Tables 1 & 2. Clinical Feasibility Data per Amplicon:** Table 1 (left) shows confirmation status per amplicon for a total of 730 amplicons. Table 2 (right) **PacBio confirmed 90.7% of variants (single amplicon assessment) vs. 58.9% by Sanger.**

**PER-VARIANT CONSENSUS**

| | | SANGER | | |
|---|---|---|---|---|
| | | Confirmed | False_Positive | No_Call |
| PACBIO | Confirmed | 205 | 2 | 36 |
| | Failed_Allele | 1 | 1 | 5 |
| | Nocall_Misamp | 0 | 0 | 2 |

**PER-VARIANT**

| Status | # |
|---|---|
| Confirmed in both | 205 |
| PB confirmed, Sanger failed | 38 |
| Sanger confirmed, PB failed | 1 |
| Failed in both | 8 |
| Total | 252 |
| PacBio Confirmed (%) | 96.4% |
| Sanger confirmed (%) | 81.7% |

**Tables 3 & 4. Clinical Feasibility Confirmation Consensus per Variant:** Table 3 (left) shows the consensus confirmation status per variant across multiple amplicons per variant for a total of 730 amplicons. Table 4 (right) **PacBio confirmed 96.4% of variants vs. 81.7% by Sanger.**

### 3. Validation

| Variant Category | Category Description |
|---|---|
| SNV #1-15 | 15 unique SNV's, representative of all possible 12 SNV variants |
| INDEL#1-5 | 5 unique ins or del ranging in size from 1-4 nt's. |
| INDEL#6-10 | 5 unique insertions or deletions ranging in size from 5-10 nucleotides. |
| INDEL#11-15 | 5 unique ins or del ranging in size from 11-15 nt's. |

**Table 5. Variant Categories and Sample Size Included for 3 Validation Runs.** A total of 30 unique SNV's & indels were run in triplicate along with a negative & positive control within each sequencing run for a batch size of 96 variants/pool/SMRTcell.

| Run | VarType | Count | Confirmed | Accuracy |
|---|---|---|---|---|
| RU#1 | INDEL | 45 | 45 | 100% |
| RU#2 | INDEL | 43 | 43 | 100% |
| RU#3 | INDEL | 45 | 45 | 100% |
| Inter-run Accuracy Summary for all Indels: | | | | 100% |
| RU#1 | SNV | 45 | 45 | 100% |
| RU#2 | SNV | 45 | 45 | 100% |
| RU#3 | SNV | 45 | 45 | 100% |
| Inter-run Accuracy Summary for SNV's: | | | | 100% |
| Inter-run Accuracy Summary for all | | | | 100% |

**Table 6. Accuracy by Run and Variant Type.** Three sequencing runs (RU's) were completed for the same batch of 96 variants (Table 5) for inter- and intra-run reproducibility. Not including 2 assay failures in RU#2 (Table 7), the accuracy for indels and SNV's was 100% for each run and across all runs.

| | Total # assays | # Failed assays | Assay failure rate (%) |
|---|---|---|---|
| SNPs | 135 | 0 | 0% |
| INDELs | 135 | 2 | 1.48% |

**Table 7. Confirmation assay failures:** Two INDEL assays failed to produce sufficient coverage (due to either mis-amplification or true low-coverage) to successfully call a variant. These two assays correspond to INDEL1 and INDEL12 in sequencing run RU#2. The other two assay replicates of INDEL1 and INDEL12 in RU#2 were successful.

## Conclusions

Modern clinical diagnostic laboratories must meet the highest standards in data quality in primary calling and confirmation of variants while enabling cost- and time-savings for enhanced patient care. We have demonstrated that the use of PacBio's RS II platform is able to meet both demands by incorporating high quality data into an entirely automated confirmation pipeline for SNV's and small indels. We have demonstrated that PacBio is not only equivalent, but superior to Sanger sequencing for confirmation purposes through the analysis of a feasibility data set of 730 amplicons containing 252 unique patient variants (96.4% vs. 81.7%). This approach was rigorously validated with a unique variant set of 30 distinct SNV's and indels representing a wide range of sequence contexts and sizes, in which we demonstrated 100% accuracy with Invitae's PacBio-driven "SMRTer-Confirmation" pipeline across three independent sequencing runs. As the volume of samples increases, we will be able to keep pace with the demands of the confirmation burden with minimal additional capital equipment costs or hand-on labor needs due to the enhanced scalability of the multiplexed PacBio workflow. Further cost and time savings will be possible through the ability to multiplex up to 384 amplicons per pool.

## References

**1.** Rehm HL1, Bale SJ, Bayrak-Toydemir P, Berg JS, Brown KK, Deignan JL, Friez MJ, Funke BH, Hegde MR, Lyon E; Working Group of the American College of Medical Genetics and Genomics Laboratory Quality Assurance Commitee. ACMG clinical laboratory standards for next-generation sequencing. Genet Med. 2013 Sep;15(9):733-47.

**2.** Rozen S, Skaletsky H (2000) Primer3 on the WWW for general users and for biologist programmers. In: Krawetz S, Misener S (eds) Bioinformatics Methods and Protocols: Methods in Molecular Biology. Humana Press, Totowa, NJ, pp 365-386