Accurate detection of small and large copy number events from targeted next-generation sequence data

K.B. Jacobs, J.S. Paul, G.B. Nilsen, M. Mikhaelian, R.K. Hart, M. Johnson, S.E. Lincoln, J.M. Sorenson.
INVITAE Corporation, San Francisco, CA, 94107
www.invitae.com

Background

Germline copy number variants (CNVs) can be detected from next-generation sequencing (NGS) data generated using targeted DNA capture technologies (e.g. exomes and other panels), however methods for doing so must overcome many technical challenges. Several algorithms have been published to detect CNVs in such data, though they may not yet be adequate for use in diagnostic testing laboratories, particularly for detection of small single-exon CNVs. Thus, diagnostic testing laboratories often resort to expensive and low-throughput methods such as MLPA to discover and confirm small CNVs. As a result, clinicians must carefully decide whether to order both a sequencing test and a deletion/duplication test for their patients. A single test that can accurately assay both types of alterations would improve patient access to comprehensive genetic testing.

We present a new method, CNVitae, which is designed to detect single-exon CNVs as well as larger regions sequenced using NGS. CNVitae is based on a statistical model for read counts and employs model-based segmentation algorithms optimized for use with sparsely distributed and highly variable targets across the genome. This framework estimates the most likely copy number for all segments, and, critically for clinical use, each called segment is assigned a robust quality score indicating confidence in the copy number determination.

Materials

A. NIBSC CNV Reference Samples: National Institute for Biological Standards and Control (NIBSC) copy number reference samples.

B. NIGMS CNV Reference Samples: The Human Genetic Cell Repository sponsored by the National Institute of General Medical Sciences (NIGMS) database[1] covers the majority of the most commonly encountered chromosomal conditions in clinical practice, as well as many rarely seen chromosomal abnormalities.

C. Stanford hereditary breast and ovarian cancer (HBOC) study: a research collaboration with Stanford University where we applied our multi-gene panel to bio-banked samples from consented breast and ovarian cancer patients with family histories of cancer.

D. Massachusetts General HBOC study: a research collaboration with Massachusetts General Hospital where we applied our multi-gene panel to bio-banked samples from consented breast and ovarian cancer patients with family histories of cancer.

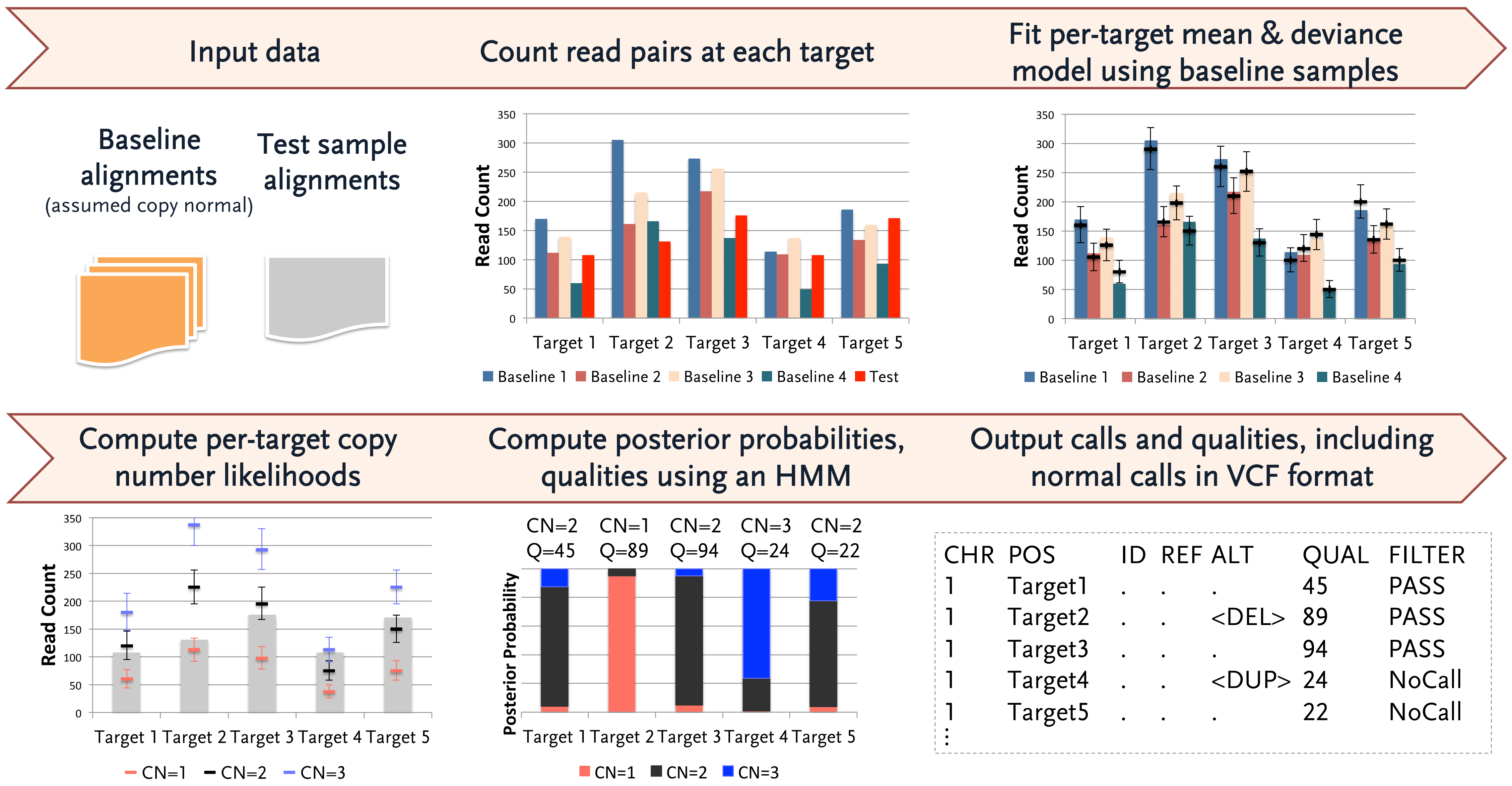
E. INVITAE research samples: 16 research samples sequenced with INVITAE's diagnostic assay at an average coverage depth of 300x at ~4,000 targets.

F. 1000 Genomes: 8 Utah Residents with Northern and Western European Ancestry (CEU). High-coverage exome data from 1000 Genomes Project[2] sequenced on the Ilumina platform (2x76 PE, 4.4-5.1 Gbp, targets with >50 read pairs). Common CNV sites were excluded from these data[3]. Samples used: NA06994, NA11840, NA12249, NA12272, NA12273, NA12275, NA12718, NA12760.

Results

CNVitae was evaluated on high-depth targeted NGS data generating using Agilent SureSelect capture and Illumina TruSeq 2x150 paired-end sequencing. In a study of 362 patient and reference samples known to carry clinically relevant CNVs, we detected all known single-exon or larger with high confidence. Several sub-exon scale CNVs were not detected by the current algorithm. Under a simulation model, altering the observed laboratory data *in silico*, we achieved a sensitivity and specificity of >99% to detect single exon hemizygous deletions at a confidence threshold of Q25 (probability of error < 0.5%). We saw 97% sensitivity and >99% specificity to detect single exon duplications (CN=3) while four exon duplications were detected with sensitivity of >99%.

Algorithm Overview



```
CHR  POS       ID  REF ALT    QUAL  FILTER
1    Target1   .   .   .      45    PASS
1    Target2   .   .   <DEL>  89    PASS
1    Target3   .   .   .      94    PASS
1    Target4   .   .   <DUP>  24    NoCall
1    Target5   .   .   .      22    NoCall
⋮
```

Validation Results

INVITAE's clinical assay was performed on samples from datasets A - D, resulting in high-depth targeted NGS data generating Agilent SureSelect capture and Illumina MiSeq 2x150 paired-end sequencing. The resulting sequence reads had an average coverage depth of over 400x and where analyzed for exon-sized or larger copy number variants using CNVitae. Shown below are the results for samples with known copy number variants in genes captured by the INVITAE assay.

Sample Source	Samples	Sensitivity	Specificity	Novel CNVs	Known CNVs present
A. NIBSC CNV Reference Samples	4	100%		4	MSH2 exon 7 deletion, MSH2 exon 1-2 deletion, MLH1 exon 13 duplication
B. NIGMS CNV Reference Samples	20	100%		4	Chromosome aneuploidies, PMP22 duplication, CFTR exons 2-3 deletion
C. Stanford HBOC study samples	6 known positive 214 known negative	100%	100%	41	BRCA1 & BRCA2 single and multi-exon deletions and duplications
D. Massachusetts General HBOC study samples	6 known positive 112 known negative	100%	100%	17	BRCA1 & BRCA2 single and multi-exon deletions and duplications

Sensitivity and specificity presented are for confirmation of known copy number variants present within each sample. Novel CNVs detected are inclusive of over 211 genes and >4,000 exon targets that are included in the INVITAE assay. These novel findings are being confirmed using orthogonal technologies. Conservatively assuming all novel findings are false findings, these represent a minimum per-exon specificity of 99.995% or a maximum false-positive call rate of 1 per 5.5 assays. Several sub-exon scale CNVs were not detected by the this algorithm (by design). These smaller CNVs were detected by read-through or split-read analysis, both standard features of INVITAE's latest analysis pipeline.

Simulation Results

We compared CNVitae with the ExomeCopy[4] software in two sets of samples under a simulation model which introduces 1 and 4 exon heterozygous deletions (del) and copy-number 3 duplications (dup). False positive and false negative rates are reported for both methods. Since CNVitae provides quality scores for normal and abnormal copy number calls, the rate of low quality "no calls" are reported separately for abnormal and normal copy number calls: "%Pos no call" and "%Neg no call", respectively.

	E. InVitae Assay, 16 research samples						F. 1000 Genomes, 8 CEU exomes					
	CNVitae				ExomeCopy		CNVitae				ExomeCopy	
TEST	%Pos no call	%Neg no call	%False Pos	%False Neg	%False Pos	%False Neg	%Pos no call	%Neg no call	%False Pos	%False Neg	%False Pos	%False Neg
Normal		2.82	0.01		0.26			9.44	0.00		0.11	
1 exon dup	2.92	2.82	0.01	0.08	0.41	13.12	12.53	9.44	0.00	0.27	0.14	29.37
4 exon dup	2.79	2.82	0.01	0.06	0.37	2.66	11.79	9.44	0.00	0.17	0.24	3.46
1 exon del	0.81	2.82	0.01	0.00	0.30	0.01	2.89	9.44	0.00	0.01	0.19	2.90
4 exon del	0.71	2.82	0.01	0.00	0.20	0.01	2.25	9.44	0.00	0.00	0.17	0.08

References

1. Tang et al. (2013) "A dynamic database of microarray-characterized cell lines with various cytogenetic and genomic backgrounds", G3 (Bethesda). 2013 Jul 8;3(7):1143-9. doi: 10.1534/g3.113.006577.
2. McVean *et al. (2012)* "An integrated map of genetic variation from 1,092 human genomes", Nature 491, 56–65. doi:10.1038/nature11632
3. Conrad et al.(2010) "Origins and functional impact of copy number variation in the human genome," Nature 464;7289;704-12. doi:10.1038/nature08516
4. Love et al. (2011) "Modeling Read Counts for CNV Detection in Exome Sequencing Data", Statistical Applications in Genetics and Molecular Biology: Vol. 10 : Iss. 1, Article 52. doi:10.2202/1544-6115.1732