VisCap: inference and visualization of germ-line copy-number variants from targeted clinical sequencing data

Genetics inMedicine | ORIGINAL RESEARCH ARTICLE Official journal of the American College of Medical Genetics and Genomics

Open

VisCap: inference and visualization of germ-line copy-number variants from targeted clinical sequencing data

Trevor J. Pugh, PhD[1,2], Sami S. Amr, PhD[3–5], Mark J. Bowser, MS, MPH[3], Sivakumar Gowrisankar, PhD[3], Elizabeth Hynes, BS[3], Lisa M. Mahanta, BA[3], Heidi L. Rehm, PhD[3–5], Birgit Funke, PhD[3–5] and Matthew S. Lebo, PhD[3–5]

Purpose: To develop and validate VisCap, a software program targeted to clinical laboratories for inference and visualization of germ-line copy-number variants (CNVs) from targeted next-generation sequencing data.

Methods: VisCap calculates the fraction of overall sequence coverage assigned to genomic intervals and computes log2 ratios of these values to the median of reference samples profiled using the same test configuration. Candidate CNVs are called when log2 ratios exceed user-defined thresholds.

Results: We optimized VisCap using 14 cases with known CNVs, followed by prospective analysis of 1,104 cases referred for diagnostic DNA sequencing. To verify calls in the prospective cohort, we used droplet digital polymerase chain reaction (PCR) to confirm 10/27 candidate CNVs and 72/72 copy-neutral genomic regions scored by VisCap. We also used a genome-wide bead array to confirm the absence of CNV calls across panels applied to 10 cases. To improve specificity, we instituted a visual scoring system that enabled experienced reviewers to differentiate true-positive from false-positive calls with minimal impact on laboratory workflow.

Conclusions: VisCap is a sensitive method for inferring CNVs from targeted sequence data from targeted gene panels. Visual scoring of data underlying CNV calls is a critical step to reduce false-positive calls for follow-up testing.

Genet Med advance online publication 17 December 2015

Key Words: copy-number variation; germ-line; molecular genetics; targeted clinical sequencing; visualization; VisCap

INTRODUCTION

Targeted DNA sequencing of disease-associated genes has long been a mainstay of clinical genetic testing to uncover causative sequence variants. Implementation of next-generation sequencing (NGS) platforms in clinical laboratories has expanded the genomic footprint of these tests,[1] and panels testing tens to thousands of genes simultaneously are now offered by academic[2–4] and commercial centers.[5–7] Initially deployed for detection of single-nucleotide variants and small insertions or deletions, many of these panels target genes for which copy-number variation is also an important source of pathogenic genome variation.

Our laboratory initially deployed targeted NGS-based tests for causative sequence variants underlying multiple cardiomyopathies[2,3] (Pan Cardiomyopathy v1, 1,016 exons in 46 genes; v2, 1,095 exons in 51 genes) and nonsyndromic hearing loss[4] (OtoGenome, v1, 1,236 exons in 71 genes; v2, 1,231 exons in 70 genes). These panels were originally launched as pooled bait sets that included probes targeting genes associated with Marfan syndrome (90 exons in 4 genes). However, these are diseases to which germ-line copy-number variants (CNVs) are also important contributors, and additional gene-specific copy-number assessments have historically been performed in parallel, resulting in increased cost and overall test complexity.

Several software tools have been developed to infer copy-number alterations from exome- and genome-scale NGS data.[8–13] These methods often use complex data-normalization methods that reduce sensitivity for exon-level copy-number alterations and provide highly segmented copy-number regions, resulting in high false-positive rates.[14] Proof-of-principle studies describing panel-based inference of CNVs have also been reported,[15,16] but these methods do not offer clear communication of exon-level data underlying these calls or native support to tune data visualization outputs to reflect data thresholds derived empirically by user-specific clinical validation experiments. Therefore, we set out to develop a method to infer constitutional (i.e., germ-line) copy-number alterations from targeted, clinical NGS data, with particular focus on data visualization and quality control suitable for deployment in clinical laboratories.

VisCap is a CNV-detection and -visualization tool that compares the relative depth of read coverage across arbitrary sets of genome coordinates (e.g., exons) targeted in a set of DNA samples using the same laboratory workflow. In addition to data normalization and identification of candidate CNVs, VisCap provides graphical outputs to enable quality control and manual review in the context of exon-level data supporting and surrounding each CNV call. Our tool was tailored for use in

[1]Princess Margaret Cancer Centre, University Health Network, Toronto, Ontario, Canada; [2]Department of Medical Biophysics, University of Toronto, Toronto, Ontario, Canada; [3]Laboratory for Molecular Medicine, Partners HealthCare Personalized Medicine, Boston, Massachusetts, USA; [4]Department of Pathology, Brigham and Women's Hospital and Massachusetts General Hospital, Boston, Massachusetts, USA; [5]Department of Pathology, Harvard Medical School, Boston, Massachusetts, USA. Correspondence: Trevor Pugh (trevor.pugh@utoronto.ca)

Submitted 10 June 2015; accepted 15 September 2015; advance online publication 17 December 2015. doi:10.1038/gim.2015.156

clinical molecular genetic laboratories with a focus on usability, clear communication of results in the context of clinically validated data thresholds, common annotation of CNVs between text and graphical outputs, and flexible, panel-specific annotation to enable downstream CNV interpretation.

MATERIALS AND METHODS

Generation of clinical sequencing data

Targeted DNA sequencing data were generated and analyzed as previous described.[2] Briefly, DNA fragments were sheared to ~150–200 bp and barcoded adapters were ligated to facilitate multiplexed capture and sequencing. Batches of 7–10 DNA samples were pooled, and regions of interest were isolated by in-solution capture using custom RNA baits (Agilent SureSelect). These baits correspond to previously described Pan Cardiomyopathy[2,3] and OtoGenome[4] panels as well as genes associated with Marfan and related syndromes. Captured fragments were purified and 50-bp paired-end sequencing reads were generated on a single lane of a HiSeq 2000 or 2500 instrument. Sequencing reads were aligned using bwa version 0.5.8c[17] and realigned around insertions and deletions using the Genome Analysis Toolkit[18] version 1.0.4705 (GATK) IndelRealigner. Quality scores were recalibrated using GATK version 1.0.4705 BaseRecalibrator and the total depth of coverage across each genomic interval was calculated using the GATK version 1.0.4705 DepthOfCoverage tool.

VisCap software code availability, dependencies, and inputs

VisCap is a publicly available (**Supplementary File S1** online, https://github.com/pughlab/viscap) CNV-detection and -visualization tool written in R (http://www.r-project.org) for analysis of targeted NGS data derived from hybrid-capture experiments. VisCap version 0.8 and R version 2.15.1 were used for the analyses in this report, as were dependent R libraries "gplots" version 2.11.0, "zoo" version 1.7–11, and "cluster" version 1.15.2. This program can be executed using the Unix command line or through a Windows graphical interface dependent on the R "winDialog()" command. As input, VisCap reads a directory containing interval summary files generated by the GATK DepthOfCoverage tool. For each sample, these files contain a summary of the total coverage of each genome region interval listed in a reference interval list. For the panels described in this report, each genomic interval corresponded to a single exon, although in practice these regions may be of any length and genomic location acceptable by the GATK DepthOfCoverage tool. A description of each output file is provided in **Supplementary Table S1** online.

Normalization and visualization of sequence coverage data

The initial step in the VisCap program is to generate a matrix of all intervals captured and the fraction of total coverage assigned to these intervals. These are derived for each sample from DepthOfCoverage "sample_interval_summary" files using total coverage values in the column "{sample name}_total_cvg". Next, the sample-specific fractional coverage of each region is divided by the median for that region across the entire batch, where the number of samples to be batched together can be any size that provides a representative median coverage across all target regions. In our clinical workflow, a batch refers to a set of 7–10 DNA samples captured in a single multiplexed pool and sequenced in a single flow-cell lane of an Illumina HiSeq 2000 or 2500 DNA sequencer. These batch median-normalized data are stored as a matrix of log2 ratios that is written to a text file. To facilitate visual review of the data for each sample, log2 ratios are plotted by relative genome order (i.e., rank order, not to scale on genome), and data points supporting CNVs are color-coded and linked to a unique identifier listed in the text output (**Figure 1**).

The X-chromosome requires further normalization because there are significant fractional coverage differences between males and females. Depending on the balance of males and females in the batch, males may display single-copy loss of chromosome X or females may display a single-copy gain. These patterns are evident from the presence of two clusters of boxplots depicting fractional coverage values across targets on the X-chromosome for each sample (**Figure 2**). These clusters are detected computationally by removing outlier probes and then partitioning all samples around two medoids, a more robust alternative to K-means clustering.[19] To enable consistent visualization of CNVs from male and females in the same batch, the log2 ratios for each sample are normalized toward zero through subtraction or addition of the cluster median. To facilitate review of this procedure, boxplots of log2 ratios are generated before and after subtraction/addition of the cluster medians (**Figure 2a**,**b**, respectively). The program also outputs predicted sex for each case as an additional source of quality control (QC) data.

To derive thresholds for detection of copy-number gains and losses, boxplots are constructed using the R graphics "boxplot()" command. This provides a visual representation of the distribution of log2 ratios from each sample, as well as the five-number summary used for subsequent thresholding and quality control. The five-number summary includes upper whisker, Q3, median, Q1, and lower whisker. Upper (lower) whiskers are the greatest (lowest) values that fall within Q3 (Q1) plus a user-defined multiplier of the interquartile range (Q3–Q1). A sample fails quality control if either of the boxplot whiskers extends beyond the expected theoretical log2 ratio for a single-copy gain (0.58) or a single-copy loss (−1). To avoid skewing of the batch median used for normalization, failed samples are identified in the boxplot summary output and removed, triggering an iterative additional run with the remaining samples (**Figure 2c**). VisCap automatically repeats the analysis until all samples pass this automated QC. In our laboratory, an entire batch is failed if fewer than three samples pass QC. Files generated by each iteration are stored in separate output folders.

Calling CNVs

The strategy to detect a gain or a loss is dependent on the distribution of log2 ratios for each sample as well as a set of user-defined thresholds. To be called, a copy-number gain must have log2 ratios that exceed (i) the user-defined gain threshold and (ii) the upper whisker of the boxplot representing the sample's data distribution.

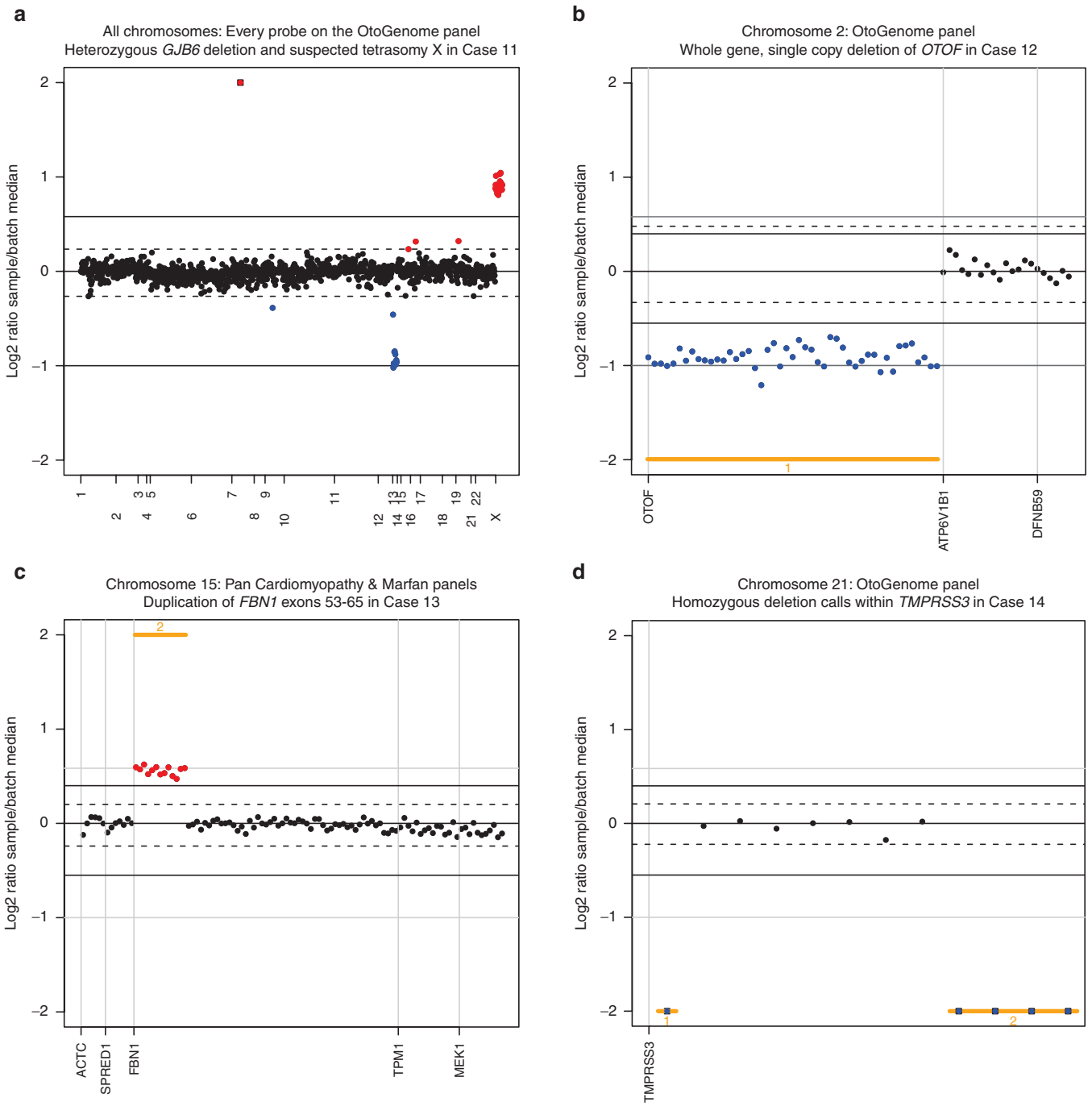

Figure 1 Example of VisCap chromosome-level outputs. Fractional depth of coverage values for each genomic interval (black dot) sequenced on a targeted panel plotted as log2 ratios against the median of a reference set of samples analyzed using the same panel and laboratory workflow. Copy-number variants supported by multiple consecutive exons with log2 ratios outside user-defined thresholds (solid black lines at –0.55 and 0.40, in this case) are color-coded: gains are red and losses are blue. An orange marker denotes the affected genomic segment with a number corresponding to the CNV identifier listed in the output text file. This plot is overlaid with guidelines depicting the two sets of user-defined thresholds used for copy-number variation (dashed = whiskers derived from the boxplot denoting the sample's overall data distribution; solid black = fixed, user-defined thresholds) and the theoretical log2 ratios (solid light gray) for single-copy gains and single-copy losses. Labels on the x-axis mark the first interval of each group of exons (commonly a gene name) as specified in the list of intervals provided to the program (e.g., TP53_Exon1 and TP53_Exon2 would be marked by a single TP53 label underneath exon 1). Panels **a–d** contain examples of VisCap output plots from four cases selected from Table 1. (**a**) Whole-genome view depicting log2 ratios from all intervals from a single patient run on a large gene panel. This individual had a *GJB6* deletion known from previous testing as well as an unexpected gain of chromosome X. The patient was phenotypically female and had median log2 ratio of X-chromosome probes twice that of other females, suggesting a potentially undiagnosed sex chromosome abnormality. (**b**) Single-chromosome view of data indicating a full gene, single-copy deletion of *OTOF* in a patient with a loss-of-function mutation on the remaining allele, illustrating compound heterozygosity resulting in nonsyndromic hearing loss. (**c**) Example of a multi-exon copy-number gain within *FBN1* leading to Marfan syndrome. (**d**) Example of two homozygous, noncontiguous exon-level deletion calls within *TMPRSS3* in a patient with nonsyndromic hearing loss. Family testing found that both of the patient's parents were heterozygous for both deletion calls, suggesting the deletions are on the same allele and may be indicative of a complex rearrangement.

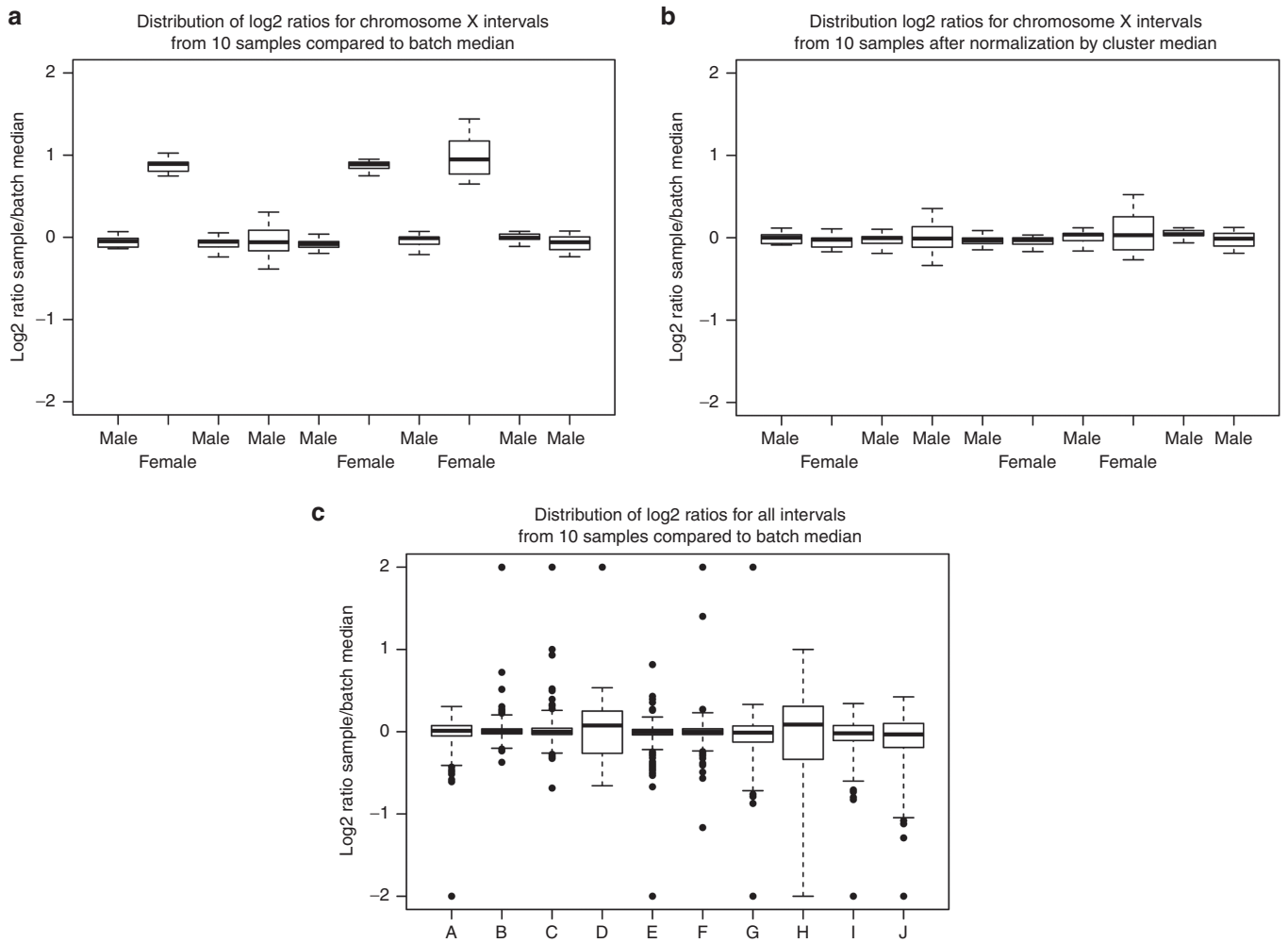

Figure 2 Normalization of log2 ratios from probes on the X chromosome. Distribution of log2 ratios from probes across all samples in a sequencing batch. Upper panels depict log2 ratios from probes on the X chromosome before (panel **a**) and after (panel **b**) inference and correction for sex composition within the sequencing batch used as a reference set. Lower panels depict log2 ratios from all probes from all samples run on a panel, including a sample that failed automated QC (panel **c**, case H) and was removed for an iterative run. Each boxplot depicts a 5-number summary dependent on the interquartile multiplier (x) set in the configuration file: 1) lower whisker is the lowest value to exceed the Q1 - x times the interquartile range; 2) lower hinge is the first quartile value (Q1); 3) middle line is the median; 4) upper hinge is the third quartile (Q3) value; 5) upper whisker is the lowest value to exceed Q3 + x times the interquartile range.

Copy-number losses must have consecutive probes below the loss threshold and the lower boxplot whisker. For our prospective study, these fixed thresholds were established based on the retrospective training set of 14 positive control samples (see Results, **Supplementary Figure S1** online), leading to a requirement for the minimum log2 ratio for gains of 0.40 and maximum log2 ratio for losses of −0.55. The boxplot IQR multiplier was set at 3.

Confirmation of candidate CNVs by genome-wide bead array

To estimate our method's ability to identify both true regions of copy-number changes and copy-neutral regions, we analyzed a subset of 16 samples from the validation set using the 2.3 million–feature Illumina Human Omni2.5–8 BeadChip array at the Princess Margaret Genomics Centre, Toronto, Canada (http://www.pmgenomics.ca). Each sample was processed following the Illumina Infinium LCG assay protocol, hybridized to two BeadChips, stained as per Illumina protocol, and scanned on an Illumina iScan. The data files were quantified and normalized in the GenomeStudio version 2010.2 genotyping module using HumanOmni25-8v1-1_C.bpm manifest.

To call CNVs, data were exported from Genome Studio and uploaded into Biodiscovery Nexus v7.5 program. The significance threshold for segmentation was set at 1×10^{-8} and also required a minimum of three probes per segment and a maximum probe spacing of 1,000 between adjacent probes before breaking a segment. The log ratio thresholds for single-copy gain and single-copy loss were set at 0.13 and −0.23, respectively. Systematic GC wave correction was applied using Linear Correction with the HumanOmni2-5-8v1-1-C_hg19_illum_correction.txt file.

Confirmation of candidate CNVs by droplet digital PCR

For more in-depth validation of VisCap, a minimum of one primer/TaqMan probe (Life Technologies, Carlsbad, CA)

combination was designed to verify candidate CNVs called by VisCap. These oligonucleotides typically targeted an exon and were designed to avoid overlap with single-nucleotide polymorphisms and nonunique homologous regions when possible. Multiple probes were designed for larger or multigene CNVs. Droplet digital polymerase chain reaction (PCR) was performed using the Bio-Rad QuantaLife Droplet Digital PCR (ddPCR) device comparing the target of interest against a control probe targeting *RPP30*. To control for potential CNV of *RPP30*, we compared the copy number of *RPP30* against a second control, *AP3B1*. As a further control, all ddPCR assays were run simultaneously against DNA from the NA12878 cell line (Corriel). Results were analyzed using the QuantaSoft program and further normalized by dividing copy-number values by a sample-specific "Reference Correction Constant," calculated as two-times the copy-number ratio of the *AP3B1* and *RPP30* reference genes. Samples with an *RPP30:AP3B1* ratio <0.9 or >1.2 were manually reviewed for quality or potential copy-number variation of reference genes. For each loci tested, a loss was called if the normalized value was <1.5, and a gain was called if the normalized value was >2.5.

RESULTS

Validation of this tool involved retrospective analysis of 14 cases with known CNVs detected by other clinical tests, as well as prospective analysis of 1,104 cases followed by confirmation of 27 candidate CNVs using ddPCR.[20] The ddPCR-confirmed CNVs were used to estimate the positive predictive value of variant calls and to assess the added value of visual scoring of the graphical VisCap output.

Retrospective analysis of 14 cases with known CNVs

To establish algorithmic thresholds to maximize the sensitivity and specificity of the VisCap program, we analyzed targeted DNA sequencing data from 10 hearing loss cases, 3 Marfan syndrome cases, and 1 hypertrophic cardiomyopathy case, with a total of 15 pathogenic CNVs known from previous testing (**Table 1**). This analysis uncovered 165 candidate CNVs at exon-level resolution, including all 15 pathogenic CNVs and 97 unique CNV calls. Of the unique CNVs, 95 were supported by data from only one or two exons and were often seen in multiple samples. We attributed many of these calls to specific baits with inconsistent capture performance. This inconsistency is probably due to extreme GC content because the small, recurrent CNV calls were enriched for exons with GC sequence <35 or >65% compared with the pathogenic and novel, larger CNVs ($P = 0.002$).

We reanalyzed all 14 cases after optimizing and implementing log2 ratio thresholds of 0.4 for gains and −0.55 for losses (see receiver-operating characteristic curve in **Supplementary**

Table 1 CNVs from 14 samples used for retrospective training of thresholds for VisCap algorithm

	Previous testing		VisCap				
Case ID	Result (heterozygous unless noted)	Method	Panel[a]	CNV call, interval range	Intervals within CNV	Median log2 ratio (copy number)	Additional CNV calls[b]
1	*CDH23* deletion of exons 13–14	Sanger, no PCR products, segregation	Oto	Loss, *CDH23* exons 12-14A	3	−0.79 (1.2)	14 > 2
2	*MYBPC3* exon 29 indel (c.2995-58_3151del215insTACCAGGCC)	MLPA	PCM	Loss, *MYBPC3* exons 28–30	3	−0.59 (1.3)	22 > 9
3	*USH2A* exon 10 deletion	aCGH	Oto	Loss, *USH2A* exon 10	1	−1.0 (1.0)	1 >1
4	*FBN1* deletion of exons 36–65	MLPA	Marfan	Loss, *FBN1* exons 36–65	30	−0.92 (1.1)	20 > 7
5	*FBN1* deletion of exons 6–65	MLPA	Marfan	Loss, *FBN1* exons 6–65	60	−0.92 (1.1)	20 > 10
6	*USH2A* deletion of exons 22–32	Microarray	Oto	Loss, *USH2A* exons 21–33	13	−0.98 (1.0)	2 > 2
7	*POU3F4* hemizygous whole-gene deletion	Sanger, no PCR products	Oto	Loss, *POU3F4* exon 1	1	−6.9[c] (0)	17 > 6
8	*USH2A* homozygous deletion of exons 63–64	Sanger, no PCR products	Oto	Loss, *USH2A* exons 63–64	2	−6.0[c] (0)	14 > 4
9	*GJB2/GJB6* regulatory region deletion	Microarray	Oto	Loss, *CRYL1* exons 2–8	4	−0.98 (1.0)	12 > 6
10	4-kb deletion on 10q22.1 including >1 exon of *CDH23*	Microarray	Oto	Loss, *CDH23* exons 13-14A	2	−1.0 (1.0)	9 > 4
11	*GJB6* D13S1830 deletion	PCR/Gel	Oto	Loss, 309kb including *GJB6* exons 1–3	5	−0.93 (1.0)	9 > 6
12	*OTOF* whole deletion	Sanger, homozygous novel LOF variant	Oto	Loss, *OTOF* exon 1-47A	47	−0.91 (1.1)	11>8
13	*FBN1* duplication	MLPA	Marfan	Gain, *FBN1* exon 53–65	13	0.57 (3.0)	10 > 10
14	*TMPRSS3* homozygous deletions of exons 2–5 and 13	Resequencing array, no PCR products	Oto	Loss, *TMPRSS3* exons 2–5; loss, *TMPRSS3* exon 13	4; 1	−4.3[c] (0.10); −4.2[c] (0.11)	4 > 2

aCGH, array comparative genomic hybridization; CNV, copy- number variant; LOF, loss-of-function; MLPA, multiplex ligation-dependent probe amplification; PCR, polymerase chain reaction; Sanger, dideoxynucleotide sequencing of PCR products; resequencing array = OtoChip; segregation, segregation with disease within family.

[a]Oto, OtoGenome; PCM, Pan Cardiomyopathy. [b]Before > after optimized thresholds applied. [c]Homozygous deletion or deletion of an X chromosome gene in a male, resulting in large, negative log2 ratios.

Figure S1 online). These thresholds enabled detection of all pathogenic variants in our training set and reduced the number of CNV calls by more than 50%, from 165 to 77 (5.5 CNVs per sample on average; **Table 1**). Given the high sensitivity for pathogenic variants and relatively few additional CNVs for follow-up interpretation per sample, we next sought to test this configuration in a larger sample set.

Prospective analysis of 1,104 cases and follow-up confirmation testing

To validate our VisCap configuration, we prospectively analyzed 1,104 cases analyzed in 113 batches using Pan Cardiomyopathy or OtoGenome panels (**Table 2**). Of these, 141 cases (12.8%) could not be scored; these consisted of 139 cases that were removed by the automated VisCap QC procedure (i.e., log2 ratio distributions were too broad to resolve single–copy number changes) and two cases that passed QC but were not analyzed due to failure of all other samples in the batch (minimum of three passing samples needed per batch). Across samples that passed QC, the median average target coverage was 724×, and 90% of cases had median target coverage between 329× and 1,684×.

From the 961 cases that passed QC, we inferred 3,005 candidate CNVs: 1,337 gains and 1,668 losses, with an average of 3.1 CNVs per sample, which is consistent with our training set. After removing single-exon CNV calls in exons frequently failing NGS analysis, 556 gains and 1,072 losses remained. Of these, 1,249 candidate CNVs corresponded to 105 unique calls seen in >1% of our cohort, likely copy-number polymorphisms or systematic artifacts; these calls were removed from further analysis. Consistent with the training cohort, the 105 recurrent calls were enriched for exons with GC content <35 or >65% compared with the remaining 378 candidate CNVs for follow-up ($P = 10^{-10}$). The CNVs for follow-up were supported by 1–37 exons and represented 0.4 CNVs per case. Of note, panels with skewed GC content are more likely to benefit from removal of recurrent artifacts calls, as illustrated by the decrease in CNV calls on the Pan Cardiomyopathy panel (3.5 reduced to 0.4 CNVs/sample) that had a median GC content of 45% compared with 52% of the OtoGenome panel (0.3 reduced to 0.2 CNVs/sample).

To confirm these candidate CNVs, and to test whether bona fide CNVs were missed due to our VisCap thresholds, we reanalyzed six samples with CNVs called by VisCap and 10 samples with no candidate CNV calls using the 2.3 million–feature Illumina Human Omni2.5–8 BeadChip array. We confirmed five of six CNVs detected by VisCap and did not identify any CNVs within the targeted panel regions of the 10 negative samples (**Supplementary Table S2** online). The sixth CNV not detected by the microarray was confirmed by digital droplet PCR, representing a likely false negative from the array platform. For further confirmation, we selected 27 of 379 candidate CNVs plus 72 negative control regions with log2 ratios between −0.55 and 0.4 (our thresholds to issue a CNV call). To assess the ability of VisCap to discriminate potentially ambiguous CNV calls, we biased our selection of copy number–neutral regions toward those with values away from log2 ratio = 0 (i.e., copy number = 2) but still below our cutoffs for calling a CNV. No CNVs were detected in the 72 negative control regions by ddPCR. Of the 27 candidate CNVs, 10 were confirmed by ddPCR (**Table 3**), with high concordance between copy number inferred by VisCap and absolute copy number detected by ddPCR ($P = 0.97$).

The 17 calls that were not confirmed by ddPCR were enriched for smaller CNVs supported by three exons on average (range, 1–8 exons). However, we were not able to differentiate false-positive and true-positive calls based on size alone, because true positive CNVs were also found in this range. Therefore, we set out to assess whether visual scoring of data surrounding candidate CNV regions could be used as an effective method to identify false-positive calls and avoid unnecessary follow-up testing.

Visual scoring of CNV calls

To determine whether visual scorers could differentiate the 10 true-positive from 17 false-positive calls, we trained four

Table 2 Summary of prospective validation cohort

Panels	Pan cardiomyopathy	OtoGenome	Total
Batches	99	14	113
Batches passing QC	98	14	112
Samples	973	131	1,104
Samples passing QC	841	122	963
Samples analyzed	839	122	961
Failure rate	13.6%	6.9%	12.8%
Mean coverage (5/50/95 percentile)	319/724/1,694	422/597/1,538	330/724/1,684
CNV calls	2,963	42	3,005
Gains	1,332	5	1,337
Losses	1,631	37	1,668
CNVs/sample	3.5	0.3	3.1
Filtered CNV calls[a]	357	22	379
Gains	110	2	112
Losses	247	20	267
CNVs/sample	0.4	0.2	0.4

[a]Removed 131 recurrent calls seen in >1% of cases.

CNV, copy- number variant.

Table 3 Candidate CNVs selected for confirmation by ddPCR and used for visual assessment training

Case ID	CNV call	Chromosome	Start interval name[b]	End interval name[b]	Intervals within CNV	VisCap visual reviews concordant with ddPCR result (four reviewers total)
Confirmed by ddPCR						
15	Gain	2	TTN Exon 233	TTN Exon 197	37	4
16	Gain	X	LAMP2 Exon 01	TAZ Exon 11	18	4
17	Gain	3	RAF1 Exon 17	RAF1 Exon 02	16	4
18	Loss	10	LDB3 Exon 01	LDB3 Exon 16	16	4
19	Gain	12	PKP2 Exon 14	PKP2 Exon 01	14	4
20	Gain	15	FBN1 Exon 65	FBN1 Exon 53	13	4
21	Gain	3	RAF1 Exon 12	RAF1 Exon 02	11	4
22	Loss	11	MYBPC3 Exon 20	MYBPC3 Exon 12	8	4
23[a]	Loss	21	TMPRSS3 Exon 05	TMPRSS3 Exon 02	4	4
24[a]	Loss	21	TMPRSS3 Exon 05	TMPRSS3 Exon 02	4	4
Not confirmed by ddPCR						
25	Loss	2	TTN Exon 112	TTN Exon 110	3	4
26	Loss	2	TTN Exon 126	TTN Exon 124	3	4
27	Loss	2	TTN Exon 170	TTN Exon 169	2	4
	Loss	12	KRAS Exon 04	KRAS Exon 04	1	4
28	Loss	2	SOS1 Exon 18	SOS1 Exon 18	1	4
	Loss	2	TTN Exon 152	TTN Exon 149	4	3
	Loss	2	TTN Exon 114	TTN Exon 112	3	3
	Loss	2	TTN Exon 170	TTN Exon 169	2	3
	Loss	12	KRAS Exon 06	KRAS Exon 04	3	2
29	Loss	2	TTN Exon 152	TTN Exon 145	8	3
	Loss	2	TTN Exon 170	TTN Exon 168	3	3
	Loss	2	TTN Exon 127	TTN Exon 125	3	3
	Loss	2	TTN Exon 114	TTN Exon 112	3	3
	Loss	12	KRAS Exon 06	KRAS Exon 04	3	3
	Loss	2	SOS1 Exon 06	SOS1 Exon 03	4	2
30	Loss	X	LAMP2 Exon 01	EMD Exon 03	4	3
31	Loss	X	LAMP2 Exon 01	EMD Exon 02	3	3

[a]Parents of case 14. [b]"Start" and "End" refer to relative genome position, whereas the exon numbers in the interval names are ordered by position on the gene transcript. Therefore, "start" exon numbers are greater than "end" exon numbers for genes on the minus strand of the genome reference.

CNV, copy-number variant; PCR, polymerase chain reaction.

laboratory technicians to read and interpret VisCap plots (training material supplied as **Supplementary File S2** online), followed by testing using a set of practice variants (**Supplementary File S3** online). The technicians next scored the VisCap plots for the 27 candidate CNVs as either true-positive or false-positive calls without knowing the outcome of the ddPCR analysis (**Table 3**). All four technicians correctly identified all 10 verified CNVs. Although two of four technicians correctly categorized all false-positive calls, the other two technicians flagged 2/17 and 12/17 false positives for follow-up, highlighting the need for training and experienced review of these data. This illustrates the added value of human review of these data to reduce the overall false-positive rate while retaining high sensitivity.

DISCUSSION

We report here an open, flexible software program (and accompanying training documents) to detect and visualize germ-line CNVs from targeted DNA sequencing data. The software was specifically designed for implementation within a clinical laboratory, including laboratory-defined thresholds, static plots amenable for routine review, and standardized procedures and training documentation. This program has been validated in a clinical diagnostic laboratory and has been used for CNV analysis of >4,000 patients in our laboratory to date (1,118 described in this report), using multiple gene panel configurations, including Pan Cardiomyopathy and OtoGenome panels. Consistent with our prospective cohort, we anticipate that this method will scale well across a wide range of coverage levels (including low-pass sequencing for CNV detection only), as long as reference samples display limited batch-to-batch technical variability across target regions (hence our focus on enabling routine manual review of VisCap data). Although NGS data used for our study were generated from genomic fragments isolated by hybrid capture followed by sequencing on the Illumina HiSeq platforms, this method is amenable to alternative target isolation and sequencing platforms that generate depth-of-coverage

data. As a proof of concept, we have previously applied this method to PCR amplicon sequencing data generated on the Ion Torrent platform (Life Technologies, Carlsbad, CA).[21]

In its current implementation, VisCap is highly sensitive (no known false negatives in our training set, 10 copy-neutral samples tested by genome-wide bead array, or 72 copy-neutral regions selected from our validation cohort) but has a relatively high false-positive rate (only 10/27 candidate CNVs targeted for verification were confirmed), a known issue in detecting small CNVs from targeted data.[14,16] Additional intronic probes may aid in identifying small CNVs, albeit potentially at an increased cost. We have addressed the issue of a high false-positive rate through manual review of all CNV calls with relatively minor impact on our overall workflow and test turnaround time (4 CNVs/case run on a 46-gene panel, <1-min review per CNV). This approach will not scale well as gene panels continue to grow, and future improvements to the software will attempt to capture features that are important for manual review but not yet part of the CNV calling algorithm. Beyond confirmation of simple gains and losses, the value of manual data review is particularly valuable to uncover potential complex structural alterations such as *STRC* deletions in hearing loss.[4] Therefore, visual interpretation of these data is currently a critical component of our clinical testing workflow. Although much of the informatics analysis is fully automated, we have demonstrated value in continued manual intervention and effective visualization for quality control and scoring of copy-number data generated by VisCap.

SUPPLEMENTARY MATERIAL

Supplementary material is linked to the online version of the paper at http://www.nature.com/gim

ACKNOWLEDGMENTS

We thank the staff and fellows of the Laboratory for Molecular Medicine and the Princess Margaret Genomics Centre for their technical assistance in generating and interpreting these data. T.J.P. is supported by funding from the Princess Margaret Cancer Foundation.

DISCLOSURE

This work was funded by internal operating funds of the Partners HealthCare Personalized Medicine. The Laboratory for Molecular Medicine is a nonprofit, fee-for-service laboratory that is a current or former employer of all authors and offers testing for cardiomyopathy, hearing loss, Marfan syndrome, and several other genetic disorders. The authors declare no other conflicts of interest.

REFERENCES

1. Rehm HL, Bale SJ, Bayrak-Toydemir P, et al.; 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;15:733–747.
2. Pugh TJ, Kelly MA, Gowrisankar S, et al. The landscape of genetic variation in dilated cardiomyopathy as surveyed by clinical DNA sequencing. *Genet Med* 2014;16:601–608.
3. Alfares AA, Kelly MA, McDermott G, 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.
4. Mandelker D, Amr SS, Pugh T, et al. Comprehensive diagnostic testing for stereocilin: an approach for analyzing medically important genes with high homology. *J Mol Diagn* 2014;16:639–647.
5. Tang W, Qian D, Ahmad S, et al. A low-cost exon capture method suitable for large-scale screening of genetic deafness by the massively-parallel sequencing approach. *Genet Test Mol Biomarkers* 2012;16:536–542.
6. Frampton GM, Fichtenholtz A, Otto GA, et al. Development and validation of a clinical cancer genomic profiling test based on massively parallel DNA sequencing. *Nat Biotechnol* 2013;31:1023–1031.
7. Kurian AW, Hare EE, Mills MA, et al. Clinical evaluation of a multiple-gene sequencing panel for hereditary cancer risk assessment. *J Clin Oncol* 2014;32:2001–2009.
8. Sathirapongsasuti JF, Lee H, Horst BA, et al. Exome sequencing-based copy-number variation and loss of heterozygosity detection: ExomeCNV. *Bioinformatics* 2011;27:2648–2654.
9. Li J, Lupat R, Amarasinghe KC, et al. CONTRA: copy number analysis for targeted resequencing. *Bioinformatics* 2012;28:1307–1313.
10. Love MI, Myšičková A, Sun R, Kalscheuer V, Vingron M, Haas SA. Modeling read counts for CNV detection in exome sequencing data. *Stat Appl Genet Mol Biol* 2011;10:.
11. Krumm N, Sudmant PH, Ko A, et al.; NHLBI Exome Sequencing Project. Copy number variation detection and genotyping from exome sequence data. *Genome Res* 2012;22:1525–1532.
12. Plagnol V, Curtis J, Epstein M, et al. A robust model for read count data in exome sequencing experiments and implications for copy number variant calling. *Bioinformatics* 2012;28:2747–2754.
13. Fromer M, Moran JL, Chambert K, et al. Discovery and statistical genotyping of copy-number variation from whole-exome sequencing depth. *Am J Hum Genet* 2012;91:597–607.
14. Samarakoon PS, Sorte HS, Kristiansen BE, et al. Identification of copy number variants from exome sequence data. *BMC Genomics* 2014;15:661.
15. Krawitz PM, Schiska D, Krüger U, et al. Screening for single nucleotide variants, small indels and exon deletions with a next-generation sequencing based gene panel approach for Usher syndrome. *Mol Genet Genomic Med* 2014;2:393–401.
16. Feng Y, Chen D, Wang GL, Zhang VW, Wong LJ. Improved molecular diagnosis by the detection of exonic deletions with target gene capture and deep sequencing. *Genet Med* 2015;17:99–107.
17. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 2009;25:1754–1760.
18. McKenna A, Hanna M, Banks E, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 2010;20:1297–1303.
19. Maechler M, Rousseeuw P, Struyf A, Hubert M, Hornik K. Cluster: Cluster Analysis Basics and Extensions. R package version. 2.0.3. 2015. https://cran.r-project.org/web/packages/cluster/citation.html.
20. Hindson BJ, Ness KD, Masquelier DA, et al. High-throughput droplet digital PCR system for absolute quantitation of DNA copy number. *Anal Chem* 2011;83:8604–8610.
21. Abou Tayoun AN, Tunkey CD, Pugh TJ, et al. A comprehensive assay for CFTR mutational analysis using next-generation sequencing. *Clin Chem* 2013;59:1481–1488.