# Accurate Prediction of Chromosome-Level CNVs from Targeted NGS

Hyunjun Nam, PhD[1] • Christophe Magnan, PhD[1] • Fernando Lopez-Diaz, PhD[1] • Ryan Bender, PhD, FACMGG[1] • Sally Agersborg, MD, PhD[2] • Segun C Jung, PhD[2] • Vincent Funari, PhD[2]
[1]NeoGenomics Laboratories, Inc., Carlsbad, CA; [2]NeoGenomics Laboratories, Inc., Aliso Viejo, CA

#3994

## Introduction

Aneuploidy and large-scale Copy Number Variations (CNVs) are prominent features of cancer cells. While Fluorescence in situ hybridization (FISH) and conventional cytogenetics (CC) are the gold standard for detecting aneuploidy and CNVs, NGS-based assays are currently used for high-resolution detection of copy number alterations assessing the whole genome. However, although an increasing number of NGS-based tools have been developed for detecting aneuploidy or CNVs from whole genome or exome sequencing data, only a limited number of options are available for targeted gene panels. Despite mechanisms provided to establish normal profiles for a specific panel, the accuracy of these tools at the chromosome level suffer when only a small number of regions are targeted on each chromosome. Here we leveraged on a custom amplicon based NGS assay designed to detect somatic alterations (SNVs and indels) in 297 hematological cancer relevant genes, previously validated in our clinical laboratory. We introduce a simple approach to accurately predict chromosome-level CNVs such as monosomy and trisomy for a targeted gene panel, commonly used in a clinical setting.

Mutation profiles, including SNVs, INDELs, and structural changes, were interrogated with an in-house bioinformatics pipeline that utilized CNVKit (Bastian et al) and PureCN (Morrissey et al) algorithms to detect structural changes. The first step consists of finding optimal panel-specific decision thresholds for gains and losses at the gene level. This step was performed using an independent set of 1,314 clinical samples sequenced with the NeoType® Heme assay developed by NeoGenomics Laboratories, Inc. for which at least one FISH test was performed in addition to the sequencing. Three genes (ATM, TP53, and NF1) were used to find optimal decision thresholds based on the FISH result for these markers. These thresholds are used afterward to predict a gain or a loss for any other gene in the panel. The second step consists of predicting the chromosome-level gain or loss based on the individual predictions at the gene level by simply observing the frequency of targeted genes on the corresponding chromosome predicted as either gained or lost by the first step approach. The 19, 7, and 18 targeted genes in the NGS panel (Table 2) were respectively used to predict monosomy 7, trisomy 8, and trisomy 12 in a second set of over 7,000 clinical samples with known ploidy for chromosomes with clinically relevant ploidy abnormalities in hematological malignancies.

## Step 1. Optimal thresholds for gains and losses of CNV at the gene level

The log2 ratio thresholds for gain and loss of CNV were optimized to discriminate the positive and negative set of Chronic lymphocytic leukemia (CLL) and Plasma cell myeloma (PCM) samples on the three genes of each FISH probes. The log2 ratio is the most commonly used to determine copy number status, which the standard formula for is log2 (observed intensity/reference intensity). The copy number of genes for CNV detection by the NGS assay is as follows: Gain of CNVs, log2 ratio ≥ 0.03; Loss of CNV, log2 ratio ≤ -0.03. Based on these criteria, a total 430 CNVs (213 gain and 207 loss) were detected in 382 (370 CLL samples and 12 PCM samples) out of 1,306 samples (1,005 CLL samples, 296 PCM samples and 5 CLL & PCM samples).

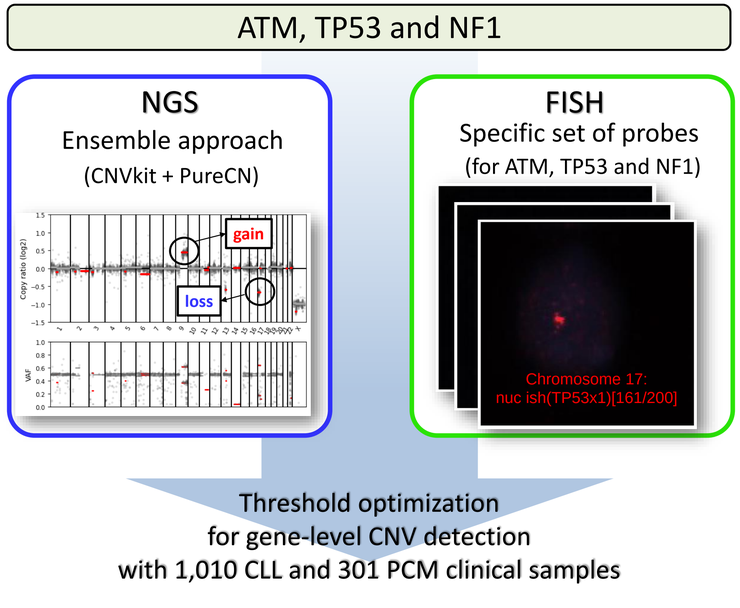


**Figure 1. Threshold optimization of gene-level CNV prediction based on FISH result for three markers, ATM, TP53 and NF1.** NGS pipeline for CNV prediction consists of two different software: CNVkit (Bastian, 2016) and PureCN (Morrissey, 2016). A custom python script initiates sample processing of each software and provides the consensus output of the CNV call for each gene with an ensemble approach. The log2 cutoffs for gain and loss was assessed with 1,306 CLL and PCM clinical samples tested with both NGS and FISH assay.



| Accuracy | |
|---|---|
| Total no: of CNVs tested | 3331 |
| No: of concordant gain of CNVs | 205/216 |
| No: of concordant loss of CNVs | 179/190 |
| No: of concordant non-gain/non-loss events | 2879/2925 |
| **Overall CNV concordance** | 3263/3331 = 97.95% |

**Table 1. Accuracy of gene-level CNV detection with orthogonal experiment FISH.** Evaluation of the first stage gene-level CNV prediction on 1,314 clinical samples shows a concordance rate of 97.95% between NGS and FISH results on ATM, TP53, and NF1

## Step 2. CNV prediction of all genes on targeted chromosomes

The second stage chromosome-level CNV prediction was evaluated in clinical samples sequenced using the same targeted panel and assessed by FISH for chromosome-level variation on chromosomes 7, 8 and 12 (Table 2). The log2 ratio was calculated for all the genes on chromosome 7, 8 and 12 through the NGS-based CNV pipeline optimized in the step 1. The heatmap of the predicted log2 ratios for each sample and targeted gene shows a clear distinctive signal between aneuploidy and diploid samples (Figure 2 and Figure 3).

| Chromosome | Samples | CNV loss | Normal | CNV gain | Targeted genes | Targeted positions |
|---|---|---|---|---|---|---|
| 7 | 5,715 | 276 | 5,404 | 35 | GNA12, CARD11 ,PMS2, RAC1, IKZF1, EGFR, SBDS, CDK6, SAMD9, SAMD9L, CUX1, MET, ATG7, SMO, LUC7L2, BRAF, EZH2, RHEB, KMT2C | 67,197 |
| 8 | 5,715 | 5 | 5,211 | 499 | FGFR1, IKBKB, ZFHX4, NBN, UBR5, RAD21, MYC | 32,898 |
| 12 | 2,441 | 4 | 1,910 | 527 | CCND2, ETV6, CDKN1B, ETNK1, KRAS, IRAK4, ARID2, KMT2D, PRPF40B, CDK2, RPS26, ERBB3, STAT6, CDK4, MDM2, ATP2A2, SH2B3, PTPN11 | 52,828 |

**Table 2. Dataset description for evaluation of the chromosome-level CNV prediction.**

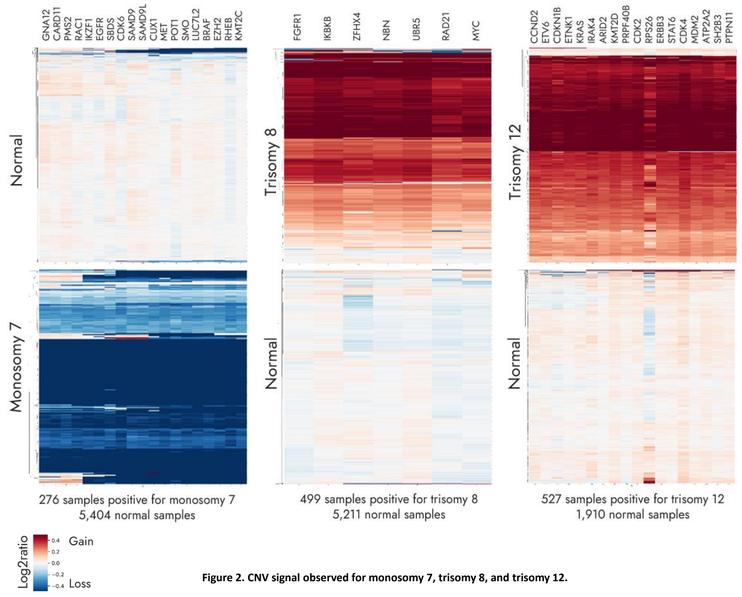


**Figure 2. CNV signal observed for monosomy 7, trisomy 8, and trisomy 12.**

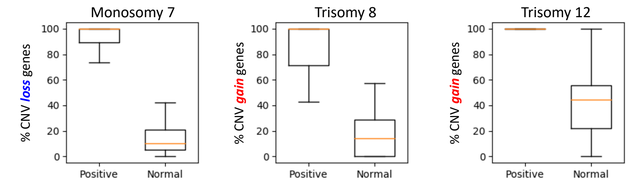


**Figure 3. Frequency of gain or loss CNV signals of genes on corresponding chromosomes.**

## Step 3. Optimal thresholds for the chromosome-level CNV prediction

The cutoff for frequency of CNV gain or loss of genes on each chromosome was optimized for predicting aneuploidy (Figure 4). At the chromosome level, the concordance rate between the final prediction and the FISH results is consistently observed above 93% (Table 3). Roughly 50% of the 12, 78, and 40 discordant calls for monosomy 7, trisomy 8, and trisomy 12, respectively captured by FISH but not by NGS can be explained by low tumor content (less than 20%) in the tested samples. The concordance rate between NGS and FISH is consistently observed above 96% when leaving these samples aside. Note that results in Table 3 are obtained using all samples to decide the optimal decision threshold for the chromosome-level prediction, but are found identical when using a leave-one-out evaluation procedure, and nearly identical when using a repeated cross-validation procedure.

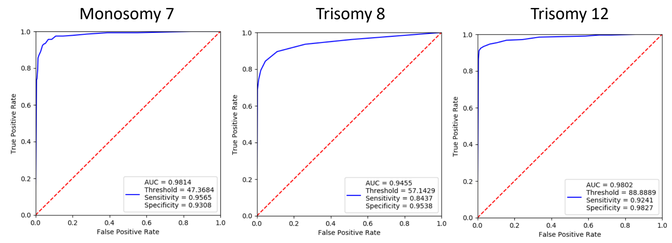


**Figure 4. AUC of chromosomal CNV prediction on monosomy 7, trisomy 8 and trisomy 12.**

| CNV | Result | Accuracy | Sensitivity | Specificity |
|---|---|---|---|---|
| **Monosomy 7** | All samples (5,680) | 0.93 | 0.96 | 0.93 |
| | TC* > 20% (5,619) | 0.96 | 0.97 | 0.96 |
| | PureCN (5,619) | 0.90 | 0.61 | 0.91 |
| **Trisomy 8** | All samples (5,710) | 0.95 | 0.84 | 0.95 |
| | TC* > 20% (5,558) | 0.97 | 0.90 | 0.98 |
| | PureCN (5,558) | 0.76 | 0.85 | 0.75 |
| **Trisomy 12** | All samples (2,437) | 0.97 | 0.92 | 0.98 |
| | TC* > 20% (2,332) | 0.99 | 0.97 | 0.99 |
| | PureCN (2,332) | 0.52 | 0.96 | 0.42 |

*TC: tumor content

**Table 3. Performance result of chromosomal-level CNV prediction.**

## Conclusion

This study demonstrates that chromosome-level CNVs can be accurately predicted in hematologic malignancies even when the number of targeted genes on a given chromosome is low. Despite the simplicity of the approach, the two stages bioinformatics pipeline based on an ensemble method allowed us to gain between 8% and 46% accuracy compared to relying only on the prediction of a single tool like PureCN (Table 3). Samples with low tumor content remain, however, a difficult case to tackle with bulk NGS as it is difficult to distinguish a CNV from the natural variability of the sequencing coverage.

## References

1. Boris C Bastian et al., CNVkit: Genome-Wide Copy Number Detection and Visualization from Targeted DNA Sequencing. PLoS Comput Biol. 2016 Apr 21;12(4):e1004873.
2. Michael P Morrissey et al., PureCN: copy number calling and SNV classification using targeted short read sequencing. Source Code Biol Med. 2016 Dec 15;11:13. doi: 10.1186/s13029-016-0060-z.

Hyunjun.nam@neogenomics.com