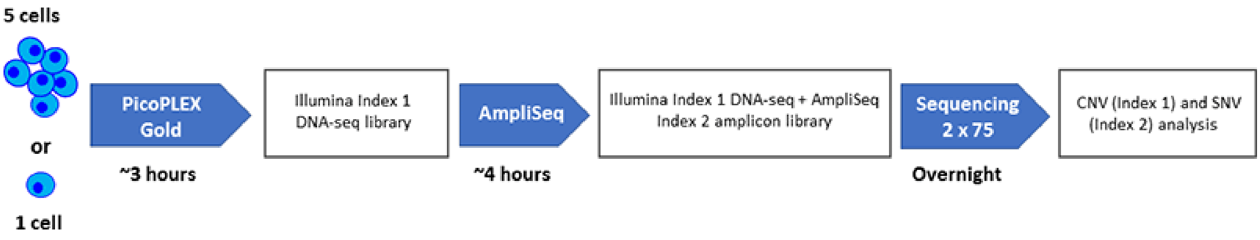
Accurate detection of SNVs and CNVs from five-cell inputs in a single, low-pass sequencing run

DATE: January 1, 2019
AUTHOR: Takara Bio Blog Team
CATEGORIES: Research News | Cancer | Single-cell

Accurate detection of both copy-number variants (CNVs) and single-nucleotide variants (SNVs) in a low-input DNA sample is often costly due to the need to perform separate library preparations for each readout. Unfortunately, readouts for both CNV and SNV are often required in order to obtain useful data to aid oncology research. One possibility is to use high-depth whole-genome sequencing to capture CNVs and SNVs, however this is quite costly and impractical. The most common approach for achieving the depth necessary to resolve low-frequency SNVs is to use targeted sequencing, which enables researchers to focus sequencing reads only on the genomic regions of interest. However, as mentioned above, a separate preparation and sequencing run is required to capture the CNVs.

By combining high-quality library prep with amplicon-based enrichment, our scientists were able to detect targeted SNVs *and* genome-wide CNVs with shallow sequencing of only 1 million reads per sample, in the same sequencing run. This is in contrast to whole-genome approaches that require additional NGS library preparation and separate sequencing runs for CNV and SNV detection using deep sequencing to detect variants. The icing on the cake—libraries can be prepared in a single day, significantly reducing turnaround time and labor costs.

As a proof-of-concept evaluation, our team used two commercial kits out of the box with no additional optimization to prepare five-cell DNA-seq libraries with PicoPLEX Gold Single Cell DNA-Seq Kit (PicoPLEX Gold) followed by AmpliSeq (Illumina®) and scored both CNVs and SNVs in the same sequencing run at shallow sequencing (0.5 million 75 bp paired-end reads). This workflow can be done in one day and is faster and more economical than any other technology available on the market.


5 cells
or
1 cell
PicoPLEX Gold
~3 hours
Illumina Index 1 DNA-seq library
AmpliSeq
~4 hours
Illumina Index 1 DNA-seq + AmpliSeq Index 2 amplicon library
Sequencing 2 x 75
Overnight
CNV (Index 1) and SNV (Index 2) analysis

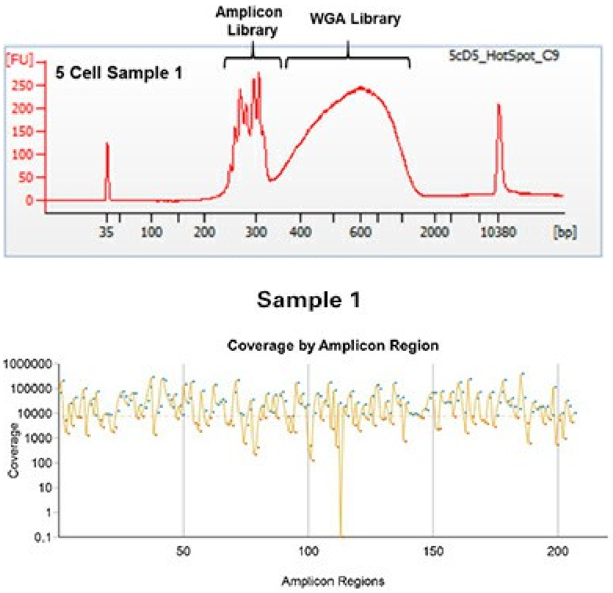
Workflow and timeline for SNV and CNV analysis using PicoPLEX Gold and the AmpliSeq Hot-Spot Cancer Panel.

PicoPLEX Gold + AmpliSeq = SNV and CNV detection in a single sequencing run

The GM12878 cell line was used for library prep from two five-cell samples with PicoPLEX Gold, followed by enrichment with the AmpliSeq for Illumina Cancer Hotspot Panel v2. The DNA-seq libraries from both samples demonstrated high amplicon-coverage reproducibility, 100% variant concordance, high bin-level genome coverage correlation, and low noise levels per chromosome (see representative figures below).

Uniform and reproducible library size distribution and amplicon coverage

The combined NGS libraries produced by PicoPLEX Gold followed by AmpliSeq panel enrichment show a high fragment-size reproducibility, demonstrated by the following representative Bioanalyzer trace of the amplicon output. Note that the DNA-seq whole-genome library is natively contained within the amplicon library and not modified or damaged by the amplicon enrichment process. 206 out of 207 amplicons were reproducibly and uniformly covered at a depth of at least 100X by both 5-cell replicate samples (a single amplicon was completely dropped). We also observed high amplicon-coverage reproducibility between the two replicate libraries (data not shown).



Bioanalyzer traces and amplicon coverage depth of representative library. Bioanalyzer trace (left) shows the expected size separation between the whole-genome NGS library and the amplicon library. Coverage plot (right) shows the read depth coverage of each individual amplicon plotted on a $\log_{10}$ scale. For this sample, amplicon mean coverage was 38,037.5 with 67.63% uniformity.

Variant detection analysis

The reference genome GM12878 contains three homozygous and eight heterozygous variants within the 22-kb amplicon panel used in the study. The two five-cell replicate samples prepared using PicoPLEX Gold followed by enrichment by AmpliSeq panel represented all eleven variants with no allele dropouts and a frequency close to the expected 50% for most of the heterozygous variants. A single false positive was detected in each sample. The detection rate of the alternate allele for the three homozygous variants was between 99–100% while the detection rate for the eight heterozygous variants was between 33–86% at an allele depth of ≥30X.

	True positives*	%True positives	Allele dropouts**	False positives***	%False positives
Reference	11	100	0	0	0
Sample 1	11	100	0	1	4.5E–3
Sample 2	11	100	0	1	4.5E–3

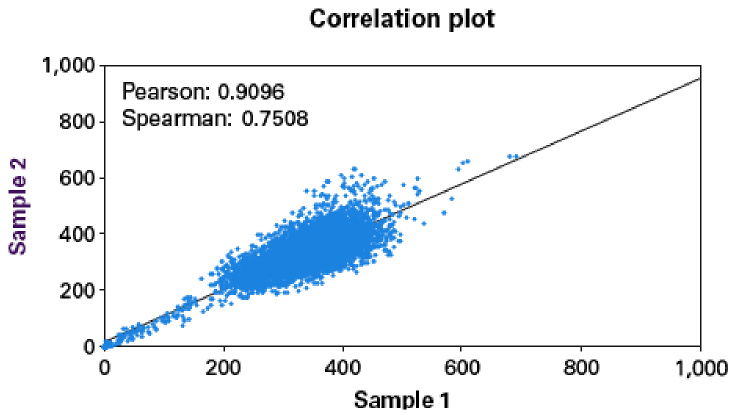
*For the GM12878 cell line, there were eleven variants contained in the Hot-Spot Cancer Panel.
**Filters: Minor allele frequency >0.3, Variant depth >10
***Total panel size=22 kb

Variant detection for the two five-cell samples from the AmpliSeq amplicon libraries. The data shows detection of three homozygous and eight heterozygous variants found in the GM12878 genome along the 22-kb amplicon panel, and concordance with the bulk DNA genotype. The false positive percentage was calculated as discordance with the reference at any position of the 22-kb amplicon panel.

CNV analysis

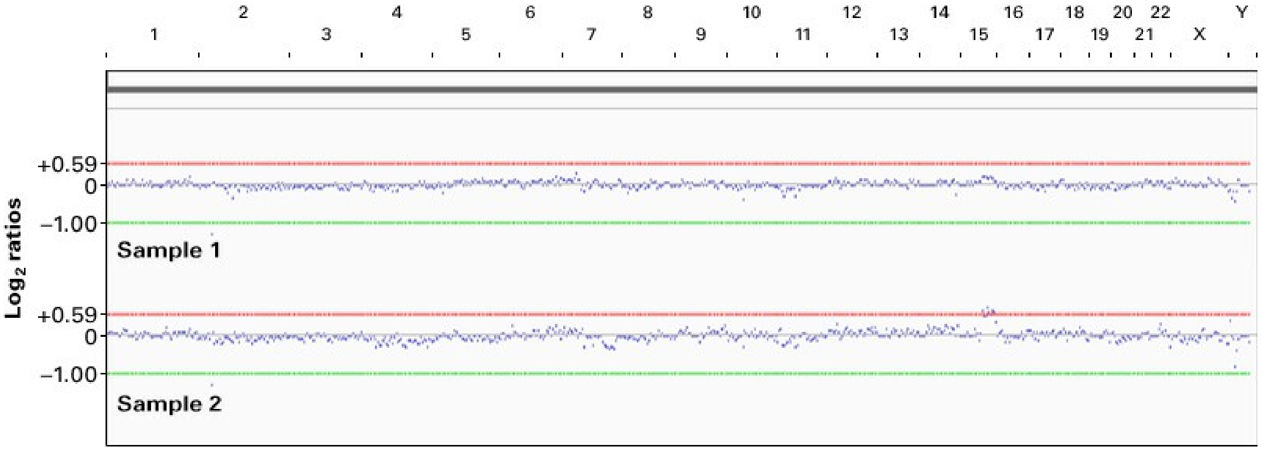
For this initial proof of concept, we did not seek to demonstrate the detection of clinically-important CNVs directly. However, it is well known that Median Absolute Pair-wise Differences (MAPD) and correlation coefficients at this shallow sequencing and high resolution (0.5 M read pairs, 1 MB bin size) are very sensitive and reliable predictors of CNV detection with real biological samples. Additional analyses of cell lines with known clinically relevant aneuploidies of different size and real biological samples are in progress and we will publish the results in a follow-up white paper. You can sign up to receive this data when it is available.

For these samples, average chromosome level MAPD noise (normalized to a PicoPLEX Gold library from bulk NA12878 gDNA, sequenced at 75 million reads as the reference) varied between 0.069–0.093 ±0.0071. Taken together with the sample-to-sample coverage correlation shown below, this data demonstrates high coverage reproducibility and low noise levels across the genome.



Coverage reproducibility for the two five-cell samples from the whole-genome NGS libraries at 0.5 million read pairs and 1 MB bin size. This correlation plot shows bin-level counts correlation between the samples with Pearson and Spearman coefficients.

The analysis of bin-level coverage of the chromosomes shows a gain on chromosome 15 in one of the samples, which is frequently observed in the GM12878 genome along with other segmental aneuploidies of various sizes across the genome in a population of non-synchronously grown tissue-culture cells.



Bin-level genome coverage for the two five-cell samples from the whole-genome NGS libraries at 0.5 million read pairs and 1 MB bin size. Log$_2$ ratios of bin-level genome coverage, normalized to a PicoPLEX Gold library from five GM12878 cells, sequenced at 75 million reads as the reference are shown.

Experimental design

PicoPLEX Gold was used to generate replicate five-cell libraries from a well-characterized cell line, GM12878, with two different barcodes. 50-ng aliquots of each PicoPLEX Gold library were then used as the input for the AmpliSeq for Illumina Cancer Hotspot Panel v2 (207 amplicons, 22 kb size) and indexed with different barcodes than the original NGS libraries. Final libraries were sequenced directly using an Illumina MiSeq® platform to a depth of 1 million read pairs per sample.

CNV analysis was performed following down-sampling to 1 million total reads (0.5 million read pairs) and binning at 1 MB using the Takara Bio Mendel R&D pipeline for generating coverage, correlation, and MAPD noise analysis. AmpliSeq analysis was performed using the Illumina DNA Amplicon Module for amplicon coverage and the Takara Bio Mendel R&D pipeline with VarDict for variant analysis.

Takara Bio USA, Inc.
United States/Canada: +1.800.662.2566 • Asia Pacific: +1.650.919.7300 • Europe: +33.(0)1.3904.6880 • Japan: +81.(0)77.565.6999

FOR RESEARCH USE ONLY. NOT FOR USE IN DIAGNOSTIC PROCEDURES. © 2023 Takara Bio Inc. All Rights Reserved. All trademarks are the property of Takara Bio Inc. or its affiliate(s) in the U.S. and/or other countries or their respective owners. Certain trademarks may not be registered in all jurisdictions. Additional product, intellectual property, and restricted use information is available at takarabio.com.