APPLICATION NOTE

Germline CNV analysis using Illumina DNA Prep with Exome 2.5 Enrichment workflow

Effective detection of copy number variants using a panel of normals

illumina®

Introduction

Copy number variations (CNVs) are a major source of genetic diversity in humans.[1] This class of genomic variation occurs when the number of copies of a gene or genomic region varies across individuals and includes large genomic duplications or deletions. CNV analysis can help researchers identify variants associated with complex traits or susceptibility to diseases, such as cancer, autoimmune diseases, inherited genetic disorders, and more.[1–4]

Compared to use of microarrays for CNV detection, next-generation sequencing (NGS) methods reveal more detail around structural variations. In addition, NGS can detect some variants < 50 kb that arrays cannot. Illumina DNA Prep with Exome 2.5 Enrichment is part of a high-performance, fast, and reliable human whole-exome sequencing (WES) solution that generates libraries that are compatible with NGS CNV analysis. Combined with Illumina sequencing systems and analysis pipelines, this library preparation kit can be used to detect single-nucleotide variants (SNV), small insertions and deletions (indels), and large structural CNVs.

This application note demonstrates a streamlined exome sequencing workflow for germline CNV analysis, from library preparation through insights (Figure 1). The WES solution integrates Illumina DNA Prep with Exome 2.5 Enrichment, proven Illumina NGS, and highly accurate DRAGEN™ secondary analysis for identifying CNVs.

Users can also apply the Emedgene™ variant interpretation research platform, powered by explainable artificial intelligence (XAI). Emedgene software enables variant analysis and insights, curation, and standard operating procedure (SOP) automation, and includes streamlined options for research report generation.

Methods

Samples

To enable germline CNV detection across the exome, users must develop a baseline reference file from a panel of putatively normal samples for comparison during analysis. This panel of normals reference is used to normalize target counts and allows for accurate detection of CNVs during evaluation of case samples. The samples for the panel of normals should be prepared and sequenced under the same conditions as the case samples.

To create the panel of normals, 54 samples (Coriell Institute for Medical Research) with no known pathogenic CNVs (Table 1) were selected to represent the diversity of natural variation across human superpopulations (based on the 1000 Genomes Project, Table 2).[5] To demonstrate the performance of the WES workflow for CNV detection, 17 case samples containing a previously detected CNV within a gene or region of interest were used in an analysis with the panel of normals.

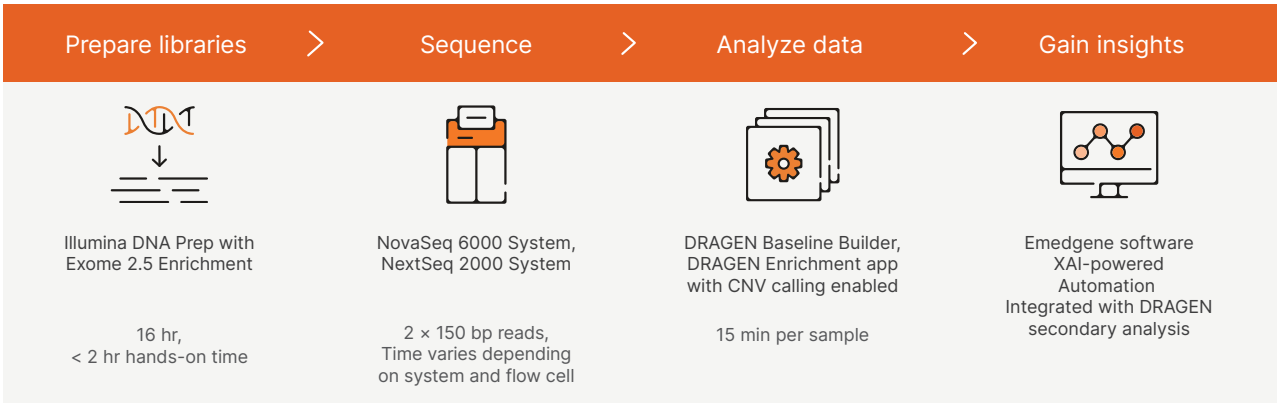

Figure 1: Comprehensive WES workflow for CNV calling—Combine Illumina DNA Prep with Exome 2.5 Enrichment, Illumina NGS platforms, rapid data analysis with DRAGEN secondary analysis, and variant interpretation with Emedgene software to achieve a user-friendly and streamlined workflow for accurate exome sequencing to interrogate CNVs.

Table 1: Coriell Institute sample IDs included in the panel of normals

HG00096	HG01393	HG01950	HG03006	NA10830	NA12877	NA18502	NA18970	NA21112
HG00190	HG01441	HG01985	HG03870	NA10831	NA12878	NA18508	NA19681	NA24143
HG00262	HG01551	HG01990	HG03882	NA10835	NA12889	NA18622	NA19720	NA24149
HG00626	HG01599	HG02013	HG03898	NA10838	NA12890	NA18637	NA20509	NA24631
HG00628	HG01896	HG02348	HG04090	NA10839	NA12891	NA18942	NA20875	NA24694
HG01392	HG01914	HG02521	HG04214	NA12249	NA12892	NA18957	NA21098	NA24695

Library preparation

Exome sequencing libraries were prepared in duplicate following the Illumina DNA Prep with Enrichment protocol (Document #1000000048041 v07) using 50 ng gDNA from Coriell Institute samples, unless otherwise stated. IDT for Illumina DNA/RNA UD Indexes Sets A-D, Tagmentation (Illumina, Catalog nos. 20027213, 20027214, 20042666, and 20042667) were used to enable unique indexing across all prepared libraries. Pre-enrichment libraries were pooled by mass (target 250 ng per sample) at 12-plex per pool and hybridized overnight using the Twist Bioscience for Illumina Exome 2.5 Panel, available as part of Illumina DNA Prep with Exome 2.5 Enrichment kits (Illumina, Catalog nos. 20077595 and 20077596). Libraries were prepared with and without spike-in of the Twist Bioscience for Illumina Mitochondrial Panel (Illumina, Catalog no. 20093180). Mitochondrial panel spike-in followed the 1:100 dilution outlined in the Illumina DNA Prep with Exome 2.5 Enrichment Reference Guide (Document #1000000157112).

Sequencing

Enriched libraries were sequenced on NovaSeq™ 6000 and NextSeq™ 2000 Systems using standard sequencing by synthesis (SBS) reagents and 150-bp paired-end runs. The highest data outputs available were used to achieve high depth of coverage per sample and allow for downsampling as part of the analysis. The panel of normals samples and the case samples were sequenced using the NovaSeq 6000 S4 Reagent Kit v1.5 (300 cycles) (Illumina, Catalog no. 20028312) and NextSeq 2000 P3 Reagents (300 cycles) (Illumina, Catalog no. 20040561). Reagent kits for the NovaSeq 6000 System are designed for ease of use with three ready-to-use cartridges prefilled with all needed reagents.

Table 2: Summary attributes of panel of normals

Sex	No. of samples
Male	27
Female	27
Superpopulation[5]	
European (EUR)	18
Mixed American (AMR)	8
African (AFR)	7
South Asian (SAS)	9
East Asian (EAS)	12

Table 3: Sequencing throughput and coverage across platforms for panel of normals and CNV samples

Platform	Samples per run	Average coverage per sample
NovaSeq 6000 System	72	~540×
NextSeq 2000 System	12	~280×

The NextSeq 2000 System provides flexibility with multiple flow cell configurations and offers onboard DRAGEN analysis. The number of samples per run and average coverage achieved per sample across all runs is shown in Table 3.

Data analysis

The DRAGEN Baseline Builder app v4.2, accessible in BaseSpace™ Sequence Hub, was used to create the panel of normals individual target counts files with the hg38 nongraph reference genome. Sequencing data for CNV case samples was downsampled to 80M reads (~120–140× average on-target coverage depth) and 50M reads (~70–90× average on-target coverage depth), and analyzed using the DRAGEN Enrichment app v4.2, also accessible in BaseSpace Sequence Hub, with CNV detection enabled. For each platform, the corresponding panel of normals was used for normalization of samples. The DRAGEN Enrichment app performs accurate, efficient secondary analysis for comprehensive variant calling, including SNV, indels, CNV, and structural variants (SV), among other applications. DRAGEN variant calling for Illumina DNA Prep with Exome 2.5 Enrichment can also be performed onboard the NextSeq 2000 System or fully integrated in the Emedgene software variant interpretation research workflow.

Results

CNV detection across a range of variation

Analysis with the DRAGEN CNV pipeline in the DRAGEN Enrichment app revealed robust CNV detection for the variants assessed in this study at as low as 50M reads, or ~70–90× average coverage depth, across all platforms. Detection of CNVs is determined by the formula: (number of regions called)/(regions overlapping with exome panel and truth). CNVs were considered detected if this metric was > 98% (Table 4).

Several of the variants tested are found in the *DMD* gene, which is an important target in genetic disease research (Table 4).[6] Using Integrated Genomics Viewer, the expected *DMD* CNV events detected can be visualized for the seven affected samples tested (Figure 2). This demonstrates the ability to detect CNVs across this gene when using the Illumina Exome 2.5 workflow. The Exome 2.5 panel BED file track shows the regions covered by the panel (Figure 2, top). Results will depend on various factors and it is recommended to sequence to > 150× mean coverage for initial testing of a lab workflow to ensure sufficient coverage for most variants.

Additional analysis possible with Emedgene software

Emedgene software streamlines and scales variant interpretation, saving 50%–75% time per research case.[7] Multiple features power user-defined interpretation, including XAI for transparent, evidence-backed, automated rankings of potentially causative variants for samples; an always up-to-date annotation and evidence graph; variant visualization; variant curation; user-defined automation; and more to promote informed variant interpretation. Emedgene software was designed for an efficient and intuitive user experience.

To streamline case review, the AI variant prioritization or "shortlist" functionality compiles variants, including CNVs, that are most likely to solve a case. This functionality includes backing evidence and provides significant time savings for the analyst. In a separate study[8] of 51 singletons previously solved by a CNV variant, the solving variant was identified in a short list of 10 variants in 92% of cases. In 6% (n = 3), the solving variant was present in the candidate list. During case review, Emedgene software allows untagging of variants selected by the shortlist or manual tagging of variants not selected by the shortlist. Automated ACMG* classification also covers CNVs.

Access panel of normals files

The panel of normals files, generated using Coriell Institute samples, provide researchers with a known data set for comparison with case samples without the need to spend the time and resources acquiring, sequencing, and analyzing this sample set. These default files are best used for initial testing of the Illumina DNA Prep with Exome 2.5 Enrichment workflow for CNV analysis and labs are encouraged to develop a panel of normals specific to their laboratory protocols. If using these files for initial testing, libraries should be prepared for case samples following the same protocol used to prepare the panel of normals. Deviations in protocol may lead to differences in coverage profiles, mitigating the effectiveness of the default panel of normals. Files and additional information can be found on the Illumina Support Site.

* ACMG, American College of Medical Genetics and Genomics.

Table 4: CNV detection ordered by approximate size of event based on expected event coordinates

Coriell ID	Chromosome	Gene (affected exons) or chromosomal location	Expected event	Approximate size of event (kb)[a]	Detected at 50M reads?
NA04315	X	*DMD* (44)	Loss	0.15	✓
NA05115	X	*DMD* (45)	Loss	0.18	✓
NA05117	X	*DMD* (45)	Loss	0.18	✓
NA23599	X	*MECP2* (3–4)	Loss	2.25	✓
NA18949	17	*BRCA1* (15–16)	Loss	3.59	✓
NA21939	15	*FBN1* (42–43)	Loss	3.70	✓
HG03857	16	*PALB2* (5–7)	Loss	4.23	✓
HG00343	22	*CHEK2* (9–10)	Loss	4.25	✓
NA23127	X	*DMD* (27–28)	Gain	7.47	✓
NA04520	16	*TSC2* (1–15)	Loss	16.28	✓
NA04327	X	*DMD* (5–7)	Gain	22.70	✓
NA10283	X	*DMD* (72–79)	Loss	54.38	✓
HG00500	2	*SPAST* (4–17)	Gain	58.84	✓
NA23675	X	*MECP2* (all)	Gain	76.14	✓
NA23087	X	*DMD* (2–30)	Gain	608.45	✓
NA06870	18	18p11.32-18p11.1	Gain	15390.21	✓
NA20125	10	10q23.1-10q26.3	Gain	52877.99	✓

a. Based on affected exon or chromosomal location expected event coordinates.

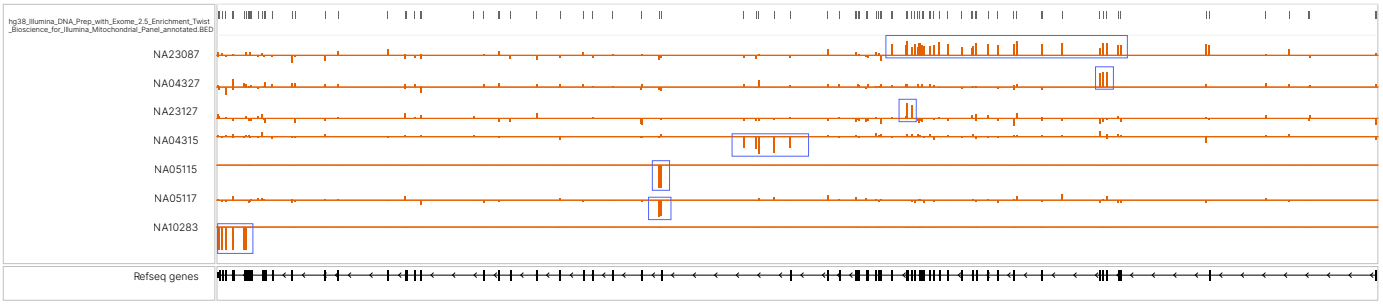

Figure 2: Visualization of detected CNVs across seven samples containing expected events in the *DMD* gene—Files providing BigWig representation of the tangent normalized signal (*.tn.bw) are output as part of the DRAGEN Enrichment app analysis and show gains and losses based on the targeted regions in the Twist Bioscience for Illumina Exome 2.5 enrichment panel. Tracks are shown autoscaled as a bar chart.

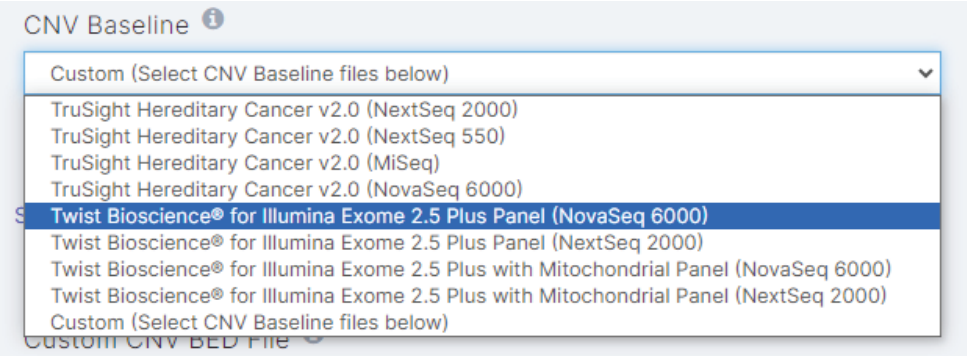

Figure 3: Drop-down menu of panel of normals in DRAGEN Enrichment app v4.3 in BaseSpace Sequence Hub.

When using BaseSpace Sequence Hub, the DRAGEN Enrichment app v4.3 provides drop-down options to perform initial CNV analysis with Twist Bioscience for Illumina Exome 2.5 enrichment panel for multiple references (graph and nongraph hg19, hg38, and hs37d5) across multiple platforms (NextSeq 2000 and NovaSeq 6000 Systems) using multiple libraries (Twist Bioscience for Illumina Exome 2.5 enrichment panel with and without Twist Bioscience for Illumina Mitochondrial Panel spike-in) (Figure 3). Case samples used with this analysis workflow should be prepared as described in the Methods section of this application note.

Summary

The NGS exome workflow for analyzing CNVs uses Illumina DNA Prep with Exome 2.5 Enrichment for library preparation, sequencing on the NovaSeq 6000 or NextSeq 2000 Systems, and DRAGEN applications for data analysis. The results demonstrate the application of a panel of normals as a reference for CNV analysis of case samples using the described WES workflow. The high recall of CNVs by DRAGEN secondary analysis software correlates well with other methods. This workflow, including use of the panel of normals, for CNV analysis can also be adapted to other Illumina sequencing platforms. Emedgene software helps labs perform additional analysis such as variant interpretation and research report generation.

Learn more

Panel of normals reference data

Copy number variant analysis

Illumina DNA Prep with Exome 2.5 Enrichment

Illumina sequencing platforms

DRAGEN secondary analysis

DRAGEN Germline application

Emedgene software

References

1. Zhang F, Gu W, Hurles ME, Lupski JR. Copy number variation in human health, disease, and evolution. *Annu Rev Genomics Hum Genet*. 2009;10:451-481. doi:10.1146/annurev.genom.9.081307.164217
2. Mehawej C, Maalouf JE, Abdelkhalik M, Mahfouz P, Chouery E, Megarbane A. CNV Analysis through Exome Sequencing Reveals a Large Duplication Involved in Sex Reversal, Neurodevelopmental Delay, Epilepsy and Optic Atrophy. *Genes (Basel)*. 2024;15(7):901. doi:10.3390/genes15070901
3. Fang X, Ma M, Rong W, et al. Exome sequencing confirms the clinical diagnosis of both joubert syndrome and klinefelter syndrome with keratoconus in a han Chinese family. *Front Genet*. 2024;15:1417584. doi:10.3389/fgene.2024.1417584
4. Zeng Y, Ding H, Wang X, et al. High positive predictive value of CNVs detected by clinical exome sequencing in suspected genetic diseases. *J Transl Med*. 2024;22(1):644. doi:10.1186/s12967-024-05468-1
5. Harrison PW, Amode MR, Austine-Orimoloye O, et al. Ensembl 2024. *Nucleic Acids Res*. 2024;52(D1):D891-D899. doi:10.1093/nar/gkad1049
6. Kozareva V, Stroff C, Silver M, Freidin JF, Delaney NF. Clinical analysis of germline copy number variation in DMD using a non-conjugate hierarchical Bayesian model. *BMC Med Genomics*. 2018;11(1):91. doi:10.1186/s12920-018-0404-4
7. Greenwood Genetic Center. GGCC reduces turn around time on genomic analysis by 75% with Emedgene's AI platform. ggc.org/in-the-news-app/ggc-reduces-turn-around-time-on-genomic-analysis-by-75-with-emedgenes-ai-platform. Published September 12, 2019. Accessed August 14, 2024.
8. Data on file. Illumina, Inc. 2024.

1.800.809.4566 toll-free (US) | +1.858.202.4566 tel
techsupport@illumina.com | www.illumina.com

M-GL-01453 v1.0