NONACUS
Advancing Non-Invasive Healthcare

Nonacus™

ExomeCG - Whitepaper

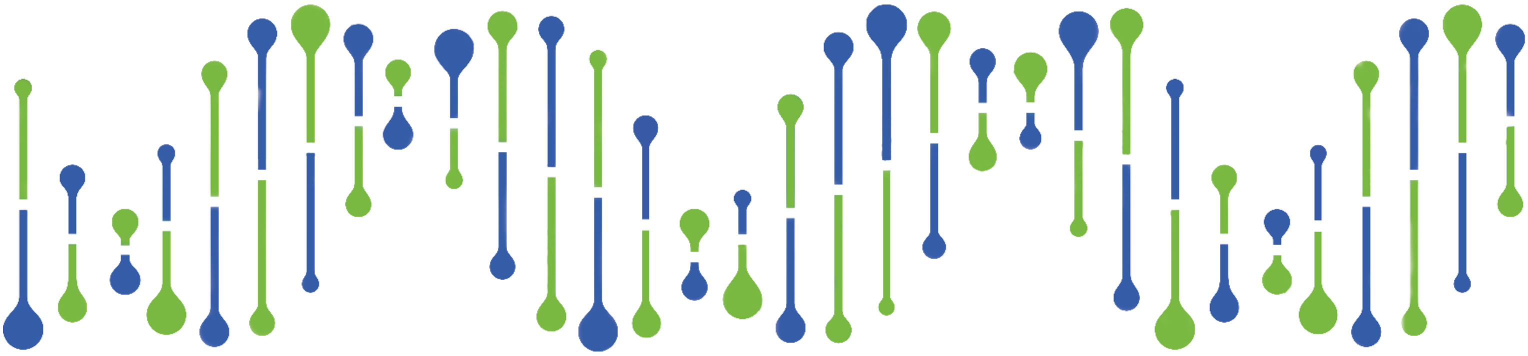

Table of Contents

ExomeCG Kit Whitepaper

Abstract

We describe a new DNA sequence capture kit, ExomeCG, which has been specifically designed to detect copy number variants (CNVs), single nucleotide variants (SNVs), indels and other structural variants across the exome. This assay has been augmented with baits providing increased coverage for genes of clinical importance for both prenatal and postnatal testing. We demonstrate that the kit is capable of detecting CNVs with sizes spanning from a single exon up to multiple contiguous genes (~100bp–40Mb) and that detection of clinically relevant variants is achieved with superior precision and recall.

Introduction

Copy number variants account for ~10% of curated disease associated variants and are identified in ~10–20% of individuals with neurodevelopmental disorders (Stenson et al. 2017; Pfundt et al. 2017). Chromosomal microarray (CMA) and multiplex ligation dependent probe amplification (MLPA®) have been the gold standards for CNV detection and intragenic del/dup events respectively. However, such approaches alone miss cases with both CNVs and pointmutations and so need to be coupled with next-generation sequencing (NGS) tests to capture single nucleotide variants (SNVs) and small indels.

The ability to combine these tests into a single assay reduces sample requirements, provides time, cost and logistical benefits as well as increasing the diagnostic yield of genomic testing.

Boosted regions of the ExomeCG include OMIM morbid genes (Online Mendelian Inheritance in Man 2018), the ACMG 59 secondary findings gene list (Kalia et al. 2017), a curated set of fetal abnormality genes (curated from DDG2P (Deciphering Developmental Disorders 2018), the BabySeq project (Ceyhan-Birsoy et al 2017) and multiple fetal exome series publications) and a curated set of early infantile epileptic encephalopathy (EIEE) genes.

The kit has also been extended to include coverage of exon level deletions and duplications currently targeted by commercially available MLPA kits. In addition, the kit is designed to detect a set of reported pathogenic non-coding SNVs, pharmacogenomic (PGx) markers and includes sample tracking variants.

The design of ExomeCG has been specifically tailored to optimise its performance when analysed using the Congenica CNV calling pipeline. In particular, we have increased the number of bait probes within target regions to better support the statistical model applied by the ExomeDepth CNV caller (Plagnol, 2012) implemented in our secondary analysis pipeline.

Exome Sequencing

Our validation dataset comprised 30 samples1 including CNVs detected by CMA and MLPA. Genome in a Bottle sample, HG-002 (Zook et al. 2016), was also included to confirm performance of the kit for SNV detection. Libraries were prepared from 100ng input gDNA using the ExomeCG kit and sequenced on an Illumina NextSeq500 (130bp paired-end high-output flow cell) yielding a median target coverage of 150x.

Data Analysis
Sequencing reads were aligned and SNVs called using the Illumina DRAGEN Bio-IT platform (http://www.illumina.com/DRAGEN).
CNVs were called using the ExomeDepth algorithm. The SNV calls generated from the HG-002 sample using the ExomeCG were compared with a set of reference SNV calls using the hap.py tool (https://github.com/Illumina/hap.py). In order to provide a true like-for-like comparison, these reference SNVs were called using DRAGEN on whole genome sequencing read data provided by the Genome in a Bottle consortium.

In addition to analysis of clinical samples to determine the sensitivity of the kit in a real-world setting, we performed a complementary analysis to determine the specificity of the kit. Currently established CNV truth sets focus on commonly occurring CNVs found in normal individuals, rather than rare, disease-causing CNVs (e.g. Zook et al. 2019). To address this, we have instead evaluated the performance of the ExomeCG assay using a synthetic data method (Sadedin et al. 2018) which selectively down samples the sequence read data from a set of individuals to mimic the effect of CNVs within these samples. The truth set created in this way was used to score whether the synthetic CNVs are detected by a given variant calling algorithm. We omit genomic regions within the samples containing previously detected native CNVs from this analysis. As shown in Figure 1, the coverage of these panels and variants is substantially improved compared with other commercially available assays, without sacrificing any coverage across the wider exome.

[1] Appropriately consented samples were provided by collaborators at South West Thames Regional Genetics Service, St. George's NHS Hospitals Trust, London, UK.

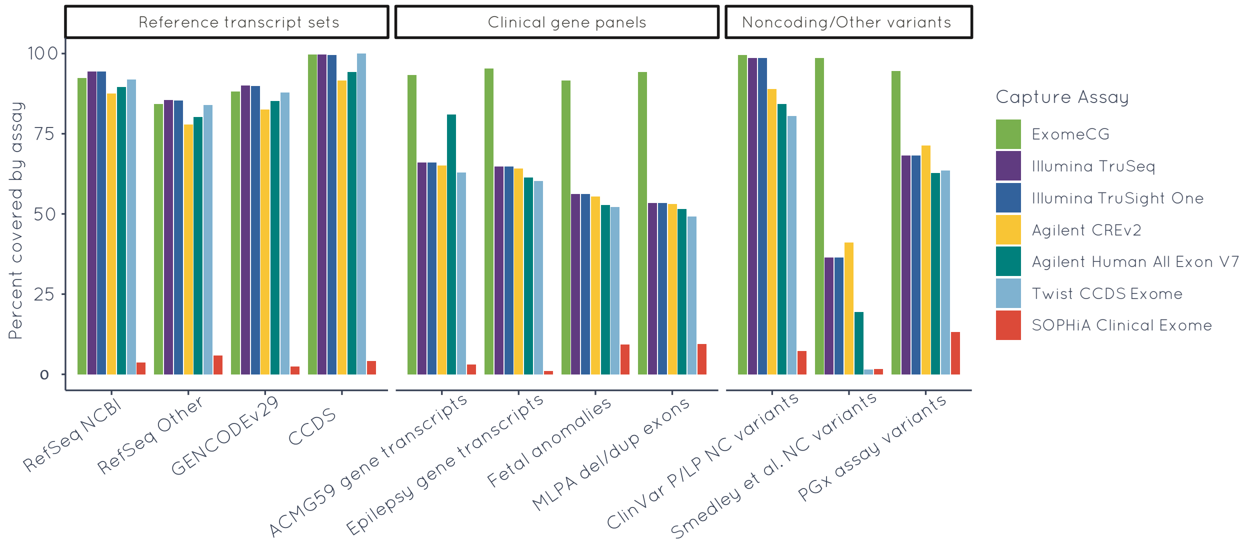

Figure 1. Design coverage of targeted gene panels and variant sets by ExomeCG compared to other commercially available kits.

Kit Performance

Observed read coverage

The ExomeCG kit provides improved read depth across clinically relevant gene panels, as shown in Figure 2. The fraction of exons yielding a high read depth (median reads per kilobase million (RPKM) > 5) is improved relative to competing capture kits, resulting in a greatly reduced set of low-coverage exons.

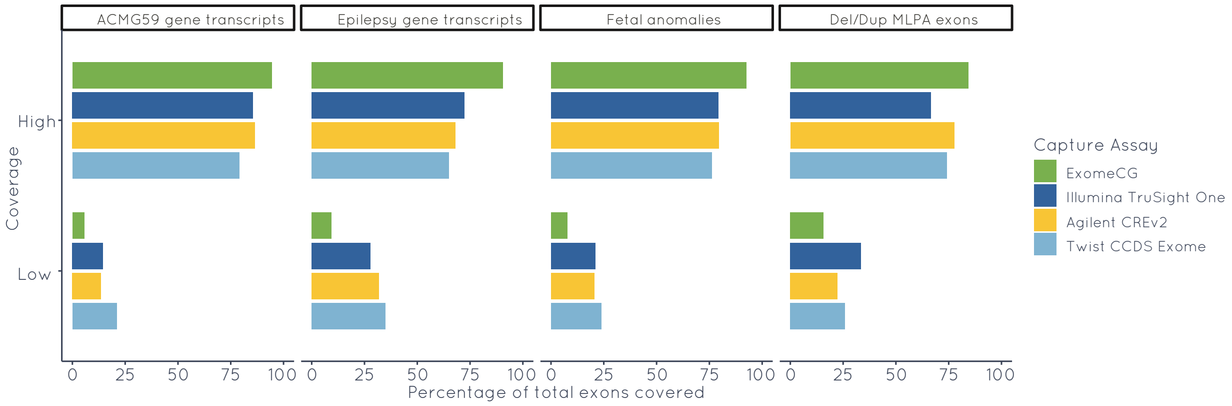

Figure 2. Actual read coverage across clinically relevant gene panels achieved using different capture kits. The threshold dividing low from high coverage was set at RPKM > 5.

Detection of MLPA- and CMA-confirmed CNVs

A key requirement for any NGS CNV assay is the ability to detect variants previously identified using CMA or MLPA technologies. We have been able to detect MLPA-confirmed

CNVs as small as 84bp using our Exome kit with the Congenica CNV calling pipeline (Table 1 and Figure 3). CNVs spanning one or more exons were detected.

Affected gene	CNV region	CNV size (bp)	CNV exons	CNV type	Bayes Factor
FBN1	exons 29-65	74632	37	deletion	320.0
BRCA1	exons 1-23	77841	24	deletion	190.0
FBN1	exons 1-17	142063	18	deletion	300.0
BRCA1	exons 1-17	57876	18	deletion	200.0
BRCA1	exons 8-13	17956	6	deletion	40.4
BRCA1	exons 8-13	17956	6	deletion	82.4
BRCA2	exons 5-7	513	3	deletion	22.1
NSD1	exons 7-9	6034	3	deletion	34.5
FBN1	exons 60-62	3934	3	deletion*	32.8
NSD1	exons 1-3	58095	3	deletion	54.8
BRCA2	exons 1-2	1054	2	deletion	28.3
BRCA1	exons 7-8	311	2	deletion	4.7
BRCA1	exons 8-9	1444	2	deletion	7.5
BRCA1	exon 16	211	1	deletion	14.5
BRCA1	exon 20	84	1	deletion	9.4

Table 1. Detection of MLPA-confirmed CNVs by the ExomeCG assay. The Bayes factor is the lo

Table 1. Detection of MLPA-confirmed CNVs by the ExomeCG kit. The Bayes factor is the log10 of the likelihood ratio, which quantifies the evidence for the CNV call divided by that for normal copy number. *FBN1 exons 60-62 deletion as visualised in Figure 3.

A Bayes factor of 10 or above is regarded as strong evidence for the presence of a CNV (Jeffreys, 1998). By examining the degrees of precision and recall achieved at a range of Bayes factor thresholds we have established an optimum lower threshold of 8 for this factor.

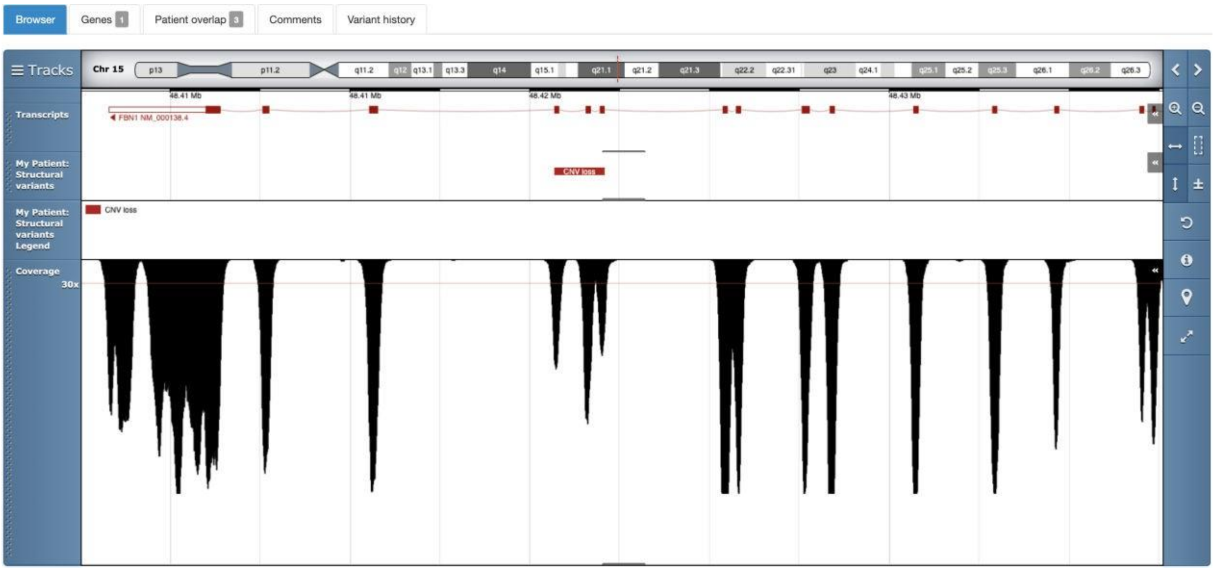

Figure 3: FBN1 deletion encompassing exons 60-62 detected by ExomeCG analysed and visualised in

Figure 3: FBN1 deletion encompassing exons 60-62 detected by ExomeCG analysed and visualised in the Congenica platform Genome Browser. CNV position and type indicated on the My Patient: Structural Variants track and flanked by the Transcript and Coverage tracks.

Table 2 lists the results of CNV calling on a series of CMA-confirmed CNV patients. These large multi-gene CNVs yield strong and easily detectable signals up to a size of at least 42Mb.

CNV region	CNV size (Mb)	CNV genes	CNV type	Bayes Factor
13q14.2q32.1	42.0	367	loss	2410
4p16.3p15.2	22.9	339	loss	4620
20q11.22q13.12	11.3	244	loss	7000
7p14.1p11.2	15.9	182	loss	5040
1p36.32	3.7	140	loss	2710
22q11.21	2.0	83	loss	2890
8q23.1q24.12	11.8	71	loss	1330
22q11.21	2.2	64	gain*	1430
11p12p11.2	2.3	54	loss	1240
7q11.23	1.4	38	loss	2080
15q11.2	0.9	31	loss	494
17p12	1.3	24	loss	275
14q22.1	0.7	20	loss	508
15q11.2	0.5	4	gain	370
13q12.11	0.2	2	loss	75

*Table 2. Detection of CMA-confirmed multi-gene CNVs by the ExomeCG assay. *22*

Table 2. Detection of CMA-confirmed multi-gene CNVs by the ExomeCG kit. *22q11.21 CNV gain as visualised in Figure 4.

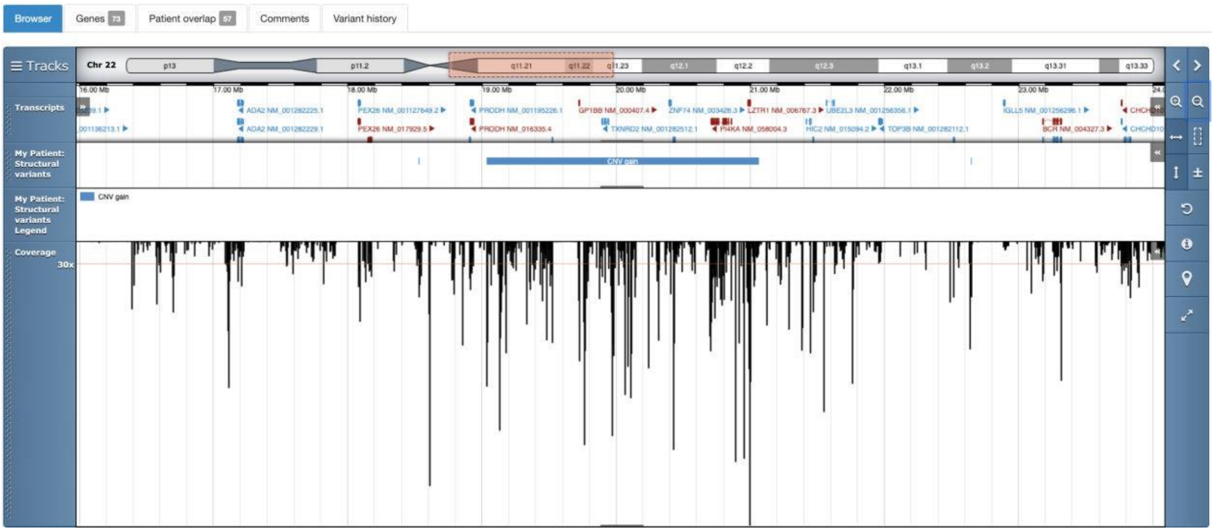

Figure 4: CMA-confirmed multi-gene 2Mb gain on chromosome 22 at 22q11.21 detected by ExomeCG, analysed and visualised in the Congenica platform Genome Browser. CNV position and type indicated on the My Patient: Structural variants track and flanked by the Transcript and Coverage tracks.

Precision and recall

Using simulated data to provide CNV truth sets, we have generated precision-recall and ROC curves for the ExomeCG kit compared to leading competitor assays. Figure 5 shows these curves for the regions targeted by clinically relevant gene panels (ACMG 59, epilepsy, fetal anomalies and MLPA del/dup exons).

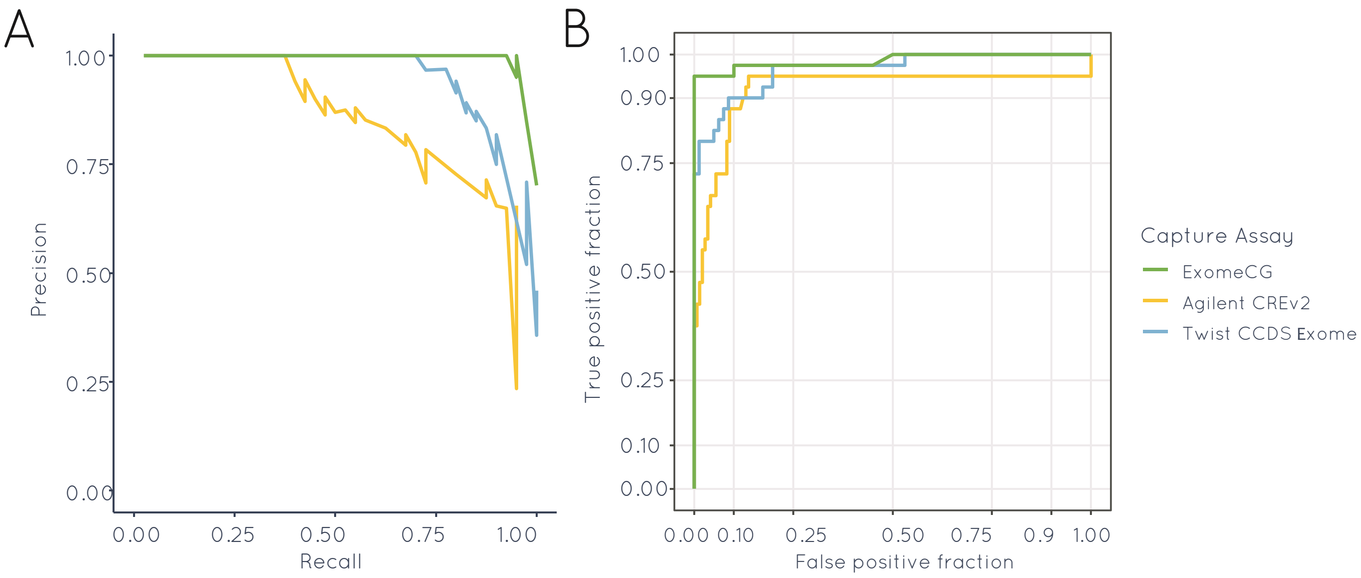

Figure 5. CNV calling performance of the ExomeCG kit in comparison to alternative providers Exome products. A: precision-recall curve; B: ROC curve.

These figures show that ExomeCG provides significantly improved CNV calling performance within these clinically important regions. This targeted improvement comes with no loss of performance across the remaining content (data not shown).

SNV Detection Performance

As shown in Table 3, ExomeCG supports the calling of SNVs and indels to a high degree of precision and recall.

Variant type	Target regions	Total variant calls		Recall %	Precision %	F1 Score
		GiaB	Exome			
SNV	All	32045	31513	96.7	98.3	0.98
SNV	Noncoding	69	70	100.0	98.6	0.99
Indel	All	2583	2676	86.8	85.3	0.86

Table 3. Detection of SNVs and indels by the ExomeCG assay. The figures given are for variants Table 3. Detection of SNVs and indels by the ExomeCG kit. The figures given are for variants passing all quality filters. The F1 score uses the harmonic mean of recall and precision to summarise the trade-off between these values. 'GiaB' represents reference calls from Genome in a Bottle WGS data.

Conclusions

Here we demonstrate that the ExomeCG DNA sequence capture kit supports sensitive and specific detection of CNVs across a wide size range. The coverage of clinically important coding and noncoding regions by ExomeCG is substantially improved in comparison with other commercially available exome capture kits, with no loss of coverage across the wider exome.

We have established that CNV calling with ExomeCG provides superior precision and recall of a simulated truth set within the clinically significant targeted gene panels. The sensitivity of the ExomeCG kit in detecting small CNVs (100bp or larger) was assessed on a representative set of real-world clinical samples and equals that of MLPA and CMA and includes the potential to detect CNVs as small as 50bp. Further we have confirmed that ExomeCG provides excellent SNV calling performance across all kit design regions.

Our new ExomeCG assay is complemented by the latest release of the Congenica clinical decision support platform, which adds validated support for CNV calling from the ExomeCG and other available kits.

References

Stenson, P. D. et al. (2017) The Human Gene Mutation Database: towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next-generation sequencing studies. Hum Genet 136:665–677 doi: 10.1007/s00439-017-1779-6

Pfundt, R. et al. (2017) Detection of clinically relevant copy-number variants by exome sequencing in a large cohort of genetic disorders. Genetics in Medicine 9:667:675 doi: 10.1038/gim.2016.163

Online Mendelian Inheritance in Man, OMIM®. McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University (Baltimore, MD), 2018. https://omim.org/

Kalia, S. S. et al. (2017) Recommendations for reporting of secondary findings in clinical exome and genome sequencing, 2016 update (ACMG SF v2.0): a policy statement of the American College of Medical Genetics and Genomics. Genet Med. 19(2):249-255. doi: 10.1038/gim.2016.190

Ceyhan-Birsoy et al. (2017) A curated gene list for reporting results of newborn genomic screening. Genet Med. 19:809-818 doi: 10.1038/gim.2016.193

Plagnol, V. et al. (2012) A robust model for read count data in exome sequencing experiments and implications for copy number variant calling. Bioinformatics 28(21):2747–54 doi:
10.1093/bioinformatics/bts526

Zook, J. M. et al. (2019) A robust benchmark for germline structural variant detection. bioRxiv, June 2019 doi: 10.1101/664623

Sadedin, S. P. et al. (2018) Ximmer: a system for improving accuracy and consistency of CNV calling from exome data. Gigascience 7(10) doi: 10.1093/gigascience/giy112

Zook, J. M. et al. (2016) Extensive sequencing of seven human genomes to characterize benchmark reference materials. Sci Data 3:160025 doi: 10.1038/sdata.2016.25

Jeffreys, H. (1998) The Theory of Probability (3rd ed.). Oxford, England.
P. 432. ISBN 9780191589676

Smedley, D. et al. (2016) A Whole-Genome Analysis Framework for Effective Identification of Pathogenic Regulatory Variants in Mendelian Disease. Am J Hum Genet 99(3):595–606 doi:
10.1016/j.ajhg.2016.07.005