2017 9th International Conference on Knowledge and Systems Engineering(KSE)

Genomedics: Whole exome analysis system for clinical studies

Le Sy Vinh, Nguyen Duc Canh, Bui Ngoc Thang
University of Engineering and Technology
Vietnam National University Hanoi
Hanoi, Vietnam

Do Thi Thu Hang
School of Medicine
Vietnam National University HCMC
Ho Chi Minh City, Vietnam

Duong Quoc Chinh, Tran Cong Hoang
National Institute of Hematology and Blood Transfusion
Hanoi, Vietnam

Le Ba Hong Minh, Pham Thi Dieu Linh
Vinmec Research Institute of Stem Cell and Gene Technology
Hanoi, Vietnam

***Abstract*—Whole exome sequencing (WES) is a widely used technique in both medical studies and clinical practice. However, a number of studies show that the results produced by different WES analysis pipelines are not always homogeneous. To this end, we propose a method (called Genomedics) using a consensus approach to expand the list of variants by combining results called from six separate pipelines with sensitive options. To evaluate the performance of the proposed method, Gemomedics was compared to seven existing methods when they were tested on two datasets and F1-score was used as an indicator of accuracy. The results showed that Genomedics has the highest score among seven methods. We also applied Genomedics to analyze whole exomes from Multiple Myeloma and Dravet syndrome patients and found interesting results. The results demonstrate the promising applications of Genomedics in clinical studies.**

Keywords: genomedics, whole exome sequence, variant calling, clinical studies, application of WES.

I. INTRODUCTION

The human genome consists of about 3 billion nucleotides. Each person has from 3 to 4 million of variants on the genome [1,2]. The exome is the part of the genome that encodes proteins. It contributes approximately 1% of human genome [3]. Using the next generation sequencing (NGS) technologies, whole exome and whole genome are sequenced with affordable cost allowing their applications for clinical studies [4].

Whole exome sequencing analysis has undeniable application potential. Although human exome only contributes a very small percentage of human genome, most of our current knowledge about functional genetic variation is on this region [5]. It is estimated that the whole exome contributes about 85% of the disease-causing mutations in Mendelian disorders [6]. Another advantage of whole exome sequencing is that its cost is significantly lower than the cost of whole genome sequencing. Therefore, whole exome sequencing is not only limited in clinical research but also is becoming a standard test to identify the underlying genetic cause of disease for a variety of indications in clinical diagnoses [4,7,8].

Analyzing whole human exomes from NGS data is a complex problem. It consists of several main steps: mapping short reads on to the reference genome, calling variants, and annotating called variants. Different methods have been proposed for each step (e.g., BWA for mapping short reads [9], GATK for calling variants [10], and SNPeff for annotating variants [11]).

It is well known that variants called from different methods can be discordant, therefore, popular methods are typically used together to call variants [1,12]. Software packages such as Seqmule [12] have been developed allowing users to call variants from a number of methods. Although these packages also combine variants called from different methods, they are not designed to evaluate the reliability of a variant based on the results from different methods.

In this study, we introduce Genomedics to analyze and annotate whole human exomes. Genomedics employs the consensus approach to combine results obtained from a number of pipelines with sensitive options. We examined the performance of Genomedics and other methods on the gold standard datasets. Finally, we applied Genomedics to study epilepsy and multiple myeloma diseases.

II. METHODS

A. Variant calling methods

Calling variants is a crucial step in the whole exome analysis. This step involves various computational methods to determine different variant types from the raw data. The variant types can be classified into short variants and structural variants. Short variants include single nucleotide variants and short insertions/deletions, called short indels. The structural variants include a wide range of variants that can affect the structure of genes, and consequently corresponding proteins. Among structural variant types, copy number of variations are the causes of a number of neuro-related diseases such as epilepsy, and autisms [13].

To call variants from the raw data, short reads are mapped to the reference sequence to create alignments. A number of alignment methods have been proposed, and two popular methods are Burrows-Wheeler Aligner (BWA) [9] and Bowtie2 [14]. The BWA aligner uses Burrows-Wheeler Transformation (BWT) of the reference genome to efficiently align short sequencing reads. The BWA not only minimises the memory needed to store the reference, but also allows a matching strategy for the reads operating in the order of the read length. The Bowtie2 aligner is an ultrafast, memory-

978-1-5386-3576-6/17/$31.00

efficient alignment program for aligning short reads. Bowtie2 extends previous Burrows-Wheeler techniques with a novel quality-aware backtracking algorithm that permits mismatches. The reference genome is indexed by using a scheme based on (BWT) and the full-text minute-space index (FM index). A Bowtie indexing for the human genome requires only few gigabytes of memories. Note that alignments obtained from BWA and Bowtie2 are then sorted, indexed by Samtools [15] and marked duplications by Picard software [16], and subsequently used for calling variants.

A number of calling methods have been proposed to call single nucleotide variants and short indels from alignments such as GATK [17,18], Platypus [10], and Freebayes [19]. GATK is a unified analytic framework to discover genotype variation among multiple samples. It consists of the following stages: (i) initial read mapping; (ii) local realignment around indels; (iii) base quality score recalibration; (iv) SNP discovery and genotyping to find all potential variants; and (v) separating true segregating variations from machine artifacts common to NGS technologies. The Platypus method is a haplotype-based variant caller for next generation sequence data. This variant caller is able to efficiently and accurately detect variants from both whole genome and whole exome data. Platypus has been extensively examined in discovering somatic mutations in cancer studies from whole human exome data. The Freebayes caller detects haplotype-based variants by applying a Bayesian statistical framework to model multiallelic loci on multiple individuals.

The combination of an aligner (i.e, BWA or Bowtie2) with a variant caller (i.e., GATK, Platypus, or Freebayes) is called a pipeline. Thus, there are six pipelines for calling single nucleotide variants and short indels: 1) BWA and GATK, 2) BWA and Platypus, 3) BWA and Freebayes, 4) Bowtie2 and GATK, 5) Bowtie2 and Platypus; and 6) Bowtie2 and Freebayes. These pipelines can be conducted with different options. In this study, we used the default options or recommended options in the best practice of the software.

Calling structure variants from whole genome data, especially from whole exome data is still a challenging problem. Determining copy number of variants from whole exome data plays an important role in clinical diagnoses and precision medicine. A number of methods have been proposed such as Conifer [20], XHMM [21], and Excavator [22]. However, an extensive evaluation of these tools with different parameter sets for properly calling structural variants is still required.

B. Genomedics variant calling method

A number of studies have shown the advantages and disadvantages of different pipelines. The most challenging problem is the discordance between variants called from different pipelines [1,12]. Genomedics applies a simple consensus strategy to determine variants from different pipelines. Genomedics includes two main steps:

- Use six pipelines with very sensitive options to call variants. This strategy enables pipelines to discover as many variants as possible, thus, increases the sensitive of Genomedics.
- Use to consensus approach to reduce the fall positive rate. Genomedics evaluates called variants from different pipelines, and only considers variants as consensus variants if they are called by at least two pipelines.

The full workflow of Genomedics is described in Figure 1. Note that Genomedics provides three different methods (i.e., Conifer, XHMM, and Excavator) to call copy number of variations from the whole exome data.

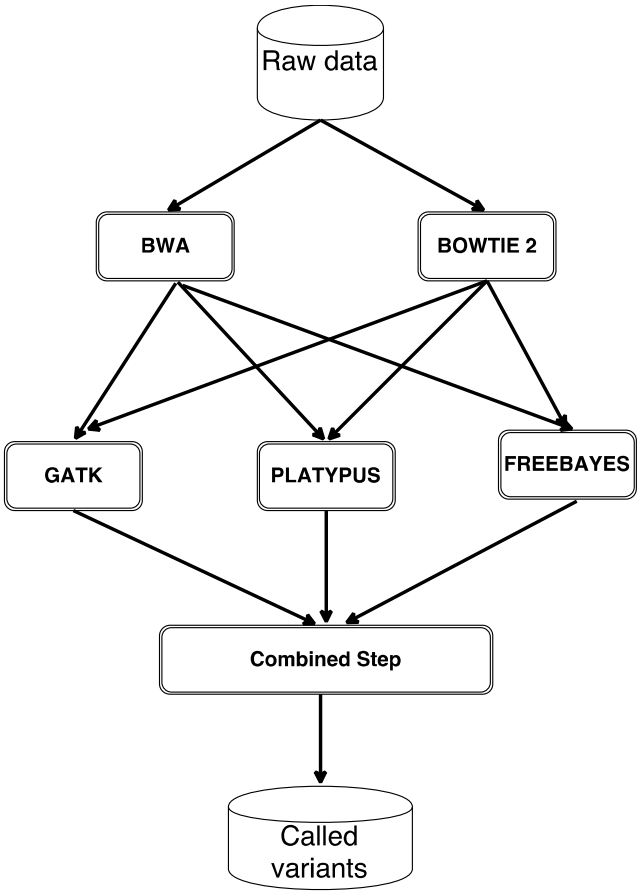

Figure 1: The workflow of Genomedics to determine consensus variants from different pipelines with high sensitive options.

C. Genomedics annotations

Variants called from Genomedics are then annotated. Each variant will be annotated with following essential information:

- The type and impact of variants on genes and on proteins are annotated by SNPeff [11].
- The deleterious level of variants are annotated by SIFT software [23]. Specifically, if the SIFT score of a variant is greater than 0.05, it is annotated as "Tolerated", otherwise, it is annotated as "Damaging".

- The minor allele frequency (MAF) of a variant is annotated by two large databases, i.e., the Exome Aggregation Consortium (ExAC)/The Genome Aggregation Database (gnomAD) [24] and the 1000 Genome Project [1]. The ExAC database is a powerful resource to filter variants that are not disease-related mutations on severe pediatric diseases. The database from the 1000 human genome project (phase 3) consists of 2504 healthy individuals from 26 populations. The MAF from the 1000 human genome project is used to filter variants that are not causing disease mutations.
- The variants called from Genomedics are finally annotated with the human gene mutation database [30]. Genomedics selects only variants with phenotypes annotated as pathogenetics or likely pathogenetics.

III. Experiments

A. Data

The Genome in a Bottle (GIAB) Consortium hosted by NIST created the highly confident small variant (SNP and Indel) calls for NA12878 sample. The data can be used as reference data to measure the performance of different whole genome/exomes analysis pipelines for single nucleotide variants and short indels [25]. We used the whole exomes of sample NA12878 sequenced from Garvan to examine Genomedics and different pipelines in this study. The sample NA12878 was sequenced in two separated runs (namely NIST7086 and NIST7035).

B. Results

We summarized the data coverage on bases with Phred quality score of at least 10. The data average coverage is 52X and 55X for NIST7086 and NIST7035, respectively. Figure 2 shows the percentage of the whole exome covered at different depth coverage levels. For example, only 87.97% of the whole exome of NIST7035 are covered by at least one read.

We measure the precision and the sensitivity to calculate F-score of Genomedics and other pipelines. The F-score is calculated as following:

$$F_a = \frac{(1+a^2).precision.sensitivity}{a^2.precision + sensitivity}$$

Where a is a positive real value. We can easily see that if $a > 1$ then the F_a–score will emphasize on precision more than on sensitivity, and reversely, if $a < 1$ then F_a–score will emphasize on sensitivity more than on precision. In this work, we want to balance weights between the precision and the sensitivity, thus, we use $F_1 -$ score as a main criteria to compare Genomedics and the six pipelines.

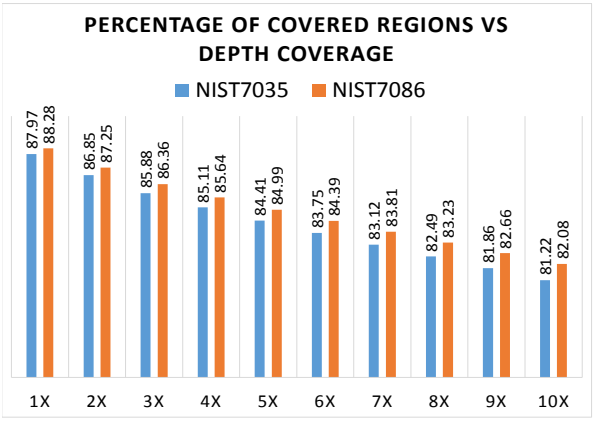

Figure 2: The percentage of data coverage on targeted regions with different depth coverages

The results on NIST7035 and NIST7036 are presented in Table 1 and Table 2, respectively. On both datasets, Genomedics has the highest $F_1 -$ score among the seven methods (93.3% on NIST7035 and 93.8% on NIST7036). It is slightly better than the combination of BWA and GATK. In addition, it is clearly better than the combination of Bowtie2 and Platypus. The results show that the sensitivity of Freebayes method is not high as other methods (i.e., only 81.4% and 77.1% on BWA and Bowtie2 alignments, respectively).

We also measured the performance of the Genomedics on positions with the coverage depth of at least 4. The $F_1 -$ score of Genomedics increases significantly to above 97% (both sensitivity and the precision are similar and equal to about 97%). The performances of all methods are similar on both NIST7035 and NIST7086 datasets.

TABLE 1: RESULTS FROM TNIST7035 DATASET. GENOMEDICS PERFORMS BETTER THAN PIPELINES.

pipelines	Precision	Sensitivity	F-Measure
BWA + GATK	98.0	87.9	92.7
BWA + Platypus	99.1	86.5	92.4
BWA + Freebayes	98.6	81.4	89.2
Bowtie2 + GATK	98.1	86.4	91.9
Bowtie2 + Platypus	99.3	82.1	89.9
Bowtie2 + Freebayes	99.4	77.1	86.9
Genomedics	**96.5**	**90.3**	**93.3**
Genomedics*	**96.9**	**97.3**	**97.1**

**positions covered by less than 4 reads are excluded from analyses*

TABLE 2: RESULTS FROM THE NIST7086 DATASET. GENOMEDICS PERFORMS BETTER THAN PIPELINES.

pipelines	Precision	Sensitivity	F-Measure
BWA + GATK	98.3	88.5	93.1
BWA + Platypus	99.1	87.8	93.1
BWA + Freebayes	99.1	87.8	93.1
Bowtie2 + GATK	98.0	87.8	92.6
Bowtie2 + Platypus	99.3	83.3	90.6
Bowtie2 + Freebayes	99.5	78.6	87.8
Genomedics	**97.0**	**90.8**	**93.8**
Genomedics*	**97.4**	**97.2**	**97.3**

*positions covered by less than 4 reads are excluded from analyses

IV. APPLICATION OF GENOMEDICS IN CLINICAL STUDIES

A. *Multiple Myeloma case study*

Multiple Myeloma is a type of blood cancer that forms in a type of white blood called a plasma cell. Multiple Myeloma relates to a large number of pathways and genes. A wide range of mutations have been reported, notably single nucleotide variants, short indels, copy number of variants, etc. [26]. In this paper, we used Genomedics to analyze the whole exomes sequenced from two multiple myeloma patients, named M1 and M2. For each patient, the DNA was extracted from peripheral blood and bone marrow. The whole exomes were sequenced using the MiSeq, and we obtained four exomes (see Table 3).

TABLE 3: NGS DATA OBTAINED FROM TWO MULTIPLE MYELOMA PATIENTS, NAMED M1 AND M2. FOR EACH PATIENT, THE WHOLE EXOMES FROM BLOOD AND MARROW DNA WERE SEQUENCED.

	M1 bone marrow	M1 blood	M2 bone marrow	M2 blood
Mapped bases (millions)	4012	3076	3895	3494
Mapped bases on targeted regions (millions)	2720	1961	2525	2241
Coverage on targeted regions	60X	43X	56X	49X

We used Genomedics to call and annotated consensus variants for four exomes. We used a multiple Myeloma gene panel of 77 genes [27] to diagnose variants from the four exomes.

Table **4** presents SNVs called from M1 samples. We found 4 SNVs that appear in the marrow bone sample, but not blood sample. We found 2 frame-shift SNVs on TP53 gene; one missense SNV on IL6ST; and one missense SNV on RIPK4 gene. The show the damage of exome exacted from marrow bone in comparison to that from blood.

TABLE 4: SINGLE NUCLEOTIDE VARIANTS AND SHORT INDELS FROM M1 PATIENT.

Chr	Pos	Gene	Variant	HGVS.p	Blood	Marrow bone
17	7578221	TP53	frameshift	p.Arg209fs	0/0	0/1
17	7578221	TP53	frameshift	p.Arg209fs	0/0	0/1
5	55247795	IL6ST	missense	p.Tyr554Phe	0/0	0/1
21	43161219	RIPK4	missense	p.Ala760Thr	0/0	0/1

e, but not in the blood sample.

TABLE **5** presents SNVs called from S2 samples. We found 4 SNVs that appear in the marrow bone sample; 3 of that also appear in blood sample. The first two SNVs are predicted as splice acceptor variants on ATM gene. The third one is annotated as a structural interaction variant on TRAF2 gene. The last variant (a missense variant) appears only in the marrow bone sample, but not in the blood sample.

TABLE 5: SINGLE NUCLEOTIDE VARIANTS AND SHORT INDELS FROM M2 PATIENT.

Chr	Pos	Gene	Variant	HGVS.p	Blood	Marrow bone
11	108121410	ATM	Splice acceptor	p.Arg209fs	0/1	0/1
11	108121410	ATM	Splice acceptor	p.Arg209fs	0/1	0/1
9	139793242	TRAF 2	Structural interaction	p.Tyr554Phe	0/1	0/1
1	115256529	NRAS	Missense variant	p.Gln61Arg	0/0	0/1

Finally, we used Excavator on BWA alignment to call CNVs from M1 and M2 samples. The blood exomes played as the control to call CNVs from marrow bone exomes. Table 6 presents the number of duplications and deletions detected from M1 and M2 marrow bone exomes. The CNVs are overlap to 20 and 15 multiple myeloma related genes on M1 and M2 marrow bone exomes, respectively.

TABLE 6: NUMBER OF DUPLICATIONS AND DELETIONS DETECTED FROM M1 AND M2 MARROW BONE EXOMES.

Sample	#Duplications	#Deletion	#Related genes
M1 marrow bone	10	5	20
M2 marrow bone	8	2	15

B. *Dravet case studies*

Dravet syndrome is a rare and severe type of epilepsy in infants. The genetic causes of Dravet syndrome are due to mutations in SCN1A and several other genes, including but not limited to PCDH19, GABRG2, SCN1B, SCN9A, CHD2 [28,29]. We also applied Genomedics to analyzed whole exome of two boys diagnosed with the Dravet syndromes, named S1 and S2.

Analyzing whole exome of S1 revealed a nucleotide deletion (c.4503delA) on SCN1A gene (at position 166852600 on the chromosome 2) that is annotated as a frame-shift variant with high impact. The variant are not reported in the ExAC, 1000 Genome Project databases. It is not reported in the Human Gene Mutation Database [30].

We also found a missense variant (c.4573C>T) on SCN1A gene (at position 166852531 on chromosome 2) on the S2 sample. The variant was annotated as stop-gained variant with high impact. The variant is not reported in the ExAC, 1000 Genome Project databases. However, it is reported as causing disease mutation in The Human Gene Mutation Database [30]. These two variants on SCN1A gene on both S1 and S2 samples were confirmed by the Sanger sequencing method.

V. DISCUSSION

Whole exome analysis is becoming more and more popular in clinical studies. Calling variants from whole exomes has been studied intensively, however, the results obtained from different variant callers are partly discordant. We have described Genomedics as a simple and efficient method to analyze whole exomes. Genomedics not only combines results from different pipelines, but also annotates variants to assist diagnosing diseases.

Since Genomedics combines six separate pipelines in one package, it is relatively computational expensive. However, considering the time and effort we spend on collecting and sequencing the samples, we think it is a reasonable trade-off to extensively work on searching for meaningful variants. In fact, we already applied Genomedics to analyze whole exomes from Multiple Myeloma and Dravet syndrome patients and found promising results. Currently, we are applying Genomedics to other large-scale clinical studies.

ACKNOWLEDGMENT

This work is financially supported by BIDV through the foundation of science and technology development, Vietnam national university Hanoi.

REFERENCES

[1] The 1000 Genomes Project Consortium. A global reference for human genetic variation. Nature. 2015. p. 68–74.

[2] Hai DT, Thanh ND, Trang PTM, Quang LS, Hang PTT, Cuong DC, et al. Whole genome analysis of a Vietnamese trio. J. Biosci. 2015;40:113–24.

[3] Directors AB of. Points to consider in the clinical application of genomic sequencing. Genet Med. 2012;14:759–61.

[4] Rabbani B, Tekin M, Mahdieh N. The promise of whole-exome sequencing in medical genetics. J. Hum. Genet. 2014;59:5–15.

[5] Jacob HJ. Next-Generation Sequencing for Clinical Diagnostics. N. Engl. J. Med. 2013;16–7.

[6] Majewski J, Schwartzentruber J, Lalonde E, Montpetit a., Jabado N. What can exome sequencing do for you? J. Med. Genet. 2011;48:580–9.

[7] Yang Y, Muzny DM, Reid JG, Bainbridge MN, Willis A, Ward P a, et al. Clinical Whole-Exome Sequencing for the Diagnosis of Mendelian Disorders. N. Engl. J. Med. 2013;369:1502–11.

[8] Retterer K, Juusola J, Cho MT, Vitazka P, Millan F, Gibellini F, et al. Clinical application of whole-exome sequencing across clinical indications. Genet Med. American College of Medical Genetics and Genomics; 2015.

[9] Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25:1754–60.

[10] Rimmer A, Phan H, Mathieson I, Iqbal Z, Twigg SRF, Wilkie AOM, et al. Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. Nat. Genet. 2014;46:1–9.

[11] Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w 1118; iso-2; iso-3. Fly (Austin). 2012;6:80–92.

[12] Guo Y, Ding X, Shen Y, Lyon GJ, Wang K. SeqMule: automated pipeline for analysis of human exome/genome sequencing data. Sci. Rep. 2015;5:14283.

[13] Marshall CR, Scherer SW. Detection and characterization of copy number variation in autism spectrum disorder. Methods Mol. Biol. 2012;838:115–35.

[14] Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9:357–9.

[15] Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25:2078–9.

[16] Broad Institute. Picard tools. https://broadinstitute.github.io/picard/. 2016.

[17] DePristo MA, Banks E, Poplin R, Garimella K V, Maguire JR, Hartl C, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat. Genet. 2011;43:491–8.

[18] McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20:1297–303.

[19] Garrison E, Marth G. Haplotype-based variant detection from short-read sequencing. arXiv Prepr. arXiv1207.3907. 2012;9.

[20] Krumm N, Sudmant PH, Ko A, O'Roak BJ, Malig M, Coe BP, et al. Copy number variation detection and genotyping from exome

sequence data. Genome Res. 2012;22:1525–32.

[21] Fromer M, Purcell SM. Using XHMM software to detect copy number variation in whole-exome sequencing data. Curr. Protoc. Hum. Genet. 2014;

[22] Magi A, Tattini L, Cifola I, D'Aurizio R, Benelli M, Mangano E, et al. EXCAVATOR: detecting copy number variants from whole-exome sequencing data. Genome Biol. 2013;14:R120.

[23] Ng PC, Henikoff S. SIFT: Predicting amino acid changes that affect protein function. Nucleic Acids Res. 2003;31:3812–4.

[24] Lek M, Karczewski KJ, Samocha KE, Banks E, Fennell T, O AH, et al. Analysis of protein-coding genetic variation in 60,706 humans. bioRxiv. 2016;536:30338.

[25] Zook J, Chapman B, Wang J. Integrating sequencing datasets to form highly confident SNP and indel genotype calls for a whole human genome. arXiv Prepr. arXiv. 2013.

[26] Walker BA, Boyle EM, Wardell CP, Murison A, Begum DB, Dahir NM, et al. Mutational spectrum, copy number changes, and outcome: Results of a sequencing study of patients with newly diagnosed myeloma. J. Clin. Oncol. 2015;33:3911–20.

[27] K. Martin Kortuem, Stewart AK, Al. E. Development and Results of a Multiple Myeloma Specific Custom 77-Gene Mutation Panel for Clinical Targeted Sequencing. Blood. 2014;124:169.

[28] Marini C, Scheffer IE, Nabbout R, Suls A, De Jonghe P, Zara F, et al. The genetics of Dravet syndrome. Epilepsia. 2011;52:24–9.

[29] Carvill GL, Weckhuysen S, McMahon JM, Hartmann C, Møller RS, Hjalgrim H, et al. GABRA1 and STXBP1: Novel genetic causes of Dravet syndrome. Neurology. 2014;82:1245–53.

[30] Stenson PD, Mort M, Ball EV, Shaw K, Phillips A, Cooper DN: The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. Hum. Genet. 2014 ;133:1-9.