Reliable detection of mosaic variants based on whole exome sequencing data.

Key words: Clinical diagnostics application, mosaic variants, whole exome sequencing.

Layman's Summary

When reading someone's DNA, every position is labeled as reference or a variant, a variant means different from the reference. Each person has wo copies of each somatic chromosome, called alleles, and a variant on each allele can either be healthy like the reference (A) or carry a variant (B). So, the combinations can be healthy AA, heterozygous variant AB or homozygous variant BB. The variant allele is thus typically found in 0, 50, or 100 percent, depending on the combination of healthy and variant alleles.

Another type of variants that are known to cause disease are Mosaic variants. These variants are different from normal variants as they are present in only a subpopulation of cells in the body, rather than in all cells. When reading the DNA of a person with a mosaic variant, the percentage of the variant will be somewhere between 0 and 50 percent. Mosaic variants are known to be involved in the development of various diseases, including autoinflammatory diseases. However, diagnosing patients with mosaic variants is currently a challenging task because it requires either manually examining of the DNA reads or using a targeted test. Current clinical practices could benefit from having a test that can find mosaic variants in all the genes. The MosaicHunter tool allows for the detection of these mosaic variants in DNA and could be of value in diagnostics. It offers a more general approach that can be used to detect mosaic variants in the exome, which is the complete set of genes in an organism.

To evaluate the effectiveness of the MosaicHunter tool, simulated datasets were used that contain a number of true positive mosaic variants. The dataset where simulated because there are no datasets available that we need. These datasets were created in two different ways: one exome-wide dataset in which the mosaic variants were randomly distributed throughout the exome, and another dataset focused on six genes known to cause disease with a mosaic variant present. The MosaicHunter tool was able to detect 80 percent of the simulated mosaic variants in these datasets, demonstrating its potential usefulness for diagnosing patients with mosaic variants.

Overall, the MosaicHunter tool is a promising development in the field of genetics because it allows for the detection of mosaic variants in a more general and efficient way than previous methods. Further research and testing will be necessary to fully assess its capabilities and potential applications in clinical settings.

Abstract

With next generation sequencing (NGS) becoming faster, cheaper and more reliable, the use of NGS has become a standard procedure to diagnose genetic conditions in clinical human genetics. Genetic testing focusses on the identification of single nucleotide variants (SNV), INDELs, and copy number alterations (CNV) in the DNA. In the UMCU, when a patient undergoes genetic testing, whole exome sequencing (WES) is often carried out and variants are detected using the GATK haplotypecaller. There is increased evidence that mosaic variants can also cause genetic diseases. Mosaic variants, which are variants that are only present in a sub population of cells in the body, are hard to distinguish from sequencing errors and artifacts. Currently, detecting mosaic variants requires manual examination or a targeted approach, and is only performed when mosaicism is suspected. We would prefer to use an exome-wide approach, such as WES, to detect mosaic variants because it would be more flexible, generic, and able to explore mosaic variants that are not included in the targeted panel. In this study, we investigate the potential of MosaicHunter software to reliably detect mosaic variants in non-paired whole exome sequencing samples (WES). MosaicHunter is a tool that uses a Bayesian genotyper and error filters to make mosaic variant calls. In this study we used simulated data and positive controls with known mosaic variants to optimize the settings of MosaicHunter and determine the sensitivity MosaicHunter on different amounts of coverage. We show that MosaicHunter has a high sensitivity, and a better sensitivity then the GATK pipeline, to detect mosaic variants between 3 and 12%. We demonstrate that the method is of value in combination with the current GATK haplotype caller because, we showed that a combined analysis of MosaicHunter and GATK haplotypecaller results in a sensitivity >80% for mosaic variants in range of 2-12%, and >95% for mosaic variant above 12%.

Introduction

With next generation sequencing (NGS) becoming faster, cheaper and more reliable, the use of NGS has become a standard procedure to diagnose genetic conditions in clinical human genetics. Genetic testing focusses on germline, inherited, or somatic, acquired, alterations of the DNA (Yohe & Thyagarajan, 2017). The results of a genetic test can rule out or confirm a suspected genetic condition or can determine the probability to develop a disease. Several genetic tests are currently in use at the genome diagnostics lab in the University Medical Center Utrecht (UMCU). These methods can have a targeted approach for single variants to a few genes (i.e., Sanger sequencing and smMIP based targeted assays), or targeted to all genes such as whole exome sequencing (WES). Within the UMCU, WES is currently the standard test for most patients with a suspected genetic disease, although some specific lab tests are still needed as WES is insufficient to detect all forms of genetic variation. Around 7000 WES samples are currently sequenced per year. With WES a sample is taken from a person, this can be blood, skin or other tissue. In the UMCU the DNA is isolated from the sample and the sample is prepped for sequencing, for WES this means enrichment of the exome and sequencing is done by Illumina short read sequencing using the Illumina Novaseq 6000. After sequencing, reads are mapped to the reference genome, and variant calling is performed to identify germline and de novo variants for single nucleotide variants (SNV), small insertion/deletions (INDELs), and copy number variants (CNV) (Ernst & Elferink, 2020). With the use of WES, we now focus on germline or *de novo* variants (figure 1, left path).

But next to germline variants which are present in virtually all cells of the body, genetic variants can be present within subpopulations of cells, referred to as mosaic or somatic variants (figure 1 right). There is increased evidence that mosaic variants can also cause genetic diseases (Vijg & Dong, 2020). Mosaic variants are acquired later after fertilization, postzygotic, during cell mitosis or DNA repair (figure 1, right figure, left path). Variants acquired later in development are located at a specific site (figure 1, right figure, right path). Mosaic variants are harder to detect with NGS methods due to the low frequencies which makes it difficult to distinguish them from sequencing errors and artefacts that frequently occur during library prep and sequencing. Moreover, due to the low frequency of these variants high sequencing depth is generally required.

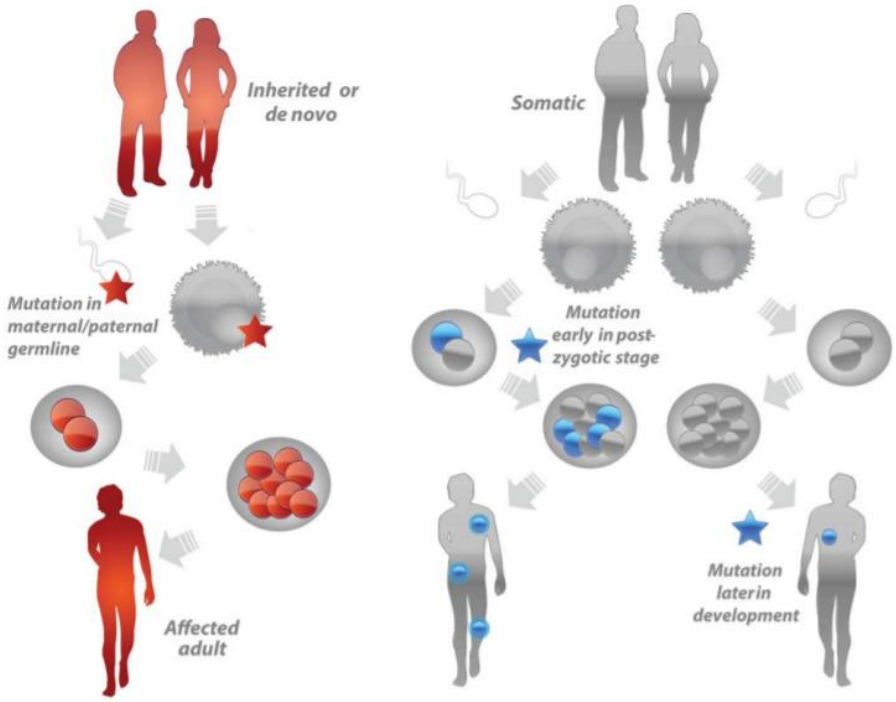

Figure 1: Tree different ways of acquiring a variant. Left, in red, an acquired germline variant, all cells of the affected person contain the variant. On the right, in blue, two different mosaic variants are depicted. A variant that occurred in postzygotic stage this variant only affects a part of the persons tissue but is seen in different sites in the body. The variant that occurred later in the development, that is only in one place at the person's body. (Hoffman & Broderick, 2017)

Several auto inflammatory diseases including VEXAS (vacuoles, E1 enzyme, X-linked, autoinflammatory, somatic) syndrome, and CAPS (cryopyrin-associated periodic syndrome) are known to be caused by mosaic variants (Ionescu et al., 2022; Labrousse et al., 2018; van der Made et al., 2022). For example, disease symptoms in CAPS are already observed when the causal variant is present in 0.5% of the blood cell population (Labrousse et al., 2018). Auto inflammatory diseases cause a broad range of symptoms for these including fever, headaches, redness of the skin and arthritis (Beck et al., 2020). Early diagnosis of these auto inflammatory diseases can provide better prognostic information and more suitable treatment options including stem cell transplantation (Beck et al., 2020; van der Made et al., 2022).

Currently to find these auto inflammatory diseases and confirm mosaicism single-molecule molecular inversion probes (smMIP) tests are being used (Eijkelenboom et al., 2016). This is an autoinflammatory gene panel, in current use of the UMCU, that consist of 5 genes that are known be pathogenic as a mosaic variant. Although the smMIP test is very sensitive to detect mosaic variants, the test is limited to a few genes and prohibits the

detection of mosaic variants in other genes. Additional disease genes can be included in the test but requires a labor-intensive procedure for optimalization and validation. Due to the inflexibility of the smMIP method it will be beneficial to use a more comprehensive and generic genetic test such as WES for the detection of mosaic variation.

Use of WES would fit us best for mosaic variant detection because of the general use in our diagnostics in the UMCU and because it gives us flexibility to explore mosaic variants in other cases. However, several challenges arise to detect mosaic variants based on WES, as WES data shows non-negligible capturing bias and over-dispersion in the distribution of alternative allele fractions (Ramu et al., 2013). This causes problems when calling variants because not all exons are covered equally. Despite the power of NGS, NGS library preparation, base-calling, and alignment introduce technical artifacts that are difficult to distinguish from true mosaic variants (Huang et al., 2017). The major challenge for bioinformatic algorithms is to find the physiologically relevant signals, coming from true mosaic variants, from these methodologically induced variation (Forsberg et al., 2017).

Several tools are developed that focus on the detection of mosaic variants in NGS data such as MosaicHunter, MosaicForecast and DeepMosaic (table 1) (Dou et al., 2020; Huang et al., 2017; Yang et al., 2021). More tools are developed for calling mosaic variants, but these need tumor-normal pairs, these tools include Varscan2, MuTect and SomVarIUS. However, all of them are focused on the detection of somatic variants in cancer samples (Koboldt et al., 2012; *Mutect2*, z.d.; Smith et al., 2016) and cannot be used when paired control samples are unavailable which is the case in our diagnostic patients.

Table 1: Overview of mosaic calling tools. With the possible inputs; whole genome sequencing (WGS) or whole exome sequencing (WES). The type of variants, single nucleotides variants (SNV) and indels. The method used and the output.

Tool	Input	Variant detection	Method	output
MosaicHunter	WES, WGS	SNV	Filters	Text file
MosaicForecast	WGS (WES)	SNV & indels	Phasing & random forest	Text file
DeepMosaic	WGS (WES)	SNV	Neural network	Text file

From the three available tools that could be used to detect mosaic variants (table1), MosaicHunter is the only tool that is developed and tested on the detection of mosaic variants in WES, unlike MosaicHunter and MosaicForecast which are based on WGS. A disadvantage of MosaicHunter is that it focused on mosaic SNVs and cannot detect mosaic INDELs.

MosaicHunter (Huang et al., 2017) is a Java-based computational tool for identification of mosaic variants in WGS and WES data without the need of having paired control samples. It is a Bayesian genotyper that uses a set of several specific error filters to detect mosaic variants. These filters include multiple filters for correcting for data quality and bias. It works in two steps. First, for WES data a beta binomial model is fitted to correct for overdispersion and capturing bias. For this an alpha and beta are fitted on this beta binomial model and the alpha and beta are input BAM specific. Secondly, with the parameters fitted to correct for over dispersion and capturing bias the suspected mosaic variants are calculated. Figure 2 gives an overview of the Bayesian genotyper and the error filters. MosaicHunter gives in the end a text file with a list of mosaic sites found in the input BAM file.

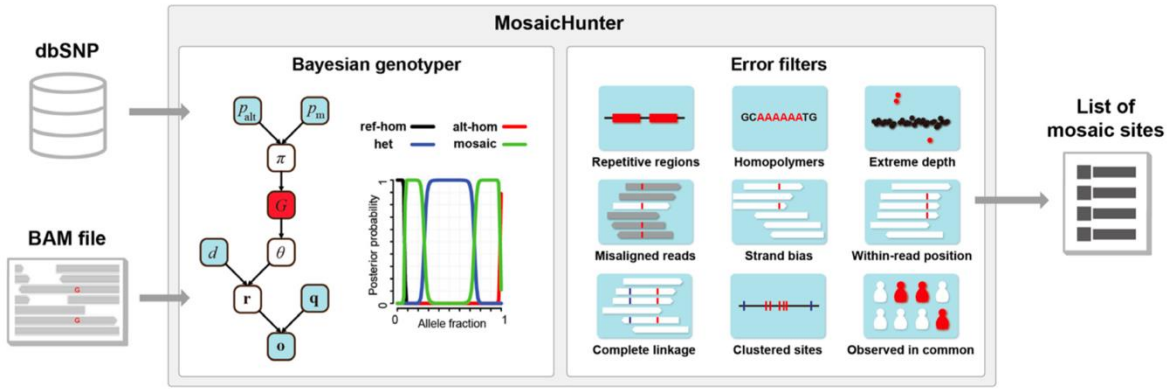

Figure 2: Overview of MosaicHunter. From left to right: input BAM file and dbSNP file with common SNPs are used for the Bayesian genotyper to call potential mosaic variants. These variants are subsequently filtered using a series of stringent error filters. Ref-hom: homozygous for reference allele. Alt-hom: homozygous for alternative allele. Het: heterozygous. Altered from (Huang et al., 2017).

In this study, we investigated the added value of MosaicHunter to reliably detect mosaic variants within our current diagnostic WES workflow. Towards this end, we used data from positive control samples with known mosaic variants, and simulated datasets to optimize settings and to determine sensitivity to detect mosaic variants within the exome and an auto-inflammatory gen-panel currently used in de smMIP diagnostic. Based on the simulated datasets, we furthermore determined the sensitivity to detect mosaic variants using the current variant caller (GATK haplotypecaller) in our diagnostic workflow.

We show that the GATK-haplotypecaller is sensitive to detect mosaic variants for frequencies above 12% (sensitivity >95%) but performs poorly below this frequency (sensitivity <5%), and the mosaic variants are called as heterozygous with GATK. MosaicHunter has a high sensitivity to detect mosaic variants from 3-22% (sensitivity >80%), including a high sensitivity in the range of 3-12% (sensitivity >70%). Adding MosaicHunter to the diagnostic workflow would therefore be beneficial, increasing the sensitivity to detect genetic variant present at 3% or more.

Material and Methods

Installation of tools.

To deploy the tools on the UMCU high-performance compute cluster (HPC) docker images were created for BAMSurgeon (supplementary 1, part 1.2) and MosaicHunter (supplementary 1, part 1.1). Dockers can be found at: hub.docker.com (https://hub.docker.com/repository/docker/anabuurs/bamsurgeon & https://hub.docker.com/repository/docker/anabuurs/mosaichunter) (*Docker Hub Container Image Library | App Containerization*, z.d.). The Dockerfiles used can also be found on the GitHub links can be found in supplementary 1, part 1.1 and 1.2. To check the installation of BAMSurgeon a random variant was simulated in a BAM file and checked in IGV. The demo of MosaicHunter https://github.com/zzhang526/MosaicHunter/tree/master/demo was performed to check whether the tool was installed properly.

For the in house developed python scripts virtual environments were used. A list of dependencies for the variant selection can be found at the git and supplementary 1, part 1.3. A list of dependencies for the tsv to vcf parser can be found at the git and supplementary 1, part 1.4.

GATK calling

The SD_Exome_Normal and SD_Exome_Deep datasets were analyzed with GATK haplotypecaller. Diagnostic settings like in the WES workflow of the UMCU were used (supplementary 1, part 3). The script used can be found at the GitHub (supplementary 1, part 3). These settings of GATK were used for all variant callings with GATK.

Creation of datasets

Five datasets were made in which mosaic variant SNVs were simulated: SD_Exome_Normal SD_Exome_Deep and SD_PID_5, SD_PID_5strand and SD_PID_10. Simulation of variants within BAM files was performed using BAMsurgeon (adamewing, 2012, https://github.com/adamewing/bamsurgeon). See supplementary 5 for more details about BAMsurgeon.

Table 2: Overview of datasets used for testing MosaicHunter.

Dataset	Sample	Settings	Goal
SD_Exome_Normal	3 100X WES	100 mosaic variants per % 1-30 (n=3000)	Find sensitivity of MosaicHunter for every percentage between 1-30 for 100X WES samples
SD_Exome_Deep	1 500X WES	100 mosaic variants per % 1-30 (n=3000)	Find sensitivity in deepseq samples for every percentage between 1-30
SD_PID_10	30 100X WES	10% mosaic	Look at target regions and the sensitivity in them
SD_PID_5	10 100X WES	5% mosaic	Look at target regions and the sensitivity in them
SD_PID_5strand	10 100X WES	5% mosaic & strandbias threshold 0.000001	Look at target regions and the sensitivity in them

TP_Regular	3 100X WES	-	True positive samples for validation
TP_Deepseq	7 500X WES	-	True positive samples for validation

Dataset SD_Exome_Normal and SD_Exome_Deep

This dataset was used to determine the mosaic detection sensitivity on Exome wide scale. The BAM file used for simulation of data: U175754PMGIAB12891beta with SureSelectV7 enrichment and 211X. Per percentage from 1 to 30, 100 SNVs were simulated. These simulations were spread over 8 times the same BAM file to prevent simulation variants in a close range and to prevent to many mosaic variants in one BAM file. SNV positions in which mosaic variant were simulated using BAMSurgeon were randomly selected from a sub selection of the genome aggregation database (gnomAD). The v2.1.1 data set (GRCh37/hg19) (https://storage.googleapis.com/gcp-public-data—gnomad/release/2.1.1/vcf/exomes/gnomad.exomes.r2.1.1.sites.vcf.bgz) all exomes of gnomAD was used to pick the SNV positions. These were selected with *creating_varfiles* a in house developed python tool; this script can be found at https://github.com/UMCUGenetics/Dx_mosaic_project/blob/main/creating_varfiles/creating_varfiles.py. SNV with a low allele frequency (AF < 0.01) and an allele count of more than 0 (AC > 0) were selected (see supplementary 2.1.1 and 2.1.2 for used command). Selected SNV position in text format made with *creating_varfiles* were used as input for BAMSurgeon to simulate mosaic variants in the BAM file (see supplementary 2.2 for used command).

Dataset SD_PID_5, SD_PID_5strand and SD_PID_10

This dataset was used to determine the mosaic detection sensitivity base pair resolution within an autoinflammatory gene-panel (PID09). The PID09 gene panel consist of five clinically relevant genes; NLRP3, NLRC4, PSTPIP1, NOD2, and TNFRSF1A and UBA1 was added because it is also clinically relevant. Target regions can be found in the supplementary, all exons apart in supplementary 2 table 1 and overlapping regions collapsed in supplementary 3 table 2. For each base pair in the target (32164 positions) a mosaic variant was simulated with a specific percentage. These variants were based on not being the reference allele in 3 iterations of 10 samples, so every non-reference allele was tested. The SD_PID_5 and SD_PID_10 datasets were created with the Settings_LH_NS_mem settings (table 3).

- Dataset SD_PID_5, mosaic variants were simulated at 5% in BAM files from 10 samples for which a regular diagnostic WES was performed (see supplementary 6, table 3).
- Dataset SD_PID_5strand, similar to SD_PID_5 but with the Settings_LH_NS_mem_SB settings (table 3).
- Dataset SD_PID_10, mosaic variants were simulated at 10% in BAM files from 30 samples for which a regular diagnostic WES was performed (see supplementary 6 table 3).

First the small variation files (varfiles) per position were written with a loop (supplementary 1, part 2.4). This python script *writing_small_varfiles* can be found at the GitHub (supplementary 1, part 2.4). An overview of varfiles was created and an overview of the BAM files with their alpha, beta, sex, location, and name was created. The alpha and beta were calculated on the whole BAM file one time with strandbias on 0.05 and one time with

strandbias on 0.000001 (supplementary 1, part 2.5.1 & 2.5.2), for the simulation of the mosaic variants sliced BAM files were used to reduce computing time. These sliced BAM files were made with samtools (supplementary 1, part 2.6). The overview of the varfiles and BAM files were used to start arrays (supplementary 1, part 2.7.1 & 2.7.2), which simulated the mosaic variant in the sliced BAM file and performed MosaicHunter on this BAM file with mosaic variant.

Simulating a variant on top of an already present variants with BAMSurgeon is not possible. These positions were ignored.

Parameter optimalisation and selection of filters MosaicHunter

Settings of MosaicHunter were optimized based on our datasets using an iterative approach (see table 3 for the used settings sets). The following changes were made on the recommended settings by the developers of MosaicHunter and were used as a starting point.

- dbSNP filtering was turned off
- The minimum depth was decreased to 25
- The maximum depth was increased to 5000 (as our WES contained high-coverage, high-reliable regions e.g., the BRCA1/2 genes)
- The minimum mosaic percentage was set to 1
- The minimum number of alternative reads was set to 2

Based on the iterations 4 more filters were altered

- Syscall filter was turned off
- The homopolymer filter was elongated. The minimum length of small homopolymers was changed to 10 and the minimum length of long homopolymer was changed to 16.
- The aligner that BAMSurgeon uses was changed to the bwa-mem aligner
- The strandbias filter p-value was changed to 0.000001

Table 3 contains an overview of the changed parameters and the configuration files used when testing the different settings.

Table 3: Configuration files used during parameter optimalisation. With the changed parameters in comparison to the recommended setting from MosaicHunter. Changed parameters are based on MosaicHunter recommended settings. Links to the configuration files can be found in supplementary 1, part 4.1 until part 4.5.

	Changed parameters	Configuration files
Settings_default	Depth min: 25 Depth max: 5000 No dbSNP file Min alt allele reads: 2 Min mosaic percentage: 0.01	Step 1) exome_default_parameters.conffile Step 2) exome_default.conffile
Settings_LH_NS	Depth min: 25 Depth max: 5000 No dbSNP file Min alt allele reads: 2 Min mosaic percentage: 0.01 Small homopolymer: 10 Long homopolymer: 16 Syscall filter off	Step 1) exome_lh_ns_parameters.conffile Step 2) Exome_lh_ns.conffile
Settings_LH_NS_	Depth min: 25	Step 1)

mem	Depth max: 5000 No dbSNP file Min alt allele reads: 2 Min mosaic percentage: 0.01 Small homopolymer: 10 Long homopolymer: 16 Syscall filter off	exome_lh_ns_parameters.conffile Step 2) exome_lh_ns.conffile
Settings_LH_NS_ mem_SB	Depth min: 25 Depth max: 5000 No dbSNP file Min alt allele reads: 2 Min mosaic percentage: 0.01 Small homopolymer: 10 Long homopolymer: 16 Syscall filter off Strandbias: 0.000001	Step 1) exome_lh_ns_sb_parameters.conffile Step 2) exome_lh_ns_sb.conffile

Testing of MosaicHunter with simulation datasets

Sensitivity of MosaicHunter on 100X WES samples was determined with SD_Exome_Normal (table 2). MosaicHunter was used to call mosaic variants for each of the 8 BAM files per sample. This analysis was performed using a connected pipeline that included two steps (see supplementary 1, part 5.1.3):

1) Determining the alpha and beta
2) Calling mosaic variants using the determined alpha and beta values

To determine the sensitivity of MosaicHunter on deep sequencing 500X WES samples, we used SD_Exome_Deep (table 2). The 8 BAM files with simulated mosaic variants were analyzed using the MosaicHunter Settings_LH_NS_mem_SB settings. To reduce the analysis time for deep sequencing samples, we performed step 1 (determining the alpha and beta values) as described above (see supplement 1, part 5.2.1). But for step 2, the analysis was split per chromosome so it could run in parallel. Chromosomes 1-22 were processed using an array-job (see supplement 1, part 5.2.2), and chromosomes X and Y were processed separately (see supplement 1, part 5.2.3).

To determine the sensitivity of MosaicHunter in the exons of PID09 panel and UBA1, we used SD_PID_5, SD_PID_5strand and SD_PID_10. The sensitivity was calculated by finding the percentage of BAM files where the simulated mosaic variant was detected at each simulated mosaic variant position. We also calculated the percentage of simulated positions found per BAM file and the average number of positions found in all BAM files.

Dataset TP_Regular & TP_Deepseq

Two datasets were available for samples with a known mosaic variant: TP_Regular and TP_Deepseq. All samples were sequenced using the Ilumina Novaseq 6000 platform (paired-end sequencing 2x150bp). Data was processed using regular diagnostic workflow including mapping to the reference genome (hg19) using bwa-mem (Li & Durbin, 2009), and variant calling with GATK haplotypecaller (Ernst & Elferink, 2020).

Dataset TP_Regular (table 4) consisted of 3 patients for which a regular diagnostic WES was performed. Exome capture was performed using the Agilent SureSelect Crev2 design and the aimed sequencing was 100X.

Table 4: TP_Regular dataset. true positive samples 100X depth WES. With suspected mosaic position and percentage, and the gene where it is located.

Samplename	Average coverage	Position	Mosaic Percentage*	Gene
Sample_1	89X	X:47058450 (A>G)	31%	UBA1
Sample_2	90X	X:47058451 (T>C)	83%	UBA1
Sample_3	136X	X:47058451 (T>C)	43%	UBA1

*) as provided by labspecialist

Dataset TP_Deepseq (table 5) consisted of 7 samples for which WES sequencing was performed aimed at high coverage (deep-sequencing or deepseq). Exome capture was performed using the Agilent SureSelect V7 design and the aimed sequencing was 5 times regular sequencing depth, thus 500X.

Table 5: TP_Deepseq dataset. true positive samples 500X depth WES. With suspected mosaic position and percentage, and the gene where it is located.

Samplename	Average coverage	Position	Mosaic Percentage**	Gene
Sample_4	560X*	1:247588457 (G>A)	17,7%	NLRP3
Sample_5	190X	1:247587669 (A>T)	11%	NLRP3
Sample_6	665X*	1:247588457 (G>A)	35,9%	NLRP3
Sample_7	785X	X:47058450 (A>G)	31%	UBA1
Sample_8	533X*	12:6442956 (C>A)	1,40%	TNFRSF1A
Sample_9	807X	X:47058451 (T>C)	17%	UBA1
Sample_10	524X*	X:47058451 (T>C)	43%	UBA1

*) 1 of 2 lanes used for analysis of this sample
**) as provided by labspecialist

To test the sensitivity of MosaicHunter, we used true positive control samples with both 100X and deep sequencing 500X coverage. For this, we used the Settings_LH_NS_mem_SB configuration files. For the 100X WES samples a connected pipeline was used *mosaichunter_best_practice_pipeline* (supplementary info 1, part 5.1.3). For the deepsequencing 500X coverage WES. The first step was done with *exome_lh_ns_sb_stetp1* (supplementary info 1, part 5.2.1) and the second step was performed in parallel for each chromosome, with *mosaichunter_deepseq_best_practice_step2* (supplementary info 1, part 5.2.2) and *mosaichunter_deepseq_best_practice_X_Y_step2* (supplementary info 1, part 5.2.3).

A known mosaic variants was scored as true positive if the position and genotype were detected in de MosaicHunter output, irrespectively of the frequency. For the false negative results MosaicHunter results were parsed to determine if and which filter was responsible for the filtering.

VCF parser

A VCF parser was written in python to convert the .tsv output of MosaicHunter to the standardized variant calling format (VCF). See supplementary info 1, part 6 for the command to run this parser.

Results & Discussion

The only data available with known true positive pathogenic mosaic variants has only one mosaic variant present in them. But datasets with several true mosaic variants are needed to test MosaicHunter, calculate statistics and optimize settings. These datasets would preferably be exome-wide and for multiple mosaic percentages, and preferably based on data similar to our own diagnostics (lab)flow. Due to this lack of datasets, it was decided to create needed datasets for validation and testing ourselves, based on our diagnostic WES data. BAMSurgeon was used to simulate mosaic variants in both GIAB WES and random diagnostic WES samples. BAMSurgeon was used because the tools consider already existing variants and can drop variants that misalign.

Installation of tools

Before creating datasets and testing MosaicHunter and BAMSurgeon we need to ensure that they could run on the high-performance cluster (HPC) of the UU. To do this, dockers and virtual environment were used, which allow us to have multiple installed tools in them without having to install them directly on the HPC. Using this it is also possible to use several versions of a tool or package next to each other in different containers or environments. The current best practice in bioinformatics almost always makes use of these virtual containers with only needed packages installed because it makes room for better version control.

Creation of datasets

For SD_Exome_Normal and SD_Exome_Deep variants were randomly simulated in the whole exome. The artificial variants made are based on low allele frequency in the world population for this the gnomAD database was used. Variants with a low frequency in the population are selected because we suspect that there are not many mosaic variants present in the database and because it prevents from simulating a mosaic variant on top of an already existing variant. For this reason, variants in the gnomAD with AF < 0.01 & AC > 0 were selected.

For SD_PID_5, SD_PID_5strand and SD_PID_10 in all the base pairs of six clinically relevant genes mosaic variants were simulated. The six genes were chosen because their known pathogenicity with a mosaic variant present.

Parameter optimalisation

Parameter optimalisation for based on simulated data was performed to get an optimal sensitivity for MosaicHunter. Base settings were determined based on what we want to detect with MosaicHunter, these settings were altered from the recommended settings. For this project we would like to find mosaic percentages as low as 1% so we changed the minimal percentage to 1, recommended by the developers of MosaicHunter was 5%. The minimum and maximum depth were changed to depths that included all regions of the exome, as our coverage often exceeded the maximum default setting (especially for deepseq). The minimum was set to 25 and the maximum was set to 5000, as our WES contained high-coverage, high-reliable regions e.g., the BRCA1/2 genes. The minimum number of reads was set to 2, to allow for low mosaic percentages. The dbSNP file was not used, to not have the results based on existing SNVs. These altered settings on top of the recommended settings of MosaicHunter gave the Settings_default settings set. After this several filters were tuned in iterations, described below.

Settings_default

With the settings set Settings_default 44% of simulated variants in dataset SD_Exome_Normal were detected (table 7), which resulted in a low sensitivity. It was noticed that high percentage of variants were filtered out by the homopolymer filter (32%), mappingquality filter (28%), and syscallfilter (20%) (table 6). Close inspection on the syscall filter revealed that this filter was based on a Hiseq dataset. As our sample were sequenced Novaseq this is different data therefore the syscall filter was turned off. The homopolymer filter was elongated to allow for mosaic variants to be in short homopolymers. These settings gave settings set Settings_LH_NS

Settings_LH_NS

With the settings set Settings_LH_NS sensitivity of dataset SD_Exome_Normal increased to 63% (table 7). Many simulated mosaic variants were filtered with the mapping quality filter (49%) (table 6). After investigating, we discovered that the aligner used by BAMSurgeon did not match the aligner used to create the BAM files that we used. As a result, we needed to adjust the aligner used by BAMSurgeon to match the aligner we use to create the BAM file in order to reduce the number of variants filter by the mapping quality filter. As a last optimalisation the aligner that BAMSurgeon uses was changed to the BWA-mem aligner that we use to create our BAM files.

Settings_LH_NS_mem

With the settings set Settings_LH_NS_mem sensitivity of dataset SD_Exome_Normal increased to 76% (table 7). The strandbias filter was filtering out the majority of mosaic variants simulated in the sample (11%) (Table 6). The strandbias filter was altered to p=0.000001 instead of 0.05.

Settings_LH_NS_mem_SB

With the settings set Settings_LH_NS_mem_SB sensitivity of dataset SD_Exome_Normal decreased to 74% (table 7). But a shift in the found percentages was observed, the low percentages were picked up better and the high worse. Between 1 and 20% now 79% were found but between 20% and 30% 61% of mosaic variants were found. This is preferable because we want to look at the low percentage mosaic variants rather than high percentage mosaic variants. It was decided these settings of MosaicHunter and BAMSurgeon would be used in further testing of the tool.

Table 6 overview of total simulated variants filtered by each filter. Based on a sample of dataset SD_Exome_Normal.

Filter	Settings_default	Settings_LH_NS	Settings_LH_NS_mem	Settings_LH_NS_mem_SB
Homopolymers filter	408 (32%)	14	20	20
Mapping quality filter	354 (28%)	416 (49%)	37	38
Mosaic filter	92	190	228	346
Syscall filter	263 (20%)	Filter off	Filter off	Filter off
Within read position filter	80	95	109	112
Complete linkage filter	2	3	3	3
Misaligned reads filter	45	80	77	77

Strand bias filter	41	52	58 (11%)	0
Common site filter	1	1	1	1
Total filtered	**1286**	**851**	**533**	**597**

Table 7: Results for MosaicHunter for the different parameter settings. Based on a sample of dataset SD_Exome_Normal.

	Settings_ default	Settings_LH_ NS	Settings_LH_ NS_mem	Settings_LH_ NS_mem_SB
Detected simulated mosaic variants (percentage of total)	1040 (44%)	1475 (63%)	1805 (76%)	1741 (74%)
Total variants simulated in BAM file	2345	2345	2372	2372
Missed simulated variants	1305	870	533	597
Extra mosaic variants found	173	288	282	327

By optimizing the settings (Settings_LH_NS_mem_SB) we increased sensitivity from 44% to 74%, an increase of 30% compared to the default setting (Settings_default). We therefore recommend using these settings in future analyses. Within this study we used Settings_LH_NS_mem_SB to analyze the simulated and true positive datasets, unless stated otherwise.

Further optimizing is still possible. The syscall filter could be trained on own data. Most reads are filtered out by the misaligned reads filter and the within read position filter, it would be interesting to examine these filters.

GATK

To determine the number of mosaic variants that could already be detected using our WES workflow genotyping based on SD_Exome_Normal and SD_Exome_Deep were performed using the GATK haplotypecaller. GATK was able to detect >95% of mosaic variants above 12% mosaic in SD_Exome_Normal (fig 3) and SD_Exome_Deep (fig 4) but genotyped these mosaic variants as heterozygous variants.

Dataset SD_Exome_Normal and SD_Exome_Deep

In the previous analysis we discovered that the GATK haplotypecaller sensitivity to detects mosaic variants >12% mosaic is high. We also want to determine the sensitivity to detect mosaic variants of MosaicHunter on variants between 1-30%. especially the variants that are not detected by GATK between 1-12%. So, an in-depth comparison of GATK and MosaicHunter performance was performed.

First, we analyzed GATK vs MosaicHunter for the SD_Exome_Normal dataset. MosaicHunter does not detect mosaic variant from 0 to 3 percent reliably (<70% sensitivity). From 3 to 22 percent MosaicHunter has a high >70% sensitivity for mosaic variants (fig 3). Above 22 percent MosaicHunter performs worse <70% sensitivity. GATK is able to detect

>95% of mosaic variants >12% but has a low sensitivity <12% (fig 3). MosaicHunter performs well 3-12% compared to GATK (fig 3). With the current settings and sequence depth MosaicHunter is therefore able a clear added value when combined with GATK.

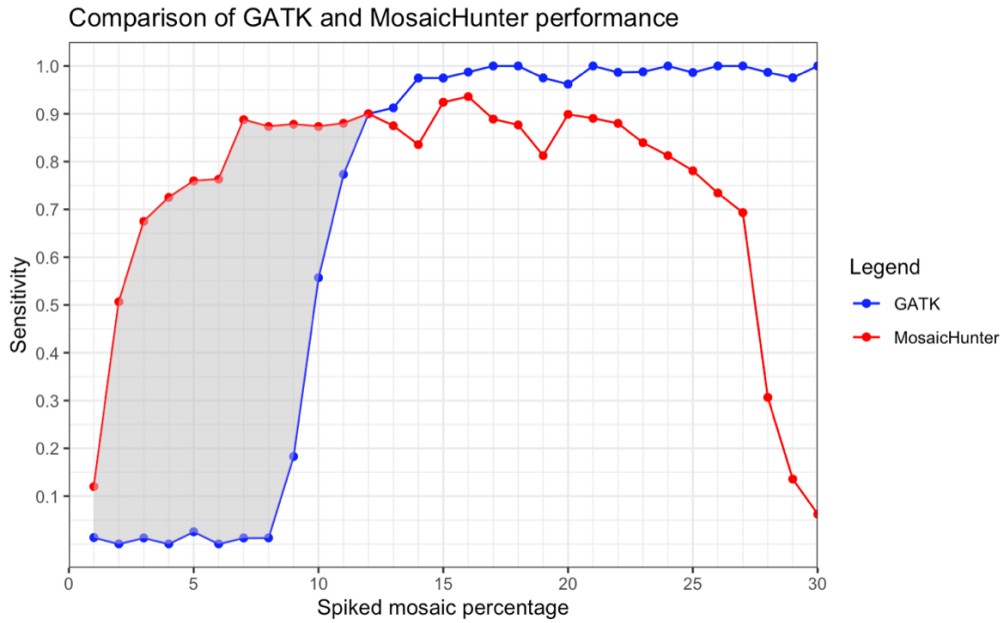

Figure 3: Comparison of the performance of GATK and MosaicHunter on 100X coverage WES sample. In Blue the sensitivity of GATK and in red the sensitivity of MosaicHunter. On the y-axis the sensitivity is depicted and, on the x-axis, the different mosaic percentages.

To investigate the potential added value of sequence coverage, we also performed the similar analysis with SD_Exome_Deep. GATK is able to detect >95% of mosaic variants >11% in the 717X coverage WES sample but has a low sensitivity <5% (fig 4). MosaicHunter has a high >80% sensitivity in the range of 2-10% (fig 4) and outperforms GATK here. In comparison with the 100X coverage samples there is a clear increase in sensitivity both with GATK (>11%) and MosaicHunter (>2%) (fig 5).

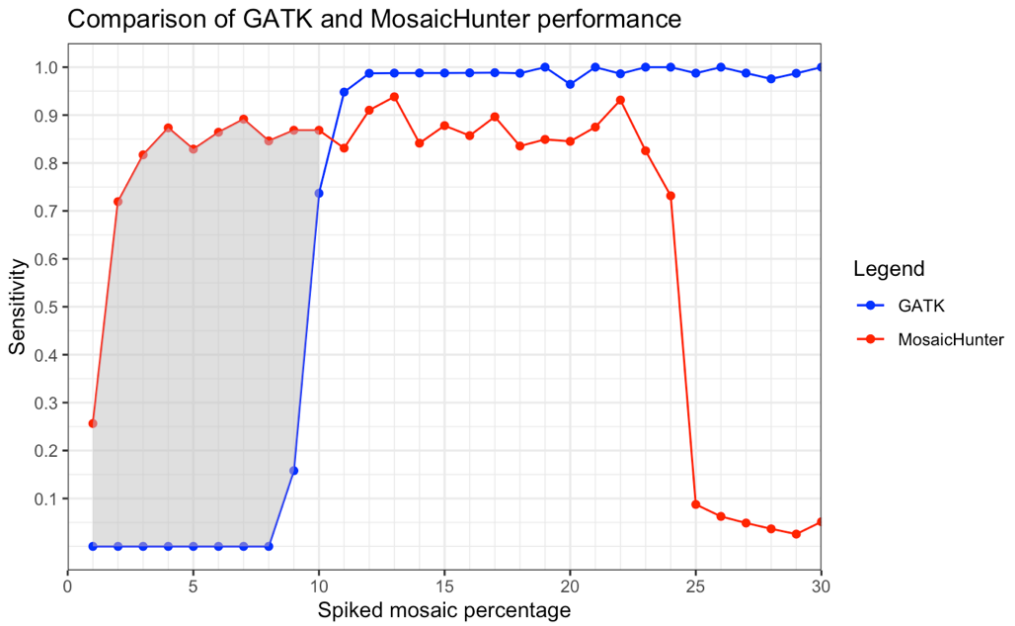

Figure 4: Comparison of the performance of GATK and MosaicHunter on 500X coverage WES sample. In Blue the sensitivity of GATK and in red the sensitivity of MosaicHunter. On the y-axis the sensitivity is depicted and, on the x-axis, the different mosaic percentages.

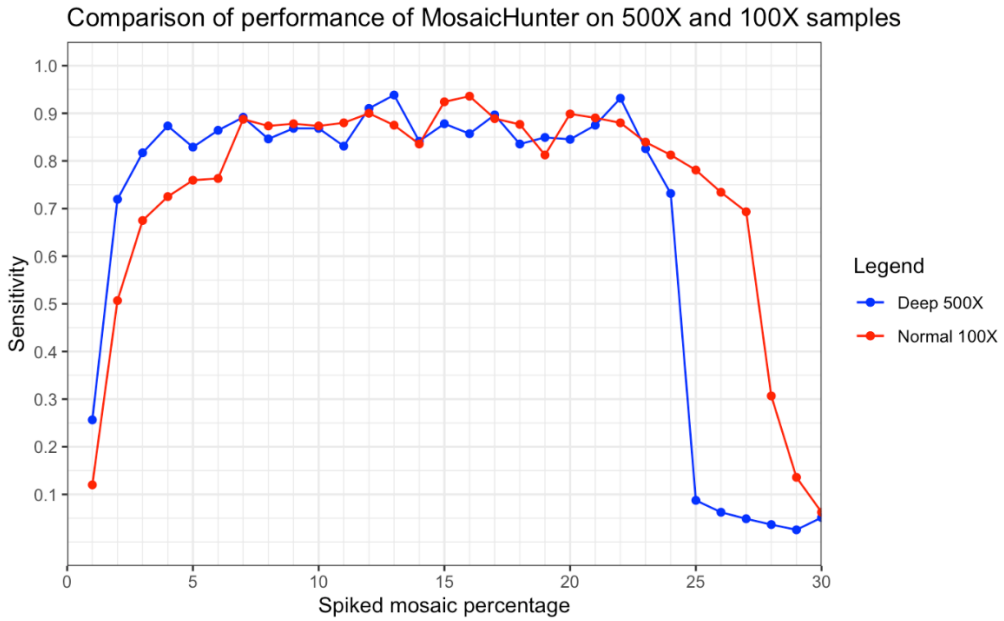

Figure 5: Comparison of sensitivity of MosaicHunter on deepsequencing and normal sequencing samples. Deepsequencing, 500X coverage, sample (blue) and a normal, 100X coverage, sample (red). On the y-axis the sensitivity is depicted and, on the x-axis, the different mosaic percentages.

MosaicHunter performs better on 500X deepsequencing data than it does on 100X data (fig 5). For deepsequencing data >80% of mosaic variants between 3 and 7% are detected where on 100X coverage between 70 and 80% is detected (fig 5). So, using high coverage would be beneficial based on our simulated data.

With GATK being very sensitive (>95%) for variants above 12% mosaic, it is reliable to detect these mosaic variants, although as heterozygous variants. MosaicHunter is reliable (>80%) to detect variants between 3-22% and will be of added value to GATK between 3-12%.

For the 100X SD_Exome_Normal dataset 3 samples are used to calculate results but for the 500X SD_Exome_Deep dataset only 1 sample was used to calculate the results. To Improve the comparison in extra deepsequencing samples mosaic variants need to be simulated.

Dataset SD_PID_5, SD_PID_5strand and SD_PID_10

In the previous analyses we determined the exome wide statistics to detect simulated variants. However, these analyses were performed using a relative low resolution (100 variants per BAM per percentage). To reach high resolution and gain insight in sensitivity we decided to perform a base pair resolution simulation for the six clinically relevant genes in which pathogenic mosaic variants are known. These six clinically genes were based on the PID09 gene panel and UBA1, so it contained six genes known to cause autoinflammatory diseases. Five out of these six genes are currently part of the autoinflammatory mosaic-detection gene panel sequenced with smMIPs, the PID09.

Because MosaicHunter outperforms GATK between 3 and 12 percent it was decided to focus on mosaic percentages of 5% (lower limit) and 10% (upper limit). For each percentage, we simulated a mosaic variant in each base pair within BAM files of 30 samples.

Results SD_PID_10 & SD_PID_05

For SD_PID_05 with the 5% mosaic percentage variants simulated on average 72,5% of simulated mosaic variants in the target region were detected by MosaicHunter. For SD_PID_10 with the 10% mosaic percentage simulated an average 81,5% of the simulated variants in the target region were detected by MosaicHunter. A positive trend to detect mosaic variants was observed with increased sequence coverage (fig 8).

Most target regions are detected even well, more than 50% of positions of every target are above 80% detected in the 30 BAM files (fig 6) for the SD_PID_10 dataset. For the SD_PID_5 dataset more than 50% of positions of every target are above 70% in the 10 BAM files (fig 7) which is a bit lower than in the SD_PID_10 dataset.

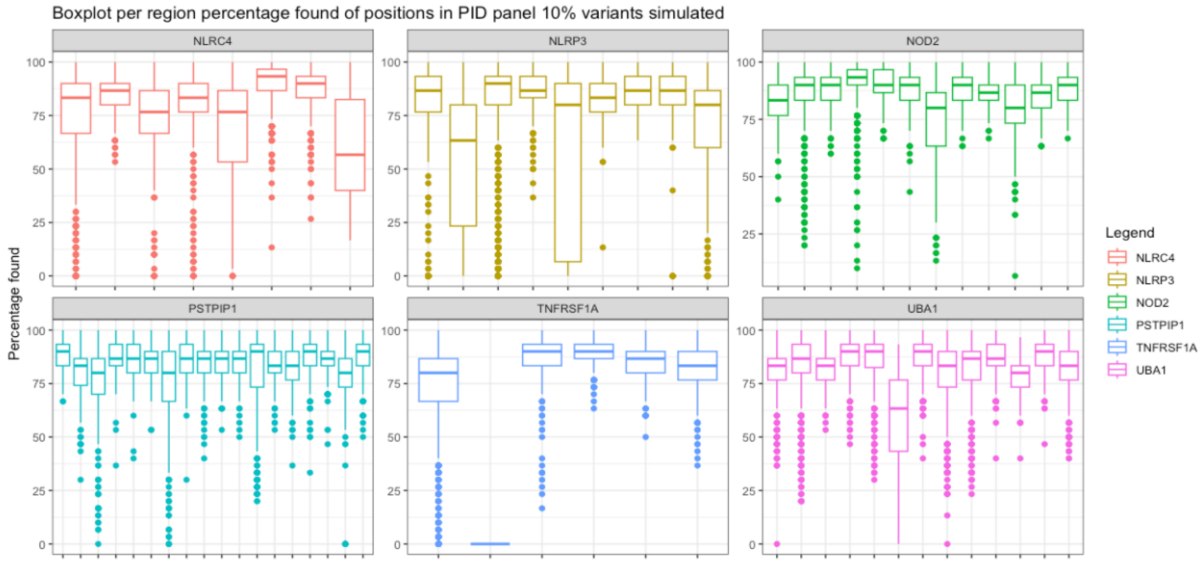

Figure 6: Overview of boxplots of 10% simulated variants per target region in PID09 + UBA1 panel. Each boxplot represents a target region (supplementary 2) which could consist of one or more exons, overlapping regions are collapsed for this analysis). For each position in each region the percentage of found simulated variants in calculated and a boxplot of these percentages was created.

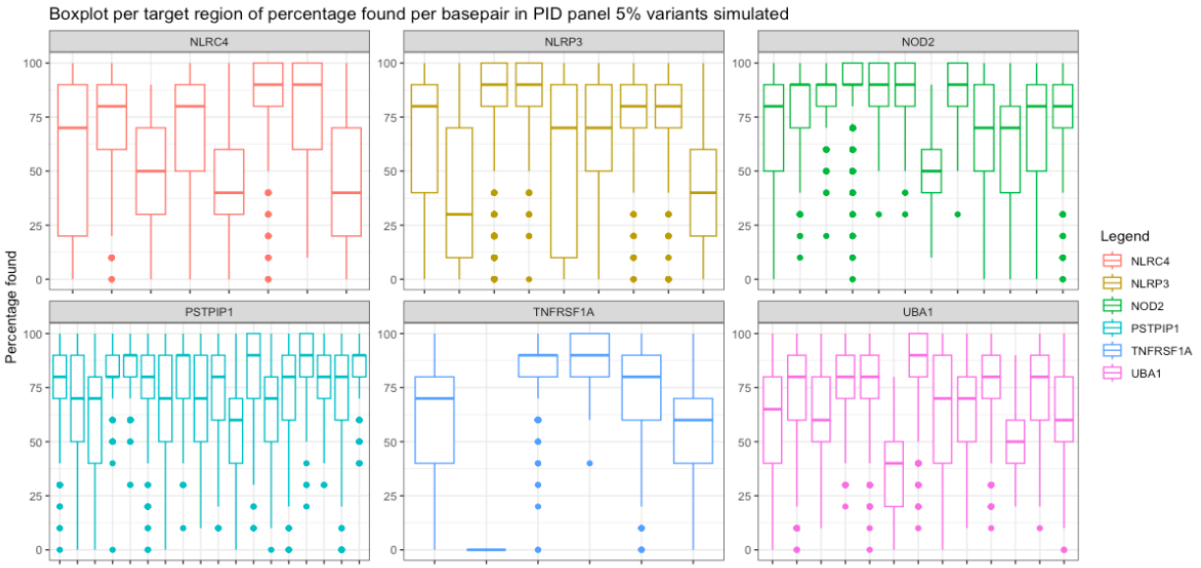

Figure 7: Overview of boxplots of 5% simulated mosaic variants per target region in PID09 + UBA1 panel. Each boxplot represents a target region (supplementary 2) which could consist of one or more exons, overlapping regions are collapsed for this analysis). For each position in each region the percentage of found simulated variants in calculated and a boxplot of these percentages was created.

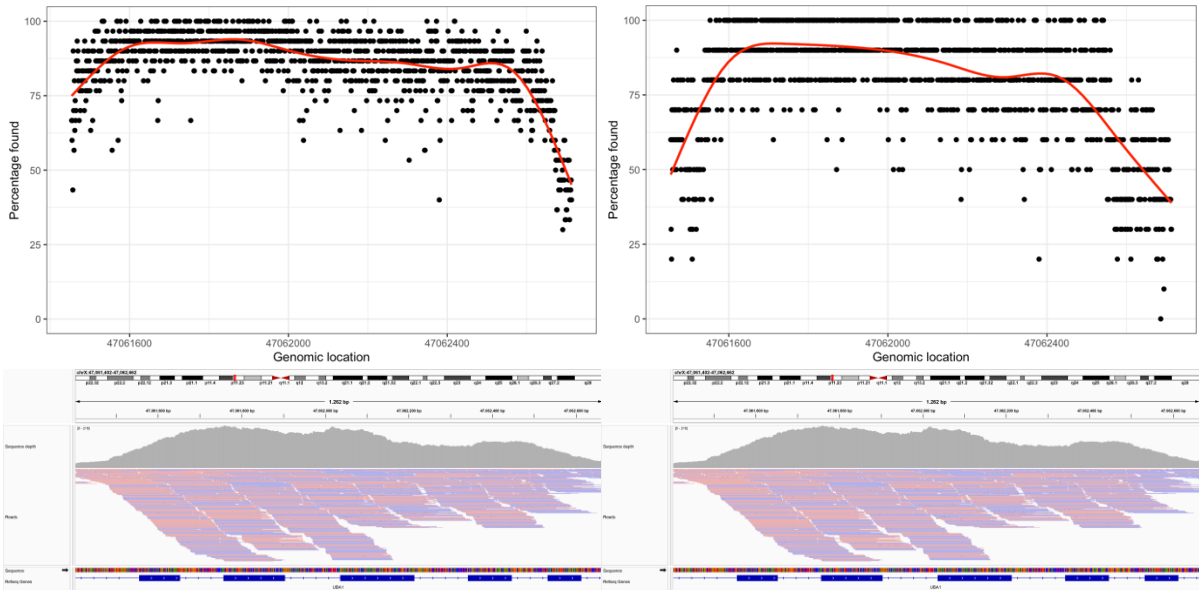

Figure 8: percentage found per position in representative target region. Top part, percentage of simulated variants detected per position. In red the line that represents the average over the positions. Bottom part, overview of IGV of the same target region. Depicted in grey is the depth and in blue and red are the forward and reverse reads. Left graph, 10% mosaic simulated variants. Right graph, 5% simulated variants.

There is a clear drop in sensitivity in the flanks of the targeted regions (fig 8) and within intronic regions (fig 8). For the coding part, the exons, it is expected that the sensitivity per position on average is higher than for the flanks. It would be interesting to look at the sensitivity of MosaicHunter on the exomes without the 100bp flanks and with 20bp flanks. Also, diagnosis of genetic disorders in the UMCU is performed on the exome with 20 bp flanks. Which means that the outer 80bp of the flanks we also simulated variants in will not be considered in routine diagnostics. This would be interesting for both the 10% and 5% mosaic variant simulations.

With tuned strandbias in the SD_PID_5strand analyses no large differences were found (Data not shown).

Dataset TP_Regular & TP_Deepseq

In the previous analyses we determined the exome wide statistics and statistics for high resolution in clinically relevant gene regions to detect simulated variants. This was all done on mosaic variants simulated in BAM files. To get an insight on the performance of MosaicHunter on true diagnostic samples with a known pathogenic mosaic variant two datasets TP_Regular and TP_Deepseq were used. TP_Regular consisting of WES data from 3 samples with a sequence coverage of 90-140X, and TP_Deepseq consisting of WES data from 7 samples with a sequence coverage of 500-1000X.

After running the TP_Deepseq for the first time (settings set Settings_LH_NS_mem) 4 of 7 samples were filtered with the strandbias filter. It was decided to alter the p-value of the strandbias filter to allow for a bigger different in strandbias between the mosaic variant and the reference allele.

Results TP_Regular

For validation of MosaicHunter three true positive 100X WES samples were available. These samples were analyzed with MosaicHunter with Settings_LH_NS_mem_SB settings. The mosaic variants of sample Sample_3 and Sample_2 were detected by MosaicHunter and with GATK. The mosaic variant of sample Sample_1 was not detected with MosaicHunter as

this variant was filtered in the mosaic filter and was thus classified as a heterozygous variant by MosaicHunter. Which is likely because the mosaic variant was present in 31% of reads. It also was called as a heterozygous variant by our GATK pipeline. Called MosaicHunter 2/3 (67%), called GATK 3/3 (100%), called by either MosaicHunter or GATK 3/3 (100%).

Results TP_Deepseq

For validation of MosaicHunter 7 samples with known mosaic variants were sequenced on 5 times sequence depth, per sample coverage is shown in table 8. In total 85.7% (6/7) of the mosaic variants were detected by either MosaicHunter (3/7) or GATK (6/7).

The mosaic variant of sample Sample_8 was not detected using either method, likely as a result of the low mosaic percentage. As shown in the results of exome wide simulation, sensitivity to detect mosaic variants <3% is low for WES samples with both normal and deep coverage. It would be interesting to see if increased coverage could lead to detection (see final discussion). Sample Sample_71 & sample_6 were not detected due to the mosaic filter because the mosaic percentages are 31% and 36% MosaicHunter filtered them as heterozygous variants. Sample Sample_4 was not detected due to the strandbias filter, there was a big difference in the proportion of forward and reverse reads between the reference allele and variants present.

This finding shows that for these samples GATK is sensitive enough to detect the mosaic variants because of the percentages that are >12% (table 8). This is in line with the results found with the exome wide analysis.

Table 8: overview of dataset TP_Regular and TP_Deepseq. Included are the positions, depth and calls of MosaicHunter and GATK.

Sample name	TP variant	Mosaic percentage*	Found with MosaicHunter	GATK call	Depth
Sample_10	X:47058451 (T>C)	43%	Yes (43%)	Heterozygous	524X
Sample_3	X:47058451 (T>C)	43%	Yes (42%)	Heterozygous	136X
Sample_2	X:47058451 (T>C)	17%	Yes (17%)	Heterozygous	90X
Sample_9	X:47058451 (T>C)	17%	Yes (15%)	Heterozygous	807X
Sample_5	1:247587669 (A>T)	11%	Yes (10%)	Heterozygous	190X
Sample_7	X:47058450 (A>G)	31%	No	Heterozygous	785X
Sample_4	1:247588457 (G>A)	18%	No	Heterozygous	560X
Sample_6	1:247588457 (G>A)	36%	No	Heterozygous	665X

Sample_1	X:47058450 (A>G)	31%	No	Heterozygous	89X
Sample_8	12:6442956 (C>A)	1.4%	No	No call	533X

*) as provided by labspecialist

One note to make is that except Sample_6, the TP_Deepseq samples failed diagnostics QC due to high contamination (6-8% based on VerifyBAM contamination results) with a yet to be determined source. This contamination could have negative influence on the results. To get the most reliable results these samples need to be sequenced again on a depth of 500X coverage and without contamination. One other note is that the 100X coverage WES samples were sequenced using an older enrichment design (CREv2). To have the results most in line with our current methods the samples have to be sequenced with the SureSelect V7 capturing kit, as this is the current diagnostic standard at the UMCU when performing WES.

Precision

In the previous chapters we focused on the sensitivity to detect mosaic variants as this is the most important for diagnostics. However, typically high sensitivity has a tradeoff for lower precision or increased false positive results.

To get a feeling about the precision the number of found mosaic variants by MosaicHunter on both TP_Regular 100X and TP_Deepseq 500X was looked at. In a 100X sample on average 350 (± 128) variants were found, in a 500X sample on average 307 (± 38) variants were found. Subtracting the variants found with GATK from the list of mosaic variants found by MosaicHunter leads to a 50% reduction of variants that need to be checked (~300 > 150) (table 9 & 10). Filtering for the six clinically relevant genes (PID09+UBA1) results in 0-2 mosaic variants per sample (table 9 & 10).

Apart from that the variant rate for mosaic variants in a person is not known (based on literature study), which makes it difficult to estimate the number of mosaic variants an individual normally has. Not knowing the truth makes it difficult to determine the false positive rate. Mosaic variants detected with MosaicHunter can either be true positive variants (which might or might not be involved in genetic diseases) but can also be artifacts that are labeled as mosaic variants resulting from i.e., sequencing mistakes or mis mapping. This does mean that you cannot label all the found mosaic variants that are not known to cause illness as false positives. Whether these mosaic variants are true positives or false positives can only be determined by another method, like for instance ddPCR.

Table 9: overview of amount of mosaic variants detected by MosaicHunter per BAM file. Number of mosaic variants called by MosaicHunter and filter for several panels and ranges.

Sample	Exome 100 bp flanks	PID09 panel + UBA1	Next to GATK
Sample_4	273	0	146
Sample_5	362	2	144
Sample_6	261	0	94
Sample_7	344	0	147
Sample_8	276	0	128
Sample_9	305	1	115
Sample_10	325	1	169

Sample_1	240	0	79
Sample_2	322	1	100
Sample_3	489	1	82

Table 10: Average number of mosaic variants detected by MosaicHunter per BAM file. Number of mosaic variants called by MosaicHunter and filter for several panels and ranges.

Depth	Exome 100bp flanks	PID09 panel + UBA1	Next to GATK
500X	307 (± 38)	0-2	135 (± 25)
100X	350 (± 128)	0-1	87 (± 12)

Future perspectives

Before mosaic variant detection using MosaicHunter can be implemented in routine diagnostics, more (validation) analyses are needed. This includes finding the optimal coverage for MosaicHunter, analyzing more samples with a true positive variant between 2 – 12% mosaic and simulating more datasets. There are currently two major challenges when using MosaicHunter: analysis time and precision.

The analysis time of MosaicHunter can be long, especially for high sequence depths. This means that when samples need to be analyzed fast MosaicHunter cannot be performed. A 100X WES sample can be analyzed in 8 hours, and by using parallelization in step 2 this can be reduced to around 4-5 hours. However, analyzing deep sequencing 500X samples will take more than a week. Even with parallelization in step 2, this will only be reduced to about 4 days. Parallelization of step 1 is not possible, due to the alpha and beta that are calculated on all the reads, making the analysis time for deep sequencing samples problematic. Possible solutions to the analysis time challenge include basing the alpha and beta on a smaller part of the exome. For instance, on one chromosome that represents the whole exome, or basing the alpha and beta on only the clinically relevant genes (~ 3000), or finding another region that can provide representative alpha and beta values. Another solution to shorten analyzing time might be to focus the whole analysis with MosaicHunter on a subset of genes, possibly the ~3000 clinically relevant diagnostic genes. This will prohibit direct mosaic variant calling in other genes, but it provides much broader resolution compared to for instance smMIP panels.

The precision of MosaicHunter is another bottleneck. There are currently hundreds of mosaic variants detected by MosaicHunter, which is too many to check manually for pathogenicity. Different filtering strategies need to be tested to reduce this number to an acceptable number of variants. One option is to subtract the variants found with GATK, but flag them as mosaic in the GATK results, from the list of mosaic variants found by MosaicHunter and examine the number of remaining variants. This reduction will lead to a 50% reduction of variants (~300 > 150) (table 9 & 10). Another option is to filter for different gene panels. These panels can be based on literature and would contain clinically relevant genes that cause disease. This will significantly reduce the number of variants found. For the six clinically relevant genes (PID09+UBA1) this results in 0-2 mosaic variants that need to be annotated and possibly validated with a second method (table 9 & 10). It is suspected hat other gene panels will yield similar results. Because MosaicHunter outperforms GATK in the 3 – 12% range, filtering for the variants in this range could reduce the number of variants that need to be checked. At last, filtering for impact and annotation on gene function and relation to the disease of interest can also be effective in reducing the list of variants, because only variants with a high pathogenic impact will be kept.

In this study, we only validated performance of MosaicHunter on mosaic variants outside the 3-12% range. Additional samples with a true positive mosaic variant in this range are required for complete validation of MosaicHunter.

More simulations like the SD_PID09_10 and SD_PID09_5 should be conducted, including every percentage between 1-15% and using both deep sequencing and normal sequencing. This will give a base pair resolution of MosaicHunter for the clinically relevant genes and will give more insight in the sensitivity of MosaicHunter on the different mosaic percentages.

To improve the statistical power of the results of MosaicHunter, it would be beneficial to find the coverage where performance is optimal. This could be done by stepwise down sampling 500X samples and assessing sensitivity for each step. Another option may be to consider sequencing with a coverage >500X, which may increase sensitivity for detection of mosaic variants below 2%. This study has shown that MosaicHunter performs better on 500X samples than on 100X samples but has not investigated other coverages. It would be interesting to examine the coverage at the six clinically relevant genes (PID09 + UBA1) and compare this to the coverage of the entire exome. If the coverage is the same, we would expect similar results of MosaicHunter for all positions of the exome as the performance for the six clinically relevant genes. We expect the performance of MosaicHunter is positively correlated with coverage, but this assumption has not been validated in this project.

One option of implementation of MosaicHunter is to use the tool alongside the current GATK germline variant calling pipeline, which is run on all whole exome sequencing (WES) samples by the UMCU. Examining the results and the number of variants called can provide insight into the performance of MosaicHunter.

Based on these considerations, it is recommended to implement MosaicHunter in diagnostics, but to also run it alongside GATK for a while as mentioned above. It would also be advisable to perform validation with another method.

References

Beck, D. B., Ferrada, M. A., Sikora, K. A., Ombrello, A. K., Collins, J. C., Pei, W., Balanda, N., Ross, D. L., Cardona, D. O., Wu, Z., Patel, B., Manthiram, K., Groarke, E. M., Gutierrez-Rodrigues, F., Hoffmann, P., Rosenzweig, S., Nakabo, S., Dillon, L. W., Hourigan, C. S., … Grayson, P. C. (2020). Somatic Mutations in UBA1 and Severe Adult-Onset Autoinflammatory Disease. *The New England journal of medicine*, *383*(27), 2628–2638. https://doi.org/10.1056/NEJMoa2026834

Docker Hub Container Image Library | App Containerization. (z.d.). Geraadpleegd 24 november 2022, van https://hub.docker.com/

Dou, Y., Kwon, M., Rodin, R. E., Cortés-Ciriano, I., Doan, R., Luquette, L. J., Galor, A., Bohrson, C., Walsh, C. A., & Park, P. J. (2020). Accurate detection of mosaic variants in sequencing data without matched controls. *Nature biotechnology*, *38*(3), 314–319. https://doi.org/10.1038/s41587-019-0368-8

Eijkelenboom, A., Kamping, E. J., Kastner-van Raaij, A. W., Hendriks-Cornelissen, S. J., Neveling, K., Kuiper, R. P., Hoischen, A., Nelen, M. R., Ligtenberg, M. J. L., & Tops, B. B. J. (2016). Reliable Next-Generation Sequencing of Formalin-Fixed, Paraffin-Embedded Tissue Using Single Molecule Tags. *The Journal of Molecular Diagnostics*, *18*(6), 851–863. https://doi.org/10.1016/j.jmoldx.2016.06.010

Ernst, R. F., & Elferink, M. (2020). *UMCUGenetics/DxNextflowWES: V1.2.0*. Zenodo. https://doi.org/10.5281/zenodo.4551799

Forsberg, L. A., Gisselsson, D., & Dumanski, J. P. (2017). Mosaicism in health and disease—Clones picking up speed. *Nature Reviews Genetics*, *18*(2), Art. 2. https://doi.org/10.1038/nrg.2016.145

Hoffman, H. M., & Broderick, L. (2017). It just takes one: Somatic mosaicism in

autoinflammatory disease. *Arthritis & rheumatology (Hoboken, N.J.)*, *69*(2), 253–256. https://doi.org/10.1002/art.39961

Huang, A. Y., Zhang, Z., Ye, A. Y., Dou, Y., Yan, L., Yang, X., Zhang, Y., & Wei, L. (2017). MosaicHunter: Accurate detection of postzygotic single-nucleotide mosaicism through next-generation sequencing of unpaired, trio, and paired samples. *Nucleic Acids Research*, *45*(10), e76. https://doi.org/10.1093/nar/gkx024

Ionescu, D., Peñín-Franch, A., Mensa-Vilaró, A., Castillo, P., Hurtado-Navarro, L., Molina-López, C., Romero-Chala, S., Plaza, S., Fabregat, V., Buján, S., Marques, J., Casals, F., Yagüe, J., Oliva, B., Fernández-Pereira, L. M., Pelegrín, P., & Aróstegui, J. I. (2022). First Description of Late-Onset Autoinflammatory Disease Due to Somatic NLRC4 Mosaicism. *Arthritis & Rheumatology*, *74*(4), 692–699. https://doi.org/10.1002/art.41999

Koboldt, D. C., Zhang, Q., Larson, D. E., Shen, D., McLellan, M. D., Lin, L., Miller, C. A., Mardis, E. R., Ding, L., & Wilson, R. K. (2012). VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Research*, *22*(3), 568–576. https://doi.org/10.1101/gr.129684.111

Labrousse, M., Kevorkian-Verguet, C., Boursier, G., Rowczenio, D., Maurier, F., Lazaro, E., Aggarwal, M., Lemelle, I., Mura, T., Belot, A., Touitou, I., & Sarrabay, G. (2018). Mosaicism in autoinflammatory diseases: Cryopyrin-associated periodic syndromes (CAPS) and beyond. A systematic review. *Critical Reviews in Clinical Laboratory Sciences*, *55*(6), 432–442. https://doi.org/10.1080/10408363.2018.1488805

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)*, *25*(14), 1754–1760. https://doi.org/10.1093/bioinformatics/btp324

Mutect2. (z.d.). GATK. Geraadpleegd 11 juli 2022, van https://gatk.broadinstitute.org/hc/en-us/articles/360037593851-Mutect2

Ramu, A., Noordam, M. J., Schwartz, R. S., Wuster, A., Hurles, M. E., Cartwright, R. A., & Conrad, D. F. (2013). DeNovoGear: De novo indel and point mutation discovery and phasing. *Nature methods*, *10*(10), 985–987. https://doi.org/10.1038/nmeth.2611

Smith, K. S., Yadav, V. K., Pei, S., Pollyea, D. A., Jordan, C. T., & De, S. (2016). SomVarIUS: Somatic variant identification from unpaired tissue samples. *Bioinformatics*, *32*(6), 808–813. https://doi.org/10.1093/bioinformatics/btv685

van der Made, C. I., Potjewijd, J., Hoogstins, A., Willems, H. P. J., Kwakernaak, A. J., de Sevaux, R. G. L., van Daele, P. L. A., Simons, A., Heijstek, M., Beck, D. B., Netea, M. G., van Paassen, P., Elizabeth Hak, A., van der Veken, L. T., van Gijn, M. E., Hoischen, A., van de Veerdonk, F. L., Leavis, H. L., & Rutgers, A. (2022). Adult-onset autoinflammation caused by somatic mutations in UBA1: A Dutch case series of patients with VEXAS. *Journal of Allergy and Clinical Immunology*, *149*(1), 432-439.e4. https://doi.org/10.1016/j.jaci.2021.05.014

Vijg, J., & Dong, X. (2020). Pathogenic mechanisms of somatic mutation and genome mosaicism in aging. *Cell*, *182*(1), 12–23. https://doi.org/10.1016/j.cell.2020.06.024

Yang, X., Xu, X., Breuss, M. W., Antaki, D., Ball, L. L., Chung, C., Li, C., George, R. D., Wang, Y., Bae, T., Abyzov, A., Wei, L., Sebat, J., Network, N. B. S. M., & Gleeson, J. G. (2021). *DeepMosaic: Control-independent mosaic single nucleotide variant detection using deep convolutional neural networks* (p. 2020.11.14.382473). bioRxiv. https://doi.org/10.1101/2020.11.14.382473

Yohe, S., & Thyagarajan, B. (2017). Review of Clinical Next-Generation Sequencing. *Archives of Pathology & Laboratory Medicine*, *141*(11), 1544–1557.

https://doi.org/10.5858/arpa.2016-0501-RA